KARISE LEARNING SYSTEM

# AI Omkostninger

## Hvad koster AI egentlig?

Bent Karise

En faktabog om
de reelle omkostninger ved AI

# AI Omkostninger

## Hvad koster AI egentlig?

En faktabog om de reelle omkostninger ved AI

Version 1.1

**Bent Karise**

Karise Learning System · Karise Media & Publishing

# Kolofon

AI Omkostninger · Hvad koster AI egentlig?

Forfatter: Bent Karise

Udgiver: Karise Media & Publishing · Karise Learning System



Digital udgave · dansk

| | |
|---|---|
| Version: | 1.1 |
| Første udgave: | 2026 |
| Senest fagligt opdateret: | 5. september 2026 |
| Aktuel udgave: | www.bentkarise.com |

Denne bog er et faktabaseret referenceværk om de økonomiske, energi-, klima-, vand-, materiale-, infrastruktur- og arbejdsrelaterede omkostninger ved AI. Kilder og beregningsforudsætninger er angivet løbende. Tal, leverandørpriser, regulering og tekniske forhold kan ændre sig, og læseren bør derfor kontrollere bogens versionsnummer og dato, når oplysninger bruges i en aktuel beslutning eller offentlig gengivelse.

Bogen erstatter ikke juridisk, økonomisk, investeringsmæssig, teknisk, miljøfaglig eller anden professionel rådgivning. Produktnavne og varemærker tilhører deres respektive ejere. Omtale af virksomheder, produkter og teknologier er redaktionel og indebærer ikke sponsorering eller tilknytning.

Den seneste offentlige version findes via www.bentkarise.com. Ved citering skal titel, forfatter, versionsnummer og år angives.

# Forord

Hvor meget koster AI egentlig?

Det spørgsmål er det der fik mig til at skrive denne bog. Jeg mente først, at det primært handlede om elektricitet, datacentre og måske prisen på de AI-værktøjer, virksomheder køber adgang til.

Jo længere jeg fulgte tallene, desto tydeligere blev det, at spørgsmålet kun ser enkelt ud, fordi mange forskellige regnskaber ofte bliver presset sammen til ét.

I den offentlige debat møder vi meget præcise udsagn. En prompt bruger et bestemt antal watt-timer. En AI-søgning bruger ti gange så meget strøm som en almindelig søgning. En prompt bruger et glas vand. AI bruger strøm som et helt land. En GPU holder tre til fem år. Hvert udsagn kan have en historie og en kilde bag sig. Problemet opstår, når systemgrænsen, geografien, tidspunktet eller de oprindelige antagelser forsvinder, mens tallet bliver stående.

Det samme gælder den positive side af regnskabet. AI kan spare tid, øge kvaliteten og gøre opgaver mulige, som tidligere var for dyre eller langsomme. Men sparet tid er ikke automatisk lig med sparet løn. Lavere energi pr. opgave er ikke automatisk lavere samlet energiforbrug. En billigere model pr. token er ikke nødvendigvis den billigste måde at få et brugbart resultat på.

Derfor har jeg forsøgt at skrive en bog, der først og fremmest hjælper læseren med at stille det rigtige spørgsmål. Hvad er det præcist, vi måler? Hvad er med i regnestykket? Hvilket land, hvilket elsystem, hvilket vandopland og hvilket tidspunkt gælder tallet for? Er data målt, estimeret eller modelleret? Og er to løsninger overhovedet sammenlignelige, når kvaliteten af resultatet tages med?

Bogen er skrevet til dig som er politiker, journalist, IT-beslutningstager eller andre, der allerede møder AI som teknologi, investering,

arbejdsredskab eller samfundsspørgsmål. Den skal samtidig kunne læses af den interesserede læser, som ikke arbejder professionelt med energi, livscyklusanalyser eller datacenterteknik. Ambitionen er ikke at gøre alle til specialister, men at gøre de vigtigste skel synlige.

Europa er bogens primære referencepunkt, når geografi ændrer resultatet. Det betyder ikke, at amerikansk eller asiatisk forskning er irrelevant. Tværtimod bygger store dele af den tekniske evidens på globale studier. Men amerikanske emissionsfaktorer, vandforhold eller tarifstrukturer bliver ikke automatisk europæiske, blot fordi teknologien er den samme.

Jeg har også valgt at lade usikkerhed stå synligt. På flere områder findes der endnu ikke et robust offentligt tal. Det gælder blandt andet den faktiske retirement-alder for moderne AI-acceleratorer og en særskilt europæisk statistik for AI-relateret e-affald. I sådanne tilfælde er manglen på data en del af konklusionen. Et tomrum skal ikke fyldes med et bekvemt tal, bare fordi et regneark ellers ser mere komplet ud.

AI udvikler sig hurtigt. Derfor er denne bog heller ikke tænkt som et øjebliksbillede, der står urørt efter udgivelsen. Nye forskningsresultater og myndighedsdata bliver løbende vurderet. Når ny dokumentation ændrer, korrigerer eller væsentligt styrker bogens konklusioner, udkommer en ny version.

Mit håb er, at bogen gør det lidt sværere at gentage et dramatisk AI-tal uden at spørge, hvad det egentlig betyder – og lidt lettere at føre en mere præcis samtale om både omkostningerne og gevinsterne.

Mit mål gennem denne research- og skriveproces har været at fremstille tallene så neutralt som muligt, uden at forholde mig til andet end det tallene viser. Der ligger derfor ingen politiske eller andre motiver bag arbejdet med denne bog.

Du er meget velkommen til at bruge denne bog i dit arbejde og citere den så meget du vil, men da data ændrer sig og AI udvikler sig meget

hurtigt, så citér venligst med Titel, Versionsnummer, Årgang og så selvfølgelig mit navn.

God læselyst.

Bent Karise
2026

# Om denne udgivelse

En levende faktabog

Bogen er tænkt som et levende dokument som jævnligt bliver opdateret og kan hentes gratis på min hjemmeside.

AI-teknologi, datacenterdrift, energisystemer, leverandørpriser og forskningen i AI's miljø- og samfundspåvirkning ændrer sig hurtigt. AI Omkostninger udgives derfor som en levende digital publikation med et synligt versionsnummer.

Nye peer-reviewede studier, myndighedsdata, tekniske rapporter og andre væsentlige kilder bliver vurderet løbende. En ny bogversion udkommer, når dokumentationen ændrer, korrigerer eller væsentligt styrker det eksisterende evidensgrundlag. En ny artikel udløser altså ikke automatisk en ny version.

Versionsnummeret gør det muligt at se, hvilken udgave en oplysning stammer fra. Væsentlige ændringer registreres i ændringsloggen bagest i bogen. Den aktuelle offentlige udgave findes via www.bentkarise.com.

| | |
|---|---|
| **1.0** | Første offentlige udgave. |
| **1.x** | Nye data, kildeopdateringer, præciseringer og mindre faglige udvidelser. |
| **2.0** | Større faglig revision, væsentligt nyt evidensgrundlag eller større ændring af bogens struktur. |

## Sådan læses tallene

Observerede data, beregnede værdier, modellerede scenarier og illustrative regneeksempler holdes adskilt. Illustrative eksempler er markeret som sådanne og skal ikke læses som målinger af en bestemt model eller virksomhed.

Hvor geografi er vigtig, anvendes princippet global evidens – europæisk fortolkning. Tal fra andre regioner kan være stærk evidens, men bruges

ikke ukritisk som direkte proxy for europæiske el-, emissions-, vand- eller infrastruktursystemer.

Bogens tilbagevendende element “Påstanden under lup” undersøger udbredte udsagn. Vurderingen kan være bekræftet, delvist korrekt, misvisende, ikke dokumenteret eller forældet afhængigt af den konkrete evidens.

# Indholdsfortegnelse

## Del 1

# Hvad koster AI egentlig?

Del 1 fastlægger bogens målemetode. Før AI's omkostninger kan sammenlignes, skal vi vide, hvilket regnskab der undersøges, hvilken systemgrænse der bruges, og hvilken funktionel enhed tallene beskriver.

## Kapitel 1

# Hvad menes med en omkostning?

Hvad koster AI?

Det lyder som et enkelt spørgsmål, men rummer flere forskellige regnskaber.

For nogle handler spørgsmålet om prisen på et abonnement til ChatGPT, Copilot eller en anden AI-tjeneste. For andre handler det om strømforbruget i de datacentre, hvor modellerne kører. I den offentlige debat handler det ofte om klima, vand, råstoffer eller behovet for nye elnet og kraftværker. I virksomheder kan den største omkostning i stedet vise sig at være medarbejdernes tid, integration med eksisterende systemer eller arbejdet med at kontrollere og rette AI-genereret indhold.

Alle disse forhold kan være reelle omkostninger ved AI.

De beskriver forskellige typer omkostninger.

Det er bogens udgangspunkt.

Hvis vi vil forstå de reelle omkostninger ved kunstig intelligens, må vi først skelne mellem de forskellige regnskaber. Ellers risikerer vi at sammenligne liter med kilowatt-timer, investeringer med driftsudgifter og potentielle samfundsomkostninger med konkrete virksomhedsregninger.

De forskellige regnskaber kan godt sammenholdes, men kun når det er tydeligt, hvad der sammenlignes, og hvilke antagelser analysen bygger på.

Det er især vigtigt i en tid, hvor AI udvikler sig hurtigt, og hvor meget præcise tal let bliver gengivet uden den sammenhæng, de oprindeligt blev beregnet i.

## Flere regnskaber på samme tid

En virksomhed, der tager et nyt AI-værktøj i brug, kan forholdsvis let se den direkte økonomiske pris. Der kommer en faktura på et abonnement eller et antal API-kald. Måske købes der nye computere, eller måske indgås der en kontrakt med en cloudleverandør.

Det er den synlige del af regningen.

Bag den ligger andre regnskaber.

AI-systemet kræver elektricitet. Den elektricitet produceres et sted og på et bestemt tidspunkt. Serverne skal fremstilles. Chips, hukommelse, strømforsyning, racks og kølesystemer kræver materialer og energi, før de overhovedet bliver installeret i et datacenter.

Datacenteret skal bygges og tilsluttes elnettet. Der kan være behov for nye transformere, transmissionsforbindelser eller anden infrastruktur. Kølingen kan kræve vand, elektricitet eller begge dele. Når hardwaren udskiftes, skal den enten bruges et andet sted, sælges videre, skilles ad eller ende som affald.

Samtidig findes virksomhedens interne omkostninger. Medarbejdere skal lære at bruge systemet. AI-output skal kontrolleres. Fejl skal findes og rettes. Data skal måske forberedes. Systemer skal integreres. Sikkerhed og adgangsrettigheder skal håndteres.

AI har derfor flere regninger på samme tid.

En vigtig opgave i denne bog bliver at holde dem adskilt længe nok til, at de hver især kan vurderes på deres egne præmisser.

Den internationale teleunion ITU har gjort denne afgrænsning til et centralt metodekrav i ITU-T L.1801 fra 2026. Anbefalingen beskriver AI-systemets miljøpåvirkning gennem en livscyklus, hvor råmaterialer, produktion, drift, støtteinfrastruktur og end-of-life kan indgå afhængigt af analysens formål og systemgrænse.

**Figur 1.1 – AI's omkostningsregnskaber**

| Energi | Klima | Vand |
|---|---|---|
| Materialer | **AI-aktivitet / anvendelse** | Infrastruktur |
| Direkte økonomi | Menneskelig tid | Gevinster / undgåede omkostninger |

*Kilde/ramme: Forfatterens oversigt med udgangspunkt i ITU-T L.1801 (2026) og de kilder, der anvendes i bogens efterfølgende kapitler.*

## Økonomiske omkostninger

Den mest intuitive omkostning er den økonomiske.

Hvad betaler virksomheden?

Det kan være licenser, abonnementer, API-forbrug, cloudkapacitet, hardware, softwareintegration, konsulenter, uddannelse og drift.

Også her bliver spørgsmålet hurtigt mere komplekst.

En model kan være billig pr. token, men kræve flere forsøg, mere kontrol og mere efterbehandling end en dyrere model. En medarbejder kan spare ti minutter på at fremstille et første udkast og bagefter bruge femten minutter på at kontrollere og rette det.

Den direkte AI-pris fortæller derfor ikke nødvendigvis, hvad opgaven reelt kostede at løse.

Senere i bogen vender vi tilbage til dette spørgsmål og ser på omkostningen ved det resultat, der faktisk kan accepteres og bruges.

## Energi er et fysisk regnskab

Elektricitet er en anden type omkostning.

Her kan vi måle energiforbruget i watt-timer, kilowatt-timer eller megawatt-timer.

Det lyder mere entydigt.

Også energiregnskabet afhænger af, hvor vi sætter grænsen.

Måler vi kun acceleratorchippen?

Hele serveren?

Kølingen?

Netværket mellem serverne?

Storage?

Datacenterets øvrige strømforbrug?

Eller går vi længere tilbage og medregner den energi, der blev brugt til at fremstille hardwaren og bygge infrastrukturen?

Alle systemgrænser kan være relevante.

De besvarer bare forskellige spørgsmål.

To analyser af det samme AI-system kan derfor komme frem til forskellige energital og stadig begge være metodisk korrekte, hvis de har målt forskellige dele af systemet.

Forskellen ligger i så fald i afgrænsningen, ikke nødvendigvis i regnefejlen.

### Klima og energi er to forskellige regnskaber

Energi og klima bliver ofte omtalt næsten som synonymer, selv om de beskriver forskellige størrelser.

En kilowatt-time er en fysisk mængde energi. Klimaeffekten afhænger blandt andet af, hvordan elektriciteten er produceret.

Den samme mængde elektricitet kan derfor have forskellig klimabelastning afhængigt af land, region, tidspunkt og metode.

Derudover findes emissionerne fra fremstillingen af servere, chips, bygninger og anden infrastruktur.

Et energieffektivt system er derfor ikke automatisk det system med det laveste samlede klimaaftryk.

Hvis en ny generation hardware f.eks. bruger mindre elektricitet pr. beregning, men kræver, at fungerende udstyr udskiftes tidligere, opstår

der et nyt regnestykke mellem lavere driftsforbrug og den miljøbelastning, der følger med fremstillingen af nyt udstyr.

Svaret afhænger af de konkrete forhold.

Det er et gennemgående tema i denne bog.

ITU-T L.1801 (2026) anbefaler netop, at klima, vand og ressourceforbrug behandles som særskilte påvirkningskategorier, når de er relevante for den konkrete AI-analyse. Det giver et mere transparent beslutningsgrundlag end at presse alle påvirkninger ind i ét samlet tal.

### Vand er et lokalt regnskab

Vanddebatten viser endnu tydeligere, hvorfor et enkelt tal sjældent kan stå alene.

Et datacenter kan trække vand fra en kommunal drikkevandsforsyning, en kanal, overfladevand, genanvendt spildevand eller havvand.

Noget vand kan blive returneret. Noget kan fordampe. Noget kan indgå i andre processer.

Derudover kan vandforbruget i den elproduktion, der leverer elektriciteten til datacenteret, være en del af en bredere analyse.

En million liter vand er derfor ikke automatisk den samme belastning to forskellige steder.

I et område med rigelige vandressourcer kan konsekvensen være én. I et område med sæsonmæssig eller permanent vandknaphed kan den være en anden.

Det Europæiske Miljøagentur, EEA, bruger derfor den sæsonopdelte Water Exploitation Index Plus (WEI+) på vandoplandsniveau. Metoden er netop udviklet til at synliggøre lokal og sæsonmæssig vandstress, som et årligt landsgennemsnit kan skjule.

Det er en af grundene til, at denne bog ikke vil behandle liter som en universel miljøvaluta.

Vi skal vide, hvilken type vand der er tale om, hvor det kommer fra, og hvad der sker med det.

## Materialerne begynder før datacenteret

Når AI diskuteres som fysisk teknologi, bliver opmærksomheden ofte rettet mod selve GPU'en.

Det er forståeligt. De avancerede acceleratorchips er centrale for moderne AI.

Men chippen er kun én del af den fysiske infrastruktur.

Serveren indeholder hukommelse, printkort, strømforsyning og andre komponenter. Datacenteret kræver racks, køling, kabler og elektrisk udstyr. Strømforsyningen kan kræve transformere og nye netforbindelser.

Når efterspørgslen efter AI-kapacitet bliver stor nok, kan materialebehovet derfor brede sig langt ud over de materialer, der findes inde i den enkelte chip.

Begrebet kritisk råstof skal derfor bruges med præcision.

Et materiale kan være kritisk, fordi leverandørkæden er koncentreret eller strategisk sårbar, uden at materialet nødvendigvis udgør en stor del af det samlede fysiske materialeforbrug.

Omvendt kan almindelige materialer som kobber og aluminium blive væsentlige, når infrastrukturen bygges i meget stor skala.

## Omkostningen ved infrastrukturen

Store AI-datacentre eksisterer ikke isoleret fra resten af energisystemet.

De skal tilsluttes et elnet, der har tilstrækkelig kapacitet.

I nogle tilfælde kan nye store belastninger fremskynde behovet for netudbygning, transformere, produktionskapacitet eller lagring.

IEA fremhæver i Energy and AI (2025), at datacentre kan etableres på få år, mens elnet, produktionskapacitet og anden energiinfrastruktur ofte har længere planlægnings- og byggetider. Tidsforskellen er en af

årsagerne til, at meget store nye datacenterbelastninger kan få betydning langt ud over selve elregningen.

Det rejser et politisk og økonomisk spørgsmål, som er langt mere kompliceret end blot at spørge, hvor meget et datacenter betaler for elektricitet.

Hvem udløser investeringen?

Hvem betaler den?

Hvem bærer risikoen, hvis den forventede efterspørgsel ikke kommer?

Hvem får gavn af infrastrukturen bagefter?

Det kan være fire forskellige svar.

En ny transmissionsforbindelse kan være udløst af en stor kunde og senere komme mange andre brugere til gode. Omvendt kan fælles infrastruktur blive bygget på forventninger om fremtidig efterspørgsel, der ikke materialiserer sig som forventet.

Derfor kan prisen på et netprojekt ikke uden videre skrives på AI's regning. Samtidig vil det være misvisende at ignorere den infrastruktur, som store nye belastninger kan være med til at udløse.

Det kræver attribution, kontrafaktiske scenarier og en klar forståelse af, hvordan omkostningerne faktisk fordeles.

## Den menneskelige regning

Der findes også en omkostning, som let forsvinder i diskussionen om chips og datacentre.

Menneskelig tid.

AI bliver ofte introduceret med et løfte om effektivisering.

Det kan være reelt.

En medarbejder kan få hjælp til research, oversættelser, analyse, tekstproduktion, kundeservice eller andre opgaver.

Men AI fjerner sjældent alle menneskelige aktiviteter omkring opgaven.

Nogen skal formulere opgaven.

Nogen skal vurdere resultatet.

Ved følsomme eller komplekse opgaver skal oplysninger kontrolleres. Fejl skal rettes. Der kan være behov for flere forsøg, fordi det første svar ikke er godt nok.

I andre tilfælde kan AI faktisk skabe nyt arbejde: mere dokumentation, flere muligheder, flere analyser og dermed også mere, der skal vurderes.

Derfor skal en påstået tidsbesparelse behandles med samme kildekritiske disciplin som en påstået miljøbelastning.

Hvis en undersøgelse viser, at medarbejdere udførte en bestemt opgave hurtigere med AI, er det værdifuld information.

En målt tidsbesparelse kan ikke direkte oversættes til en tilsvarende besparelse på lønkontoen.

Den frigjorte tid kan være brugt på andre opgaver. Produktionen kan være steget. Kvaliteten kan være forbedret. Der kan være opstået nye opgaver.

Tidsbesparelsen kan altså være værdifuld uden at være en direkte kontant besparelse. Analysen skal vise, hvilken gevinst der faktisk er målt.

### Omkostning for hvem?

En af de vigtigste skelnen i denne bog bliver spørgsmålet om perspektiv.

Hvad koster AI for hvem?

For virksomheden kan den relevante omkostning være licenser og medarbejdertid.

For datacenteroperatøren kan det være hardware, elektricitet, køling og kapital.

For elnettet kan det være kapacitet og investeringer.

For lokalsamfundet kan vand, areal, støj eller arbejdspladser være relevante forhold.

For samfundet samlet kan spørgsmålet også handle om produktivitet, skatteindtægter, energisikkerhed og miljøpåvirkning.

Ingen af disse perspektiver er i sig selv mere rigtige end de andre.

Problemet opstår, hvis et resultat fra ét perspektiv præsenteres som svaret for alle.

En virksomheds lave direkte AI-regning fortæller ikke nødvendigvis noget om de samlede fysiske ressourcekrav.

Et højt fysisk ressourceforbrug afgør omvendt heller ikke i sig selv, om investeringen er samfundsøkonomisk god eller dårlig. Det kræver yderligere analyse.

En politisk neutral analyse begynder med at få regnskaberne frem, før konklusionen drages.

### Det, vi kan måle, behøver ikke at kunne lægges sammen

Når flere omkostningsregnskaber bliver synlige, opstår fristelsen til at samle dem.

Det kan i visse analyser være nyttigt at sætte økonomiske værdier på klima, vand eller andre eksterne effekter.

I det øjeblik bevæger analysen sig fra fysisk måling til værdisætning.

Det kræver nye antagelser.

Hvad koster et ton $CO_2$?

Hvad koster en kubikmeter vand?

Er værdien den samme i Danmark og i et område med alvorlig vandmangel?

Hvordan prissættes belastningen af et elnet?

Hvad er værdien af en medarbejders frigjorte tid, hvis den bruges på mere produktion i stedet for at reducere bemandingen?

Der findes metoder til at arbejde med disse spørgsmål.

Resultaterne afhænger af værdisætning og antagelser og kan derfor ikke betragtes som neutrale naturkonstanter.

Derfor vil denne bog så vidt muligt først præsentere de fysiske og økonomiske størrelser i deres egne enheder.

Hvis der senere foretages en monetær eller samlet vurdering, skal antagelserne være synlige.

### Gevinster hører også med

En neutral analyse skal også registrere dokumenterbare gevinster.

AI kan have gevinster.

Et AI-system kan reducere energiforbruget i en bygning. Det kan optimere transport. Det kan hjælpe medarbejdere med at udføre opgaver hurtigere eller med færre fejl. Det kan reducere spild eller forbedre udnyttelsen af eksisterende kapacitet.

IEA's Energy and AI (2025) behandler tilsvarende begge sider af energiregnskabet: AI kan øge datacentrenes elbehov og samtidig skabe effektiviseringsmuligheder i energisystemet. I denne bog anvender vi samme princip bredere og kræver dokumentation, uanset om en påstand handler om belastning eller gevinst.

Hvis disse effekter kan dokumenteres, skal de med.

Gevinsterne skal behandles med præcis samme krav til evidens som omkostningerne.

Potentiale, realiseret bruttogevinst og nettogevinst er tre forskellige niveauer, som skal holdes adskilt.

Hvis et AI-system reducerer et andet energiforbrug med 100 enheder, men selv kræver 30 enheder for at skabe besparelsen, er den relevante nettoeffekt ikke 100.

Det samme gælder økonomisk.

Hvis AI sparer tid på en arbejdsopgave, men kræver licenser, kontrol, implementering og korrektion, må begge sider af regnskabet med.

Det afgørende spørgsmål bliver derfor ofte:

Hvad ville der realistisk være sket uden AI?

Det kontrafaktiske spørgsmål kommer til at følge os gennem hele bogen.

## Regnskaber før domme

AI-debatten bliver let polariseret.

Den ene fortælling beskriver kunstig intelligens som en teknologi, der vil effektivisere samfundet, accelerere forskning og reducere ressourceforbrug.

Den anden beskriver en industri med voksende energiforbrug, store datacentre, vandforbrug og omfattende behov for hardware og infrastruktur.

Begge fortællinger kan indeholde dokumenterbare dele.

De kan også begge blive overdrevne.

Bogens formål er at undersøge, hvilke omkostninger der faktisk kan dokumenteres, hvordan de er beregnet, hvor de opstår, hvem der bærer dem, og hvilke usikkerheder der følger med.

En samlet dom over AI eller ét stort facittal ville skjule de forskelle, som beslutningstagere har brug for at kunne se.

Det samme gælder gevinsterne.

Når evidensen er stærk, skal konklusionen være tydelig.

Når resultatet afhænger af bestemte antagelser, skal antagelserne være synlige.

Når vi ikke ved det endnu, skal det også fremgå.

Usikkerhed er en del af beslutningsgrundlaget. For beslutningstagere er forskellen mellem det, vi ved, det vi estimerer, og det vi endnu ikke kan dokumentere, ofte vigtigere end endnu et tilsyneladende præcist tal.

## Det første spørgsmål er: Hvilken omkostning?

Før vi kan vurdere størrelsen på AI's omkostning, må vi vide, hvilket regnskab der undersøges.

Hvilken omkostning taler vi om?

Er det virksomhedens faktura?

Elektriciteten?

Klimaeffekten?

Vandet?

Materialerne?

Infrastrukturen?

Medarbejdernes tid?

Eller en kombination?

Først når det er tydeligt, bliver tallene meningsfulde.

I næste kapitel begynder vi med en af de mest udbredte måder at gøre AI's omkostninger konkrete på: prisen på den enkelte prompt.

Det lyder som en enkel måleenhed.

Den viser samtidig, hvor hurtigt et præcist tal kan blive til en uprækis fortælling.

### Kilder nævnt i kapitlet


- ITU-T L.1801 (2026): Guidelines for assessing the environmental impact of artificial intelligence systems
- IEA (2025): Energy and AI
- EEA (2025): Water scarcity conditions in Europe / Water Exploitation Index Plus (WEI+)

## Kapitel 2

# Hvorfor den enkelte prompt fortæller så lidt

Hvis man vil gøre AI's omkostninger konkrete, er den enkelte prompt et oplagt sted at begynde.

Hvor meget strøm bruger den? Hvor meget vand? Hvor meget $CO_2$? Hvad koster den?

Spørgsmålene inviterer til et enkelt tal. Et tal kan sammenlignes, citeres og huskes. Netop derfor har målinger som watt-timer pr. prompt og liter vand pr. forespørgsel fået en fremtrædende plads i debatten om AI's ressourceforbrug.

Udfordringen opstår, når måleenheden bliver behandlet som mere ensartet, end den er.

En prompt er først og fremmest en handling i en brugerflade. Den fortæller meget lidt om, hvor meget beregningsarbejde der ligger bag. En kort tekstforespørgsel, analyse af et langt dokument, et avanceret reasoning-forløb, billedgenerering og en agent, der foretager flere modelkald og bruger eksterne værktøjer, kan alle begynde med noget, brugeren oplever som én prompt.

De fysiske omkostninger kan alligevel være vidt forskellige.

### Samme prompt - to forskellige energital

Google offentliggjorde i august 2025 en analyse af energiforbruget ved en median tekstprompt i Gemini Apps. Målingen byggede på produktionsdata fra maj 2025 og gav et estimat på 0,24 watt-timer pr. medianprompt.

Tallet er interessant i sig selv, men metodikken bag er endnu mere interessant.

Google medregnede ikke kun energien i de acceleratorer, der var aktive under selve modelkørslen. Metoden omfattede også provisioneret kapacitet, som stod klar til at håndtere trafikspidser og fejl, energi til CPU og RAM samt datacenterets øvrige energiforbrug gennem PUE.

Google viste samtidig, hvad resultatet ville blive med en smallere metode, hvor kun aktive TPU'er og GPU'er blev talt med. Her faldt estimatet til omkring 0,10 watt-timer.

Selv med samme tjeneste, samme type prompt og samme periode blev resultatet markant forskelligt afhængigt af systemgrænsen.

Den bredere driftsgrænse løftede energiestimatet fra 0,10 til 0,24 watt-timer - en faktor 2,4.

Vandberegningen ændrede sig på samme måde. Den smallere metode gav 0,12 milliliter pr. medianprompt, mens den bredere driftsmetode gav 0,26 milliliter. Begge vandtal var stadig afgrænset til Googles driftsmæssige beregning og udgjorde ikke et fuldt livscyklus-vandfodaftryk med hardwareproduktion og alle upstream-effekter.

**Tabel 2.1 – Samme Gemini-prompt, to systemgrænser**

| Måling | Smal grænse | Bred driftsgrænse | Forskel |
|---|---|---|---|
| **Energi** | 0,10 Wh | 0,24 Wh | 2,4× |
| **Vand** | 0,12 ml | 0,26 ml | ca. 2,2× |
| **$CO_2e$** | 0,02 g | 0,03 g | 1,5× |

*Kilde: Google, "Measuring the environmental impact of AI inference", 21. august 2025. Googles egne data; ikke uafhængigt tredjepartsverificeret. Vandtallene er operationelle og ikke et fuldt livscyklus-vandfodaftryk.*

Eksemplet viser, hvorfor et præcist tal bør læses sammen med sin systemgrænse. To decimaler kan give indtryk af stor sikkerhed, selv om den største forskel måske ligger i, hvad analysen har valgt at tælle med.

## En prompt er ikke en standardiseret arbejdsenhed

I energisammenhæng ville det være bekvemt, hvis en prompt svarede til en nogenlunde ensartet mængde beregningsarbejde. Så kunne man måle en repræsentativ prompt og bruge den som omregningsfaktor.

Sådan fungerer moderne AI-systemer ikke.

En bruger kan skrive: "Oversæt denne sætning til engelsk." Svaret kan være få ord og kræve et relativt begrænset inferensforløb.

Den næste bruger kan uploade et langt dokument og bede modellen sammenligne kontraktvilkår, finde uoverensstemmelser og skrive en rapport. En tredje kan bede en reasoning-model løse en kompleks analyse, hvor systemet bruger langt mere beregning før svaret afleveres. En fjerde kan generere billeder eller video. En femte kan aktivere en agent, der søger efter oplysninger, kalder flere værktøjer og vender tilbage til modellen adskillige gange undervejs.

For brugeren kan alle fem forløb begynde på samme måde: en besked sendes til et AI-system.

Som fysisk måleenhed er de meget forskellige.

Det er også en af forklaringerne på, at nyere analyser viser store spænd mellem AI-opgaver. IEA's Key Questions on Energy and AI (2026) beskriver korte tekstforespørgsler som brøkdele af en watt-time på moderne systemer, mens video kan kræve hundreder til tusinder af gange mere energi pr. forespørgsel afhængigt af længde og opløsning. Reasoning og agentiske forløb kan ligeledes løfte beregningsarbejdet markant. IEA understreger samtidig, at tallene er indikative og stærkt afhængige af modelvalg og opgavetype.

Det samlede energiforbrug ved AI bestemmes derfor af flere bevægelige størrelser på én gang: effektiviteten pr. opgave, antallet af opgaver og hvilke typer opgaver der bliver populære.

En lavere pris pr. simpel tekstprompt fortæller kun noget om den første af disse størrelser.

### Modellen ændrer regnestykket

Selv inden for tekst er "en AI-prompt" en uprækcis kategori.

Modeller har forskellig størrelse, arkitektur og specialisering. Nogle er optimeret til hurtige, relativt enkle opgaver. Andre bruger langt mere beregningskraft for at levere højere kvalitet på komplekse spørgsmål. Quantization, batching, caching og serving-arkitektur kan flytte energiforbruget yderligere.

Hardwaregenerationen spiller også ind. Den samme type model kan køre mere energieffektivt på nyere acceleratorer, mens udnyttelsesgraden i datacenteret afgør, hvor godt den installerede kapacitet faktisk bruges.

Et modelnavn er derfor heller ikke nok som dokumentation. En god måling bør så vidt muligt fortælle, hvilken tjeneste eller model der er undersøgt, hvilken opgave der blev udført, hvilken periode målingen vedrører, og hvilken del af infrastrukturen der er medregnet.

Når klima eller vand indgår, bliver geografi og tidspunkt desuden vigtige. Den samme elektriske energimængde kan have forskellig emissionsprofil, og den samme vandmængde kan have vidt forskellig lokal betydning.

### Outputtet er en del af opgaven

Diskussionen om prompts fokuserer ofte på det, brugeren skriver. For ressourceforbruget er det mindst lige så vigtigt, hvad systemet skal producere.

Et kort spørgsmål kan udløse et langt svar. Et andet kan besvares med én sætning. Nogle systemer genererer interne mellemtrin eller bruger ekstra beregning til reasoning. Andre kalder værktøjer eller søger i dokumenter, før svaret dannes.

To identiske inputtekster kan derfor føre til forskelligt beregningsarbejde, hvis systemindstillinger, model eller outputkrav er forskellige.

Det gør også sammenligninger på tværs af tjenester vanskelige. En model, der leverer et kort og upræcist svar, kan bruge mindre energi end en model, der leverer en omfattende analyse. Hvis den korte version efterfølgende kræver flere nye prompts, menneskelig kontrol eller et skift til en stærkere model, har den første måling kun fanget en del af arbejdsforløbet.

Senere i bogen vender vi tilbage til dette problem gennem begrebet pris pr. accepteret resultat. Her er pointen blot, at en prompt er et teknisk hændelsespunkt - ikke nødvendigvis den relevante enhed for den opgave, brugeren faktisk vil have løst.

### Når et historisk tal bliver en nutidig sandhed

Et af de mest udbredte energital i AI-debatten illustrerer et andet problem.

I 2024 udgav Electric Power Research Institute, EPRI, rapporten Powering Intelligence. Rapporten gengav et tidligt estimat på omkring 2,9 watt-timer for en ChatGPT-forespørgsel og cirka 0,3 watt-timer for en traditionel Google-søgning.

Herfra var vejen kort til formuleringen, at en AI-forespørgsel bruger omtrent ti gange så meget elektricitet som en Google-søgning.

Forholdet 10:1 er let at forstå og let at citere. Det er også blevet gentaget meget bredt.

EPRI beskrev selv tallene som estimater for tidlige anvendelser og fremhævede den betydelige usikkerhed omkring AI's fremtidige energiforbrug. Rapporten blev skrevet i en periode, hvor de offentlige data om produktionstjenester var langt mere begrænsede end i dag.

Googles måling fra 2025 viste 0,24 watt-timer for en median Gemini Apps-tekstprompt med virksomhedens bredere driftsmetode. Det tal kan ikke sættes direkte op mod de 0,3 watt-timer for en ældre Google-søgning og bruges til at konkludere, at AI nu er mere energieffektivt end

søgning. Målemetoderne, tjenesterne, perioderne og funktionerne er forskellige.

Sammenstillingen viser noget andet og mere nyttigt: et forholdstal, der blev rimeligt gengivet som et tidligt estimat i 2024, bør ikke leve videre som en tidløs fysisk konstant.

**Figur 2.1 – Fra afgrænset estimat til generaliseret påstand**

| |
|---|
| **1. Kilde:** EPRI, Powering Intelligence (2024) |
| **2. Tidlige estimater:** ca. 2,9 Wh pr. ChatGPT-forespørgsel og ca. 0,3 Wh pr. traditionel Google-søgning |
| **3. Forholdstal:** ca. 10× under de konkrete antagelser |
| **4. Generaliseret formulering:** "En AI-forespørgsel bruger 10 gange så meget strøm som en Google-søgning." |

*Pointen er ikke, at EPRI's historiske estimat var værdiløst, men at årstal, metode og funktionel sammenligning skal følge med, når forholdstallet citeres videre.*

### Påstanden under lup

**Påstand:** En AI-forespørgsel bruger cirka 10 gange så meget strøm som en Google-søgning.

**Vurdering:** Forældet som generel nutidig påstand.

**Kort svar:** Sammenligningen kan spores til et tidligt estimat, hvor en ChatGPT-forespørgsel blev sat til cirka 2,9 Wh og en traditionel Google-søgning til cirka 0,3 Wh. Tallene beskrev bestemte antagelser og en bestemt teknologisk periode. De kan ikke bruges som en fast omregningsfaktor for moderne AI-tjenester.

Hvor kommer påstanden fra? EPRI's rapport Powering Intelligence fra 2024 gengav de to værdier og beskrev generative AI-forespørgsler som omtrent ti gange mere elkrævende end traditionelle Google-søgninger.

Hvad siger nyere evidens? Google estimerede i 2025 en median Gemini Apps-tekstprompt til 0,24 Wh med en bred driftsgrænse. IEA beskrev i 2026 samtidig meget hurtige effektivitetsforbedringer for simple AI-opgaver, mens reasoning, video og agentiske opgaver kan ligge langt højere.

**Det afgørende forbehold:** Der findes fortsat ingen universel energifaktor for hverken en AI-prompt eller en Google-søgning. En retvisende sammenligning kræver samme funktionelle opgave, kendt systemgrænse og sammenlignelige målemetoder.

## Et lille tal kan stadig blive et stort system

At energien pr. simpel tekstprompt falder, gør målingen relevant, men den siger ikke alene noget om AI's samlede energiefterspørgsel.

Hvis en tjeneste bliver ti gange mere energieffektiv pr. opgave og samtidig bruges hundrede gange så meget, vokser det samlede energiforbrug stadig. Nye anvendelser kan desuden være langt mere

beregningstunge end de tekstprompts, som de laveste per-request-tal typisk beskriver.

Det er en klassisk forskel mellem effektivitet og samlet forbrug.

For beslutningstagere er begge størrelser relevante. Effektivitet pr. opgave fortæller noget om teknologiens udvikling. Det samlede forbrug fortæller noget om belastningen på energisystem, datacentre og infrastruktur.

De to målinger kan bevæge sig i hver sin retning på samme tid.

### Vandglasset følger den samme logik

Den ofte citerede påstand om, at en AI-prompt bruger et glas vand, bygger på en tilsvarende forenkling.

Li, Yang, Islam og Ren estimerede i studiet Making AI Less "Thirsty" – senere publiceret i Communications of the ACM – at en 500 ml flaske vand under deres GPT-3-scenarie svarede til omtrent 10-50 mellemstore responses afhængigt af tid og sted. Senere forskning fra Lawrence Berkeley National Laboratory (Lei et al., 2025) finder mere end 10.000-fold variation i workload-relateret vandforbrug på tværs af kombinationer af servereffektivitet, elproduktion, køling, klima, udnyttelse og andre forhold. Det understreger, hvorfor et fast "vand pr. prompt"-tal kræver en meget præcis afgrænsning.

Da historien senere blev gengivet som "et glas vand pr. prompt", forsvandt både intervallet og systemgrænsen.

Vi behandler vandpåstanden fuldt i kapitel 10. Her fungerer den som endnu et eksempel på, hvordan et konkret forskningsresultat kan ændre betydning, når den metodiske kontekst falder bort.

### Per-prompt-tal har stadig en funktion

Kritikken af universelle prompttal bør ikke føre til den modsatte fejl, hvor måleenheden afvises helt.

Målinger pr. prompt eller request kan være meget nyttige inden for en tydeligt defineret ramme. En leverandør kan følge, om den samme tjeneste bliver mere energieffektiv over tid. En virksomhed kan sammenligne to modeller på den samme opgave. En forsker kan undersøge effekten af outputlængde, hardware eller batching. En systemoperatør kan bruge per-request-data til at skalere en kendt workload til et større trafikniveau.

Værdien ligger i den kontrollerede sammenligning.

Jo længere tallet flyttes væk fra den oprindelige model, opgave, periode og systemgrænse, desto større bliver risikoen for, at det mister sin forklaringskraft.

Et forsvarligt per-prompt-tal bør derfor ledsages af nok metadata til, at andre kan forstå, hvad det repræsenterer. For AI-energi vil det som minimum typisk være tjeneste eller model, opgavetype, periode og systemgrænse samt en markering af, om tallet er målt, estimeret eller modelleret. Klima- og vandtal kræver yderligere oplysninger om geografi og metode.

### Fra én besked til et helt workflow

I praktisk AI-brug er én prompt ofte kun begyndelsen.

Et første svar kan være utilstrækkeligt. Brugeren præciserer spørgsmålet, tilføjer et dokument eller beder om en ny version. Et system kan selv kalde modellen flere gange. En agent kan hente data, kontrollere et resultat og prøve igen. Mennesker kan efterfølgende bruge tid på verifikation og korrektion.

En analyse, der kun tæller den første modelkørsel, kan derfor være teknisk korrekt og samtidig utilstrækkelig som mål for den samlede opgave.

Det bliver især vigtigt, når modeller sammenlignes økonomisk. En billig og energieffektiv model pr. kald kan være en god løsning på simple

opgaver. På en mere krævende opgave kan flere mislykkede forsøg, ekstra modelkald og længere reviewtid ændre regnestykket.

Derfor vender bogen senere tilbage til den funktionelle enhed: det resultat, der faktisk opfylder det nødvendige kvalitetskrav.

## Det præcise spørgsmål før det præcise tal

Når en rapport, pressehistorie eller politisk debat præsenterer et tal for energien eller vandet i én AI-prompt, er det værd at undersøge tallets ledsagende oplysninger, før det bruges videre.

Hvilken model eller tjeneste blev målt? Hvilken type opgave? Hvor langt var svaret? Hvilket år eller hvilken måned? Hvilken hardware og hvilken serving-arkitektur? Var idle capacity, CPU, RAM og datacenter-overhead med? Er tallet observeret direkte eller beregnet ud fra andre størrelser? Hvilken geografi gælder klima- eller vandberegningen for?

Jo flere af disse oplysninger der mangler, desto mindre sikkert er det at flytte tallet til en anden sammenhæng.

Den enkelte prompt er derfor ikke værdiløs som måleenhed. Den er blot en langt dårligere universel enhed, end den umiddelbart ser ud til at være.

Et præcist svar kræver først et præcist spørgsmål.

I næste kapitel går vi et skridt længere og ser på det værktøj, der afgør, hvilke poster der overhovedet kommer med i regnskabet: systemgrænsen.

### Kilder nævnt i kapitlet


- Google (2025): Measuring the environmental impact of AI inference
- EPRI (2024): Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption
- IEA (2026): Key Questions on Energy and AI
- Li et al. (2025): Making AI Less “Thirsty”
- Lei et al. / Lawrence Berkeley National Laboratory (2025): The water use of data center workloads

## Kapitel 3

# Systemgrænsen bestemmer svaret

I kapitel 2 så vi, at den samme Gemini-tekstprompt kunne få et energital på 0,10 eller 0,24 watt-timer, alt efter hvilke dele af driften der blev regnet med.

Forskellen opstod ikke, fordi modellen pludselig udførte mere arbejde. Den opstod, fordi analysen tegnede grænsen omkring systemet forskellige steder.

Den grænse kaldes systemgrænsen, og den er et af de vigtigste begreber i denne bog.

Når et tal om AI skal bruges i journalistik, politik, virksomhedsbeslutninger eller offentlig debat, er det sjældent nok at kende selve tallet. Man skal også vide, hvad der ligger inden for regnskabet, og hvad der ligger udenfor.

### Tegn en kasse omkring det, du måler

En systemgrænse kan forstås ganske konkret. Forestil dig, at du tegner en kasse rundt om den del af AI-systemet, du ønsker at undersøge.

Den mindste kasse kan ligge omkring selve acceleratorens beregningsarbejde. En større kasse kan omfatte hele serveren med CPU, hukommelse, storage og netværksudstyr. Udvider vi grænsen igen, kommer køling, strømforsyning og øvrig datacenterdrift med.

En livscyklusanalyse kan gå endnu længere og medtage fremstillingen af chips og servere, bygningen, den elektriske infrastruktur, udskiftning af hardware og behandling efter endt brug.

ITU-T L.1801 (2026) formaliserer samme princip: systemgrænsen skal passe til analysens formål og den funktionelle enhed og skal gøre det

tydeligt, hvilke livscyklusfaser og processer der er medtaget. Den regel er central, når resultater fra forskellige AI-studier skal sammenlignes.

Ingen af disse grænser er automatisk den rigtige i alle situationer. Valget afhænger af spørgsmålet.

Vil en leverandør optimere energien i selve modelkørslen, kan en relativt smal teknisk grænse være nyttig. Skal en virksomhed vurdere den samlede driftsomkostning ved en løsning, må flere poster med. Skal en myndighed forstå konsekvenserne af stor ny datacenterkapacitet, bliver elnet, produktion, lokalitet og langsigtede investeringer relevante.

Problemerne opstår, når et resultat fra én systemgrænse bliver præsenteret, som om det besvarer et spørgsmål fra en anden.

### AI-workload, AI-datacenter og datacentersektor er ikke det samme

En af de mest almindelige sammenblandinger i AI-debatten opstår mellem AI og datacentre.

Datacentre er den fysiske infrastruktur, hvor en stor del af moderne AI trænes og anvendes. De driver samtidig en lang række andre digitale tjenester: cloudsystemer, databaser, streaming, almindelig webhosting, virksomhedsapplikationer, lagring og klassiske beregningsopgaver.

Derfor er et tal for hele datacentersektoren ikke automatisk et tal for AI.

Det er nyttigt at skelne mellem mindst fire niveauer. Det første er den konkrete AI-workload: træning eller inference for en bestemt model og opgave. Det næste er den server eller acceleratorplatform, der udfører arbejdet. Derefter kommer et AI-fokuseret eller AI-optimeret datacenter, hvor en stor del af kapaciteten er bygget til AI. Det fjerde niveau er hele datacentersektoren, som også rummer betydelige ikke-AI-workloads.

De fire niveauer kan alle optræde i seriøse analyser. De må bare ikke få samme etiket.

### Et regneeksempel: samme datacenter, fire forskellige tal

Forskellen mellem systemgrænser bliver tydeligere med et konkret regnestykke. Eksemplet nedenfor er illustrativt og bruger konstruerede tal. Det viser metoden og må ikke læses som data for et bestemt datacenter.

Antag et datacenter med et målt årligt elforbrug på 120 GWh og en PUE på 1,20. PUE er forholdet mellem datacenterets samlede energiforbrug og energien til IT-udstyret. IT-forbruget kan derfor beregnes som 120 GWh / 1,20 = 100 GWh.

Antag videre, at dokumenterede AI-workloads står for 60 GWh af IT-forbruget, og at acceleratorerne inden for disse workloads står for 45 GWh. Datacenterets øvrige 20 GWh går til køling, strømkonvertering og anden støtteinfrastruktur. Hvis denne overhead fordeles proportionalt efter IT-forbruget, tilskrives AI 60 procent af de 20 GWh, altså 12 GWh. Den bredere AI-driftsgrænse bliver dermed 60 + 12 = 72 GWh.

| Systemgrænse | Hvad tælles med? | Beregning | Årligt elforbrug |
|---|---|---|---|
| Accelerator | Kun acceleratorernes energiforbrug | Givet i eksemplet | 45 GWh |
| AI-workload/server | AI-servernes samlede IT-forbrug | Givet i eksemplet | 60 GWh |
| AI + allokeret datacenter-overhead | AI-IT + 60 % af 20 GWh overhead | 60 + 12 | 72 GWh |
| Hele datacenteret | AI og ikke-AI samt al overhead | Målt siteforbrug | 120 GWh |

Tabel 3.1 – Illustrativ beregning. Tallene er konstruerede og bruges kun til at demonstrere betydningen af systemgrænsen.

*Kilde: Illustrativ beregning af forfatteren. PUE anvendes som forholdet mellem samlet datacenterenergi og IT-energi, i tråd med den almindelige datacenterdefinition anvendt bl.a. af IEA.*

Hvis hele siteforbruget på 120 GWh i dette eksempel blev omtalt som AI-forbrug, ville AI få tilskrevet 48 GWh mere end den bredere, proportionelt allokerede AI-driftsgrænse på 72 GWh. Det svarer til cirka 67 procent højere energiforbrug. Samtidig ville et accelerator-only-tal på

45 GWh undervurdere den samme brede driftsgrænse med 27 GWh. Ingen af tallene er i sig selv meningsløse; de beskriver forskellige dele af systemet.

Eksemplet viser også, at selve allokeringen er en metodebeslutning. Proportional fordeling af overhead er én mulig metode, men i virkelige analyser kan mere direkte målinger eller andre kausale fordelingsnøgler være bedre. Metoden skal derfor fremgå sammen med tallet.

IEA's Key Questions on Energy and AI (2026) estimerer det globale datacenterforbrug til omkring 485 TWh i 2025 og fremskriver cirka 950 TWh i 2030. I samme fremskrivning vokser elforbruget i AI-fokuserede datacentre langt hurtigere og omtrent tredobles fra 2025 til 2030. Forskellen er præcis den sondring, dette kapitel kræver: sektorens samlede elforbrug og den AI-fokuserede del er relaterede størrelser, men ikke den samme størrelse.

### Påstanden under lup

**Påstand:** AI bruger lige så meget strøm som Japan.

**Vurdering:** Misvisende som generel påstand.

**Kort svar:** IEA sammenlignede i 2025 en fremskrivning af hele verdens datacentre i 2030 med Japans nuværende elforbrug. Det var en skalaillustration for datacentersektoren, ikke en måling af AI alene og heller ikke et udsagn om AI-forbruget i dag.

Hvor kommer påstanden fra? IEA's Energy and AI fra 2025 fremskrev globalt datacenterforbrug til omkring 945 TWh i 2030, lidt mere end Japans samlede elforbrug på sammenligningstidspunktet. IEA beskrev AI som den vigtigste driver af væksten, men rapporten omfattede alle datacentre.

Hvad siger nyere evidens? IEA's 2026-opdatering fastholdt størrelsesordenen med cirka 950 TWh for alle datacentre i 2030. Samtidig forventes elforbruget i AI-fokuserede datacentre at vokse betydeligt hurtigere og omtrent tredobles fra 2025 til 2030.

**Det afgørende forbehold:** Landesammenligningen siger noget om størrelsesorden. Den må ikke omskrives til, at AI alene bruger samme mængde strøm som Japan, og prognosens årstal skal følge med, når tallet citeres.

## Et tal kan være korrekt og alligevel få den forkerte etiket

Det afgørende er derfor ikke kun, om et tal er regnet korrekt. Kategorien skal også være korrekt.

Hvis en rapport måler hele et datacenters strømforbrug, er resultatet et datacentertal. Hvis forskeren derefter kan dokumentere, hvor stor en del af belastningen der skyldes en bestemt AI-workload, kan denne del attribueres til AI. Uden den attribution bliver oversættelsen fra datacenter til AI en antagelse.

Det samme gælder i den anden retning. Et målt acceleratorforbrug er et vigtigt datapunkt, men det beskriver ikke automatisk hele serverens eller datacenterets energiforbrug.

Denne præcision kan virke teknisk, men den har direkte betydning for beslutninger. En politiker, der vurderer behovet for netkapacitet, har brug for en anden systemgrænse end en softwareudvikler, der sammenligner to modeller. En journalist, der videreformidler et nationalt eller globalt tal, bør kunne se, om tallet gælder AI, AI-fokuserede datacentre eller datacentre generelt.

### Målt, estimeret, modelleret og fremskrevet

Systemgrænsen fortæller, hvad analysen omfatter. Den næste oplysning er, hvilken type evidens tallet bygger på.

Et målt tal bygger på en fysisk observation. Det kan være en elmåler, en vandmåler, servertelemetri eller en anden direkte registrering.

Et estimeret tal er beregnet ud fra observerede inputs. Man kan f.eks. kende antallet af servere, deres belastning og en dokumenteret effektprofil og derfra estimere energiforbruget.

Et modelleret resultat afhænger i højere grad af en modelstruktur og dens antagelser. Det kan være en livscyklusmodel, hvor produktionsdata, levetid, udnyttelsesgrad og andre parametre sættes sammen til et samlet resultat.

En fremskrivning beskriver fremtiden under bestemte forudsætninger. Prognoser for datacenterforbrug i 2030 er derfor ikke observationer af et forbrug, der allerede findes. De er scenarier eller centrale projektioner baseret på den viden, der var tilgængelig, da analysen blev lavet.

Alle fire typer evidens kan være værdifulde. De giver læseren forskellige grader af direkte observation og forskellige typer usikkerhed.

IEA's Energy and AI (2025) er et godt eksempel: rapporten kombinerer historiske data med scenarier og modeller for 2030. Kilden kan derfor være meget stærk uden at alle dens fremtidstal er observationer. Den korrekte betegnelse – måling, estimat, model eller fremskrivning – er en del af selve dokumentationen.

For beslutningstagere er mærkningen vigtig. En prognose fra en anerkendt myndighed kan være det bedste tilgængelige grundlag for langsigtet planlægning, men den bør stadig omtales som en prognose. Et leverandørmålt produktionsresultat kan være mere direkte observeret, men have en smallere systemgrænse og begrænset generaliserbarhed.

### Attribution og konsekvens er to forskellige spørgsmål

Når AI's omkostninger skal vurderes, opstår endnu en sondring: Vil vi fordele belastningen i et eksisterende system, eller vil vi undersøge, hvad en ændring i AI-efterspørgslen medfører?

Det første er et attributionelt spørgsmål. Her forsøger analysen at fordele et eksisterende energiforbrug, en emission eller en investering mellem aktiviteter. Hvor stor en del af et datacenters elforbrug kan f.eks. knyttes til AI-workloads?

Det andet er et konsekvensspørgsmål. Her undersøger vi, hvad der sker, hvis efterspørgslen ændres. Hvis en ny AI-installation kræver 100 MW yderligere kapacitet, kan effekten være nye netinvesteringer, ændret elproduktion eller andre systemændringer.

De to analyser kan pege på forskellige tal, fordi de besvarer forskellige spørgsmål. Den gennemsnitlige emissionsintensitet i et elnet kan være relevant for ét regnskab, mens den marginale eller langsigtede systemeffekt kan være mere relevant for en beslutning om ny stor belastning.

Denne forskel bliver udfoldet i kapitel 9 om elektricitet og kapitel 15 om infrastruktur. Her er det nok at fastholde, at fordeling af et eksisterende

regnskab og konsekvensen af en ny beslutning ikke bør blandes sammen.

### Geografi og tidspunkt kan ændre resultatet

En systemgrænse har også en geografisk og tidsmæssig side.

En kilowatt-time er samme mængde energi i Danmark, Irland, Texas og Singapore. Klimaeffekten af den anvendte elektricitet kan være forskellig, og den marginale effekt af en ny belastning kan variere endnu mere.

Vand gør lokaliteten endnu tydeligere. Den samme mængde vand kan være en begrænset lokal belastning ét sted og en væsentlig ressourcekonflikt et andet sted. Årstiden kan ændre vurderingen yderligere.

Derfor kan amerikanske datacentertal være relevante som amerikansk evidens uden at være en direkte proxy for europæiske forhold. Når resultatet afhænger af elmix, vandstress, tarifstruktur eller netforhold, må den geografiske kontekst følge med.

Bogen har derfor europæiske forhold som sit primære referencepunkt. Global forskning og data fra USA, Asien og andre regioner bruges, når de tilfører relevant metode eller empirisk dokumentation. Når geografi påvirker resultatet – f.eks. gennem elmix, vandstress, netforhold eller regulering – overføres globale tal ikke direkte til Europa uden en særskilt begrundelse.

Tidspunktet er også vigtigt, fordi AI-teknologien ændrer sig hurtigt. Hardwaregenerationer, modelarkitektur, serving, energisystem og regulering kan flytte sig på få år. Et korrekt 2024-tal kan være historisk relevant i 2026 uden at være en passende nutidig standardværdi.

### Den funktionelle enhed: hvad er det, vi sammenligner?

Systemgrænsen fortæller, hvad vi tæller med. Den funktionelle enhed fortæller, hvad de sammenlignede systemer skal levere.

Det bliver vigtigt, når to AI-løsninger skal vurderes mod hinanden. En lille lokal model kan bruge mindre energi end en stor cloudmodel, men hvis den ikke kan løse den krævede opgave med tilstrækkelig kvalitet, er de to energital ikke en fuld sammenligning af samme funktion.

Det samme gælder økonomisk. En model, der koster mindre pr. kald, kan kræve flere forsøg og mere menneskelig kontrol. En dyrere model kan i en konkret arbejdsproces levere det accepterede resultat første gang.

En retvisende sammenligning kræver derfor en fælles opgave og et defineret kvalitetsniveau. I livscyklusanalyser kaldes dette ofte den funktionelle enhed. Senere i bogen bliver princippet oversat til virksomhedens arbejdsproces gennem begrebet cost per accepted outcome - omkostningen pr. resultat, der faktisk opfylder kravene.

### Et lille kontrolværktøj til store tal

Når et AI-tal dukker op i en rapport, en pressemeddelelse, et debatindlæg eller en politisk diskussion, kan få kontrolspørgsmål afsløre, hvor meget tallet egentlig kan bære.

#### Når du møder et AI-omkostningstal

**1.** Hvad er måleenheden - energi, vand, $CO_2$, penge, materialer eller noget andet?

**2.** Hvad ligger inden for systemgrænsen, og hvad er udeladt?

**3.** Gælder tallet en AI-workload, AI-fokuseret infrastruktur eller hele datacentersektoren?

**4.** Er resultatet målt, estimeret, modelleret eller fremskrevet?

**5.** Hvilken model, opgave, hardware og periode vedrører det?

**6.** Hvilken geografi gælder resultatet for?

**7.** Hvad er sammenligningsgrundlaget eller den kontrafaktiske baseline?

**8.** Hvilke usikkerheder og antagelser kan ændre konklusionen?

Spørgsmålene gør ikke komplekse analyser enkle. De gør det lettere at se, om et tal bliver brugt uden for den sammenhæng, hvor det blev beregnet.

Det er især værdifuldt i AI-debatten, fordi præcise tal ofte cirkulerer hurtigere end deres metodebeskrivelser.

### Fra metode til den fysiske regning

De første tre kapitler har etableret bogens metode. En AI-omkostning skal have en tydelig kategori, en synlig systemgrænse og en klar beskrivelse af, om resultatet er målt, estimeret, modelleret eller fremskrevet. Sammenligninger skal så vidt muligt gælde den samme funktion, og geografi, tidspunkt og usikkerhed skal følge med, når de påvirker resultatet.

Det er de regler, resten af bogen arbejder efter.

Nu kan vi bevæge os fra målemetoden til den fysiske virkelighed bag AI.

Før en model kan trænes, og før den første prompt kan sendes, skal der udvindes og forarbejdes materialer, fremstilles chips og hukommelse, bygges servere og etableres strøm- og datacenterinfrastruktur.

I næste del følger vi regningen tilbage til begyndelsen af denne kæde.

### Kilder nævnt i kapitlet


- ITU-T L.1801 (2026): Guidelines for assessing the environmental impact of artificial intelligence systems
- IEA (2025): Energy and AI
- IEA (2026): Key Questions on Energy and AI
- Google (2025): Measuring the environmental impact of AI inference

## Del 2

# Regningen før den første prompt

Før AI kan bruge elektricitet, skal den fysiske kapacitet eksistere. Det kræver råstoffer, forarbejdning, halvledere, servere, køling og elektrisk infrastruktur. Del 2 følger denne regning baglæns gennem systemet.

## Kapitel 4

# Råstofferne bag AI

Når AI omtales som en digital teknologi, er det let at glemme, hvor fysisk infrastrukturen er.

Modellerne består af software, men den beregningskapacitet, der træner og driver dem, er bygget af materialer. Silicium indgår i halvlederne. Kobber og aluminium leder strøm og varme. Stål og beton indgår i bygninger og tekniske installationer. Batterier kan understøtte backup og fleksibilitet, mens transformere, kabler og netforbindelser gør det muligt at levere store mængder elektricitet til datacenteret.

Det gør råstofsporet bredere end spørgsmålet om, hvilke metaller der findes i en GPU.

IEA fremhæver i Energy and AI fra 2025, at datacenterudbygning trækker på både bulk-materialer som stål og beton og på kobber, aluminium, silicium, gallium, sjældne jordarter og batterimineraler. US Geological Survey, USGS, kortlægger tilsvarende mineralerne på tværs af semiconductors, varmehåndtering, magneter, storage og den øvrige datacenterinfrastruktur.

Det første analytiske valg er derfor det samme, som vi mødte i kapitel 3: Hvor langt ud i systemet går materialeregnskabet?

### Tre spørgsmål, der bør holdes adskilt

Råstofdebatten bliver hurtigt uklar, fordi tre forskellige spørgsmål ofte blandes sammen.

Det første handler om fysisk masse: Hvilke materialer bruges der meget af? Her kan kobber, aluminium, stål og beton fylde langt mere end de specialmaterialer, der typisk får størst opmærksomhed.

Det andet handler om forsyningsrisiko: Hvilke materialer er vanskelige at skaffe, fordi produktion eller raffinering er koncentreret hos få lande eller virksomheder, fordi markedet er lille, eller fordi materialet produceres som biprodukt af et andet råstof? Her kan et materiale være strategisk vigtigt, selv om der kun bruges små mængder.

Det tredje handler om miljø- og samfundspåvirkning. Udvinding og raffinering kan kræve energi og vand og kan påvirke arealer, lokalmiljø, arbejdsvilkår og lokalsamfund. Størrelsen afhænger af det konkrete mineral, malmens kvalitet, processen, energimixet, vandforholdene og den lokale regulering.

En tonnageanalyse, en forsyningssikkerhedsanalyse og en miljøanalyse kan derfor rangordne de samme materialer forskelligt, fordi de måler forskellige risici.

### Fra chip til elnet

USGS' oversigt over datacentermineraler illustrerer, hvor bred materialekæden er. Silicium og specialmaterialer indgår i semiconductors og elektronik. Kobber og aluminium findes i elektriske forbindelser, varmehåndtering og strømforsyning. Sjældne jordarter kan indgå i permanente magneter til motorer, ventilatorer og pumper. Batterier og energilagring kan trække på lithium, nikkel, kobolt, grafit eller andre materialer afhængigt af kemien.

Når systemgrænsen udvides til selve datacenteret og den nødvendige elinfrastruktur, kommer yderligere materialer til. Kabler, busbars, transformere, substations og transmissions- og distributionsnet kan kræve betydelige mængder kobber og aluminium. Bygningen og de mekaniske systemer tilføjer blandt andet stål og beton.

Tabel 4.1 viser derfor ikke en opskrift på et bestemt datacenter. Den viser, hvor forskellige materialetyper typisk optræder i den samlede kæde, og hvilken type risiko de især kan repræsentere.

**Tabel 4.1 – Materialer i forskellige dele af AI-infrastrukturen**

| Systemlag | Eksempler på materialer | Typisk funktion | Analytisk pointe |
|---|---|---|---|
| Halvledere og hukommelse | Silicium; gallium; germanium; tantalum; hafnium m.fl. | Logic, memory, power- og specialkomponenter | Små mængder kan være funktionelt kritiske; proces- og raffineringsevne kan være vigtigere end rå geologisk forekomst. |
| Server og elektronik | Kobber; aluminium; silicium; stål; specialmetaller | Printkort, forbindelser, chassis, heat sinks og strømfordeling | Materialefodaftrykket er bredere end GPU-dien. |
| Køling og mekanik | Kobber; aluminium; stål; sjældne jordarter | Heat exchangers, rør, pumper, ventilatorer og motorer | Materialebehov afhænger af køleteknologi og design. |
| Strøm på sitet | Kobber; aluminium; transformer-stål; batterimineraler | Busbars, UPS, transformere, backup og evt. lagring | Power delivery kan være en stor del af materialemassen. |
| Net og bygning | Kobber; aluminium; stål; beton | Substations, kabler, transmission/distribution og bygning | AI-attribution kræver et but-for-spørgsmål: hvilken ekstra infrastruktur blev faktisk udløst af belastningen? |

*Kilde/ramme: Sammenstillet af forfatteren på baggrund af IEA, Energy and AI – AI and energy security (2025), USGS, Key Minerals in Data Centers (2025/2026), samt Amoah et al. (2026). Tabellen viser funktioner og systemlag, ikke materialemængder i et bestemt datacenter.*

## Kobber kan fylde mere end de sjældne materialer

Et peer-reviewed studie af Macdonald Amoah og medforfattere, offentliggjort i Resources Policy i 2026, gør systemgrænseproblemet meget konkret. Forskerne opstiller en bottom-up-model for 20 mineraler og følger materialerne gennem compute, thermal management og power- og gridinfrastruktur i AI-datacentre.

I studiets referencecases udgør kobber omkring 82–83 procent af den modellerede mineralmasse. Omkring 64 procent af den modellerede kobberefterspørgsel ligger i transmission og distribution af elektricitet.

Tallene er markante, men de skal læses som det, de er: resultater fra en bestemt model med bestemte antagelser om kapacitetsvækst, elnet,

batterilagring og teknologivalg. De er hverken en universel materialefaktor for AI eller en europæisk standardværdi.

Den robuste pointe ligger i strukturen. Når power delivery og netinfrastruktur kommer med i regnskabet, kan bulk-materialer i den elektriske kæde fylde mere fysisk end de mere eksotiske materialer i semiconductors.

### Et regneeksempel: normaliser studiets model til 1.000 enheder

Studiets procentandele kan normaliseres for at vise størrelsesforholdet uden at foregive, at vi kender materialemassen i et konkret datacenter. Regnestykket nedenfor bruger 1.000 fiktive masseenheder. Det ændrer ikke studiets procenter og må ikke læses som 1.000 ton i et faktisk anlæg.

**Tabel 4.2 – Illustrativ normalisering af Amoah et al.s modelresultat**

| Størrelse | Studiets andel | Normaliseret til 1.000 enheder | Beregning |
|---|---|---|---|
| Samlet modelleret mineralmasse | 100 % | 1.000 | Referenceenhed |
| Kobber | ca. 82–83 % | ca. 820–830 | 1.000 × 0,82–0,83 |
| Øvrige mineraler samlet | ca. 17–18 % | ca. 170–180 | 1.000 minus kobber |
| Kobber i transmission/distribution | ca. 64 % af kobberet | ca. 525–531 | 820–830 × 0,64 |
| T&D-kobber som andel af totalen | afledt | ca. 52,5–53,1 % | 525–531 / 1.000 |

*Kilde: Forfatterens beregning ud fra Amoah et al. (2026), som rapporterer ca. 82–83 % kobber af den modellerede mineralmasse og ca. 64 % af kobberefterspørgslen i transmission og distribution. Tabellen er en matematisk normalisering – ikke en måling af et konkret datacenter og ikke en materialekoefficient, der kan overføres direkte til Europa.*

Normaliseringen viser, hvorfor en meget smal fortælling om "sjældne metaller i GPU'en" kan give et skævt billede af den fysiske masse. I denne model svarer kobberet i transmission og distribution alene til lidt over halvdelen af den samlede modellerede mineralmasse.

Det er samtidig vigtigt at fastholde attributionen. Et nyt datacenter kan udløse en ny transformer, substation eller netforstærkning, men et delt

elnet har ofte lang levetid og mange brugere. Hele nettets materiale kan derfor ikke automatisk skrives på AI's konto. Den relevante andel afhænger af, hvilken ekstra kapacitet AI-belastningen faktisk udløser, og hvordan den efterfølgende anvendes.

### Kritisk er ikke det samme som sjælden

Begrebet kritiske mineraler bliver ofte læst, som om det betyder mineraler, der næsten er ved at slippe op. Det er en upræcis fortolkning.

USGS definerer i sin amerikanske 2025-metode kritikalitet ud fra materialets væsentlige funktion og sårbarheden i forsyningskæden. Andre jurisdiktioner kan bruge andre lister og vægtninger, men princippet er det samme: kritikalitet handler om konsekvensen af en forstyrrelse og mulighederne for at erstatte eller skaffe materialet – ikke blot om geologisk sjældenhed.

Kobber er et godt eksempel. Det er et stort globalt råvaremarked og et almindeligt grundstof sammenlignet med flere minor minerals. Alligevel kan kobber blive strategisk vigtigt, fordi elektrificering, netudbygning, datacentre, transport og energiteknologier konkurrerer om den samme forsyningskæde.

IEA's Global Critical Minerals Outlook 2026 projekterer omkring 7 millioner ton ekstra kobberefterspørgsel frem mod 2040 og peger i sit STEPS-forløb på et muligt forsyningsgab omkring 25 procent i 2035 ud fra annoncerede projekter. Det er en fremskrivning for det brede elektrificeringsmarked, ikke et AI-tal. AI-datacentre er én af flere efterspørgselskilder.

Den modsatte mekanisme findes blandt de små markeder. Gallium anvendes i avancerede high-tech-komponenter og produceres ofte som biprodukt. IEA vurderede i Energy and AI, at datacentres galliumefterspørgsel i 2030 kan nå op mod 10 procent af det globale udbud på det daværende niveau, samtidig med at raffineringen er stærkt geografisk koncentreret. I Global Critical Minerals Outlook 2026 angiver

IEA, at Kina står for over 90 procent af raffineringen af blandt andet gallium og magnetiske sjældne jordarter.

Her er den fysiske masse lille sammenlignet med kobber. Forsyningsrisikoen kan alligevel være høj, fordi markedet er koncentreret og mindre fleksibelt.

IEA's offentlige galliumtal gælder datacentre bredt. Kilden specificerer ikke, hvor meget der går direkte til accelerator-GPU'er, og hvor meget der ligger i power electronics, optik, netværk eller andre komponenter. Derfor er formuleringen "AI-GPU'er bruger store mængder gallium" mere præcis, end evidensen kan bære.

**Påstanden under lup**

**Påstand: AI's råstofproblem ligger primært i GPU'en.**

**Vurdering: Misvisende.**

Kort svar: Acceleratorer og hukommelse rummer vigtige og undertiden forsyningskritiske materialer, men den fysiske materialekæde omfatter også servere, køling, bygning, strømforsyning, transformere, kabler og eventuel netudbygning. Når systemgrænsen udvides, kan bulk-materialer som kobber dominere den modellerede masse.

Hvor kommer forestillingen fra? Den avancerede GPU er AI-systemets mest synlige og værdifulde komponent, og diskussioner om chips og kritiske mineraler gør det naturligt at fokusere på selve acceleratorboardet.

Hvad siger evidensen? IEA og USGS kortlægger datacentermaterialer på tværs af IT, køling og strømforsyning. Amoah et al. finder i deres 2026-model, at power- og gridinfrastrukturen kan dominere den samlede mineralmasse, og at kobber alene står for omtrent 82–83 procent i referencecases.

Det afgørende forbehold: Studiets procentandele er modelspecifikke. De viser systemgrænsens betydning, men må ikke bruges som universelle materialekoefficienter for alle AI-datacentre.

## Forsyningsrisiko kan blive en økonomisk omkostning før fysisk mangel

Et materiale behøver ikke være fysisk udsolgt, før en koncentreret forsyningskæde påvirker økonomien.

IEA's 2026-outlook viser f.eks., at priser på gallium og visse tunge sjældne jordarter i Europa i den beskrevne periode lå omkring fem gange de kinesiske hjemmemarkedspriser. Tallet er tidsfølsomt, men mekanismen er vigtig: eksportkontrol, raffinering, handelsbarrierer eller kapacitetsknaphed kan påvirke pris og leveringstid længe før en global geologisk mangel opstår.

For en datacenterudvikler eller hardwareproducent kan forsyningsrisiko derfor optræde som højere priser, længere leveringstider, behov for alternative designs eller forsinkede projekter. For en politisk beslutningstager kan den samme risiko handle om strategisk afhængighed og behov for diversificering, recycling eller ny forarbejdningskapacitet.

Disse økonomiske konsekvenser er heller ikke det samme som materialets miljøpåvirkning. Et materiale med lav fysisk masse kan være strategisk kritisk, mens et bulk-materiale med relativt robust forsyning stadig kan stå for betydelig energi- og arealbelastning i udvinding og forarbejdning.

### Miljøregningen er lokal og procesafhængig

Udvinding og raffinering af råstoffer har ikke én fælles miljøprofil.

IEA's analyse Sustainable and Responsible Critical Mineral Supply Chains fremhæver miljø-, arbejds- og samfundsrisici på tværs af mining, processing og refining. Belastningen varierer med blandt andet malmkvalitet, energiforsyning, vandforhold, kemiske processer, affaldshåndtering og lokal governance.

Det gør det metodisk problematisk at bruge en generel liste over "AI-mineraler" som miljøregnskab. To materialer med samme masse kan have meget forskellige produktionsforløb. Det samme materiale kan have forskellig påvirkning afhængigt af mine, raffinaderi og energimix.

I kapitel 5 går vi tættere på halvlederproduktionen, hvor denne forskel bliver særlig tydelig. Her er rå silicium kun begyndelsen. Ultrapure water, elektricitet, proceskemikalier, yield og avanceret packaging kan være vigtigere for den samlede produktionsbelastning end selve massen af silicium i den færdige chip.

### Kan recycling løse råstofregningen?

Recycling er et vigtigt modtræk, især for materialer med etablerede sekundære markeder. Kobber og aluminium kan i høj grad cirkulere tilbage i materialekæden, og genanvendelse kan reducere behovet for ny udvinding.

IEA's Recycling of Critical Minerals fra 2024 estimerer i et stærkere recycling-forløb, at behovet for nye mineinvesteringer frem mod 2040 kan reduceres med omkring 30 procent. IEA finder også betydeligt lavere drivhusgasudledninger for genanvendt nikkel, kobolt og lithium end for tilsvarende primære materialer i de analyserede kæder.

Det er et vigtigt potentiale, men ikke en universel reduktionsfaktor. Effekten afhænger af materialet, produktets levetid, hvor stor en andel der indsamles, hvor effektiv recovery-processen er, og om den genvundne kvalitet kan erstatte primært materiale i den ønskede anvendelse.

Hurtig vækst skaber samtidig et tidsproblem. Selv med høj genanvendelse kan materialer først komme tilbage i kredsløbet, når de produkter, de sidder i, tages ud af brug. Hvis den installerede kapacitet vokser hurtigere end mængden af udtjent hardware, er der fortsat behov for betydelige mængder primære materialer.

Recycling ændrer derfor råstofregningen, men fjerner den ikke.

### Det relevante beslutningsspørgsmål

For beslutningstagere er råstofsporet mest nyttigt, når spørgsmålet gøres præcist.

Hvis spørgsmålet er fysisk materialemasse, bør analysen favne mere end chips og inkludere de dele af strøm- og datacenterinfrastrukturen, som faktisk er nødvendige for den pågældende kapacitet. Hvis spørgsmålet er forsyningssikkerhed, bør koncentration, raffinering, substitution og markedsstørrelse vægtes. Hvis spørgsmålet er miljøpåvirkning, skal udvindings- og procesforhold vurderes mineral for mineral og sted for sted.

AI konkurrerer desuden om mange af de samme ressourcer som den bredere elektrificering. Kobber bruges også i elnet, vedvarende energi, bygninger, industri og transport. Batterimineraler bruges langt ud over datacentre. Derfor kan et globalt pres på et materiale ikke tilskrives AI alene, selv om AI kan være en voksende efterspørgselskilde.

Råstofregningen skal med andre ord læses som en kæde af fysiske behov og afhængigheder, ikke som en liste over eksotiske grundstoffer.

## Fra råstof til chip

Dette kapitel har udvidet materialebilledet fra acceleratorchippen til det system, der gør AI-kapaciteten mulig. Næste kapitel går den modsatte vej og zoomer ind på den mest teknologisk avancerede del af kæden.

En færdig AI-accelerator vejer relativt lidt sammenlignet med datacenterets samlede infrastruktur. Fremstillingen af dens semiconductors og high-bandwidth memory kan alligevel kræve avancerede fabrikker, store mængder ultrapure water, elektricitet, kemikalier og specialiseret packaging.

Kapitel 5 undersøger derfor ikke kun, hvilke materialer chippen indeholder, men hvad der kræves for at fremstille dem i den kvalitet og kompleksitet, moderne AI-hardware forudsætter.

### Kilder nævnt i kapitlet


• IEA (2025): Energy and AI – AI and energy security
• IEA (2026): Global Critical Minerals Outlook 2026 – Executive summary og Outlook
• USGS (2025/2026): Key Minerals in Data Centers Infographic

• USGS (2025): About the 2025 List of Critical Minerals
• IEA (2024): Recycling of Critical Minerals
• IEA (2023): Sustainable and Responsible Critical Mineral Supply Chains
• Amoah, Brown, Simon, Bazilian & Matisek (2026): Mineral demand from AI data centers: Infrastructure intensity, processing bottlenecks, and supply competition, Resources Policy

## Kapitel 5

# Chippen, HBM og semiconductor-fabrikken

Når en AI-accelerator ankommer til et datacenter, er en stor del af dens fysiske regning allerede afholdt. Før den første modelkørsel har råstoffer været udvundet og raffineret, siliciumwafers er blevet fremstillet og bearbejdet, hukommelse og integrerede kredsløb er produceret, komponenterne er pakket, testet og samlet, og store mængder elektricitet, vand og proceskemikalier har været igennem fabrikkerne.

Det gør halvlederproduktionen til et vigtigt led i AI's omkostningskæde. Samtidig er det et af de vanskeligste led at beskrive med ét enkelt tal. Moderne acceleratorsystemer består af flere avancerede komponenter, og fabrikkernes miljødata offentliggøres ofte på virksomheds- eller fabriksniveau frem for som et fuldt sporbar ressourceforbrug pr. H100-, B200- eller anden konkret AI-chip.

Kapitlets opgave er derfor ikke at opfinde et liter- eller kilowatt-timetal pr. GPU. Målet er at vise, hvilke processer der ligger bag produktet, hvilke belastninger vi kan dokumentere, og hvor produkt-specifikke data stadig mangler.

## En chip er et produktionssystem, ikke bare et stykke silicium

Ordet chip får let produktionen til at lyde enkel. Den færdige komponent er lille, men dens fremstilling består af mange gentagne procestrin med ekstremt høje krav til renhed og præcision.

Wafer fabrication – den del af produktionen, hvor kredsløbene bygges lag for lag på en siliciumwafer – omfatter blandt andet deposition, lithography, etching, cleaning og andre kemiske og termiske processer. Hertil kommer cleanrooms, luftbehandling, køling, vakuum, abatement og behandling af procesvand og spildevand. Efter front-end-produktionen følger packaging, interconnects, test og i avancerede systemer integration af flere chips og hukommelsesstakke.

NIST's programmatic environmental assessment for moderne semiconductor-fabs fra 2024 beskriver netop fabrikken som et system med betydeligt forbrug af elektricitet, ultrapure water, specialgasser, syrer, opløsningsmidler og andre proceskemikalier. NIST-materialet er amerikansk og kan derfor ikke bruges som direkte proxy for en europæisk fab, men det giver en nyttig teknisk kortlægning af, hvor ressourcerne bruges.

Qi Wang og medforfattere sammenstillede i 2023 offentliggjorte 2021-data fra 28 semiconductor-selskaber. I dette datasæt lå det samlede vandudtag på 7,89×10^8 $m^3$, energiforbruget på 1,49×10^11 kWh – svarende til 149 TWh – og de rapporterede drivhusgasudledninger på 7,15×10^7 ton $CO_2e$. Tallene er en sammenstilling af de inkluderede selskabers rapportering og må ikke læses som et komplet globalt industriregnskab. De viser først og fremmest størrelsesordenen på den upstream-industri, som også leverer til AI-hardware.

## Vandet i semiconductor-fabrikken

Vand spiller en særlig rolle, fordi waferprocesser kræver en renhed langt ud over almindeligt drikkevand. Ultrapure water, UPW, fremstilles gennem flere rensningstrin, så ioner, partikler, organisk materiale og

andre urenheder fjernes. Selve rensningen kræver både energi, kemikalier og vand, og en del af det indtagne vand går tabt under behandlingen.

NIST's tekniske syntese fordeler et typisk fab-vandbehov på omtrent 48 procent procesvand, 23 procent køling, 20 procent abatement, 9 procent tab ved UPW-behandling og under 1 procent ikke-industrielt brug. Fordelingen bygger på sekundært citerede branchedata og skal derfor læses som en generel procesfordeling, ikke som en profil for en bestemt fabrik.

**Tabel 5.1 – Illustrativ normalisering af NIST's fab-vandfordeling**

| Anvendelse | NIST-syntese | Hvis samlet vandbehov sættes til 100 $m^3$ |
|---|---|---|
| Procesvand | ca. 48 % | ca. 48 $m^3$ |
| Køling | ca. 23 % | ca. 23 $m^3$ |
| Abatement / emissionskontrol | ca. 20 % | ca. 20 $m^3$ |
| Tab ved UPW-behandling | ca. 9 % | ca. 9 $m^3$ |
| Ikke-industrielt brug | <1 % | <1 $m^3$ |

*Kilde: NIST (2024), teknisk syntese. Tredje kolonne er forfatterens illustrative normalisering. Den beskriver ikke en konkret fab og må ikke bruges som fabriks- eller chipkoefficient.*

Den simple normalisering er nyttig, fordi den viser, at halvledervand ikke kan reduceres til køletårne. Næsten halvdelen af den citerede fordeling ligger i selve procesvandet, mens både køling, emissionskontrol og produktionen af ultrapure water bidrager til resten.

Vandkilden varierer også. Wang et al. fandt i deres 2021-datasæt, at de rapporterende semiconductor-selskabers vandkilder fordelte sig på omkring 47,0 procent overfladevand, 35,3 procent municipal supply, 8,5 procent grundvand, 5,8 procent tredjepartsforsyning og 3,2 procent eksternt reclaimed water. Fordelingen er hverken et nutidigt globalt

gennemsnit eller et AI-specifikt tal. Den understreger, at chipproduktion – ligesom datacenterdrift – kræver information om vandtype og lokalitet, før et volumetal kan fortolkes.

### Et kildekritisk eksempel: når en faktor 10 gemmer sig i en myndighedsrapport

Halvlederdata giver samtidig et konkret eksempel på, hvorfor selv stærke myndighedskilder skal kontrolleres matematisk.

NIST-rapporten angiver median UPW-brug blandt 29 fabs som 4,25 milliarder gallons i 2021. Senere i samme dokument bliver den samme årlige størrelse omregnet til 1,16 millioner gallons om dagen. De to tal kan ikke begge være rigtige: 4,25 milliarder divideret med 365 er cirka 11,64 millioner gallons om dagen. Den daglige omregning er dermed omtrent en faktor 10 for lav. Omregningsfejlen isoleres derfor her, mens de øvrige dokumenterede oplysninger i kilden behandles særskilt.

Eksemplet er vigtigt for resten af bogen. En myndighedsrapport kan være en stærk kilde til metode, procesbeskrivelse og en række andre data, selv om ét tal er forkert. Kildekritik handler derfor om at vurdere påstandsniveau og kontrollere enheder, tidsbasis og størrelsesorden.

### Et europæisk proceseksempel: mindre kemi, strøm og vand

En europæisk case fra imec og KU Leuven viser, hvordan selve procesdesignet kan flytte ressourceforbruget mærkbart. Et peer-reviewed studie fra 2026 analyserede backside wet cleaning med primærdata fra imec's 300 mm pilotlinje i Belgien.

Forskerne sammenlignede en baseline-proces med en alternativ single-wafer-proces baseret på ozonerede kemier. I den analyserede proces faldt det samlede miljøaftryk med 55 procent, elforbruget til cleaning med 67 procent, HF-forbruget med 59 procent og UPW-forbruget med 31 procent. Når processen blev skaleret til en model-facility med 50.000 wafer starts om måneden og 46 backside-clean-trin pr. wafer, gav

beregningen omtrent 4 millioner kWh – 4 GWh – lavere elforbrug og 28 millioner liter lavere tap-water-forbrug om året.

Det er et stærkt datapunkt for procesforbedring, men ikke en H100- eller B200-koefficient. Casen vedrører en N28 logic-model og et bestemt wet-cleaning-forløb. Dens værdi ligger i at vise, at semiconductor-footprint ikke er statisk: procesvalg, kemi og forsyningsenergi kan ændre resultatet betydeligt.

### HBM er ikke en detalje ved siden af GPU'en

Når fokus flyttes fra fabrikken til det færdige acceleratorsystem, bliver high-bandwidth memory, HBM, central. Moderne AI-acceleratorer flytter enorme datamængder mellem beregningsenheder og hukommelse, og HBM er blevet en tæt integreret del af high-end acceleratorsystemet.

NVIDIA offentliggjorde i 2025 tredjepartsreviewede product carbon footprint-sammenfatninger for HGX H100- og HGX B200-baseboards. Begge er cradle-to-gate-analyser: de dækker råmaterialer, komponentproduktion, transport og assembly frem til det færdige baseboard, mens use-phase og end-of-life er udeladt. NVIDIA oplyser samtidig, at PCF-modellerne bruger primær leverandørdata for komponenter, der repræsenterer over 92 procent af H100-produktets vægt og over 90 procent af B200-produktets vægt.

Resultaterne er interessante, fordi hukommelsen er den største enkeltkategori i begge analyser. For HGX H100 udgør HBM 42 procent af cradle-to-gate-aftrykket, mens integrerede kredsløb – en kategori der omfatter GPU'er, netværksprocessorer og andre IC'er – udgør 25 procent. For HGX B200 er de tilsvarende andele 49 og 28 procent.

**Tabel 5.2 – NVIDIA/WSP's cradle-to-gate PCF for HGX H100 og HGX B200**

| Størrelse | HGX H100 | HGX B200 | Analytisk pointe |
|---|---|---|---|
| Produktvægt | 24 kg | 32 kg | Forskellige produktgenerationer og fysisk størrelse |
| Samlet PCF | 1.312 kg $CO_2$e | 2.274 kg $CO_2$e | B200 er 962 kg $CO_2$e højere pr. baseboard |
| HBM / memory | 546 kg / 42 % | 1.123 kg / 49 % | Største komponentkategori i begge PCF'er |
| IC'er inkl. GPU'er | 332 kg / 25 % | 643 kg / 28 % | GPU'en indgår sammen med øvrige IC'er |
| Thermal components | 230 kg / 18 % | 267 kg / 12 % | Absolut aftryk stiger mindre end andelen antyder |
| Materialer + komponenter | 91 % | 94 % | Hovedparten ligger før assembly og drift |
| Assembly | 8,6 % | 5,6 % | Andel af cradle-to-gate PCF |
| Transport | 0,4 % | 0,3 % | Meget lille andel i disse PCF-modeller |

*Kilde: NVIDIA/WSP (2025), Product Carbon Footprint Summary for NVIDIA HGX H100 og HGX B200. Leverandørkommissionerede, tredjepartsreviewede produkt-PCF'er. Use-phase og end-of-life er udeladt.*

**Figur 5.1 – HBM fylder mere end IC-kategorien i begge produkt-PCF'er**

*Kilde: NVIDIA/WSP (2025). "Øvrige" samler de mindre komponentkategorier, så figuren summerer til 100 %. Figuren viser andele af cradle-to-gate klimaaftryk – ikke materialemasse, vandforbrug eller driftsenergi.*

### Et regneeksempel: større absolut produktionsaftryk er ikke samme spørgsmål som dårligere effektivitet

De to PCF'er kan bruges til en enkel, men metodisk vigtig beregning. H100-baseboardet er opgjort til 1.312 kg $CO_2$e cradle-to-gate, mens B200-baseboardet er opgjort til 2.274 kg $CO_2$e. Forskellen er 962 kg $CO_2$e. Divideret med H100-værdien svarer det til cirka 73 procent højere absolut cradle-to-gate-aftryk pr. B200-baseboard.

HBM-posten vokser endnu mere i absolutte tal: fra 546 til 1.123 kg $CO_2$e, en stigning på 577 kg eller cirka 106 procent. IC-kategorien vokser fra 332 til 643 kg, omtrent 94 procent. Thermal components går fra 230 til 267 kg, cirka 16 procent.

Her bør analysen stoppe, før den drager en forkert konklusion. Det højere absolutte produktionsaftryk dokumenterer ikke i sig selv, at B200 er dårligere pr. nyttig AI-opgave. B200 leverer en anden hukommelseskapacitet og compute-performance, og de to PCF'er udelader selve driftsfasen. En funktionel sammenligning kræver derfor et fælles workload, en defineret levetid og samme kvalitetskrav. Det spørgsmål vender vi tilbage til i kapitlerne om levetid, drift og cost per accepted outcome.

**Påstanden under lup**

**Påstand:** Det meste af acceleratorens produktionsaftryk kommer fra selve GPU-chippen.

**Vurdering:** Misvisende som generel påstand.

**Kort svar:** I NVIDIA/WSP's produkt-PCF'er for HGX H100 og HGX B200 er high-bandwidth memory den største enkeltkategori med henholdsvis 42 og 49 procent af cradle-to-gate klimaaftrykket. Kategorien "ICs", som inkluderer GPU'er, netværksprocessorer og andre integrerede kredsløb, udgør 25 og 28 procent.

**Hvad viser det?** Avanceret AI-hardware er et integreret system, hvor hukommelse, packaging, thermals og andre komponenter kan være væsentlige produktionsposter. GPU'en alene er derfor en for smal systemgrænse.

**Det afgørende forbehold:** Tallene gælder to konkrete NVIDIA-baseboards og et klimaaftryk fra cradle-to-gate. De er ikke universelle procentandele for alle acceleratorer og kan ikke omregnes direkte til vandforbrug eller materialemasse.

### Packaging har sin egen regning

Selv når wafer fabrication er afsluttet, er semiconductorens produktionsregning ikke færdig. Packaging beskytter komponenten, etablerer de elektriske forbindelser og gør det muligt at forbinde flere dies, hukommelse og interconnects i avancerede konfigurationer.

Et peer-reviewed LCA-studie i Cleaner Waste Systems fra 2026 undersøgte lead-frame- og flip-chip-packaging med guld- eller kobberbaserede bonding wires. Studiet brugte primærdata fra en stor semiconductor assembly- og testvirksomhed i Taiwan og vurderede både carbon footprint og water scarcity footprint efter ISO-baserede metoder.

Systemgrænsen omfattede packaging og udelod upstream IC-design og midstream IC manufacturing.

Med den valgte funktionelle enhed på 1 $mm^3$ packaging gav LF-Cu den laveste carbon footprint på 4,80×10^-4 kg $CO_2$e, mens FC-Au lå højest med 2,72×10^-3 kg $CO_2$e. Det er omtrent 5,7 gange så højt i netop dette study setup. For volumetrisk vandforbrug var LF-Cu 1,32×10^-3 $m^3$ pr. funktionel enhed mod 3,62×10^-3 $m^3$ for FC-Au – cirka 2,7 gange højere.

Forskellen skyldes både packaging-teknologi, materialevalg og produktionens energibehov. Guld-baserede løsninger blev især belastet af råmaterialetrinnet, mens elektricitet fyldte mere i de kobberbaserede konfigurationer. Det gør packaging til et godt eksempel på, hvorfor et enkelt ord som "chipproduktion" kan skjule flere forskellige hotspots.

## Det samme vandvolumen kan få en anden betydning, når fabrikken flytter

Packaging-studiet giver også en bro til bogens senere vandkapitel. Når forskerne kombinerede vandforbruget med AWARE-faktorer for regional vandknaphed, varierede water scarcity footprint med mere end en faktor 20 mellem de analyserede lokationer. Den samme produktionsproces kunne dermed få meget forskellig lokal vandbetydning afhængigt af, hvor den blev placeret.

Det er vigtigt at skelne mellem volumetrisk vandforbrug og scarcity-weighted impact. Et lavt AWARE-resultat gør ikke literne usynlige, og et højt scarcity-resultat betyder heller ikke, at processen nødvendigvis bruger mere vand i fysisk volumen. De to størrelser besvarer forskellige spørgsmål: hvor meget vand der forbruges, og hvor presset den lokale ressource er.

## Europæisk semiconductorpolitik gør upstream-regningen mere relevant – ikke mindre

Semiconductorproduktionen er samtidig blevet et strategisk industripolitisk spørgsmål i Europa. Europa-Kommissionens Chips Act 2.0

Impact Assessment fra juni 2026 beskriver semiconductors som en strategisk nøgleteknologi og analyserer en styrkelse af Europas semiconductor-økosystem og reduktion af strukturelle afhængigheder.

Mere europæisk produktion kan ændre forsyningsrisici, transportafstande, energimix og lokal regulering. Den fysiske produktionsregning forsvinder dog ikke, fordi fabrikken flytter geografisk. Den ændrer karakter. For en europæisk analyse bliver spørgsmål om elforsyning, vandkilde, lokal vandstress, kemikaliehåndtering og procesinnovation derfor mere – ikke mindre – relevante.

imec-casen illustrerer netop denne pointe fra europæisk produktionsmiljø: procesdesign kan reducere ressourceforbruget, mens geografisk placering og energiforsyning fortsat påvirker det samlede resultat.

### Producentdata viser skalaen i semiconductor-leddet

TSMC's 2025 Sustainability Report giver et sjældent producentblik på energiregningen før acceleratoren når datacenteret. Selskabets samlede globale elforbrug steg fra cirka 25,55 TWh i 2024 til 28,77 TWh i 2025, svarende til cirka 12,6 procent. Omkring 5,78 TWh kom fra vedvarende energi, svarende til 20,1 procent.

Tallene kan ikke kaldes AI-forbrug. TSMC producerer chips til mange markeder, og rapporten giver ikke en transparent allokering til AI-acceleratorer, HBM eller bestemte produktfamilier. Værdien er i stedet, at produktionsleddet nu er bedre dokumenteret med primære producentdata: et stort og voksende elforbrug ligger upstream, før GPU'en tændes.

For denne bog styrker det evidensen for semiconductor-leddet. Selve produktionssystemets energi- og vandkrav er bedre dokumenteret, mens den robuste AI-specifikke allokering fra fab, HBM, packaging og test til én moderne accelerator fortsat mangler.

*Kilde: TSMC, 2025 Sustainability Report, offentliggjort 20. august 2026. Tallene dækker TSMC's samlede globale produktion og må ikke fortolkes som AI-specifikt forbrug.*

## Hvorfor "liter pr. AI-chip" stadig er et usikkert tal

Efter alle disse datapunkter kan det være fristende at samle vandregningen til et enkelt tal pr. accelerator. Det er præcis her, datagrundlaget endnu ikke er stærkt nok.

Et fabriks samlede vandudtag omfatter flere produkter, procesgenerationer og støttefunktioner. For at allokere vand til en konkret accelerator skal man kende waferstørrelse, node, antal procestrin, yield, die-areal, hvor mange gode dies en wafer leverer, HBM-produktion, packaging, test og den valgte allokeringsmetode. Hertil kommer spørgsmålet om, hvor upstream-grænsen sættes.

Produkt-PCF'erne for H100 og B200 kan ikke løse dette hul, fordi de opgør klimaaftryk og ikke et fuldt vandfootprint. Omvendt kan fab-vanddata ikke uden videre omregnes til et specifikt baseboard. Vores hardware-matrix fastholder derfor produkt-specifik embodied water som et datagap.

Den korrekte konklusion er dermed mere afgrænset: semiconductor- og HBM-produktion er dokumenterbart vand-, energi- og procesintensiv, men et universelt liter-tal pr. moderne AI-accelerator er ikke dokumenteret med tilstrækkelig transparens.

## Hvad bør en beslutningstager kræve af et hardwaretal?

Når et produktionsaftryk bruges som beslutningsgrundlag, bør tallet ledsages af sin funktionelle enhed og sin systemgrænse. Er det én GPU-chip, et acceleratorboard, et helt baseboard eller en server? Dækker analysen kun cradle-to-gate, eller også brug og end-of-life? Er HBM, packaging, thermals og networking med?

For vand bør dokumentationen skelne mellem withdrawal og consumption, angive vandkilde og lokation og beskrive, hvordan fab-data er allokeret til produktet. For klima bør det fremgå, om data bygger

på leverandørmålinger, LCA-databaser eller modeller, og hvor stor en del af produktets masse eller forsyningskæde der dækkes af primærdata.

For sammenligning mellem hardwaregenerationer bør både absolut produktionsaftryk og funktionel ydelse være synlige. Et lavere kg $CO_2$e pr. fysisk enhed er én måling. Et lavere kg $CO_2$e pr. leveret beregning over en realistisk levetid er en anden.

Det er samme disciplin, som bogen bruger på promptenergi og datacenterforbrug: en præcis måleenhed er først nyttig, når læseren ved, hvad den repræsenterer.

## Fra produktion til levetid

Kapitel 4 viste, at AI's materialebehov strækker sig fra chippen til datacenteret og elnettet. Dette kapitel har zoomet ind på den avancerede semiconductor-kæde og vist, at en betydelig del af regningen ligger før hardwaren tændes: i wafer fabrication, ultrapure water, kemi, HBM, packaging og komponentproduktion.

Det næste spørgsmål handler om tid. Et højt embodied footprint får forskellig betydning, hvis hardwaren bruges intensivt i mange år, hvis den udskiftes hurtigt, eller hvis den får et nyt liv i en mindre krævende funktion.

Kapitel 6 undersøger derfor forskellen mellem fysisk levetid, teknisk forældelse, økonomisk retirement og regnskabsmæssig useful life – og hvorfor en afskrivningsperiode aldrig bør forveksles med datoen, hvor en GPU bliver til affald.

### Kilder nævnt i kapitlet


• NIST / CHIPS Program Office (2024): Final Programmatic Environmental Assessment for Modernization and Expansion of Existing Semiconductor Fabrication Facilities

• Wang, Qi et al. (2023): Environmental data and facts in the semiconductor manufacturing industry: An unexpected high water and energy consumption situation

• NVIDIA / WSP (2025): Product Carbon Footprint Summary for NVIDIA HGX H100

• NVIDIA / WSP (2025): Product Carbon Footprint Summary for NVIDIA HGX B200

• imec / KU Leuven (2026): Reducing the Environmental Impact of Wet Chemical Processes for Advanced Semiconductor Manufacturing
• Ouattara et al. / Cleaner Waste Systems (2026): Towards cleaner semiconductor packaging: Life cycle assessment of material and location decisions on carbon and water scarcity footprints
• European Commission (2026): SWD(2026) 504 final – Chips Act 2.0 Impact Assessment Report
• TSMC (2026): 2025 Sustainability Report.

## Kapitel 6

# Serveren, levetiden og den teknologiske forældelse

Når en ny generation af AI-hardware kommer på markedet, bliver ældre servere hurtigt omtalt som forældede. Det kan være rigtigt i én betydning og forkert i en anden. En GPU kan fortsat fungere fysisk, selv om den ikke længere er den mest attraktive platform til et bestemt workload. En server kan være fuldt afskrevet i regnskabet og samtidig fortsætte i drift. Udstyr kan også forlade sin primære rolle og få et nyt liv som reserve, i et mindre krævende system eller på et sekundært marked.

For AI's reelle omkostninger er denne sondring vigtig. Levetiden bestemmer, hvor mange års drift fremstillingsbelastningen fordeles over, hvor ofte ny hardware skal produceres, og hvornår energieffektiviteten i en ny generation kan opveje den embodied belastning ved at fremstille den.

Der findes derfor ikke én "GPU-levetid", som kan indsættes ukritisk i alle beregninger. Vi må først være tydelige om, hvilken levetid vi mener.

### Flere ure tikker samtidig

Hardware har flere tidsforløb på én gang. De hænger sammen, men de må ikke behandles som synonymer.

**Tabel 6.1 – Fem forskellige tidspunkter i hardwareens liv**

| Begreb | Hvad afgør det? | Hvad fortæller det ikke? |
|---|---|---|
| Fysisk levetid | Om komponenten fortsat kan fungere med acceptabel fejlrate og vedligeholdelse. | Siger ikke, om den stadig er økonomisk eller teknisk attraktiv. |
| Teknisk relevans | Om hardwaren kan levere den nødvendige performance, hukommelse, interconnect, softwarestøtte og servicekvalitet. | Siger ikke, at udstyret fysisk er slidt op. |
| Økonomisk levetid / primary-service retirement | Om det fortsat kan betale sig at bruge hardwaren i den primære rolle sammenlignet med alternativer. | Siger ikke, at hardwaren bliver til affald ved retirement. |
| Regnskabsmæssig useful life | Den periode virksomheden anvender til afskrivning af aktivets bogførte værdi. | Er et regnskabsmæssigt estimat, ikke en fysisk holdbarhedstest. |
| Second life / end-of-life | Hvad der sker efter primær drift: genplacering, reservedel, videresalg, recycling eller endeligt affald. | Primary retirement og end-of-life kan ligge langt fra hinanden. |

*Kildegrundlag: syntese af ORNL's GPU-reliabilitydata, regulerede årsrapporter og cloudoperatørernes reuse- og circularity-data.*

### Fysisk levetid kan være længere end markedets teknologiske tempo

Oak Ridge National Laboratory fulgte 18.688 NVIDIA K20X-GPU'er i Titan-supercomputeren og analyserede driftsdata gennem en næsten syvårig systemlevetid. Studiet fra SC20 er ældre end den nuværende H100/B200-generation, men det giver noget, der ellers er sjældent i den offentlige debat: flerårige feltdata fra en stor GPU-flåde.

ORNL fandt, at reliability ikke alene var et spørgsmål om kalenderalder. Fejlrisikoen hang sammen med varmeafledning, den konkrete kølearkitektur og brugsmønstre knyttet til job scheduling. Det understøtter et mere præcist syn på fysisk levetid: hardwareens driftstilstand afhænger af miljø, belastning og vedligeholdelse, ikke blot af hvor mange år der er gået siden indkøbet.

Titan-data kan ikke fortælle os, hvor længe en moderne H100 eller B200 typisk bliver i primær hyperscale-drift. K20X-generationen, workloads og driftsmiljøet er anderledes. Til gengæld viser studiet, hvorfor et rundt tal som "tre år" eller "fem år" ikke bør præsenteres som en fysisk naturlov for datacenter-GPU'er.

### Regnskabslevetid er en ledelsesvurdering

Virksomhedernes årsrapporter viser tydeligt, hvor farligt det er at læse depreciation schedules som fysisk levetid. Regnskabsmæssig useful life er ledelsens estimat af den periode, aktivet forventes at skabe økonomisk nytte under virksomhedens aktuelle planer og teknologiske forhold.

Amazon er et særligt illustrativt eksempel. Selskabet forlængede servernes estimerede useful life fra fem til seks år fra 2024, men ændrede allerede fra 2025 en delmængde af server- og netværksudstyr tilbage fra seks til fem år. Amazon begrundede den kortere vurdering med det øgede tempo i teknologiudviklingen, især inden for AI og machine learning. Selskabet rapporterede også accelerated depreciation i forbindelse med tidlig retirement af visse server- og netværksaktiver.

Meta gik samme år i den anden retning og forlængede den estimerede useful life for de fleste server- og network assets til 5,5 år. Microsofts 2025-årsrapport anvender en bredere kategori for computer equipment på to til seks år. Tallene kan ikke sammenlignes som identiske aktivklasser, men netop forskellene viser pointen: regnskabsmæssig levetid er virksomhedsspecifik og kan ændres, når forventninger til brug, teknologi og økonomi ændres.

**Tabel 6.2 – Regnskabsmæssige levetider er ikke et branchetal**

| Virksomhed / kilde | Rapporteret useful life | Hvad eksemplet viser |
|---|---|---|
| Amazon, 2025 Form 10-K | 5–6 år for servers and networking equipment; en delmængde ændret 6 → 5 år fra 2025. | AI/ML-tempo kan forkorte den økonomiske/regnskabsmæssige forventning. |
| Meta, 2025 Form 10-K | 5–5,5 år; de fleste server/network assets forlænget til 5,5 år fra 2025. | Samme teknologiske periode kan føre til en længere regnskabsmæssig vurdering hos en anden operatør. |
| Microsoft, Annual Report 2025 | Computer equipment: 2–6 år. | Bred aktivkategori og virksomhedsspecifik regnskabspraksis; ikke et GPU-lifetime benchmark. |

*Kilder: Amazon.com, Inc. Annual Report 2025 (Form 10-K); Meta Platforms, Inc. Annual Report 2025 (Form 10-K); Microsoft Annual Report 2025.*

## AI gør faste refresh-cyklusser mindre pålidelige

Microsoft Research undersøgte i 2026 AI-datacenterets lifecycle fra bygning og IT-provisionering til drift. Forskerne modellerede refresh-strategier på tværs af GPU-generationer og workloads og fandt, at en fast kalenderregel fungerer dårligt for AI. I deres scenarier var det økonomisk fordelagtigt at skifte tidligt, når en ny generation gav et stort

reelt effektivitetsløft, mens ældre hardware i andre situationer forblev konkurrencedygtig i mere end seks år.

Studiet er en TCO-model, ikke en observation af den faktiske retirement-alder i Microsofts GPU-flåde. Det er derfor ikke dokumentation for, at H100-servere "holder seks år" eller bør gøre det. Værdien ligger i mekanismen: optimal refresh afhænger af modelarkitektur, workload, performance pr. watt, hardwarepris, kapacitetsbehov og den eksisterende datacenterinfrastruktur.

Forskerne viser også, at nyere generationer ikke forbedrer alle workloads lige meget. For mindre modeller kunne ældre A100- og endda V100-hardware i deres test være konkurrencedygtig målt som performance pr. watt pr. dollar, mens større modeller havde langt større fordel af nyere acceleratorer. Dermed bliver "nyere" en utilstrækkelig beslutningsregel; den relevante sammenligning er den konkrete funktion, der skal leveres.

### Virkelige hyperscale-flåder består af flere generationer samtidig

Meta og University of Texas at Austin offentliggjorde ved OSDI 2026 en analyse af power planning baseret på live produktionstrafik fra millioner af servere og flere hardwaregenerationer. Forfatterne beskriver kommercielle hyperscale-datacentre som heterogene flåder, hvor nye servere indsættes for at møde voksende compute-behov, samtidig med at ældre servere beholdes af økonomiske og miljømæssige grunde.

Det er vigtig empirisk triangulering. Moderne datacentre fungerer ikke nødvendigvis som et transportbånd, hvor hele flåden udskiftes, så snart en ny GPU-generation udkommer. Flere generationer kan eksistere parallelt, og workloads kan placeres dér, hvor den konkrete kombination af performance, kapacitet, strøm og økonomi giver mening.

### En A100-kontrakt viser forskellen mellem nyeste generation og økonomisk end-of-life

CoreWeave oplyste på sit Q2 2026 earnings call, at selskabet har indgået en ny kontrakt på Nvidia A100-kapacitet, der løber ind i 2029. A100-

generationen blev lanceret i 2020. Selskabet beskrev samtidig fortsat høj efterspørgsel på både ældre og nyere Nvidia-generationer og mulighed for at genkontrahere ældre Ampere- og Hopper-installationer.

Det er observeret markedsadfærd, ikke en levetidsmodel. Casen viser, at teknologisk forældelse ikke automatisk er det samme som økonomisk end-of-life. En accelerator kan fortsat have kommerciel værdi, selv om nyere hardware leverer bedre performance eller energieffektivitet.

En kontrakt til 2029 dokumenterer dog ikke, at alle konkrete A100-GPU'er har kørt kontinuerligt siden 2020, og den fortæller ikke utilisation, workload eller fysisk fejlrate. Evidensen dækker økonomisk anvendelighed, ikke en industriwide fysisk levetid.

*Kilde: CoreWeave, Q2 2026 earnings call, 11. august 2026. Oplysningen bruges som dokumentation for økonomisk anvendelighed, ikke som en måling af fysisk GPU-levetid.*

### Nyere hardware kan spare strøm og stadig koste mere $CO_2$ at fremstille

Kapitel 5 viste, at en ny acceleratorgeneration kan have et højere absolut cradle-to-gate-aftryk pr. fysisk enhed. Samtidig kan den levere langt mere arbejde pr. watt eller pr. server. Derfor kræver et klimamæssigt refresh-valg et break-even-regnestykke.

Et europæisk peer-reviewed lifecycle-studie af scientific computing centres fra Wadenstein og Vanderbauwhede viser netop, at den optimale replacement-strategi ændrer sig med elnettets emissionsintensitet. I lavemissionssystemer fylder embodied carbon relativt mere, så længere hardwarelevetid kan være fordelagtig. I højere emissionssystemer kan den løbende strømbesparelse fra nyere og mere effektiv hardware hurtigere opveje fremstillingsbelastningen.

Det er endnu et sted, hvor bogens europæiske referencepunkt er afgørende. Det samme hardware-refresh kan få forskelligt $CO_2$-resultat i forskellige elsystemer. Globale hardwaredata kan være de samme, mens use-phase-regningen ændres med geografi og tidspunkt.

## Et illustrativt break-even-regnestykke

Antag, at en ny server medfører 2.000 kg $CO_2$e i ekstra embodied belastning sammenlignet med at fortsætte med den eksisterende hardware. Den nye server leverer samme nyttige service med 4 MWh lavere elforbrug om året. Tallene er konstruerede og bruges kun til at vise mekanismen.

**Tabel 6.3 – Samme hardwarebeslutning, forskellig $CO_2$-break-even**

| Illustrativ emissionsintensitet | Årlig besparelse ved 4 MWh | Tid til at "betale" 2.000 kg $CO_2$e embodied tilbage |
|---|---|---|
| 50 g $CO_2$e/kWh | 200 kg $CO_2$e/år | 10 år |
| 200 g $CO_2$e/kWh | 800 kg $CO_2$e/år | 2,5 år |
| 500 g $CO_2$e/kWh | 2.000 kg $CO_2$e/år | 1 år |

*Illustrativ beregning: break-even = 2.000 kg $CO_2$e / (4.000 kWh/år × emissionsintensitet). Tallene er ikke landefaktorer og beskriver ikke H100/B200 direkte. Formålet er at vise følsomheden over for elproduktionens $CO_2$-intensitet.*

Regneeksemplet viser, hvorfor en kalenderbaseret klimaregel er utilstrækkelig. Ved 50 g $CO_2$e/kWh tager det ti år at opveje den antagne fremstillingsbelastning gennem den årlige elbesparelse. Ved 500 g $CO_2$e/kWh sker det på ét år. Hardwarebeslutningen er den samme; elsystemet ændrer resultatet med en faktor ti.

I en virkelig analyse skal tallene erstattes med produktets faktiske embodied footprint, den målte eller realistisk modellerede energibesparelse for samme workload, datacenterets PUE, den relevante geografiske emissionsfaktor og den forventede resterende levetid. Kapitel 9 går dybere ned i forskellen mellem gennemsnitlig, location-based, market-based og marginal elektricitet.

### Påstanden under lup

**Påstand:** En AI-GPU holder typisk tre til fem år.

**Vurdering:** Ikke dokumenteret som en generel fysisk levetid.

**Kort svar:** Offentlige regnskabskilder bruger ofte useful-life-estimater omkring fem til seks år for server- og netværksaktiver, men disse er økonomiske/regnskabsmæssige vurderinger. ORNL's feltdata viser flerårig fysisk GPU-drift, og nyere TCO-modeller viser, at nogle GPU-generationer kan være økonomisk konkurrencedygtige i mere end seks år. Der findes fortsat ikke en robust offentlig fordeling for faktisk primary-service retirement af moderne H100/B200-lignende acceleratorer.

**Hvor kommer tallet fra?:** Runde 3-, 5- eller 6-årstal optræder ofte i afskrivningsregler, lifecycle-modeller og antagelser i miljøstudier. De beskriver ikke nødvendigvis samme type levetid.

**Hvad siger evidensen?:** Fysisk reliability, teknisk relevans, økonomisk retirement og accounting useful life må behandles som separate størrelser.

**Det afgørende forbehold:** Hvis et studie antager tre års GPU-levetid, er det en modelparameter, medmindre faktisk fleet retirement dokumenteres. Resultater om materialer og embodied impact kan være meget følsomme over for denne antagelse.

#### Retirement er begyndelsen på næste spørgsmål

Når en server forlader sin primære rolle, er den stadig ikke nødvendigvis ved end-of-life. Google rapporterer millioner af komponenter høstet fra decommissioned datacenterhardware til intern genbrug, og Microsoft rapporterer både reuse og recycling gennem sine Circular Centers. Disse leverandørdata fortæller ikke, hvor gamle de enkelte GPU'er var ved retirement, men de dokumenterer, at primary-service retirement og affald ikke er samme hændelse.

Denne forskel er vigtig for lifecycle-analyser. Hvis en accelerator får en reel second-life-funktion, kan en del af værdien fra den oprindelige fremstilling fortsætte i et nyt system. Hvor stor kredit der bør gives, afhænger af den faktiske funktion, levetidsforlængelse og hvilket nyt udstyr der realistisk undgås.

### Hvad bør en beslutningstager kræve af en levetidsantagelse?

Når en AI-rapport bruger en hardwarelevetid, bør det fremgå, hvilken type levetid tallet repræsenterer. Er det en afskrivningsperiode, en planlagt refresh-cycle, en antaget LCA-parameter, observeret physical reliability eller faktisk primary-service retirement?

Dernæst bør analysen vise følsomheden. Hvis resultatet ændrer sig markant ved tre, fem eller syv års brug, er levetiden ikke en detalje; den er en bærende antagelse. I miljøberegninger bør replacement desuden kobles til den regionale elproduktion, fordi embodied og operational emissions vægter forskelligt i forskellige elsystemer.

For moderne acceleratorer er den brede, faktiske industriwide retirement-fordeling fortsat ikke robust offentligt dokumenteret. CoreWeave-casen viser dog, at en automatisk 2-3-årig udskiftningsantagelse er for simpel. Det resterende datagap handler derfor om fleet-brede retirement-data, ikke om hvorvidt ældre GPU-generationer overhovedet kan have økonomisk værdi efter få år.

### Fra levetid til genbrug og e-affald

Hardwareens første liv afsluttes altså ikke nødvendigvis, når den er afskrevet eller fjernet fra et high-end AI-workload. Den kan flyttes til en anden opgave, indgå som reservedel, videresælges, skilles ad eller materialegenindvindes.

Kapitel 7 følger denne kæde videre. Her bliver spørgsmålet ikke længere kun, hvor længe hardwaren bruges, men hvad der faktisk sker med den bagefter – og hvorfor retirement, reuse, recycling og e-affald skal holdes som separate størrelser.

### Kilder nævnt i kapitlet


• Amazon.com, Inc. (2026): Annual Report 2025 – Form 10-K

• Microsoft (2025): Microsoft Annual Report 2025

• Meta Platforms, Inc. (2026): Annual Report 2025 – Form 10-K

• Ostrouchov, George et al. / Oak Ridge National Laboratory (2020): GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability

• Wadenstein, Mattias & Wim Vanderbauwhede (2025): Life cycle analysis for emissions of scientific computing centres, European Physical Journal C

• Stojkovic, Jovan et al. / Microsoft Research (2026): Rearchitecting the Datacenter Lifecycle for AI

• Li, Ruihao et al. / Meta & The University of Texas at Austin (2026): Hardware Lifecycle-Aware Power Planning in Commercial Hyperscale Datacenters, OSDI 2026

• Google Sustainability (2026): Our operations – waste reduction and data center hardware reuse

• Microsoft (2025–2026): Circular Centers / zero-waste cloud hardware reporting

• CoreWeave (2026): Q2 2026 earnings call, 11. august 2026.

## Kapitel 7

# Genbrug, secondary use og e-affald

Når en server eller accelerator forlader sin første rolle, er det fristende at beskrive den som udtjent. I praksis begynder der først et nyt fordelingsspørgsmål. Udstyret kan fortsætte internt i en mindre krævende workload, blive brugt som reservedel, repareres, videresælges, demonteres til komponenter eller sendes videre til materialegenanvendelse. Først når udstyret eller komponenterne faktisk har affaldsstatus, giver betegnelsen e-affald mening.

Den sondring er central for AI's materialeregnskab. Hvis retirement fra en high-end AI-rolle automatisk bogføres som e-affald, bliver både affaldsmængden og behovet for nyproduktion let overvurderet. Hvis alt udskiftet udstyr omvendt behandles som cirkulært, risikerer analysen at kreditere genbrug, der aldrig fandt sted. Derfor skal retirement, reuse, refurbishment, resale, recycling og residualt affald registreres som forskellige hændelser.

### Hvornår bliver elektronik til e-affald?

EU's WEEE-direktiv definerer waste electrical and electronic equipment som elektrisk eller elektronisk udstyr, der er affald efter affaldsrammedirektivets definition, inklusive de komponenter og delsystemer, som er del af produktet på kassationstidspunktet. Den juridiske nøgle er dermed affaldsstatus, ikke alder eller tidspunktet for en teknisk refresh.

Det giver en praktisk konsekvens for datacenterhardware: en server, der tages ud af sin primære produktion og sendes til direkte genbrug som fungerende udstyr, er ikke det samme som en server, der kasseres og behandles som WEEE. WEEE-reglerne indeholder ligefrem

dokumentationskrav, der skal kunne skelne brugt, funktionsdygtigt udstyr bestemt til direkte genbrug fra affald ved grænseoverskridende transport.

## Retirement er et fordelingspunkt

Kapitel 6 viste, hvorfor primary-service retirement ikke er det samme som fysisk end-of-life. Det næste led handler om, hvor udstyret faktisk går hen. En robust lifecycle-analyse bør derfor følge strømmen efter retirement frem for at afslutte regnskabet på den dato, hvor hardwaren forlader sin første AI-opgave.

**Figur 7.1 – Retirement er et fordelingspunkt, ikke automatisk e-affald**

*Analytisk oversigt. Kildegrundlag: EU WEEE Directive 2012/19/EU; Google Sustainability 2026; Microsoft Circular Centers 2025–2026.*

## Fire forskellige cirkulære udfald

Intern genbrug er den mest direkte forlængelse af komponentens funktion. En harddisk, hukommelsesmodul, strømforsyning eller anden komponent kan blive høstet fra decommissioned hardware og gå tilbage i operatørens egen beholdning til nye builds, upgrades eller reservedele. Refurbishment og resale flytter i stedet funktionelt udstyr eller komponenter til en ny bruger eller et sekundært marked. Materialegenanvendelse ligger længere nede i hierarkiet, fordi den oprindelige funktion går tabt, mens dele af materialeværdien bevares. Residualt affald er den del, der hverken får en ny funktion eller genvindes som materiale.

De fire strømme har forskellig betydning i et miljøregnskab. Direkte reuse kan forlænge den nyttige levetid og i nogle tilfælde undgå eller udsætte fremstilling af et nyt produkt. Recycling kan reducere behovet for visse primære råmaterialer, men den kræver behandling og leverer ikke nødvendigvis materialer med samme kvalitet eller i samme mængde

som det oprindelige produkt. En samlet "circularity rate" kan derfor skjule afgørende forskelle.

**Tabel 7.1 – De vigtigste hændelser efter primary-service retirement**

| Hændelse | Funktionen bevares? | Typisk resultat | Hvad bør dokumenteres? |
|---|---|---|---|
| Intern reuse | Ja | Komponent eller system bruges igen internt | Antal/masse, ny funktion, ekstra levetid og hvad der realistisk undgås |
| Refurbishment / resale | Ja eller delvist | Udstyr repareres, opgraderes eller videresælges | Funktionalitet, modtager, faktisk second-life og varighed |
| Recycling | Nej som oprindeligt produkt | Materialer genvindes til nye produktstrømme | Masse, materialefraktioner, recovery rate og faktisk downstream-behandling |
| Residualt affald | Nej | Forbrænding, deponi eller anden slutbehandling | Masse, affaldstype og behandlingsvej |

*Kilde: Egen syntese baseret på EU's WEEE-begreber og de dokumenterede reverse-supply-chain-praksisser hos Google og Microsoft. Kategorien "retirement" er en driftsmæssig hændelse og er bevidst ikke sidestillet med affald.*

## Google: millioner af komponenter får et nyt internt liv

Google Sustainability rapporterer, at virksomheden i 2025 høstede mere end 7,5 millioner komponenter fra decommissioned datacenterhardware til intern genbrug i builds, upgrades og reservedelsprogrammer. Google beskriver samtidig reverse-logistics-processer, hvor decommissioned servere demonteres, og dele kan gå til re-inventory, reuse eller secondary market.

Det er direkte evidens for, at decommissioning ikke automatisk ender som affald. Tallet har samtidig klare begrænsninger. Google offentliggør ikke en denominator, der viser hvor stor en andel af alle decommissioned komponenter de 7,5 millioner udgør, og oplysningen

isolerer hverken GPU'er eller moderne AI-acceleratorer. Ét høstet kabel og én accelerator tæller desuden begge som én komponent, selv om masse, værdi og embodied impact er meget forskellige.

### Microsoft: reuse og recycling skal holdes adskilt

Microsofts Circular Centers giver et andet nyttigt eksempel. Virksomheden rapporterede for FY2024 en combined reuse-and-recycling rate på 90,9 procent for servers og cloud hardware components og mere end 3,2 millioner komponenter genbrugt gennem interne og eksterne kanaler. Microsoft beskriver Circular Centers som reverse-supply-chain-faciliteter, der sorterer decommissioned hardware til intern genbrug, refurbishment og recycling.

Det afgørende ord er combined. En 90,9-procents reuse-and-recycling rate må ikke omskrives til, at 90,9 procent af hardwaren bliver genbrugt. Genbrug bevarer en komponent eller et produkts funktion; recycling nedbryder produktet for at genvinde materialer. Begge kan være positive cirkulære strømme, men de har forskellig miljømæssig betydning.

### Et europæisk datagap: WEEE-statistikken kan ikke isolere AI

For et europæisk referenceværk er Eurostats WEEE-statistik den naturlige hovedkilde til observerede elektronikaffaldsstrømme. Den giver officielle data for udstyr sat på markedet, indsamling, behandling, recovery, recycling og preparation for reuse under WEEE-direktivet.

Statistikken løser imidlertid ikke spørgsmålet om, hvor meget AI-hardware der bliver til e-affald. Fra referenceåret 2019 klassificeres elektrisk og elektronisk udstyr i seks brede kategorier. IT- og telekommunikationsudstyr indgår i kategorier som large equipment eller small IT and telecommunication equipment afhængigt af produktets fysiske størrelse. Der findes ingen særskilt Eurostat-kategori for server-GPU'er, AI-acceleratorer eller AI-datacenterhardware.

Eurostat fremhæver også, at oplysninger om produkternes levetid fra tidspunktet, hvor de sættes på markedet, til de bliver affald, ikke findes i

de data, der bruges til at overvåge WEEE-indsamlingen. Det er netop den information, der skulle til for at koble en bestemt generations salg eller installation til et senere affaldsflow.

**Tabel 7.2 – Hvad de centrale circularity- og e-affaldstal faktisk fortæller**

| Kilde / periode | Rapporteret størrelse | Det kan vi bruge tallet til | Det må tallet ikke gøres til |
|---|---|---|---|
| Google, 2025 | >7,5 mio. komponenter høstet til intern reuse | Dokumentere reel intern second-life-praksis i hyperscale-drift | Andel af alle retirement flows eller GPU-specifik reuse-rate |
| Microsoft, FY2024 | 90,9 % combined reuse + recycling; >3,2 mio. komponenter reused | Dokumentere stor reverse-supply-chain-aktivitet | 90,9 % reuse eller accelerator-specifik rate |
| Eurostat, EU 2023 | 14,4 mio. ton EEE sat på markedet; 5,2 mio. ton WEEE indsamlet; officiel collection rate 37,5 % | Beskrive observerede EU-WEEE-flows og officiel collection-performance | AI-e-affaldsrate eller samme-års "manglende" affald |
| UNITAR/ITU, globalt 2022 | 62 mio. ton e-affald; 22,3 % dokumenteret formelt indsamlet/recycled | Sætte e-affaldsproblemet i global størrelsesorden | AI's andel eller europæisk server-/GPU-rate |

*Kilder: Google Sustainability (2026 current operations page); Microsoft Circular Centers / zero-waste reporting; Eurostat "Waste statistics – electrical and electronic equipment" (data 2023, artikel 2025); UNITAR/ITU Global E-waste Monitor 2024. Tallene har forskellige geografi, denominator og systemgrænse og må derfor ikke sammenlignes som én fælles rate.*

## Et regneeksempel på en forkert denominator

Eurostat rapporterer for EU i 2023, at 14,4 millioner ton elektrisk og elektronisk udstyr blev sat på markedet, mens 5,2 millioner ton WEEE blev indsamlet. En hurtig division giver 5,2 / 14,4 = 36,1 procent. Det ligger tæt på Eurostats officielle collection rate på 37,5 procent og kan derfor se ud som en rimelig genvej.

Genvejen bruger imidlertid den forkerte denominator. Den officielle WEEE collection rate beregnes mod den gennemsnitlige vægt af EEE sat på markedet i de tre foregående år. En computer, server eller

husholdningsmaskine sat på markedet i 2023 forventes normalt ikke at blive affald samme år. Derfor kan forskellen på 9,2 millioner ton mellem samme års market placement og collection heller ikke kaldes "uindsamlet e-affald fra 2023". En stor del befinder sig fortsat i den aktive produktbestand, mens andet kan være oplagret eller bevæge sig gennem andre strømme.

Eksemplet er relevant langt ud over WEEE. Når AI-e-affald fremskrives fra antal solgte GPU'er eller servere, bliver levetidsantagelsen og denominator afgørende. Et salgstal i år 0 kan først omregnes til affald i år 3, 5 eller 8, hvis en faktisk eller modelbaseret retirement-fordeling er angivet.

### Global e-affaldsstatistik er vigtig – men ikke et AI-tal

UNITAR og ITU's Global E-waste Monitor 2024 estimerer 62 millioner ton globalt e-affald i 2022 og en dokumenteret formel collection-and-recycling rate på 22,3 procent. Europa ligger i rapporten højere med 42,8 procent dokumenteret collection og recycling. Tallene viser, at e-affald er en betydelig global ressource- og affaldsudfordring.

De kan samtidig ikke bruges til at sige, hvor stor en del der stammer fra AI. De 62 millioner ton omfatter hele spektret af elektriske og elektroniske produkter, fra køleskabe og skærme til telefoner, værktøj og IT-udstyr. Selv kategorien small IT and telecommunication equipment omfatter f.eks. laptops, mobiltelefoner, GPS-enheder og routere. Et globalt e-affaldstal er derfor baggrundskontekst, ikke en AI-attributionsfaktor.

### Hvornår kan second life krediteres?

En lifecycle-analyse kan blive lige så misvisende ved at overvurdere genbrug som ved at overse det. Hvis en decommissioned accelerator sælges videre, bør analysen spørge, hvilken funktion den faktisk udfører i det nye system, hvor længe den bruges, og hvilket alternativ den erstatter.

Hvis second-life-hardwaren udfører en nyttig workload, som ellers ville have krævet ny hardware, er der et reelt argument for, at den oprindelige fremstillingsbelastning fordeles over en længere funktionel levetid. Hvis hardwaren blot ligger på lager eller erstatter andet eksisterende brugt udstyr uden at undgå nyproduktion, bliver kreditten mindre. Den funktionelle enhed og det kontrafaktiske alternativ afgør resultatet.

Samme princip gælder recycling. Et kilogram indsamlet elektronik er ikke automatisk et kilogram genvundet kritisk materiale. Recovery varierer mellem materialer og processer, og kvalitetstab kan begrænse closed-loop-anvendelsen. Derfor bør bogens materialeregnskab skelne mellem indsamlet masse, behandlet masse, faktisk recovered materiale og den mængde primært materiale, som realistisk undgås.

### Påstanden under lup

**Påstand:** Når AI-hardware udskiftes, bliver den til e-affald.

**Vurdering:** Misvisende som generel påstand.

**Kort svar:** Primary-service retirement og affald er forskellige hændelser. Google og Microsoft dokumenterer store reverse-supply-chain-strømme med intern reuse, refurbishment og recycling. Først når udstyr eller komponenter faktisk har affaldsstatus, er WEEE-begrebet relevant.

**Hvad siger evidensen?:** Google rapporterer mere end 7,5 millioner komponenter høstet til intern reuse i 2025. Microsoft rapporterer 90,9 procent combined reuse and recycling i FY2024 og mere end 3,2 millioner reused komponenter. Eurostats WEEE-data kan ikke isolere AI-hardware.

**Det afgørende forbehold:** Leverandørdata giver ingen robust accelerator-specifik retirement- eller reuse-rate. De dokumenterer, at cirkulære veje findes i stor skala, men de kan ikke bruges til at beregne en universel andel af AI-hardware, der genbruges eller bliver e-affald.

## Hvad bør en beslutningstager kræve af et e-affaldstal?

Når en rapport opgiver "AI-e-affald", bør første spørgsmål være, om tallet er observeret eller modelleret. Observerede data bør angive affaldskategori, geografi, periode og om massen er collected, treated, recycled, prepared for reuse eller endeligt disponeret. Modellerede data bør derudover vise hardwarebestand, salg eller installationer, levetidsfordeling, reuse-antagelser og den valgte retirement-model.

Et troværdigt AI-specifikt tal kræver også en klar attribution. Hvis kilden bygger på hele kategorien IT- og telekommunikationsudstyr, skal den forklare, hvordan servere eller acceleratorer er isoleret. Hvis den bruger en hyperscaler-disclosure, skal denominator være synlig: komponentantal, masse, økonomisk værdi og antal servere beskriver forskellige størrelser.

For europæiske beslutningstagere er WEEE-statistikken et stærkt grundlag for den samlede elektronikstrøm og de lovregulerede behandlingsmål. Den nuværende statistik er for grov til at levere en særskilt AI-rate. Den begrænsning bør stå åbent frem for at blive udfyldt med en global procentsats eller et rundet levetidstal.

## Fra hardwarestrøm til driftsregning

Kapitel 4 til 7 har fulgt regningen før og efter den første anvendelse: råmaterialer, chipproduktion, embodied belastning, levetid, retirement og cirkulære strømme. Dermed er Del 2's hardwarekæde samlet.

Næste del flytter fokus til den løbende drift. Kapitel 8 undersøger træning og inference: den koncentrerede regning ved at udvikle og træne en model over for den gentagne regning, der opstår hver gang modellen anvendes. Her bliver trafik, modellens levetid og antallet af forespørgsler afgørende for, hvilken fase der fylder mest i det samlede regnskab.

## Kilder nævnt i kapitlet


• European Parliament and Council (2012, amended): Directive 2012/19/EU on waste electrical and electronic equipment (WEEE)

• Eurostat (2025): Waste statistics – electrical and electronic equipment, data through 2023

• Eurostat (2025): WEEE metadata – env_waselee / open-scope six product categories

• Google Sustainability (2026): Our operations – waste reduction and data center hardware reuse

• Google Sustainability (2026): Bridging the Gap: Operationalizing Circularity in Data Centers

• Microsoft (2025–2026): Circular Centers / progress towards zero waste

• UNITAR & ITU (2024): Global E-waste Monitor 2024

## Del 3

# Regningen når AI kører

Del 3 følger den løbende driftsregning fra modeltræning og inference til elektricitet, vand, køling, netværk, storage og placeringen af selve beregningen.

## Kapitel 8

# Træning og inference

Når en stor AI-model omtales i energidebatten, bliver modeltræningen ofte det første billede: tusindvis af acceleratorer, der arbejder intensivt gennem en afgrænset periode. Det er en reel og synlig energipost. Efter træningen begynder imidlertid den fase, som kan fortsætte hver dag i måneder eller år: inference, hvor den trænede model bruges til at producere svar, klassifikationer, billeder, analyser eller andre resultater.

For en model med begrænset anvendelse kan den oprindelige træning dominere regnskabet. For en tjeneste med meget stor trafik kan de mange gentagne inferenser over tid nærme sig eller overstige træningsfasens samlede belastning. Fine-tuning, retraining og kontinuerlig læring kan lægge nye træningsaktiviteter oven i driftsperioden.

Derfor er spørgsmålet "hvad bruger mest – træning eller inference?" kun meningsfuldt, når modellens livscyklus, trafik, hardware, geografiske placering og systemgrænse er kendt.

### Modeltræning er en koncentreret aktivitet

Modeltræning omdanner store datamængder og en valgt modelarkitektur til et sæt parametre, der senere kan bruges til inference. Ved store foundation models kan den centrale træningskørsel være meget compute-intensiv, men energiregnskabet begynder før den afsluttende run. Databehandling, eksperimenter, testkørsler, hyperparameterarbejde, validering og mislykkede eller gentagne runs kan også kræve beregning.

ITU-T L.1801 fra februar 2026 er særlig nyttig her, fordi standarden kræver, at træningspåvirkningen rapporteres særskilt og at initial modeltræning holdes adskilt fra continuous learning eller retraining.

Standarden placerer samtidig AI-systemet i en fuld livscyklus, hvor datahåndtering, hardware, datacenterinfrastruktur, lagring og transmission kan indgå, når de ligger inden for systemgrænsen.

Denne opdeling passer godt til bogens metode. En oplysning om energien i "træningen" bør derfor ledsages af en beskrivelse af, om tallet kun dækker den afsluttende træningskørsel eller også de aktiviteter, der førte frem til den.

### Inference er den gentagne driftsaktivitet

Inference er den fase, hvor live-data køres gennem den trænede model, og systemet producerer et resultat. Én inference kan være lille i forhold til en stor træningskørsel. Millioner eller milliarder af inferenser kan samlet blive en betydelig livscykluspost.

ITU-T L.1801 behandler driftsfasen bredere end selve modelkaldet. Her kan energien til inference suppleres af continuous learning, data storage, data transmission, installation, vedligeholdelse og site-specifikke støtteaktiviteter. Det er samme systemgrænseproblem, vi så i kapitel 2 og 3: et per-request-tal for acceleratorerne er ikke nødvendigvis et per-request-tal for hele tjenesten.

Inference er desuden ikke én ensartet arbejdsbyrde. Inputlængde, outputlængde, modelstørrelse, batching, caching, hardware, hukommelsesadgang og graden af reasoning kan flytte energiforbruget. I analytiske estimater bør prompt prefill og autoregressiv decoding derfor skilles ad, og sådanne estimater må ikke forveksles med direkte effektmålinger.

**Tabel 8.1 – Træning og inference er forskellige omkostningsprofiler**

| Dimension | Initial modeltræning | Inference | Hvad skal dokumenteres? |
|---|---|---|---|
| Tidsprofil | Koncentreret i én eller flere træningsperioder | Gentages gennem hele driftsperioden | Periode og livscyklus |
| Aktivitet | Optimering af modelparametre på træningsdata | Kørsel af nye input gennem den trænede model | Model, opgave, input/output og hardware |
| Skal holdes adskilt fra | Fine-tuning, test og continuous learning, hvis disse opgøres separat | Retraining, storage, transmission og øvrig tjenesteinfrastruktur | Systemgrænse og allokering |
| Geografi | Kan i princippet placeres eller tidsforskydes til bestemte datacentre | Kan være geografisk distribueret tæt på brugerne | Lokation, tidspunkt og relevant elmix |
| Skalering | Påvirkes af modelstørrelse, træningsdata, antal runs og hardwareudnyttelse | Påvirkes af antal requests, tokens, kvalitet, batching og tjenestens levetid | Aktivitetsniveau og funktionel enhed |

*Kilde: Egen syntese med udgangspunkt i ITU-T L.1801 (2026). Tabellen beskriver metodiske forskelle og er ikke en kvantitativ fordeling af energiforbruget.*

## Fra udvikling til gentagen brug

En livscyklusbetragtning gør det lettere at se, hvorfor træning og inference ikke bør sammenlignes som to isolerede tal. Den første modeltræning ligger tidligt, mens inference gentages efter deployment. Retraining eller continuous learning kan sende systemet tilbage i en ny træningsaktivitet under driftsperioden.

**Figur 8.1 – Modeltræning er koncentreret, inference gentages over tid**

*Analytisk oversigt baseret på AI-livscyklusfaserne i ITU-T L.1801 (2026). Figuren viser rækkefølge og gentagelse, ikke relative energimængder.*

## Der findes ingen fast training-to-inference-ratio

Det er fristende at søge én procentfordeling, der kan bruges på tværs af modeller. Den findes ikke som en stabil fysisk konstant. Fordelingen ændres af, hvor ofte modellen anvendes, hvor længe den er i drift, hvor meget compute hver request kræver, om modellen retraines, og hvilket hardware- og datacentermiljø der anvendes.

En reviewartikel i Nature Reviews Clean Technology fra juli 2026 sammenfatter eksisterende forskning og angiver, at inference i de analyserede livscyklusser samlet kan bidrage med omkring 40–60 procent af en models livstidsudledning af $CO_2$e. Det er et vigtigt signal om, at inference ikke bør behandles som en ubetydelig restpost. Intervallet er samtidig en syntese på tværs af studier og må ikke omskrives til, at "inference altid er 40–60 procent".

Applied Energy offentliggjorde i 2026 en fuld lifecycle-analyse af generativ AI på tværs af forskellige datacenterarkitekturer. Studiet inkluderer både træning og inference og viser samtidig, at modelvalg, datacenterarkitektur og geografisk elintensitet påvirker resultatet. Det understøtter samme hovedregel: livscyklusfordelingen kan ikke løsrives fra den konkrete tekniske og geografiske kontekst.

## Et konkret LCA-studie viser, hvorfor træning ikke altid dominerer

Alexandroff og kolleger analyserede i 2026 et komplet generativt AI-system til automatisk radio og brugte én times generering, udsendelse og streaming som funktionel enhed. I basisscenariet kom omkring 70 procent af klimaaftrykket fra produktionen af computerhardware, omkring 10 procent fra Tortoise text-to-speech inference, omkring 9 procent fra streaming og omkring 3 procent fra modeltræning.

I et alternativt scenario steg træningens andel til cirka 14 procent. Pointen er ikke, at træning generelt kun fylder 3 eller 14 procent. Studiet viser tværtimod, hvor kraftigt fordelingen kan ændres med model, anvendelse, hardwarelevetid, brugerantal, strømforsyning, systemgrænse og funktionel enhed.

Streaming og netværk bidrog også i casen, men forfatterne fremhæver usikkerhed i netværksdata og brug af delvist ældre datasæt. Studiet styrker derfor argumentet for at medtage netværk og transmission, når de er materielle, uden at levere en universel netværksfaktor.

*Kilde: Alexandroff, Ronja et al. (2026): Life cycle assessment of a generative artificial intelligence system: the case of a radio application, International Journal of Life Cycle Assessment, publiceret 7. august 2026. Procentsatserne er case-specifikke.*

## Et illustrativt break-even-regnestykke

Følsomheden kan vises med et enkelt regneeksempel. Antag en fiktiv model, hvor den initiale modeltræning bruger 1.000 MWh elektricitet. Det er et rundt, konstrueret tal og repræsenterer ikke en bestemt kommerciel model. Spørgsmålet er, hvor mange inferenser der samlet skal til, før inferenceenergien bliver lige så stor som de 1.000 MWh.

1.000 MWh svarer til 1.000.000 kWh eller 1 milliard Wh. Break-even-antallet beregnes derfor som træningsenergien i Wh divideret med energien pr. inference.

**Tabel 8.2 – Illustrativt break-even mellem modeltræning og inference**

| Illustrativ inferenceenergi | Break-even requests | Ved 1 mio. requests/dag | Beregning |
|---|---|---|---|
| 0,25 Wh/request | 4,0 mia. | ca. 11,0 år | 1 mia. Wh / 0,25 Wh |
| 1 Wh/request | 1,0 mia. | ca. 2,7 år | 1 mia. Wh / 1 Wh |
| 5 Wh/request | 200 mio. | ca. 0,55 år | 1 mia. Wh / 5 Wh |

*Illustrativ beregning. Træningsenergi = 1.000 MWh og request-volumen = 1 mio. pr. dag er konstruerede antagelser. Tabellen må ikke bruges som energiestimat for en konkret AI-model.*

Regneeksemplet viser, hvor meget konklusionen flytter sig, når per-request-forbruget ændres. Ved 0,25 Wh kræves fire milliarder inferenser for at matche den fiktive træningsenergi. Ved 5 Wh er break-even 200 millioner inferenser. Trafikken er lige så vigtig: ved en million requests om dagen nås de tre break-even-punkter efter henholdsvis omkring elleve år, 2,7 år og et halvt år.

Tallene fortæller ikke, hvad en virkelig model bruger. De viser, hvorfor en diskussion om "træning versus inference" mangler et afgørende input, hvis request-volumen og request-energi ikke oplyses.

### Continuous learning og retraining gør livscyklussen mindre enkel

Mange AI-systemer stopper ikke med at udvikle sig efter den første træning. Modeller kan fine-tunes, opdateres eller retraines, og nogle systemer arbejder med continuous learning. ITU-T L.1801 kræver netop, at initial training og continuous learning rapporteres separat, fordi den første aktivitet ofte kan måles som en afsluttet træningskørsel, mens fremtidig continuous learning typisk bygger på estimater og antagelser.

For beslutningstagere er sondringen vigtig. Hvis en tjeneste opdateres løbende, kan "træningsenergien" ikke nødvendigvis behandles som én historisk engangspost. Omvendt bør enhver softwareopdatering heller ikke automatisk bogføres som en ny fuld modeltræning. Aktiviteten skal beskrives konkret.

### Foundation models skaber et allokeringsproblem

Et foundation model kan trænes én gang og derefter bruges i mange tjenester, virksomheder og afledte modeller. Hvis hele pre-training-regningen lægges på én downstream-anvendelse, overvurderes den. Hvis pre-training ignoreres helt, undervurderes den.

ITU-T L.1801 anbefaler derfor en begrundet allokering af foundation-models embodied og pre-training impacts, blandt andet ud fra estimerede inferencevolumener og antallet af afledte modeller, når sådanne data er tilgængelige. Det er metodisk nødvendigt, men kan

være praktisk vanskeligt, fordi leverandører sjældent offentliggør alle de aktivitetsdata, en præcis allokering kræver.

For en bog som denne er konsekvensen enkel: en per-request-livscyklusværdi, der inkluderer en andel af pre-training, skal fortælle, hvordan denne andel er beregnet. Ellers er tallet ikke reproducerbart.

### Geografi kan være forskellig i træning og drift

Modeltræning og inference behøver heller ikke finde sted på samme geografiske lokation. En foundation model kan trænes i ét land eller én region og senere betjene brugere fra datacentre i flere andre. Elmix, vandforhold og netbelastning kan dermed være forskellige mellem livscyklusfaserne.

ITU-T L.1801 kræver, at energiforbruget knyttes til de mest repræsentative energimix og impact factors for de steder og faser, hvor energien faktisk anvendes. Det er særlig relevant for denne bog, hvor Europa er det primære referencepunkt, mens globale data bruges, når de giver relevant evidens. En amerikansk træningskørsel kan være en værdifuld case, men dens $CO_2$-faktor kan ikke uden videre flyttes til europæisk inference – eller omvendt.

Kapitel 9 går derfor videre til selve elektriciteten og viser, hvorfor den samme kWh kan give meget forskellige $CO_2$-resultater afhængigt af produktionsmix, geografi, tidspunkt og om analysen bruger gennemsnitlig eller marginal emissionsintensitet.

**Påstanden under lup**
**Påstand:** Det er modeltræningen, der står for AI's største energi- og klimaaftryk; inference er kun en mindre restpost.
**Vurdering:** For kategorisk.
**Kort svar:** En stor træningskørsel kan være den dominerende post tidligt i modellens liv. Ved høj trafik og lang levetid kan gentagen inference blive lige så stor eller større. Retraining og continuous learning kan desuden tilføje nye træningsposter under driften.
**Hvad siger evidensen?** ITU-T L.1801 kræver separat rapportering af initial training, continuous learning og inference. En 2026-review i Nature Reviews Clean Technology vurderer på tværs af analyserede livscyklusser, at inference samlet kan bidrage med omkring 40–60 % af en models lifetime $CO_2$e. Applied Energy 2026 viser samtidig, at model, datacenterarkitektur og geografi ændrer footprintet.
**Det afgørende forbehold:** 40–60 % er ikke en universel koefficient. Den relevante fordeling skal beregnes for den konkrete model, trafik, levetid, systemgrænse og geografiske kontekst.

### Hvad bør en beslutningstager kræve af et training/inference-tal?

Et troværdigt livscyklusregnskab bør først identificere, hvilke aktiviteter der ligger i training-begrebet: pre-training, fine-tuning, eksperimenter, test, retraining eller continuous learning. Derefter bør inference opgøres med en kendt aktivitetsmængde – f.eks. requests, tokens eller en anden funktionel enhed – sammen med model, hardware, periode og systemgrænse.

Hvis $CO_2$ eller vand beregnes, skal lokation og metode følge med. Hvis foundation-model-træning fordeles på downstream-brug, skal allokeringsreglen være synlig. Hvis data er estimerede eller modellerede, bør de ikke præsenteres som direkte målinger.

Den Europæiske Kommission arbejder i 2026 fortsat med en praktisk måleramme for energiforbrug og emissioner fra AI-modeller og

-systemer. Det er et vigtigt signal om metodens modenhed: der findes allerede stærke standarder og forskningsmetoder, men systemniveau-målingen er fortsat under udvikling og harmonisering i europæisk sammenhæng.

## Fra modelaktivitet til elektricitetens klimaaftryk

Træning og inference fortæller, hvornår og hvorfor AI bruger compute. De fortæller endnu ikke, hvor stort klimaaftrykket bliver. To identiske workloads kan bruge samme mængde elektricitet og alligevel få meget forskellige $CO_2$-resultater, hvis de køres i forskellige elsystemer eller på forskellige tidspunkter.

Det næste kapitel går derfor fra AI-aktiviteten til energisystemet omkring den. Her bliver forskellen mellem kWh, produktionsmix, location-based og market-based regnskaber samt gennemsnitlige og marginale emissionsfaktorer afgørende.

### Kilder nævnt i kapitlet


• ITU-T (2026): Recommendation L.1801 – Guidelines for assessing the environmental impact of artificial intelligence systems

• Chien, Andrew A. et al. (2026): Strategies and design for increasing AI sustainability, Nature Reviews Clean Technology

• Applied Energy (2026): Generative AI impact assessment through a life cycle analysis of multiple data center typologies

• European Commission, DG CONNECT (2026): Targeted consultation on measuring energy consumption and emissions of AI models and systems

• Vartziotis, Tina et al. (2026): From Tokens to Watt-hours: Analytical Energy Estimation for LLM Inference on Modern GPUs (technical preprint; used for method triangulation)

• Alexandroff, Ronja et al. (2026): Life cycle assessment of a generative artificial intelligence system: the case of a radio application, International Journal of Life Cycle Assessment.

Kapitel 9

# Elektriciteten: en kWh er ikke bare en kWh

Et datacenter kan bruge præcis den samme mængde elektricitet i Sverige, Danmark, Tyskland eller Polen. Elregningen kan opgøres i de samme kilowatt-timer. Klimaaftrykket fra elektriciteten kan alligevel være meget forskelligt.

Forskellen skyldes først og fremmest, hvordan elektriciteten produceres, hvor forbruget finder sted, hvornår det finder sted, og hvilket emissionsregnskab analysen forsøger at besvare. Det er en af de vigtigste kilder til misvisende tal i debatten om AI og datacentre.

Et energital og et $CO_2$-tal er derfor to forskellige oplysninger. Først måles energien. Derefter vælges en emissionsfaktor, som passer til det spørgsmål, analysen skal besvare.

## Fem lag, der ofte blandes sammen

Diskussionen bliver langt klarere, når fem begreber holdes adskilt: elforbrug, produktionsmix, teknologispecifik livscyklusudledning, gennemsnitlig emissionsintensitet og marginal emissionsfaktor. Hertil kommer virksomhedens regnskabsmetode, som kan være location-based eller market-based.

| Spørgsmål | Relevant størrelse | Hvad tallet fortæller |
|---|---|---|
| Hvor meget elektricitet bruges? | kWh / MWh / GWh | Den fysiske energimængde. Ingen $CO_2$-konklusion endnu. |
| Hvordan produceres en bestemt type elektricitet? | Teknologispecifik livscyklusudledning | Emissioner fra fx kul, gas, vind eller atomkraft over den valgte livscyklus. |

| Spørgsmål | Relevant størrelse | Hvad tallet fortæller |
|---|---|---|
| Hvor $CO_2$-intensiv er elproduktionen i et land eller en region i gennemsnit? | Gennemsnitlig emissionsintensitet | Et gennemsnit for den konkrete geografi og periode. |
| Hvad rapporterer en virksomhed for indkøbt elektricitet? | Location-based / market-based Scope 2 | Et regnskab efter GHG Protocols regler; de to metoder bruger forskellige datagrundlag. |
| Hvad ændres, hvis en ekstra belastning lægges på elsystemet? | Marginal emissionsfaktor | Et konsekvensspørgsmål om den generation og kapacitet, der reagerer på ændret efterspørgsel. |

*Metoderamme: GHG Protocol Scope 2 Guidance; EEA; UNECE; IEA; Brander et al. 2025. Tabellen er en redaktionel syntese.*

## Produktionsform og nationalt elmix er to forskellige ting

En meget stor del af forvirringen opstår, når en emissionsfaktor for én produktionsteknologi bliver blandet sammen med emissionsintensiteten i et helt lands elmix.

UNECE har gennemført livscyklusanalyser af forskellige elproduktionsteknologier. I rapporten ligger kulkraft uden $CO_2$-fangst omkring 751–1.095 g $CO_2$e/kWh og kombineret naturgas omkring 403–513 g $CO_2$e/kWh. Lavemissionsteknologier ligger langt lavere: atomkraft omkring 5,5 g $CO_2$e/kWh i den globale gennemsnitscase, repræsentativ vandkraft omkring 6,1–11 g, onshore vind under 16 g og offshore vind under 23 g $CO_2$e/kWh. For solceller ligger hovedparten af de analyserede teknologier under 35 g $CO_2$e/kWh, med regionale og teknologiske undtagelser.

*Kilde: UNECE (2022), Life Cycle Assessment of Electricity Generation Options. Tallene er teknologispecifikke livscyklusresultater og må ikke bruges som nationale elmixfaktorer.*

| Elproduktionsteknologi | UNECE livscyklusudledning | Vigtig afgrænsning |
|---|---|---|
| Kul uden CCS | 751–1.095 g $CO_2$e/kWh | Spænd mellem analyserede kulteknologier og regioner. |
| Naturgas, combined cycle | 403–513 g $CO_2$e/kWh | Livscyklus inkl. brændselsforsyning. |

| Elproduktionsteknologi | UNECE livscyklusudledning | Vigtig afgrænsning |
|---|---|---|
| Atomkraft | ca. 5,5 g $CO_2$e/kWh | Global gennemsnitscase i UNECE-analysen. |
| Vandkraft | ca. 6,1–11 g $CO_2$e/kWh | Repræsentativ fossil GHG; biogene effekter er stærkt siteafhængige. |
| Vind, onshore | under 16 g $CO_2$e/kWh | Hovedparten ligger i infrastruktur frem for drift. |
| Vind, offshore | under 23 g $CO_2$e/kWh | Hovedparten ligger i infrastruktur frem for drift. |
| Solceller | typisk under 35 g $CO_2$e/kWh | Variation med teknologi, produktion og region. |

*Tabel 9.1. Teknologispecifikke livscyklusresultater fra UNECE. Intervaller og formuleringer følger rapportens egne afgrænsninger.*

Et nationalt elsystem består imidlertid af flere teknologier på samme tid. Polen er ikke et kulkraftværk, og Sverige er ikke et vindkraftværk. Når et datacenters location-based driftsudledning skal beskrives med et nationalt gennemsnit, er det landets samlede produktionsmix og den anvendte emissionsmetode, der er relevant.

## Store forskelle inden for Europa

European Environment Agency, EEA, opgør den gennemsnitlige GHG-intensitet af elproduktionen i de europæiske lande. For 2024 lå Sverige på 6 g $CO_2$e/kWh, Frankrig på 36, Danmark på 64, Tyskland på 291 og Polen på 566 g $CO_2$e/kWh.

Forskellen er stor nok til at ændre resultatet markant, selv når datacenterets fysiske elforbrug holdes fuldstændig konstant.

**Figur 9.1 – Gennemsnitlig GHG-intensitet af elproduktion i fem europæiske lande, 2024**

*Kilde: European Environment Agency, 2024-data. EEA-indikatoren beskriver direkte GHG-emissioner fra elproduktionen efter EEA/UNFCCC-metoden. Den er ikke en fuld livscyklusfaktor.*

## Et illustrativt regneeksempel: samme 100 GWh, fem forskellige resultater

Antag et hypotetisk datacenter, som bruger 100 GWh elektricitet om året. Vi holder alt andet konstant og ganger alene elforbruget med EEA's nationale gennemsnitsintensitet for 2024. Beregningen viser dermed geografieffekten i én bestemt gennemsnitsmetode; den beskriver hverken et konkret datacenter, market-based Scope 2 eller den marginale klimaeffekt af et nyt load.

| Land | EEA 2024 gennemsnitsintensitet | 100 GWh × faktor |
|---|---|---|
| Sverige | 6 g $CO_2e$/kWh | 600 ton $CO_2e$ |
| Frankrig | 36 g $CO_2e$/kWh | 3.600 ton $CO_2e$ |
| Danmark | 64 g $CO_2e$/kWh | 6.400 ton $CO_2e$ |
| Tyskland | 291 g $CO_2e$/kWh | 29.100 ton $CO_2e$ |
| Polen | 566 g $CO_2e$/kWh | 56.600 ton $CO_2e$ |

*Tabel 9.2. Illustrativ beregning: 100 GWh = 100.000.000 kWh. Resultaterne er simple location-based gennemsnitseksempler baseret på EEA's 2024-landefaktorer.*

I eksemplet giver 100 GWh cirka 600 ton $CO_2e$ med Sveriges 2024-gennemsnit og 56.600 ton med Polens. Det er omtrent 94 gange så meget for den samme energimængde. Forskellen kommer fra den anvendte geografiske emissionsfaktor, ikke fra et højere elforbrug.

## Hvorfor EEA- og UNECE-tallene ikke må blandes sammen

Her er en afgørende metodisk detalje. EEA's nationale indikator og UNECE's teknologital måler ikke det samme.

EEA beregner et gennemsnit for elproduktionen på landeniveau ud fra rapporterede drivhusgasemissioner og den producerede elektricitet. I denne indikator anvendes nul driftsudledning på atomkraft og vedvarende energikilder, og upstream-emissioner som brændselsforsyning, anlægsbyggeri og andre livscyklusemissioner er ikke med. UNECE's analyse er derimod en livscyklusvurdering af konkrete

produktionsteknologier, hvor blandt andet infrastruktur og upstream-led indgår.

*Kilder: EEA, Greenhouse gas emission intensity of electricity generation in Europe, metodeafsnit; UNECE (2022), Life Cycle Assessment of Electricity Generation Options.*

Derfor er det forkert at tage UNECE's cirka 1.000 g $CO_2$e/kWh for kul, sammenligne direkte med EEA's 6 g for Sverige og omtale forskellen som to værdier fra samme type regnskab. Forskellen mellem kul og vind eller atomkraft er reel og stor, men teknologispecifik livscyklusudledning og national gennemsnitsintensitet skal have hver sin etiket.

### Tidspunktet kan ændre resultatet inden for det samme land

Et årsgennemsnit skjuler variationen mellem timer. Solproduktion er koncentreret i bestemte timer, vind varierer med vejret, efterspørgslen ændrer sig gennem døgnet, og fleksible kraftværker reagerer på markedet. Et datacenter, der kan flytte en del af sit compute i tid, kan derfor møde forskellige fysiske emissionsprofiler uden at skifte land.

Google Cloud offentliggør f.eks. både en regional grid carbon intensity og en timebaseret CFE-procent for sine cloudregioner. Google fremhæver selv, at elrelaterede emissioner afhænger af, hvilke kraftværker der producerer i regionen, og hvornår elektriciteten bruges. I 2024-dataene varierer den oplyste grid-intensitet blandt europæiske regioner fra meget lave værdier i Stockholm og Paris til langt højere værdier i blandt andet Warszawa.

*Kilde: Google Cloud, Carbon free energy for Google Cloud regions, 2024 regional data, side opdateret 11. august 2026. Google oplyser, at de underliggende Electricity Maps-data ikke er assureret; eksemplet bruges derfor som leverandørens driftsindikator, ikke som officiel national statistik.*

### Location-based og market-based svarer på forskellige spørgsmål

Virksomhedens Scope 2-regnskab tilføjer endnu et lag. GHG Protocols gældende Scope 2 Guidance skelner mellem en location-based metode, som tager udgangspunkt i emissionsprofilen for det elsystem, hvor forbruget finder sted, og en market-based metode, som anvender

kvalificerede kontraktuelle instrumenter og leverandør- eller produktdata.

En virksomhed kan derfor rapportere et market-based resultat, der afspejler indkøb af vedvarende eller anden lavemissions-elektricitet, samtidig med at den fysiske elproduktion i de timer, hvor datacenteret kører, består af et bredere mix. De to tal er ikke rivaler. De beskriver forskellige regnskabslag.

*Kilde: GHG Protocol, Scope 2 Guidance. Den gældende vejledning kræver dual reporting, hvor market-based metoden er relevant og kvalitetskriterierne opfyldes.*

### Årsmatching og 24/7-matching

Også begrebet "100 procent vedvarende elektricitet" kræver en tidsafgrænsning. IEA beskriver, hvordan årlig matching kan understøtte udbygning af ren elproduktion, men samtidig kan skabe afstand mellem det årlige regnskab og den elektricitet, der fysisk er tilgængelig på det tidspunkt og i det net, hvor forbruget finder sted. Timebaseret matching inden for samme net bringer produktion og forbrug tættere sammen i tid og geografi.

Årlig matching har fortsat en funktion, men dens betydning skal beskrives præcist: et årligt certifikat- eller indkøbsregnskab er ikke dokumentation for, at datacenterets belastning i hver enkelt time fysisk blev dækket af kulstoffri produktion.

*Kilde: IEA (2022), Advancing Decarbonisation through Clean Electricity Procurement.*

GHG Protocol er samtidig under revision. Efter den offentlige Scope 2-konsultation i 2025–2026 offentliggjorde GHG Protocol i juli 2026 feedback og en ny samlet standardtidsplan sammen med ISO. Den eksisterende Scope 2 Guidance er derfor den gældende reference i denne bog, mens de kommende regler skal genkontrolleres før hver ny udgave.

*Tidsfølsom kilde: GHG Protocol, 29. juli 2026. Den konsoliderede fælles corporate standard er på den aktuelle plan forventet i Q4 2028.*

### Gennemsnittet svarer ikke på marginalspørgsmålet

Et nationalt gennemsnit er nyttigt, når man vil beskrive den gennemsnitlige emissionsprofil for eksisterende elproduktion. Det er mindre egnet til spørgsmålet: Hvad sker der med emissionerne, hvis et datacenter øger sit forbrug med én ekstra MWh?

Her er en marginal emissionsfaktor mere relevant. Den forsøger at beskrive, hvilken produktion og hvilke emissioner der ændres, når efterspørgslen ændres. På kort sigt kan det være de kraftværker, der regulerer op og ned i den konkrete time. Ved en stor, permanent ny datacenterbelastning kan spørgsmålet række længere og omfatte investeringer i ny produktion, lagring og netkapacitet.

Et europæisk working paper fra Brander, Haxhimusa, Liebensteiner og Taschini estimerer marginale emissionsfaktorer på tværs af 29 europæiske elmarkeder og finder, at gennemsnitsfaktorer kan give væsentligt misvisende resultater for blandt andet load shifting og investeringsbeslutninger. Studiet er et working paper og bruges her som metodeunderstøttelse, ikke som et nyt universelt europæisk marginaltal.

*Kilde: Brander et al. (2025), From Average to Marginal: Estimating Emission Factors in Europe's Electricity Markets, SSRN working paper.*

#### Kort eller langt marginalperspektiv?

Et ekstra AI-job, der flyttes fra kl. 18 til kl. 03, er et andet beslutningsproblem end etableringen af et nyt 300 MW datacenter med mange års forventet drift. Den første situation handler primært om den kortsigtede systemreaktion. Den anden kan påvirke investeringsbehovet og kræver derfor et long-run eller capacity-expansion-perspektiv.

I denne bog bruges derfor en fast beslutningsregel: short-run eller operating-margin-metoder til mindre og kortsigtede loadændringer; long-run/build-margin eller kapacitetsudvidelsesscenarier til store, varige ændringer. Metodevalget skal stå synligt ved resultatet.

*Kilde/ramme: GHG Protocol Scope 2 Guidance og den europæiske litteratur om marginale emissionsfaktorer omtalt ovenfor.*

**Påstanden under lup**

**Påstand:** Hvis to datacentre bruger lige mange kWh, har de samme $CO_2$-aftryk fra elektriciteten.

**Vurdering:** Forkert som generel påstand.

**Kort svar:** Elforbruget kan være identisk, mens $CO_2$-resultatet varierer kraftigt med geografi, produktionsmix, tidspunkt og den emissionsmetode, der anvendes.

**Hvad siger evidensen?** EEA's 2024-landefaktorer spænder i vores fem europæiske eksempler fra 6 g $CO_2$e/kWh i Sverige til 566 g i Polen. UNECE viser samtidig store teknologiforskelle mellem fossil og lavemissionsproduktion.

**Det afgørende forbehold:** Et nationalt gennemsnit, en teknologispecifik livscyklusfaktor, market-based Scope 2 og en marginal emissionsfaktor er fire forskellige størrelser. De må ikke byttes rundt, blot fordi alle udtrykkes i g $CO_2$e/kWh.

## Hvad bør følge med et $CO_2$-tal?

Når et $CO_2$-tal for AI eller datacentre bruges i en rapport, pressehistorie eller politisk beslutning, bør mindst fem oplysninger kunne findes tæt på tallet: det fysiske elforbrug, geografien, perioden, emissionsfaktoren og den valgte metode. Hvis analysen handler om ændret efterspørgsel, skal det desuden fremgå, om effekten er beregnet med gennemsnitlig eller marginal elektricitet.

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Elforbrug | Hvor mange kWh/MWh/GWh ligger bag $CO_2$-tallet? |
| Geografi | Hvilket land, prisområde eller net gælder faktoren for? |
| Tid | Hvilket år – og eventuelt hvilke timer – beskrives? |
| Systemgrænse | Er faktoren drift, livscyklus eller et andet afgrænset regnskab? |

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Regnskabsmetode | Location-based, market-based, gennemsnitlig fysisk eller marginal? |
| Kildestatus | Målt, officielt gennemsnit, leverandørdata eller modelleret konsekvens? |

*Tabel 9.3. Kontrolramme for elrelaterede $CO_2$-påstande i bogen.*

Geografi er dermed ikke en fodnote til AI's klimaaftryk. Den er en del af selve beregningen. Den samme kWh kan indgå i flere forskellige emissionsregnskaber, og hvert regnskab skal navngives, før tallet kan fortolkes.

I næste kapitel følger vi den samme disciplin over i vandregnskabet. Her bliver forskellen mellem mængde, vandkilde, withdrawal, consumption, returflow og lokal vandstress mindst lige så vigtig som forskellen mellem kWh og $CO_2$ har været i dette kapitel.

### Kilder nævnt i kapitlet


• European Environment Agency (EEA): Greenhouse gas emission intensity of electricity generation in Europe – 2024 country data and methodology

• UNECE (2022): Life Cycle Assessment of Electricity Generation Options

• GHG Protocol: Scope 2 Guidance – location-based and market-based accounting

• GHG Protocol (29 July 2026): Scope 2 consultation feedback and corporate standards development update

• IEA (2022): Advancing Decarbonisation through Clean Electricity Procurement

• Google Cloud (2026): Carbon free energy for Google Cloud regions – 2024 regional data

• Brander, Haxhimusa, Liebensteiner & Taschini (2025): From Average to Marginal: Estimating Emission Factors in Europe's Electricity Markets, SSRN working paper


## Kapitel 10

# Vandet: en liter er ikke bare en liter

Vand er blevet et af de mest citerede symboler på AI's fysiske omkostninger. Fortællingen er let at forstå: datacentre skal køles, køling kan bruge vand, og derfor må hver AI-forespørgsel have et bestemt vandforbrug.

Den grundlæggende sammenhæng er reel. Datacentre og den elektricitet, de bruger, kan have et vandaftryk. Men ordet "vandforbrug" bliver ofte brugt om flere forskellige størrelser på én gang. Det kan betyde vand, der trækkes ind på et site, vand der fordamper, vand der udledes igen, drikkevand, overfladevand, havvand eller indirekte vand knyttet til elproduktionen.

En liter er derfor en fysisk mængde, men den fortæller ikke alene, hvor stor belastningen er. For beslutningstagere er mindst tre spørgsmål nødvendige: Hvor kommer vandet fra? Hvor meget bliver reelt forbrugt frem for returneret? Og hvor presset er den lokale vandressource på det tidspunkt, hvor vandet bruges?

### Withdrawal, discharge og consumption er forskellige størrelser

GRI 303 definerer water withdrawal som alt vand, der trækkes fra overfladevand, grundvand, havvand eller en tredjepart til brug i organisationen. Water discharge er det vand, der efter brug frigives til en modtager, mens water consumption beskriver den del af det indtagne vand, som ikke bliver tilgængelig igen for andre brugere eller økosystemet i rapporteringsperioden.

I et komplet vandregnskab kan consumption ofte beregnes som withdrawal minus discharge. Beregningen kræver dog, at systemgrænsen omfatter alle relevante vandkilder, returstrømme og ændringer i lagring. Hvis man kun kender én indtagsstrøm og én udledning, er differencen ikke automatisk lig med fordampning.

*Kilde: Global Reporting Initiative, GRI 303: Water and Effluents 2018 samt GRI Standards Glossary 2025. Definitionerne anvendes her som generel vandregnskabsramme.*

| Begreb | Hvad måles? | Hvad må man ikke antage? |
|---|---|---|
| Withdrawal / indtag | Alt vand, der trækkes ind fra en vandkilde eller tredjepart. | At hele mængden er "forbrugt" eller forsvundet. |
| Discharge / retur | Vand, der efter brug frigives til overfladevand, grundvand, hav eller tredjepart. | At returneret vand nødvendigvis har samme temperatur eller kvalitet som før. |
| Consumption / forbrug | Den del af withdrawal, der ikke returneres til andre brugere eller økosystemet i perioden, fx ved fordampning eller indbygning i processer. | At consumption kan beregnes sikkert, hvis vandbalancen kun dækker enkelte strømme. |
| Water stress | Presset på en lokal vandressource i forhold til tilgængelighed og efterspørgsel. | At samme liter har samme lokale betydning overalt. |

*Tabel 10.1. Redaktionel syntese af GRI's vanddefinitioner og EEA's vandstressmetode.*

## EU kræver allerede vanddata fra større datacentre

EU's fælles rapporteringsordning for datacentre gør vand til et eksplicit bæredygtighedsparameter. Kommissionens delegerede forordning (EU) 2024/1364 kræver blandt andet rapportering af total water input, WIN, samt den del af inputtet der er potable water. WIN skal omfatte alle vandmængder, som passerer datacenterets grænse og bruges til datacenterfunktioner.

Forordningen definerer samtidig Water Usage Effectiveness, WUE, som WIN divideret med IT-udstyrets energiforbrug. Det giver et nyttigt effektivitetsmål, fordi vandinput kan sættes i forhold til den leverede IT-energi.

Her er terminologien vigtig. EU's WIN er et inputmål ved datacentergrænsen. Det bør ikke automatisk sidestilles med consumptive water use i en bredere vandregnskabsdefinition. Hvis man vil vide, hvor meget vand der reelt forlader det lokale kredsløb gennem f.eks. fordampning, skal returstrømme og øvrige dele af vandbalancen med.

*Kilde: Commission Delegated Regulation (EU) 2024/1364, Annex II og Annex III. WIN måles efter EN 50600-4-9 WUE Category 2 eller Category 1, og WUE beregnes som WIN/EIT.*

| Måling | Eksempel | Fortolkning |
|---|---|---|
| Water input | 10.000 m³/år | Mængden, der går ind over den valgte datacentergrænse. |
| IT-energi | 20.000 MWh/år | Den årlige energi til IT-udstyret. |
| WUE | 0,50 m³/MWh = 0,50 L/kWh | Et effektivitetsmål for input pr. IT-energi - ikke i sig selv et mål for lokal vandknaphed. |

*Illustrativ beregning. Tallene er konstruerede og bruges kun til at vise WUE-formlen. De beskriver ikke et konkret datacenter.*

## Vandkilden ændrer betydningen af det samme volumen

Et datacenter kan bruge drikkevand, kommunalt procesvand, grundvand, overfladevand, reclaimed water eller havvand. Kilden ændrer både den lokale ressourcekonkurrence og den behandling, der kræves før og efter brug.

Googles europæiske datacentre giver to tydelige eksempler. I Saint-Ghislain i Belgien køles serverne med vand fra en industrikanal. I Hamina i Finland bruger kølesystemet havvand fra Den Finske Bugt. De to sites viser, hvorfor "datacenteret bruger vand" er for upræcist som miljøbeskrivelse, hvis vandkilden udelades.

*Kilder: Google Data Centers, St. Ghislain, Belgium; Google Data Centers, Hamina, Finland. Kilderne dokumenterer vandkilden og køleprincipperne, men bruges ikke som fuldt site-vandregnskab.*

## Saint-Ghislain: tilladelse, faktisk indtag og retur er tre forskellige tal

Saint-Ghislain er en særlig nyttig europæisk case, fordi en officiel parlamentarisk besvarelse fra Wallonien giver konkrete tal for både tilladelsen og den rapporterede vandstrøm.

Crystal Computing, Googles operatørselskab på sitet, har en tilladelse til at pumpe op til 18.000 m³ om dagen fra Nimy-Blaton-kanalen. Det er et maksimalt tilladt niveau - ikke en oplysning om det faktiske daglige forbrug.

For 2022 blev der til de wallonske myndigheder rapporteret 1.393.522 m³ ikke-drikkevands-egnet overfladevand trukket ind. Af denne vandstrøm blev 373.458 m³ rapporteret returneret til vandløb.

*Kilde: Parlement de Wallonie, svar af 26. oktober 2023 om Crystal Computing/Google i Saint-Ghislain. Myndighedssvaret oplyser både pumpetilladelsen og den rapporterede 2022-indtag/retur.*

| Størrelse | 2022 / tilladelse | Hvad tallet betyder |
|---|---|---|
| Maksimal pumpetilladelse | 18.000 m³/dag | Et juridisk loft. Ikke faktisk dagligt indtag. |
| Rapporteret overfladevandsindtag | 1.393.522 m³/år | Faktisk rapporteret indtag fra den ikke-potable overfladevandskilde i 2022. |
| Rapporteret retur til vandløb | 373.458 m³/år | Den del af den oplyste overfladevandsstrøm, der blev rapporteret returneret. |
| Difference | 1.020.064 m³/år | Matematisk difference mellem de to oplyste størrelser. Ikke automatisk lig med fordampning eller samlet consumption. |

*Tabel 10.2. Beregnet fra den officielle wallonske myndighedsbesvarelse. Differencen er 73,2 % af det rapporterede overfladevandsindtag; returstrømmen er 26,8 %. Den fulde site-vandbalance kræver også viden om andre vandtyper og relevante udledninger.*

Det årlige indtag svarer til et gennemsnit på cirka 3.818 m³ om dagen. Det er omkring 21 procent af den maksimale daglige pumpetilladelse, hvis årsvolumen blot fordeles over 365 dage. Beregningen viser, hvorfor et tilladelsesloft ikke må omtales som et faktisk forbrugstal.

*Illustrativ omregning baseret på den rapporterede 2022-årsmængde: 1.393.522 m³ / 365 dage ≈ 3.818 m³/dag. Drift og indtag kan variere betydeligt fra dag til dag.*

## Den lokale vandstress skal med i regnskabet

Selv et korrekt consumption-tal fortæller kun en del af påvirkningen. Vandets lokale betydning afhænger af, hvor stor en del af den tilgængelige vedvarende vandressource der allerede bruges, og hvordan situationen ændrer sig gennem året.

European Environment Agency, EEA, bruger Water Exploitation Index Plus, WEI+, til at beskrive presset på de vedvarende ferskvandsressourcer. Indekset sætter vandforbrug i forhold til de

vedvarende vandressourcer i den konkrete geografi og periode. EEA betragter værdier over 20 procent som vandstress.

Den seneste EEA-dataserie, offentliggjort i januar 2026 med data til og med 2023, findes på sub-basin-niveau og opdeles i årets fire kvartaler. Det er metodisk vigtigt: en national årsaverage kan skjule en lokal sommerperiode med langt større pres.

EEA understreger samtidig, at det paneuropæiske datasæt kræver gap filling og kan afvige fra nationale opgørelser. WEI+ bør derfor bruges som en konsistent europæisk screening og, hvor beslutningen kræver det, suppleres med lokale myndighedsdata.

*Kilde: European Environment Agency, Water Exploitation Index Plus (WEI+) dataset, publiceret 29. januar 2026, tidsdækning 2000-2023. EEA angiver vandstress ved WEI+ over 20 %.*

| Spørgsmål | Utilstrækkeligt svar | Bedre svar |
|---|---|---|
| Hvor meget vand? | "1 mio. liter" | Withdrawal, discharge og consumption angivet separat. |
| Hvilken ressource? | "Vand" | Drikkevand, grundvand, overfladevand, reclaimed water eller havvand. |
| Hvor? | Landegennemsnit | Lokalt catchment/sub-basin hvor det er relevant. |
| Hvornår? | Årstal alene | Sæson/kvartal ved områder med væsentlig variation. |
| Hvor presset er ressourcen? | Ingen kontekst | WEI+ eller relevant lokal vandstressindikator. |

*Tabel 10.3. Kontrolramme for vandpåstande i bogen.*

## Datacenterets vand og elproduktionens vand er to forskellige led

Vandaftrykket kan også ligge uden for datacenterets egen grund. Elektricitetsproduktion kan bruge eller forbruge vand, og den indirekte vandkomponent afhænger derfor af elmixet - på samme måde som $CO_2$-aftrykket i kapitel 9 afhænger af geografi og produktionsteknologi.

Det er vigtigt at holde direkte site-vand og indirekte elektricitet-relateret vand adskilt, før de eventuelt summeres i en livscyklusanalyse. Et

datacenter med lavt onsite-vandinput kan stadig have et indirekte vandaftryk gennem strømforsyningen. Omvendt fortæller et højt onsite-vandinput ikke alene noget om vandstress, hvis kilden er f.eks. havvand eller en lokal ressource med lavt pres.

Det geografiske princip fra elektricitetskapitlet gælder derfor også her: globale vandkoefficienter må ikke anvendes som direkte proxy for et europæisk site, når vandkilde, klima, køling eller elmix ændrer resultatet.

*Metoderamme: Li et al., Making AI Less "Thirsty"; EEA WEI+.*

### Fra forskningsresultat til "et glas vand pr. prompt"

Den mest kendte AI-vandpåstand kan spores til studiet Making AI Less "Thirsty" af Pengfei Li, Jianyi Yang, Mohammad A. Islam og Shaolei Ren. Forskerne modellerede GPT-3's operationelle vandaftryk og fandt, at en flaske på 500 ml under de analyserede forhold svarede til omtrent 10-50 mellemstore responses, afhængigt af hvor og hvornår modellen blev kørt.

Det svarer matematisk til cirka 10-50 ml pr. response i netop den modelberegning. Studiet skrev ikke, at hver enkelt ChatGPT-prompt bruger 500 ml.

Forskellen er væsentlig. Når "500 ml for 10-50 responses" bliver til "500 ml pr. prompt", ændres det oprindelige resultat med en faktor 10-50, før diskussionen overhovedet når spørgsmål om nyere hardware, køleteknologi eller systemgrænse.

Nyere operatørdata viser samtidig, hvor følsomt per-query-tallet er over for metode. Google estimerede i 2025 onsite water consumption for en median Gemini Apps-tekstprompt til 0,26 ml med virksomhedens bredere produktionsgrænse. Tallet er ikke et fuldt livscyklus-vandaftryk og er derfor heller ikke direkte sammenligneligt med Li et al.'s model, som inkluderede både onsite- og elektricitet-relateret operationelt vand.

*Kilder: Li et al., Making AI Less "Thirsty" (oprindeligt arXiv 2023, revideret version 2025 / Communications of the ACM 2025); Google (2025), Measuring the environmental impact of AI inference. De to per-response-tal har forskellige systemgrænser og må ikke behandles som samme måling.*

**Påstanden under lup**
**Påstand:** En AI-prompt bruger et glas eller en halv liter vand.
**Vurdering:** Misvisende som generel påstand.
**Kort svar:** Den kendte 500 ml-reference gjaldt cirka 10-50 mellemstore GPT-3-responses under bestemte model- og lokationsantagelser - ikke én prompt. Nyere målinger og estimater ligger anderledes, fordi model, datacenter, køling og systemgrænse er ændret.
**Hvor kommer påstanden fra?** Li et al. modellerede GPT-3 og beskrev omtrent en 500 ml flaske for 10-50 responses, afhængigt af hvornår og hvor modellen blev kørt.
**Hvad siger nyere evidens?** Google rapporterede 0,26 ml onsite water consumption for en median Gemini Apps-tekstprompt i maj 2025 med sin brede driftsgrænse. Det er en anden tjeneste og en smallere vand-systemgrænse end Li et al.'s samlede operationelle model.
**Det afgørende forbehold:** Der findes ikke én stabil liter-per-prompt-faktor for AI. Et forsvarligt tal skal angive model/tjeneste, opgave, geografi, periode, vandkilde og systemgrænse.

### "Zero water for cooling" er heller ikke det samme som "zero water"

Nye datacenterdesigns kan reducere eller fjerne løbende vandforbrug til bestemte køleprocesser. Det er en reel teknisk forbedring, men formuleringen skal knyttes til den konkrete systemgrænse.

Microsoft beskriver f.eks. nye datacenterdesigns, der er beregnet til at forbruge nul vand til køling. Det siger noget om kølesystemets onsite-driftsfase. Det siger ikke, at datacenteret eller AI-tjenesten har et samlet vandaftryk på nul, fordi blandt andet semiconductorproduktion, byggeri og elproduktion kan ligge uden for denne grænse.

*Kilde: Microsoft Environmental Sustainability Report 2025 og Microsoft Datacenters sustainability disclosures. Formuleringen "zero water for cooling" behandles her efter den systemgrænse, Microsoft selv angiver.*

## Hvad bør følge med et vandtal?

Når et vandtal for AI eller datacentre bruges som oplysningsgrundlag, bør læseren kunne se mere end selve volumenet. I praksis er et vandtal først beslutningsrelevant, når vandstrøm, kilde, geografi, periode og systemgrænse fremgår.

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Vandstrøm | Er tallet withdrawal/input, discharge/return eller consumption? |
| Kilde | Er det drikkevand, grundvand, overfladevand, reclaimed water, havvand eller tredjepartsforsyning? |
| Geografi | Hvilket catchment, sub-basin eller lokalområde påvirkes? |
| Tid | Hvilket år og, ved sæsonvariation, hvilket kvartal eller hvilken periode? |
| Vandstress | Hvor presset er den lokale vedvarende vandressource? |
| Systemgrænse | Er det onsite datacenterdrift, elektricitet-relateret vand eller et fuldt livscyklusaftryk? |
| Attribution | Er tallet for hele datacenteret, en konkret AI-workload eller en modelberegnet andel? |

*Tabel 10.4. Bogens kontrolramme for vanddata.*

Et stort liter-tal kan være relevant. Et lille liter-tal kan også være relevant. Hverken størrelsen eller den retoriske effekt fortæller i sig selv, hvor alvorlig påvirkningen er. Først når vandstrømmen, kilden og den lokale knaphed er kendt, kan volumenet fortolkes.

Det gør vandkapitlet parallelt med kapitel 9 om elektricitet. Samme antal kWh kan give forskellige $CO_2$-resultater. På samme måde kan samme antal liter have meget forskellig lokal betydning.

I næste kapitel udvider vi systemgrænsen igen. Køling er kun én del af datacenterets fysiske system. Strømforsyning, bygning, netværk, storage og den øvrige støtteinfrastruktur har også omkostninger, som kan blive usynlige, hvis analysen stopper ved GPU'en.

### Kilder nævnt i kapitlet


• Global Reporting Initiative (GRI): GRI 303 Water and Effluents 2018 samt GRI Standards Glossary 2025

• Commission Delegated Regulation (EU) 2024/1364 om fælles EU-ratingordning og rapportering for datacentre

• European Environment Agency (EEA): Water Exploitation Index Plus (WEI+) dataset, 2000-2023, publiceret 29. januar 2026

• Parlement de Wallonie (26. oktober 2023): myndighedssvar om Crystal Computing/Google Saint-Ghislain, pumpetilladelse, indtag og retur

• Google Data Centers: St. Ghislain, Belgium; Hamina, Finland

• Li, Pengfei; Yang, Jianyi; Islam, Mohammad A.; Ren, Shaolei: Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

• Google (2025): Measuring the environmental impact of AI inference

• Microsoft (2025): Environmental Sustainability Report og datacenter-water disclosures


## Kapitel 11

# Køling, bygningen, strømforsyningen, netværk og storage

Når energiforbruget i AI omtales, ender opmærksomheden ofte ved GPU'en. Det er forståeligt, fordi acceleratorerne står for en stor del af den beregning, der gør moderne generativ AI mulig. GPU'en arbejder imidlertid som en del af et større fysisk system. Serveren har CPU, hukommelse og lokale drev. Mange acceleratorer kobles sammen gennem højhastighedsnetværk. Modeller, checkpoints og data ligger på storage. Strømmen passerer gennem transformere, UPS-anlæg og strømfordeling, og den varme, som udstyret afgiver, skal transporteres væk.

Kapitel 3 introducerede dette som et spørgsmål om systemgrænse. Her bliver grænsen konkret. Et GPU-tal, et server-tal, et IT-tal og et site-tal

kan alle være korrekte samtidig, men de beskriver forskellige dele af infrastrukturen.

ITU-T L.1801 fra 2026 placerer netop storage, datatransmission og site-specifik støtteinfrastruktur inden for AI-systemets potentielle livscyklusgrænse, når de er nødvendige for den funktion, der vurderes. Et europæisk LCA-studie af Lucie 7B på Jean Zay-supercomputeren går tilsvarende længere end acceleratorerne og inkluderer fremstilling og drift af compute, midlertidig storage, systemadministration, strømforsyningskæde og køling.

**Figur 11.1 – Fra accelerator til datacenterets elmåler**

*Kilde/ramme: Forfatterens systemgrænseoversigt med udgangspunkt i ITU-T L.1801 (2026), Lucie 7B/Jean Zay-LCA (2026) og EU's PUE-definition i Delegated Regulation (EU) 2024/1364.*

## PUE er nyttigt – men fortæller kun én del af historien

Power Usage Effectiveness, PUE, er den mest udbredte indikator for datacenterets energioverhead. EU's fælles rapporteringsordning definerer PUE som datacenterets samlede energiforbrug, E_DC, divideret med energien til IT-udstyret, E_IT.

En PUE på 1,20 betyder derfor, at datacenteret bruger 1,20 energienheder ved site-grænsen for hver 1,00 energienhed, der registreres som IT-energi. De ekstra 0,20 enheder ligger i facility-laget, f.eks. køling, strømkonvertering og andre støttefunktioner.

PUE er nyttig til at se, hvor stor en facility-overhead der ligger oven på IT-energien. Den fortæller derimod ikke, hvor meget af IT-energien der ender som nyttigt AI-arbejde. CPU'er, hukommelse, storage, netværk og idle kapacitet ligger allerede inde i E_IT. PUE siger heller ikke noget om elmix, $CO_2$, vandstress eller embodied impacts fra bygningen og hardwaren.

**Tabel 11.1 – Hvad PUE fortæller, og hvad den ikke fortæller**

| Spørgsmål | Kan PUE besvare det? | Forklaring |
|---|---|---|
| Hvor meget site-energi bruges pr. IT-energi? | Ja | Det er indikatorens kerne: E_DC / E_IT. |
| Hvor stor er facility-overheaden? | Ja, indirekte | PUE-1 angiver overhead i forhold til IT-energien. |
| Hvor meget energi bruger GPU'en? | Nej | GPU'en er kun én del af IT-energien. |
| Hvor meget nyttigt AI-arbejde produceres pr. kWh? | Nej | Kræver workload-, kvalitet- og udnyttelsesdata. |
| Hvor stort er $CO_2$-aftrykket? | Nej | Kræver geografi, tidspunkt og emissionsmetode. |
| Hvor stort er vandaftrykket? | Nej | Kræver WUE/vandstrømme, kilde og lokal vandstress. |
| Hvor stort er livscyklusaftrykket? | Nej | Kræver også hardware, bygning, netinfrastruktur og end-of-life. |

*Kilde: Commission Delegated Regulation (EU) 2024/1364, Annex III; egen metodisk syntese. Tabellen bruger PUE i den EU-definerede betydning.*

## Et gennemregnet PUE-eksempel

Antag et datacenter med et konstant IT-load på 10 MW. Over et år svarer det til 87,6 GWh IT-energi, hvis belastningen holdes konstant: 10 MW × 8.760 timer.

Ved PUE 1,20 bliver site-forbruget 105,12 GWh om året. Facility-overheaden er 17,52 GWh. Hvis samme IT-load kan drives ved PUE 1,10, falder site-forbruget til 96,36 GWh, og overheaden til 8,76 GWh. Forskellen er 8,76 GWh om året.

Eksemplet viser også en lille matematisk detalje, som ofte forsvinder i kommunikationen. Ved PUE 1,20 er facility-overheaden 20 procent af IT-energien, men kun 16,7 procent af det samlede site-forbrug. Nævneren ændrer sig.

**Tabel 11.2 – Illustrativt PUE-regneeksempel ved 10 MW konstant IT-load**

| Størrelse | PUE 1,20 | PUE 1,10 |
|---|---|---|
| IT-energi pr. år | 87,60 GWh | 87,60 GWh |
| Samlet site-energi | 105,12 GWh | 96,36 GWh |
| Facility-overhead | 17,52 GWh | 8,76 GWh |
| Overhead som andel af site-energi | 16,7 % | 9,1 % |

*Illustrativ beregning. Tallene er konstruerede for at vise PUE-matematikken og beskriver ikke et konkret datacenter. Beregning: E_DC = PUE × E_IT.*

## Køling flytter energi- og vandregnskabet

Køling er en af de mest synlige støttefunktioner, men den kan ikke vurderes isoleret fra klima, vand og udstyrstæthed. Luftkøling, evaporativ køling og direkte væskekøling flytter belastningerne forskelligt. En løsning kan reducere elforbruget til ventilation og samtidig ændre vandforbruget. En anden kan bruge meget lidt onsite-vand, men kræve mere elektricitet til kompressorer eller andre komponenter.

Applied Energy-studiet fra 2026 sammenligner flere komplette datacenterarkitekturer frem for kun servere. I forfatternes simulationsramme giver høj-densitets, liquid-cooled AI-designs det laveste miljøaftryk pr. eFLOP blandt de sammenlignede GPU-baserede arkitekturer. Resultatet er værdifuldt som arkitektursammenligning, men det er ikke en universel regel om, at væskekøling altid er miljømæssigt bedst. Model, udnyttelse, elmix, vandforhold og systemgrænse kan ændre rangordningen.

Lucie 7B-casen viser en anden europæisk mekanisme. Jean Zay H100-partitionen rapporteres med WUE på 0,07 L/kWh og en Energy Reuse Factor på 0,37, fordi en del af overskudsvarmen leveres til et urbant varmenet. Dermed bliver varme ikke kun en overhead, der skal fjernes; under bestemte lokale forhold kan den også få en sekundær funktion. Om varmegenvinding giver reel nettogevinst afhænger blandt andet af temperatur, sæson, alternativ varmeproduktion og om der findes en modtager tæt på datacenteret.

### Målte driftsdata kan forværres under kraftig udbygning

NEXTDC's FY26-rapport giver en aktuel measured-data-case fra en datacenterportefølje i kraftig ekspansion. Selskabets gennemsnitlige PUE steg fra 1,44 til 1,49. Reuters rapporterede samtidig en WUE på 2,40 liter/kWh mod 2,25 året før.

Casen er et nyttigt korrektiv til forestillingen om, at nyere datacentre automatisk giver stadigt bedre observerede effektivitetsnøgletal. Commissioning, delvist udnyttede anlæg, ændrede workloads og måleforhold kan påvirke PUE og WUE under en udbygningsfase. NEXTDC-casen er australsk og skal derfor ikke bruges som direkte proxy for europæiske driftsforhold.

NEXTDC beskriver også en vand-energi-afvejning: mindre afhængighed af drikkevand og mere tør eller vanduafhængig køling kan reducere onsite-vandbehovet, men kan samtidig kræve højere kapitaludgifter eller større energiforbrug. Det understreger, at én forbedret indikator kan flytte belastningen til en anden del af regnskabet.

Intensitetsmål og absolut forbrug bør derfor vises hver for sig. En bedre PUE, WUE eller Wh pr. token fortæller, hvordan ressourceforbruget udvikler sig pr. enhed; den fortæller ikke alene, om det samlede el- eller vandforbrug falder, når kapaciteten og aktiviteten samtidig vokser.

*Kilder: NEXTDC, FY26 Annual Report, offentliggjort 28. august 2026; Reuters, 28. august 2026. PUE og WUE er driftsindikatorer og må ikke læses som mål for effektiviteten af selve AI-beregningen.*

### Strømforsyningskæden ligger mellem nettet og GPU'en

Elektriciteten går sjældent direkte fra det offentlige net til acceleratorerne. Datacenteret har transformering, koblingsanlæg, UPS, strømfordeling og ofte redundans, så kritiske workloads kan fortsætte under fejl. Hvert led har materialer, kapitalomkostninger og energitab.

Dette er en del af forklaringen på, at site-forbrug og IT-forbrug er forskellige. Det er også en livscykluspost. Lucie 7B-LCA'en medregner power chain som en særskilt del af infrastrukturen, mens den

europæiske whole-systems analyse af 36 datacentre i Danmark, Tyskland og Norge inkluderer både bygningskonstruktion og IT-infrastruktur frem for kun den løbende serverenergi.

I den europæiske 36-site-model står driftsfasen for cirka 65–91 procent af det samlede modellerede klimaaftryk. I de renere elsystemer bliver fremstillingsfasen relativt vigtigere og kan i studiets cases nå op mod 35 procent. Det er samme mekanisme, vi så i kapitel 6: når driftsstrømmen bliver mindre $CO_2$-intensiv, fylder embodied impacts relativt mere.

Studiet er vigtigt som europæisk whole-system-evidens, men det er ikke et målt AI-only datasæt. Forskerne bruger reelle datacenterlokationer og offentlig area/power-kapacitet, mens flere interne parametre som PUE, rack power og udnyttelse må antages, fordi operatørerne ikke offentliggør dem.

### Netværket kan være en reel del af inferenceenergien

Store AI-modeller kører ofte på mange acceleratorer samtidig. Det kræver, at data og modeltilstand flyttes mellem GPU'er gennem højhastighedsforbindelser. Ved tensor parallelism, expert parallelism og andre distribuerede konfigurationer bliver kommunikation en del af det samlede energiregnskab.

EnergyLens-studiet fra 2026 er nyttigt, fordi det modellerer compute og communication samlet i multi-GPU inference og validerer modellen på Llama 3 og Qwen3-MoE. Forfatterne finder betydelige forskelle mellem konfigurationer og viser, at compute-communication overlap ikke kan optimeres sikkert alene ud fra intuition eller latency.

Det giver os en klar metodekonklusion, men ikke en fast netværksprocent. Kommunikationens andel afhænger af modelarkitektur, paralleliseringsstrategi, antal GPU'er, interconnect, batch, sekvenslængde og graden af overlap. Et generelt tillæg som 'læg 10 procent oven i GPU-forbruget for netværk' ville derfor være dårligt dokumenteret.

### Storage og databevægelse kan ikke altid ignoreres

Storage optræder flere steder i AI-livscyklussen. Træningsdata skal lagres og læses. Checkpoints kan skrives gentagne gange under træning. Modeller distribueres til serving-systemer. Retrieval-baserede løsninger læser dokumenter og embeddings, og logs kan gemmes til kvalitet, sikkerhed og drift.

Energien til storage er ofte mindre synlig end acceleratorens effekt, og dens betydning varierer kraftigt med workload. Derfor bør storage ikke automatisk antages at være stor, men heller ikke sættes til nul uden begrundelse. ITU-T L.1801 behandler både data storage og data transmission som potentielle driftsaktiviteter, når de er relevante for den analyserede AI-funktion.

Lucie 7B-studiet medtager midlertidig storage og systemadministration i training-casen. Det er metodisk vigtigt, fordi et træningsjob ellers kan fremstå som om GPU-timerne eksisterer uden data, checkpoints og den systemdrift, som gør træningen mulig.

### En europæisk case med hele infrastrukturen: Lucie 7B på Jean Zay

Lucie 7B er særlig relevant for denne bog, fordi den er en europæisk LLM-case med en usædvanligt bred dokumenteret systemgrænse. Modellen blev trænet på H100-partitionen af Jean Zay-supercomputeren hos IDRIS/CNRS i Frankrig, og LCA'en omfatter både fremstilling, drift og end-of-life for den relevante hardwareinfrastruktur.

Forfatterne opgør H100-partitionens årlige klimaaftryk til 417,5 ton $CO_2e$, omtrent ligeligt fordelt mellem fremstilling og drift i deres model. Selve Lucie 7B-træningen brugte 574.564 H100 GPU-timer og blev beregnet til cirka 21 ton $CO_2e$ inklusive amortiseret hardwarefremstilling. Onsite-vandet til træningskampagnen blev opgjort til cirka 76 $m^3$.

Tallene kan ikke overføres direkte til et kommercielt hyperscale-datacenter. Jean Zay er et konkret fransk HPC-system med eget elmix, hardwaremix, WUE, varmegenvinding og allokeringsmetode. Casens

styrke er systemgrænsen: den viser, hvordan en AI-LCA kan gå fra GPU-timer til et mere komplet fysisk system uden at påstå, at alle datacentre har samme profil.

| **Påstanden under lup** |
| --- |
| **Påstand: Et datacenter med PUE 1,1 er cirka 91 procent energieffektivt.** |
| **Vurdering: Misvisende uden en præcis definition.** |
| Kort svar: Matematisk betyder PUE 1,1, at IT-energien udgør cirka 90,9 procent af site-energien. Det er ikke det samme som, at 90,9 procent af elektriciteten bliver til nyttigt AI-arbejde. |
| Hvad siger evidensen? EU definerer PUE som E_DC/E_IT. E_IT omfatter hele IT-laget, ikke kun acceleratorens aktive beregning. CPU, hukommelse, storage, netværk og idle kapacitet kan ligge inden for IT-energien. PUE indeholder heller ingen information om modelkvalitet, serverudnyttelse, elmix, vand eller embodied impacts.<br><br>Det afgørende forbehold: PUE er en facility-effektivitetsindikator. Den bør ikke omdøbes til en samlet AI-effektivitetsprocent. |

### Hvad bør følge med et systemenergital?

Et beslutningsrelevant energital bør derfor fortælle, hvor i kæden måleren eller modellen stopper. Hvis tallet er GPU-energi, skal det kaldes GPU-energi. Hvis det er serverenergi, skal CPU, hukommelse og andre komponenter fremgå. Hvis PUE bruges til at skalere til site-niveau, bør både PUE og IT-grundlaget være synligt.

Ved større distribuerede workloads bør analysen desuden oplyse, om network/fabric og storage er medregnet. Ved livscyklusanalyser bør bygning, power chain og hardwarefremstilling beskrives særskilt, så en læser kan se, om drift eller embodied impacts dominerer i den valgte geografi.

Det er sjældent nødvendigt at medtage alle poster med samme detaljeringsgrad. Materialitetsprincippet er vigtigere: de dele af

infrastrukturen, der kan ændre konklusionen, skal med eller begrundet afgrænses.

**Tabel 11.3 – Bogens kontrolramme for AI-systemets driftsinfrastruktur**

| Lag | Spørgsmål der bør kunne besvares |
|---|---|
| Accelerator | Er energien aktiv GPU/TPU-energi, TDP-estimat eller fysisk målt forbrug? |
| Server | Er CPU, HBM/DRAM, lokale drev og idle-kapacitet med? |
| Network/fabric | Er inter-GPU- og inter-node-kommunikation relevant og medregnet? |
| Storage | Indgår data, checkpoints, modeldistribution, retrieval og logs, når de er materielle? |
| Facility | Hvilken PUE, køleteknologi, UPS og strømfordeling anvendes? |
| Bygning/ livscyklus | Er konstruktion og hardwarefremstilling med, hvis analysen hævder et livscyklusaftryk? |
| Geografi | Hvilket elmix, vandforhold og klima gælder for det konkrete site? |
| Funktionel enhed | Hvad produceres: GPU-time, token, request, træningsrun eller accepteret resultat? |

*Kilde/ramme: Egen syntese baseret på ITU-T L.1801 (2026), EU 2024/1364, Lucie 7B/Jean Zay (2026), EnergyLens (2026) og Applied Energy 406 (2026).*

## Fra datacenteret til placeringen af AI

Kapitel 11 har udvidet driftsregnskabet fra accelerator til hele datacenterets tekniske kæde. PUE kan beskrive facility-overhead, men siger ikke, hvor effektiv selve AI-opgaven er. Netværk og storage kan være materielle, men deres bidrag varierer med arkitekturen. Køling kan flytte både energi- og vandregnskabet, mens embodied infrastructure får større relativ betydning i renere elsystemer. Samtidig skal intensitetsmål som PUE og WUE holdes adskilt fra det absolutte forbrug, især når kapaciteten udbygges hurtigt.

Det næste spørgsmål er, om hele denne infrastruktur overhovedet skal ligge i et centralt cloud-datacenter. Små modeller kan i nogle tilfælde

køre på telefoner, laptops eller edge-servere tæt på brugeren. Andre workloads kræver store distribuerede clusters. Kapitel 12 undersøger derfor cloud, edge og lokal AI med samme funktionelle regel som resten af bogen: systemerne skal sammenlignes på den opgave og kvalitet, de faktisk leverer.

### Kilder nævnt i kapitlet


- International Telecommunication Union, ITU-T L.1801 (2026): Guidelines for assessing the environmental impact of artificial intelligence systems.
- European Commission / EUR-Lex: Commission Delegated Regulation (EU) 2024/1364, Annex III – PUE, WUE, ERF og REF.
- Léobet, Marc; Lavallée, Pierre-François; Lorré, Jean-Pierre (2026): Life Cycle Assessment of Pre-training the Lucie 7B Open-Source Large Language Model on the Jean Zay Supercomputer.
- d'Orgeval, Alexandre et al. (2026): Generative AI impact assessment through a life cycle analysis of multiple data center typologies, Applied Energy 406, 127288.
- Song, Zhiye et al. (2026): EnergyLens: Predictive Energy-Aware Exploration for Multi-GPU LLM Inference Optimization.
- Hankendi, Can; Coskun, Ayse K.; Sovacool, Benjamin K. (2026): Beyond the carbon emissions of Artificial Intelligence (AI): A whole-systems energy and environmental sustainability analysis of datacenters in Denmark, Germany and Norway, Energy Research & Social Science 138, 104839.
- NEXTDC (2026): FY26 Annual Report.
- Reuters (28. august 2026): NEXTDC FY26 reporting, herunder WUE-udviklingen.

## Kapitel 12

# Cloud, edge og lokal AI

De fleste generative AI-tjenester opleves som cloudtjenester. Brugeren sender en forespørgsel, beregningen udføres i et datacenter, og resultatet kommer tilbage over nettet. Den model er fortsat nødvendig for de største modeller og de mest krævende workloads, men den er ikke længere den eneste tekniske mulighed.

Mindre sprogmodeller kan i stigende grad køre på telefoner, laptops, industrielle gateways og lokale servere. Andre systemer placerer en del af beregningen tæt på brugeren og resten i cloud. Dermed opstår et kontinuum fra on-device AI over lokal og regional edge til hyperscale-cloud.

Miljømæssigt er det fristende at gøre placeringen til et simpelt valg: lokal AI bruger mindre strøm, eller cloud er mere effektivt, fordi store datacentre udnytter hardware og køling bedre. Begge udsagn kan være rigtige i konkrete tilfælde. Ingen af dem kan bruges som generel regel.

Den relevante sammenligning er den samme som i resten af bogen: Hvilket system leverer den samme nyttige funktion ved det nødvendige kvalitetsniveau, og hvilke ressourcer kræver hele processen for at nå dertil?

**Figur 12.1 – AI kan placeres på et kontinuum fra enheden til cloud**

*Kilde/ramme: Forfatterens syntese baseret på Yan & Ding (2025), Li, Islam & Ren (2025), Tummalapalli et al. (2026), SparKV (2026) og Applied Energy 406 (2026).*

## Placeringer langs kontinuummet - og de systemgrænser der følger med

| Placering | Styrker | Typiske begrænsninger | Poster der let overses |
|---|---|---|---|
| On-device | Privatliv, offline, lav netafhængighed og lav latency | Hukommelse, varme, batteri og modelstørrelse | Embodied impact, opladnings-el, modeldownload og cloud-fallback |
| Lokal edge | Mere compute, lokal kontrol og lav latency | Dedikeret hardware og lokal drift | Idle kapacitet, køling, strøm og udnyttelse |
| Regional edge | Delt hardware tættere på brugeren | Mindre skala og batching end hyperscale | Netværk, elmix, PUE og redundans |
| Cloud | Store modeller, elasticitet og høj udnyttelse | Netafhængighed og central infrastruktur | PUE, netværk, storage, køling samt el- og vandprofil |
| Hybrid | Matcher opgaven til en tilstrækkelig platform | Routing og fallback kan give overhead | Lokal prøve, cloud-fallback og orchestration |

*Kilde/ramme: Egen syntese af den gennemgåede edge-, cloud- og hybridlitteratur. Tabellen beskriver typiske egenskaber, ikke faste rangordninger.*

## Et stærkt cloud-versus-edge-resultat - med en vigtig begrænsning

Li, Islam og Ren sammenlignede i 2025 generativ AI-inference på cloud-GPU'er og en Samsung Galaxy S24. I hovedforsøget kørte de samme modelarkitekturer og identiske input på Nvidia A100- eller L4-GPU'er i Google Colab og på telefonen. For de modeller, der kunne køre begge steder, fandt studiet mere end 90 procent lavere inferenceenergi på edge-platformen og mere end 80 procent lavere modelleret carbon- og vandaftryk i de analyserede scenarier.

Resultatet er vigtigt, fordi det viser, at on-device inference under bestemte forhold kan være langt mere energieffektivt end en cloudkørsel. Det er samtidig et case-studie, ikke en universel edge-faktor.

Forfatterne skriver eksplicit, at nøjagtighed og latency sættes til side i sammenligningen, og at både cloud- og edge-implementeringen kan optimeres yderligere.

Studiet kan derfor understøtte påstanden om, at edge kan være markant mere energieffektivt for visse modeller og workloads. Det kan ikke understøtte en regel om, at al lokal AI bruger 90 procent mindre energi end cloud.

**Tabel 12.1 – Hvad cloud-versus-edge-casen dokumenterer**

| Element | Li, Islam & Ren (2025) | Hvad vi kan bruge det til |
|---|---|---|
| Compute-platforme | Samsung Galaxy S24 versus Nvidia A100/L4 på Google Colab | Direkte case på on-device versus cloud inference |
| Model/input | Samme generative modeller og identiske input for sammenlignelige runs | Stærkere end at sammenligne helt forskellige modeltyper |
| Energi | Over 90 % lavere edge-energi i flere af de analyserede cases | Dokumentation for stort potentiale i den konkrete opsætning |
| Carbon/vand | Over 80 % lavere modelleret footprint i de analyserede cases | Viser at energiforskel kan forplante sig til miljøresultat |
| Begrænsning | Accuracy og latency holdes uden for hovedkonklusionen; cloud var ikke nødvendigvis optimalt tunet | Resultatet må ikke gøres til en universel edge-procent |

*Kilde: Li, Pengfei; Islam, Mohammad J.; Ren, Shaolei (2025), A Case Study of Environmental Footprints for Generative AI Inference: Cloud versus Edge, ACM SIGMETRICS Performance Evaluation Review 53(2), 21-26, DOI 10.1145/3764944.3764950.*

## Den samme model kan stadig opføre sig meget forskelligt på forskellige enheder

Et 2026-benchmark af Qwen 2.5 1.5B på fire edge-platforme viser, hvorfor peak-specifikationer ikke er nok. Tummalapalli og medforfattere testede en Raspberry Pi 5 med Hailo-10H NPU, Samsung Galaxy S24

Ultra, iPhone 16 Pro og en laptop med RTX 4050 under gentagne varme inferenceforløb.

RTX 4050-systemet leverede i studiet cirka 131,7 tokens/s ved gennemsnitligt 34,1 W, mens Raspberry Pi/Hailo-systemet leverede cirka 6,9 tokens/s ved 1,87 W. Energiforbruget pr. token var i de målte systemer omtrent samme størrelsesorden - cirka 297 mJ/token på laptop-GPU'en og 271 mJ/token på Hailo-systemet - selv om throughput varierede med en faktor omkring 19.

Telefonerne viste en anden begrænsning: varme. iPhone 16 Pro faldt fra en indledende throughput omkring 40 tokens/s til et sustained niveau omkring 24 tokens/s, efterhånden som den termiske tilstand gik fra normal til hot. Samsung-testen ramte en OS-styret GPU-frekvensbegrænsning under vedvarende belastning.

Det er en vigtig korrektion til forestillingen om, at lokal inference blot handler om lavere watt. En løsning skal også kunne levere den ønskede svartid, outputlængde og stabilitet under reel brug.

*Kilde: Tummalapalli, Pranay et al. (2026), LLM Inference at the Edge: Mobile, NPU, and GPU Performance Efficiency Trade-offs Under Sustained Load, arXiv:2603.23640. Resultaterne gælder Qwen 2.5 1.5B og de konkrete softwarestakke og enheder.*

### Modelkvalitet kan være den afgørende begrænsning

Hardware er kun den ene side af sammenligningen. Yan og Ding målte i 2025 mobile, edge- og cloudbaserede LLM-deployments i en smartphone-applikation. I deres setup kunne kun mindre modeller under cirka 4 milliarder parametre køre succesfuldt på de kraftige mobile enheder, og kompression kunne sænke hardwarekravet på bekostning af outputkvalitet. Mobile runs med meningsfuldt output havde i deres case latency over 30 sekunder, mens cloud lå under 10 sekunder.

Det gør kvalitetskravet centralt. Hvis en lokal model bruger meget lidt energi, men leverer et resultat, der ikke kan accepteres, er den lave

energi pr. inference ikke den relevante funktionelle enhed. Den efterfølgende retry eller cloud-fallback skal med i regnskabet.

For virksomhedsanvendelser kan lokal AI derfor være særligt attraktiv til afgrænsede, veldefinerede og gentagne opgaver, hvor en lille model eller specialiseret model faktisk opfylder kvalitetskravet. Større og mere åbne analyseopgaver kan fortsat kræve cloudmodeller med større kapacitet.

*Kilde: Yan, Xiao; Ding, Yi (2025), Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices, ACM FMSys 2025, DOI 10.1145/3722565.3727192.*

### Lokale systemer kan selv optimeres betydeligt

On-device energi er heller ikke en fast hardwareegenskab. EnerInfer-studiet fra 2026 viser, at den mest energieffektive kombination af NPU- og hukommelsesfrekvenser varierer med model, inference engine, platform og runtime-forhold. Frameworket forbedrede i forfatternes forsøg energieffektiviteten med op til 65 procent på telefoner uden at bryde det definerede quality-of-experience-krav.

Formuleringen 'op til' er vigtig. Resultatet er et systemoptimeringsstudie, ikke et løfte om 65 procent på alle telefoner. Dets værdi for denne bog ligger i en anden pointe: selv efter at placeringen er valgt, kan software- og runtimeindstillinger flytte energiregnskabet mærkbart.

*Kilde: Zou, Bohua et al. (2026), EnerInfer: Energy-Aware On-Device LLM Inference, arXiv:2606.23001.*

### Hybrid AI kan være mere interessant end et enten-eller

Cloud og lokal AI behøver ikke være gensidigt udelukkende. Et hybridt system kan vælge placering efter opgaven eller splitte beregningen mellem enhed og cloud.

SparKV fra 2026 er et konkret eksempel. Frameworket vurderer, om dele af KV-cache-arbejdet skal beregnes lokalt eller streames fra cloud, og tilpasser planen efter netværk og edge-ressourcer. På tværs af de testede modeller, datasæt og enheder reducerede SparKV per-request energi

med 1,5 til 3,3 gange sammenlignet med studiets baselines, samtidig med lavere time-to-first-token og ubetydelig påvirkning af svarkvaliteten.

Andre hybridstrategier bruger routing frem for split-compute: en lille model håndterer de lette forespørgsler, mens sværere opgaver sendes til en stærkere cloudmodel. Det kan reducere antallet af dyre cloudkald, men routeren, lokale forsøg og eventuelle fallback-kørsler er selv en del af systemet.

*Kilde: Liu, Hongyao et al. (2026), SparKV: Overhead-Aware KV Cache Loading for Efficient On-Device LLM Inference, arXiv:2604.21231. Hybrid-routing-princippet er desuden dokumenteret i Ding et al. (2024), Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing.*

### Et regneeksempel: den lokale prøve skal tælles med, når cloud bruges som fallback

Et simpelt illustrativt regnestykke viser, hvorfor success rate og fallback er nødvendige i sammenligningen.

Antag, at en lokal model bruger 0,20 Wh pr. forsøg, mens cloudmodellen bruger 1,00 Wh for den samme opgavetype. Systemet prøver altid lokalt først og sender kun opgaven til cloud, hvis det lokale resultat ikke opfylder kvalitetskravet.

Den gennemsnitlige energi pr. accepteret resultat bliver da 0,20 Wh plus fallback-andelen ganget med 1,00 Wh. Hvis 80 procent af de lokale svar accepteres, er fallback-andelen 20 procent, og gennemsnittet bliver 0,40 Wh. Hvis kun 20 procent accepteres lokalt, bliver gennemsnittet 1,00 Wh - det samme som cloud-only, fordi den lokale prøve i 80 procent af tilfældene blot bliver ekstra arbejde.

**Tabel 12.2 – Illustrativ hybridberegning ved lokal prøve før cloud-fallback**

| Lokal accept-rate | Fallback til cloud | Lokal energi | Forventet cloudenergi | Energi pr. accepteret resultat |
|---|---|---|---|---|
| 80 % | 20 % | 0,20 Wh | 0,20 Wh | 0,40 Wh |
| 50 % | 50 % | 0,20 Wh | 0,50 Wh | 0,70 Wh |
| 20 % | 80 % | 0,20 Wh | 0,80 Wh | 1,00 Wh |

*Illustrativ beregning med konstruerede tal. Forudsætning: hvert request forsøges lokalt først; cloud-fallback bruger 1,00 Wh; lokal kørsel bruger 0,20 Wh. Netværk, router og embodied impacts er udeladt. Eksemplet viser metoden, ikke et bestemt produkt.*

I det konstruerede eksempel ligger break-even ved 20 procent lokal accept-rate, men grænsen flytter sig, så snart energi, modelkvalitet, latency eller netværksforbrug ændres. Det relevante er, om den lokale model faktisk erstatter tilstrækkeligt mange cloudkørsler ved det krævede kvalitetsniveau.

## Geografien følger beregningen med ud på enheden

Kapitel 9 viste, at samme kWh kan få meget forskelligt $CO_2$-aftryk afhængigt af elproduktionens geografi og tidspunkt. Det gælder også, når AI flyttes fra cloud til enheden.

Cloudens elektricitet bruges i datacenterets region. En smartphone eller laptop oplades derimod hos brugeren. Hvis en opgave flyttes fra et datacenter i ét elsystem til millioner af enheder i andre elsystemer, kan energibesparelsen og $CO_2$-besparelsen derfor bevæge sig forskelligt.

Li, Islam og Ren modellerer netop carbon og vand ud fra de konkrete cloud- og edge-scenarier i deres studie. Resultatet må ikke omregnes direkte til et europæisk $CO_2$-tal uden at erstatte de geografisk følsomme faktorer med relevante europæiske data.

Den samme regel gælder vand. Cloudens onsite-køling forsvinder fra den enkelte inference, når beregningen flyttes lokalt, men elektriciteten til enheden har stadig et upstream-vandaftryk, som afhænger af det lokale elsystem.

## Den eksisterende enhed er en anden systemgrænse end en ny AI-enhed

Embodied impact gør on-device-sammenligningen endnu mere følsom. Hvis AI kører på en telefon eller laptop, som allerede er købt af andre grunde, kan det være rimeligt kun at allokere en begrænset del af enhedens produktionsaftryk til den ekstra AI-funktion. Hvis

organisationen derimod køber nye AI-PC'er, edge-servere eller acceleratorbokse specifikt for at flytte workload ud af cloud, bliver hardwareproduktionen en direkte del af beslutningen.

Derfor bør en analyse skelne mellem marginal brug af eksisterende hardware og investering i ny hardware. Det er samme kontrafaktiske princip, som vi bruger andre steder i bogen: Hvad ville være sket uden AI-beslutningen?

| **Påstanden under lup** |
| --- |
| **Påstand: Lokal AI er grønnere end cloud.** |
| **Vurdering: Kan være korrekt i konkrete cases, men er ikke en generel regel.** |
| Kort svar: Samme lille model kan i nogle målinger bruge markant mindre inferenceenergi på en moderne edge-enhed end på en cloud-GPU. Systemgevinsten afhænger samtidig af modelkvalitet, latency, fallback, hardwareudnyttelse, netværk, geografi og embodied impact. |
| Hvad siger evidensen? Li, Islam & Ren (2025) finder over 90 procent lavere inferenceenergi på edge i flere af deres sammenlignelige modelcases. Yan & Ding (2025) viser samtidig, at mobile enheder i deres setup kun kunne køre mindre modeller og havde markante latency- og kvalitetsbegrænsninger. Tummalapalli et al. (2026) viser desuden termisk throttling ved sustained mobil inference.<br><br>Det afgørende forbehold: Et edge-resultat må kun sammenlignes direkte med cloud, når de to systemer leverer den samme relevante funktion og et tilsvarende kvalitetsniveau. Hvis lokal inference kræver retry eller fallback, skal den ekstra kørsel tælles med. |

### En beslutningsramme for placering af inference

| Spørgsmål | Hvorfor det ændrer omkostningen |
| --- | --- |
| Kan den lokale model opfylde kvalitetskravet? | Hvis nej, er lav lokal energi ikke den relevante funktionelle enhed. |
| Hvor mange requests kan afsluttes lokalt? | Bestemmer hvor mange cloudkørsler der faktisk undgås. |

| Spørgsmål | Hvorfor det ændrer omkostningen |
|---|---|
| Hvilken latency kræves? | Cloud, edge og device har forskellige netværks- og compute-profiler. |
| Er hardwaren allerede til stede? | Ny dedikeret edge-hardware tilfører embodied impact og kapitalomkostning. |
| Hvor bruges elektriciteten? | Cloud og device kan ligge i forskellige el- og vandregimer. |
| Hvad er netværksbehovet? | Lokalt arbejde kan spare datatransmission; hybrid kan omvendt tilføre streaming/orchestration. |
| Er workload bursty eller konstant? | Delt cloudkapacitet kan udnyttes anderledes end dedikeret lokal hardware. |
| Hvad sker der ved fejl eller svære opgaver? | Retry, escalation og fallback skal med i energi og TCO. |

*Kilde/ramme: Egen syntese baseret på kapitlets empiriske studier og bogens princip om samme funktionelle enhed og cost per accepted outcome.*

### Cloud, edge og lokal AI er et placeringsproblem - ikke en rangliste

On-device AI har et dokumenteret potentiale til at reducere inferenceenergi for modeller og workloads, som passer til enheden. Edge kan samtidig give lav latency, privatliv og mindre netafhængighed. Cloudens styrke er adgang til større modeller, høj kapacitet, batching og delt infrastruktur.

Hybridarkitekturer gør grænsen endnu mindre binær. De kan placere hver opgave dér, hvor den kan løses billigst eller mest effektivt under et givent kvalitetskrav. Men hver ekstra lokal prøve, routingbeslutning og fallback er en del af systemet og skal tælles med.

Den relevante sammenligning er derfor, hvilken placering eller kombination der leverer den konkrete funktion med lavest samlet ressourceforbrug og acceptable krav til kvalitet, latency, sikkerhed og drift.

Med kapitel 12 afsluttes bogens del om den løbende tekniske driftsregning. I næste del skifter perspektivet fra den fysiske infrastruktur

til organisationens økonomi: først de omkostninger, der står på fakturaen, og derefter den arbejds- og kontroltid, som langt sjældnere står dér.

### Kilder nævnt i kapitlet


- Li, Pengfei; Islam, Mohammad J.; Ren, Shaolei (2025): A Case Study of Environmental Footprints for Generative AI Inference: Cloud versus Edge. ACM SIGMETRICS Performance Evaluation Review 53(2), 21-26. DOI 10.1145/3764944.3764950.
- Yan, Xiao; Ding, Yi (2025): Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices. FMSys 2025. DOI 10.1145/3722565.3727192.
- Tummalapalli, Pranay; Arayakandy, Sahil; Pal, Ritam; Kundan, Kautuk (2026): LLM Inference at the Edge: Mobile, NPU, and GPU Performance Efficiency Trade-offs Under Sustained Load. arXiv:2603.23640.
- Zou, Bohua et al. (2026): EnerInfer: Energy-Aware On-Device LLM Inference. arXiv:2606.23001.
- Liu, Hongyao et al. (2026): SparKV: Overhead-Aware KV Cache Loading for Efficient On-Device LLM Inference. arXiv:2604.21231.
- Ding, Dujian et al. (2024): Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing. arXiv:2404.14618.
- d'Orgeval, Alexandre et al. (2026): Generative AI impact assessment through a life cycle analysis of multiple data center typologies. Applied Energy 406, 127288.

## Del 4

# Den økonomiske regning

Del 4 flytter perspektivet fra de fysiske ressourcer til organisationens økonomi: først de synlige udgifter til AI-tjenester og infrastruktur, derefter den arbejdstid og kontrol, som sjældnere står på leverandørens faktura.

## Kapitel 13

# Den synlige pris

Efter tolv kapitler om materialer, chips, hardware, datacentre, elektricitet, vand og placering skifter perspektivet. Nu ser vi på den regning, som virksomheden faktisk kan se i budgettet, kontrakten eller fakturaen.

Den direkte økonomiske pris kan virke enklere end de fysiske omkostninger. Et abonnement har en pris pr. bruger. En API har en pris pr. token eller request. En cloudplatform viser forbruget i et cost dashboard. Alligevel kan to organisationer med samme AI-værktøj få meget forskellige direkte omkostninger.

Forskellen ligger i prismodellen, brugsmønstret og den tekniske arkitektur. Nogle betaler pr. sæde, andre efter forbrug. Nogle har standard-inference, andre prioriteret eller reserveret kapacitet. Data kan lagres, flyttes, indekseres og søges. Agentiske workflows kan kalde både modellen og eksterne værktøjer flere gange, før én opgave er afsluttet.

Den synlige pris er derfor bedst forstået som en omkostningsstak. Modelprisen er et centralt lag, men sjældent det eneste.

### Den direkte AI-regning består af flere lag

For en mindre organisation kan regningen være næsten identisk med et antal brugerlicenser. I en større eller mere integreret løsning kan den direkte regning fordele sig på modeladgang, cloudtjenester, data, integration, sikkerhed, observability, support og eventuel dedikeret kapacitet.

**Tabel 13.1 – Typiske direkte omkostningslag i en AI-løsning**

| Omkostningslag | Typisk fakturering | Hvad driver beløbet? | Hvad bør virksomheden følge? |
|---|---|---|---|
| Sædelicenser / abonnementer | Pris pr. bruger pr. måned/år | Antal tildelte licenser, plan og kontrakt | Tildelte sæder, aktiv brug, relevant brug, fornyelser |

| Omkostningslag | Typisk fakturering | Hvad driver beløbet? | Hvad bør virksomheden følge? |
|---|---|---|---|
| API / model-inference | Tokens, requests, billeder, lyd eller anden usage | Model, input/output, kontekst, caching, service tier | Usage pr. workload, model, team og accepteret resultat |
| Tools og agentfunktioner | Kald, søgninger, execution, knowledge base mv. | Hvor mange ekstra handlinger ét workflow udløser | Modelkald, tool calls, retries og workflow-frekvens |
| Cloud, storage og data transfer | Compute, lagring, database, netværk, throughput | Datamængde, retention, region, latency og kapacitet | Cost allocation pr. løsning, miljø og afdeling |
| Integration, sikkerhed og drift | Projekt, abonnement, support eller løbende service | Antal systemer, krav til adgang, logging og governance | Engangs- versus recurring cost; ændrings- og supportbehov |
| Egen hardware / reserveret kapacitet | Capex, leasing eller fast kapacitetsbetaling | GPU/servervalg, udnyttelse, commitment og levetid | Udnyttelsesgrad, idle capacity og alternativ anvendelse |

*Kilde/ramme: Egen syntese af de aktuelle prismodeller hos større AI- og cloudleverandører. Tabellen er en omkostningsstruktur, ikke en universel kontoplan.*

## Sædelicenser: købt adgang er ikke det samme som økonomisk udnyttelse

En licensmodel er enkel at budgettere: antal sæder ganget med pris pr. sæde. Den økonomiske fortolkning kræver imidlertid mere end at tælle, hvor mange licenser der er købt.

I det randomiserede feltstudie Early Impacts of M365 Copilot fra 2025 blev licenser tildelt medarbejdere i 56 store virksomheder. Produkttelemetrien viste, at den gennemsnitlige ugentlige brug blandt de behandlede lå omkring 38 procent efter studiets definition af tilbagevendende brug. I et senere seksmåneders feltstudie på tværs af 66 virksomheder og 7.137 vidensarbejdere brugte 80 procent af de behandlede medarbejdere værktøjet i anden halvdel af forsøget.

*Kilder: Dillon, Wiske et al. (2025), Early Impacts of M365 Copilot; Dillon, Jaffe, Immorlica & Stanton (2026), Shifting Work Patterns with Generative AI.*

Forskellen mellem de to procenter skyldes især, at studierne bruger forskellige definitioner og tidsvinduer. ”Brugt mindst én gang i perioden” og ”aktiv i en gennemsnitlig uge” er to forskellige nævnere. Vores evidensgrundlag behandler derfor licenstildeling, aktivering, ugentlig/månedlig aktiv brug, relevant opgavebrug og accepterede resultater som separate KPI’er.

Det er også muligt at have en licens, der bruges sjældent, men skaber stor værdi i få højværdite opgaver. Omvendt kan en medarbejder være registreret som aktiv uden at brugen skaber en målbar forretningsværdi. En universel "shelfware-procent" er derfor et dårligt mål for AI-økonomi.

### API-prisen er en formel, ikke et enkelt tal

API-priser bliver ofte gengivet som én sats pr. million tokens. I praksis kan input, output, cached input, cache writes, lange kontekster, service tiers og regional behandling have forskellige priser. Et agentisk workflow kan desuden udløse flere modelkald, søgninger eller værktøjskørsler under én brugerhandling.

OpenAI's aktuelle offentlige API-prisside opdeler blandt andet priser i input, cached input, cache writes og output og angiver samtidig et pristillæg for kvalificerede regional-processing-endpoints. Amazon Bedrocks cost- og usage-dokumentation viser tilsvarende, at en enkelt løsning kan generere særskilte fakturalinjer efter token-type, service tier og routing. For en økonomisk analyse er det derfor nødvendigt at forstå den konkrete billing-enhed frem for blot at gemme ét modelnavn og én listepris.

### Et illustrativt regneeksempel: output kan fylde mest på regningen

Antag en fiktiv API med 2 USD pr. million inputtokens og 10 USD pr. million outputtokens. En virksomhed kører 100.000 opgaver om måneden. Hver opgave bruger i gennemsnit 2.000 inputtokens og 800 outputtokens. Satserne er konstruerede for at vise regnemetoden; de er ikke et pristilbud fra en leverandør.

**Tabel 13.2 – Illustrativ månedlig API-regning**

| Post | Mængde | Illustrativ sats | Beregnet direkte pris |
|---|---|---|---|
| Input | 200 mio. tokens | 2 USD / 1 mio. | 400 USD |
| Output | 80 mio. tokens | 10 USD / 1 mio. | 800 USD |
| I alt – første forsøg | 280 mio. tokens | — | 1.200 USD |
| 25 % fuldt ekstra forsøg | 70 mio. tokens ekstra | samme tokenmix | 300 USD |
| I alt med ekstra forsøg | 350 mio. tokens | — | 1.500 USD |

*Illustrativ beregning med konstruerede satser. Antagelse: et fuldt ekstra forsøg har samme gennemsnitlige tokenmix som første forsøg. Human review, tool calls, storage, latencyværdi, kontraktpriser og øvrige cloudomkostninger er ikke medregnet.*

I eksemplet udgør output kun cirka 29 procent af tokens på første forsøg, men to tredjedele af selve modelregningen. Hvis man kun følger samlet tokenvolumen, overser man derfor en væsentlig del af prisdynamikken. Når 25 procent af opgaverne kræver et helt ekstra forsøg, stiger den direkte modelpris fra 1.200 til 1.500 USD om måneden, før menneskelig efterbehandling overhovedet er talt med.

## Listepriser er øjebliksbilleder

AI-priser bevæger sig hurtigt nok til, at et bogtal kan blive forældet længe før bogens metode bliver det. I juli 2026 offentliggjorde OpenAI nye lavere API-priser for GPT-5.6 Terra og Luna. Anthropic lancerede samme sommer Claude Sonnet 5 med en tidsbegrænset introduktionspris frem til udgangen af august og en højere standardpris derefter.

Prisændringerne illustrerer en vigtig regel for denne bog: aktuelle provider-priser kan bruges som datostemplede eksempler, men de bør ikke behandles som permanente benchmarkværdier. Ved en konkret business case skal den gældende officielle prisside eller den faktiske kontrakt anvendes.

Microsoft Azure gør samme forbehold eksplicit på sine AI-prissider: viste priser er estimater, mens den faktiske pris kan variere med aftaletype, købsdato og valutakurs. For en europæisk virksomhed kan region, databehandling og valuta derfor blive en reel del af den direkte økonomi.

*Kilder: OpenAI (30. juli 2026), Advancing the price-performance frontier with GPT-5.6; Anthropic (2026), Introducing Claude Sonnet 5; Microsoft Azure, Foundry Models pricing (aktuel prisside, 2026).*

## Region og service tier kan også være en prisvariabel

Kapitel 9 viste, at geografi kan ændre $CO_2$-regnestykket. I den økonomiske dimension kan geografi også påvirke prisen. Nogle tjenester

tager forskellige priser efter region eller processing scope, og data residency kan påvirke, hvilke endpoints en organisation vælger.

OpenAI oplyser aktuelt et pristillæg på 10 procent for kvalificerede regional-processing-endpoints for nyere modeller. Amazon Bedrock viser tilsvarende regionale prisforskelle for flere modeludbud, mens andre modeller eller routingformer har andre regler. De regionale prisforskelle dokumenterer ikke en generel europæisk merpris; de viser, at region er en prisparameter, som skal kontrolleres i den konkrete løsning.

*Kilder: OpenAI API Pricing (aktuel prisside, 2026) og Amazon Bedrock Pricing (aktuel prisside, 2026).*

### On-demand, batch og reserveret kapacitet besvarer forskellige behov

En løsning med ujævn eller lav trafik passer ofte naturligt til forbrugsbaseret on-demand-prissætning. Et stabilt og stort workload kan i nogle tilfælde få en anden økonomi med batch, provisioned throughput eller reserveret kapacitet. Til gengæld opstår risikoen for at betale for kapacitet, som ikke bliver udnyttet.

Amazon Bedrock dokumenterer f.eks. både on-demand, batch og reserverede service tiers, hvor reserveret kapacitet faktureres som fast kapacitet over en periode. Det gør udnyttelsesgraden til en direkte økonomisk variabel: fast kapacitet kan give forudsigelighed og ydeevne, men uudnyttet kapacitet er stadig en omkostning.

### Cloud-regningen kan ligge ved siden af modelregningen

En AI-applikation består sjældent kun af inferenskald. Data skal måske ligge i et objektlager (object storage) eller en database. Dokumenter kan indekseres til retrieval. Embeddings kan genereres. Logs og traces gemmes. Agenten kan kalde søgning, code execution eller andre tjenester. Data kan flyttes mellem regioner eller systemer.

Amazon Bedrock beskriver f.eks., at Cost and Usage Report kan indeholde separate linjer for forskellige token-typer, tiers og routingformer. AWS' egne referencearkitekturer for agentiske løsninger viser desuden, at model, knowledge base, guardrails og andre

komponenter faktureres som separate dele. En modelregning på 1.000 USD er derfor ikke nødvendigvis en samlet cloudregning på 1.000 USD.

### Integration er en direkte omkostning, selv når den ikke faktureres af modelleverandøren

Virksomheden kan også betale direkte for udvikling, integration, sikkerhedsopsætning, observability, testmiljøer, support og løbende platformdrift. Disse poster kan komme fra en systemintegrator, en intern IT-afdeling med særskilt projektbudget eller en anden cloudleverandør end selve modelleverandøren.

Kapitlets afgrænsning er stadig den synlige pris. Medarbejdernes almindelige arbejdstid til at formulere prompts, kontrollere output og rette fejl hører til den skjulte arbejdspris i kapitel 14. Hvis den samme aktivitet derimod købes som en ekstern konsulentydelse eller ligger som en identificerbar projektafregning, bliver den direkte synlig i virksomhedens økonomiske regnskab.

### Budgettet bør kunne følges fra faktura til workload

En samlet AI-faktura er vanskelig at styre, hvis den ikke kan fordeles tilbage på de workloads, der skabte den. Cost allocation bør derfor følge de faktiske beslutningsenheder: afdeling, use case, applikation, model, miljø eller agent.

AWS anbefaler f.eks. cost allocation tags i Bedrock Cost and Usage Reports, så omkostninger kan forbindes med applikationer. Det samme princip gælder uafhængigt af leverandør: en organisation bør kunne forklare, hvilke workloads der driver stigningen i forbrug, før den forsøger at vurdere ROI.

*Kilde: AWS, Understanding your Amazon Bedrock Cost and Usage Report data (2026).*

**Påstanden under lup**

**Påstand:** Virksomhedens AI-omkostning er abonnementet eller API-regningen.

**Vurdering:** Misvisende som fuld direkte omkostning.

**Kort svar:** Abonnementet eller modelregningen er ofte den mest synlige post, men en AI-løsning kan samtidig have direkte udgifter til cloud, storage, data transfer, tools, integration, sikkerhed, support, regional processing, reserveret kapacitet eller egen hardware.

Hvad siger evidensen? De aktuelle provider- og cloudprismodeller opdeler faktureringen i flere enheder og tiers. De arbejdspladsstudier og prisdata, der gennemgås i kapitlet, viser samtidig, at licenstildeling, faktisk brug og accepterede resultater er forskellige økonomiske mål.

**Det afgørende forbehold:** Den menneskelige tid til review, korrektion og integration er ikke fuldt dækket af dette kapitel. Den behandles i kapitel 14 og indgår senere i cost per accepted outcome.

## Fra direkte pris til skjult arbejdspris

Den direkte økonomi kan i princippet spores i kontrakter, fakturaer, cloud billing og interne projektkonti. Det gør den mere synlig end vand-, materiale- eller $CO_2$-regnskabet, men ikke nødvendigvis enklere.

En lav listepris kan kombineres med højt usage. En høj sædepris kan være billig, hvis få brugere løser dyre opgaver. En fast kapacitetsaftale kan være fordelagtig ved høj udnyttelse og dyr ved lav udnyttelse. Et regionalt endpoint kan koste mere, men være nødvendigt for organisationens valgte databehandlingsramme.

Det afgørende er derfor at forbinde den direkte pris med den aktivitet, der skaber den, og at datostemple de prisforudsætninger, som business casen bygger på.

I næste kapitel går vi videre til den del af AI-regningen, som sjældent står på leverandørfakturaen: den menneskelige tid til at formulere opgaven, kontrollere resultatet, håndtere fejl, forsøge igen og få outputtet sikkert ind i den virkelige arbejdsproces.

### Kilder nævnt i kapitlet

• OpenAI (2026): API Pricing samt Advancing the price-performance frontier with GPT-5.6. Officielle provider-priser og prisændringer; aktuelle priser skal genkontrolleres ved hver ny udgave.

• Anthropic (2026): Introducing Claude Sonnet 5 samt Anthropic List Prices. Bruges som eksempel på tidsbegrænset introduktionspris og flere prisdimensioner.

• Amazon Web Services (2026): Amazon Bedrock Pricing; Understanding your Amazon Bedrock Cost and Usage Report data; Service tiers for optimizing performance and cost.

• Microsoft Azure (2026): Microsoft Foundry / Foundry Models pricing. Officiel prisinformation med forbehold for aftaletype, købsdato og valutakurs.

• Dillon, Eleanor Wiske et al. (2025): Early Impacts of M365 Copilot. Randomiseret feltstudie; licenstildeling og ugentlig anvendelse.

• Dillon, Eleanor W.; Jaffe, Sonia; Immorlica, Nicole; Stanton, Christopher T. (2026): Shifting Work Patterns with Generative AI. AER: Insights forthcoming; 66 virksomheder og 7.137 knowledge workers.

## Kapitel 14

# Den skjulte arbejdspris

En AI-løsning kan være billig på fakturaen og dyr i medarbejdertid. Det modsatte kan også være tilfældet. Den forskel bliver synlig, når regnskabet følger hele arbejdsprocessen fra opgaven formuleres, til resultatet er kontrolleret, rettet og klar til at blive brugt.

Kapitel 13 beskrev de direkte udgifter til licenser, API-kald, cloud, integration og drift. Her flytter vi blikket til den arbejdstid, som allerede ligger i organisationens lønbudget. Den tid bliver let overset, fordi den sjældent udløser en særskilt faktura. Den er alligevel en reel ressourceomkostning.

AI kan forkorte den del af arbejdet, der består i at producere et første udkast, finde information, skrive kode eller formulere et svar. Samtidig kan nye arbejdsled opstå omkring styring, verification, integration, korrektion og godkendelse. Det relevante spørgsmål er derfor ikke, hvor hurtigt modellen producerer sit første svar, men hvor meget menneskelig tid der kræves for at nå frem til det resultat, virksomheden faktisk kan acceptere.

### Arbejdstiden kan flytte i stedet for at forsvinde

Microsoft Research undersøgte i 2025 319 knowledge workers og 936 konkrete eksempler på brug af generativ AI i arbejdet. Studiet fandt, at den oplevede kritiske tænkning ændrede karakter: fra informationssøgning mod informationskontrol, fra problemløsning mod integration af AI-svaret og fra selve udførelsen mod styring og ansvar for det samlede resultat.

Det er en vigtig økonomisk observation. Hvis AI skriver et udkast på 20 sekunder, er de 20 sekunder kun én del af arbejdsgangen. Medarbejderen kan stadig skulle kontrollere fakta, sammenholde med interne regler, tilpasse tonen, teste kode, verificere beregninger, dokumentere beslutningen eller håndtere de dele, som modellen ikke kunne løse sikkert.

Denne arbejdsforskydning er ikke i sig selv et argument imod AI. Flere studier viser betydelige gevinster på bestemte opgaver. Noy og Zhang fandt i et randomiseret forsøg med 453 professionelle, at ChatGPT reducerede tiden på de undersøgte skriveopgaver med 40 procent og samtidig løftede den bedømte kvalitet. Brynjolfsson, Li og Raymond fandt i kundeservice en gennemsnitlig stigning på 15 procent i løste

sager pr. time blandt 5.172 medarbejdere. Dell'Acqua og medforfattere fandt ligeledes markante gevinster på konsulentopgaver, der lå inden for den anvendte models capability frontier.

De samme forskningsspor viser samtidig, hvorfor en universel produktivitetsprocent er uforsvarlig. I Dell'Acqua-studiet faldt sandsynligheden for et korrekt resultat på en kompleks opgave uden for AI-frontieren. METR fandt i et randomiseret studie af erfarne open-source-udviklere med early-2025-værktøjer, at AI-adgang øgede den målte completion time med 19 procent, selv om udviklerne efter forsøget vurderede, at AI havde gjort dem omkring 20 procent hurtigere.

*Kilder: Lee et al. (CHI 2025), Noy & Zhang (Science 2023), Brynjolfsson, Li & Raymond (QJE 2025), Dell'Acqua et al. (Organization Science 2026) og METR (2025). Studierne måler forskellige opgaver, værktøjer, perioder og kvalitetskrav og bruges derfor som dokumentation for variation - ikke som én samlet produktivitetsfaktor.*

## Fem menneskelige tidsled bør holdes adskilt

**Tabel 14.1 – De menneskelige tidsled omkring et AI-resultat**

| Tidsled | Hvad medarbejderen gør | Hvorfor det bør måles |
|---|---|---|
| Opgavestyring / prompt | Afgrænser opgaven, samler kontekst, giver instruktioner og vælger værktøj. | En hurtig model hjælper lidt, hvis forberedelsen er lang eller gentages ofte. |
| Review / verification | Kontrollerer fakta, beregninger, compliance, tone, kode eller anden quality bar. | Review kan være kort ved lav risiko og omfattende ved høj konsekvens. |
| Korrektion / retry | Retter output, ændrer prompt, indsamler mere kontekst eller kører et nyt forsøg. | Fejl og lav first-pass-kvalitet omsætter modelkald til ekstra mennesketid. |
| Integration | Flytter resultatet ind i dokument, CRM, kodebase, beslutningsproces eller andet system. | Et godt svar skaber først værdi, når det kan bruges i den virkelige proces. |
| Eskalering / human completion | Mennesket overtager hele eller dele af opgaven, når AI ikke når kvalitetskravet. | Residualarbejdet kan dominere økonomien i svære eller risikofyldte cases. |

*Kilde/ramme: Egen syntese af de arbejdsled, der dokumenteres i kapitlets studier. Tidsleddene er en målestruktur; de angiver ingen universel fordeling eller standardtid.*

## Reviewtid skal måles - der findes ingen universel standard

Det kan være fristende at løse den skjulte arbejdspris med en fast tommelfingerregel: læg f.eks. fem minutters review til hvert AI-output. Evidensen understøtter ikke en sådan universel sats.

Et peer-reviewed studie af AI-genererede assessment-items fandt, at GPT-4-genererede items i gennemsnit krævede længere ekspertrevision

end menneskeskrevne items i studiets egen tidsmåling: 3,78 mod 2,82. Samtidig blev AI-items ofte vurderet positivt, og selve generationen var hurtig. Studiet viser dermed en arbejdsforskydning, hvor hurtig produktion kan efterfølges af mere specialistreview. Tallene tilhører denne konkrete high-stakes assessment-opgave og kan ikke overføres til almindelig kontorbrug.

Et randomiseret studie med 30 farmaceuter og 6.000 medicin-verifikationsopgaver viste en anden del af samme mekanisme. Når AI-rådet var uhensigtsmæssigt, brugte deltagerne længere kognitiv behandlingstid på de originale reference- og produktbilleder. Også her er konteksten specialiseret, og eye-tracking-dwell time er ikke det samme som samlet arbejdstid eller kroner. Resultatet dokumenterer derimod, at dårlig AI-hjælp kan øge den menneskelige verification-belastning.

Den fagligt forsvarlige konklusion er derfor lokal måling. Reviewtid bør registreres på den konkrete opgavetype og helst opdeles efter risiko, first-pass-kvalitet og om resultatet accepteres, korrigeres eller eskaleres.

*Kilder: Kowal et al. (2025), International Journal of Selection and Assessment; Tsai et al. (2025), Journal of Medical Internet Research. Begge bruges her som opgavespecifik dokumentation; deres reviewtider overføres ikke til et generelt virksomhedsbenchmark.*

## Et lille AI-beløb kan skjule en stor arbejdsomkostning

Et simpelt regnestykke viser, hvorfor modelregningen alene kan give et skævt billede. Antag et fiktivt dokumentworkflow med 100 opgaver. Den loaded medarbejderomkostning sættes til 450 kr. pr. time. Hver opgave bruger i gennemsnit tre minutter på styring og fire minutter på review. En fjerdedel af opgaverne kræver syv ekstra minutters korrektion, og ti procent eskaleres til 18 minutters menneskelig completion. AI- og toolforbruget koster samlet 43,75 kr.

Alle satser og fordelinger i eksemplet er konstruerede. Formålet er alene at vise regnskabsmetoden.

**Tabel 14.2 – Illustrativ skjult arbejdspris i et AI-workflow**

| Post | Illustrativ beregning | Omkostning |
|---|---|---|
| Prompt/styring + review | 100 × 7 min × 450 kr./time | 5.250,00 kr. |
| Korrektion | 25 × 7 min × 450 kr./time | 1.312,50 kr. |
| Eskalering / human completion | 10 × 18 min × 450 kr./time | 1.350,00 kr. |
| AI + tools | Konstrueret samlet usage | 43,75 kr. |
| Samlet workflow | Human tid + AI/tools | 7.956,25 kr. |
| Pris pr. accepteret resultat | 7.956,25 / 100 | 79,56 kr. |

*Illustrativ beregning. Der er ingen empirisk markedsbenchmark i disse minutter, andele eller kronebeløb. Alle 100 opgaver antages at ende som accepterede resultater efter eventuel korrektion eller eskalering.*

I dette konstruerede eksempel udgør AI- og toolregningen under én procent af den samlede workflowomkostning. Resten er menneskelig tid. En optimering, der halverer modelprisen, vil derfor næsten ikke flytte totalen, mens et bedre first-pass-resultat eller kortere review kan gøre en stor forskel.

Det modsatte kan forekomme i fuldautomatiske eller meget compute-tunge workflows. Pointen er ikke, at mennesketid altid dominerer. Pointen er, at dens størrelse skal måles, før virksomheden kan vide, hvad der dominerer.

## Sparet tid er ikke automatisk sparede lønkroner

Når et studie finder en tidsbesparelse, opstår et nyt spørgsmål: Hvad bliver den frigivne tid til? Den kan give flere producerede enheder, kortere kø, bedre kvalitet, mere læring, mindre overarbejde, mere slack eller i nogle tilfælde en faktisk reduktion i løn-, overtime- eller contractor-omkostninger. Disse værdikanaler er forskellige og kan ikke lægges oven i hinanden uden dokumentation.

Et seksmåneders randomiseret feltforsøg på tværs af 66 virksomheder og 7.137 knowledge workers giver et konkret eksempel. Blandt de 80 procent af de behandlede medarbejdere, der brugte værktøjet i anden halvdel af forsøget, blev der brugt omkring to færre timer på email om ugen, og arbejdet uden for normal arbejdstid faldt. Forskerne fandt samtidig ingen målbar ændring i mængden eller sammensætningen af medarbejdernes opgaver som følge af individuel AI-adgang.

Den danske arbejdsmarkedsanalyse fra Humlum og Vestergaard peger i samme metodiske retning fra et andet niveau. Ved at koble repræsentative adoptionsundersøgelser til administrative data fandt de i den tidlige adoptionsperiode ingen tydelige effekter på registrerede arbejdstimer eller indtjening og kunne statistisk afvise effekter større end omkring to procent over to år i deres design. Resultatet kan godt sameksistere med gevinster på konkrete opgaver; pointen er, at sådanne gevinster ikke automatisk bliver til lavere registrerede arbejdstimer eller højere løn.

En business case bør derfor holde mindst tre størrelser adskilt: sparet task-tid, realiseret produktions- eller kvalitetsgevinst og faktisk ændring i cash labour cost. Den samme frigivne time må ikke samtidigt bogføres som både en lønbesparelse og værdien af ekstra output, medmindre begge effekter faktisk realiseres og kan afstemmes.

*Kilder: Dillon, Jaffe, Immorlica & Stanton (2026), Shifting Work Patterns with Generative AI, AER: Insights forthcoming; Humlum & Vestergaard (2025), Large Language Models, Small Labor Market Effects. Begge kilder bruges til værdikonvertering og siger ikke, at AI mangler task-level produktivitetsgevinster.*

### Mål hele cyklussen - og mål samme kvalitetsniveau

Den enkleste robuste metode er at vælge en konkret opgave og måle både baseline og AI-workflow frem til samme accepterede resultat. Hvis baseline er et godkendt kundesvar, skal AI-siden også ende i et godkendt kundesvar. Et råt AI-udkast og et færdigt menneskeskrevet dokument er ikke samme funktionelle enhed.

Målingen bør registrere cycle time og active human time separat, når det er relevant. En medarbejder kan f.eks. vente på et langt agentflow uden selv at arbejde aktivt hele tiden. Omvendt kan flere små AI-kald kræve konstant supervision. Begge dele kan påvirke økonomien, men på forskellige måder.

First-pass-kvalitet bør også stå alene. Et workflow, der accepteres i første forsøg, har en anden arbejdsprofil end et workflow, hvor 40 procent kræver korrektion. I bogens målemetode behandles derfor first-pass,

correction og escalation som separate outcomes, som skal defineres før målingen begynder.

Til sidst skal værdien af den frigivne tid bestemmes efter, hvad organisationen faktisk gør med den. Herfra kan kapitel 18 senere samle modelpris, værktøjer, retries og mennesketid i én cost per accepted outcome. Kapitel 14 leverer den del af tælleren, der ellers let forsvinder fra regnskabet.

**Påstanden under lup**

**Påstand:** Hvis AI sparer en medarbejder én arbejdstime, har virksomheden sparet én times løn.

**Vurdering:** Misvisende som generel økonomisk konklusion.

**Kort svar:** En målt tidsbesparelse er først en løn- eller cash-besparelse, når den faktisk reducerer betalte timer, overarbejde, contractor-forbrug, headcount eller en anden identificerbar udgift. Den samme time kan i stedet blive til mere output, kortere ventetid, højere kvalitet eller mindre arbejde uden for normal arbejdstid.

**Hvad siger evidensen?** Task-studier dokumenterer reelle tids- og throughput-gevinster i bestemte workflows, mens det nyere 66-firm-feltforsøg viser, at frigivet tid blandt brugere bl.a. kunne vise sig som mindre email og mindre after-hours work uden målbar ændring i task quantity/composition. Danske administrative data viser samtidig, at tidlige AI-gevinster ikke automatisk slog igennem i registrerede arbejdstimer eller indtjening.

**Det afgørende forbehold:** Monetariseringen afhænger af den konkrete kapacitetsbegrænsning og værdikanal. Sparet tid, ekstra output, lønbesparelse og medarbejdervelfærd skal rapporteres separat, så den samme gevinst ikke dobbeltregnes.

## Den skjulte pris er målbar

Den skjulte arbejdspris behøver ikke være et skøn baseret på mavefornemmelse. Den kan måles med de samme principper som andre procesomkostninger: definer opgaven, fastlæg kvalitetskravet, registrer den menneskelige tid i de vigtigste arbejdsled og følg, hvor mange resultater der accepteres i første forsøg, efter korrektion eller efter eskalering.

Metoden gør det også muligt at se, hvor en forbedring faktisk har værdi. Måske er den billigste model dyr, fordi medarbejderne retter for meget. Måske er den dyre model unødvendig på en lavrisikoopgave. Måske er

selve AI-delen velfungerende, mens integrationen til organisationens systemer er den reelle tidsrøver.

Den økonomiske analyse bliver dermed mindre afhængig af generelle AI-produktivitetsprocenter og mere afhængig af den arbejdsproces, virksomheden selv kan observere.

I næste kapitel udvider vi igen systemgrænsen. AI's økonomiske regning stopper ikke nødvendigvis ved virksomheden, der køber tjenesten. Store datacenterbelastninger kan kræve nettilslutning, energiinfrastruktur og andre investeringer, og spørgsmålet bliver derfor, hvem der udløser omkostningen, hvem der betaler den, og hvem der får værdien.

### Kilder nævnt i kapitlet


• Lee, Hao-Ping (Hank) et al. (2025): The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. CHI 2025. 319 knowledge workers og 936 work-use-eksempler.

• Noy, Shakked; Zhang, Whitney (2023): Experimental evidence on the productivity effects of generative artificial intelligence. Science 381. Randomiseret forsøg med 453 professionelle på afgrænsede skriveopgaver.

• Brynjolfsson, Erik; Li, Danielle; Raymond, Lindsey R. (2025): Generative AI at Work. Quarterly Journal of Economics 140(2). Workplace deployment blandt 5.172 kundeservicemedarbejdere.

• Dell'Acqua, Fabrizio et al. (2026): Navigating the Jagged Technological Frontier. Organization Science. Præregistreret eksperiment med 758 knowledge workers.

• METR (2025): Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. Randomiseret feltstudie med 16 erfarne udviklere og 246 reelle issues.

• Kowal, Jaclyn Martin et al. (2025): Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items. International Journal of Selection and Assessment 33(3).

• Tsai, Chuan-Ching et al. (2025): Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Pharmacists: Randomized Controlled Trial. Journal of Medical Internet Research 27:e59946.

• Dillon, Eleanor W.; Jaffe, Sonia; Immorlica, Nicole; Stanton, Christopher T. (2026): Shifting Work Patterns with Generative AI. American Economic Review: Insights, forthcoming. 66 virksomheder og 7.137 knowledge workers.

• Humlum, Anders; Vestergaard, Emilie (2025): Large Language Models, Small Labor Market Effects. Dansk survey- og registerbaseret arbejdsmarkedsanalyse.

## Kapitel 15

# Hvem betaler infrastrukturen?

Når et stort datacenter kobles til elsystemet, opstår en anden type regning end den, der står på virksomhedens el- eller cloudfaktura. En ny eller større belastning kan kræve tilslutningsanlæg, transformerstationer, transmissionskapacitet, distributionsforstærkninger, systemydelser og i nogle tilfælde ny produktions- eller lagringskapacitet.

Det er fristende at kalde hele denne regning for "datacenterets infrastrukturomkostning". Den betegnelse bliver for upræcis, hvis den bruges uden at vise, hvordan omkostningerne faktisk fordeles.

Den neutrale analyse kræver fire separate spørgsmål: Hvem udløser behovet? Hvem investerer? Hvem betaler gennem tilslutningsbidrag, kontrakter og tariffer? Hvem får nytte af infrastrukturen over dens levetid? De fire svar kan være forskellige.

## Fire spørgsmål skal holdes adskilt

**Tabel 15.1 – Fire forskellige spørgsmål i infrastrukturanalysen**

| Spørgsmål | Hvad der skal undersøges | Typisk dokumentation |
|---|---|---|
| Hvem udløser behovet? | Om en ny belastning er den marginale årsag til en konkret forstærkning, ny station eller anden kapacitet. | Netplan, connection study, investeringsbeslutning og counterfactual. |
| Hvem investerer? | Hvilken netvirksomhed, systemoperatør, kunde eller tredjepart der anlægger og finansierer aktivet i første omgang. | Projekt- og regnskabsdata, regulatoriske afgørelser. |
| Hvem betaler? | Hvordan omkostningen opkræves: tilslutningsbetaling, effektbetaling, energitarif, abonnement, minimumsbetaling eller andre mekanismer. | Tarifmetode, kontrakt og fakturering. |
| Hvem får nytte? | Om aktivet alene betjener kunden, eller senere kan bruges af andre kunder, produktion eller systemet som helhed. | Netplanlægning, aktivets funktion og levetid. |

*Kilde/ramme: Egen analysemodel baseret på de danske tarif- og large-load-kilder i kapitlet. Tabellen er ikke en påstand om én bestemt dansk tarifordning.*

## Et direkte tilslutningsbidrag løser kun en del af spørgsmålet

En stor kunde kan betale for en konkret tilslutning og stadig påvirke omkostningerne i det fælles net. Omvendt kan en investering, som oprindeligt udløses af én stor kunde, senere blive en del af et net, som også betjener andre. Derfor fortæller et tilslutningsbidrag alene hverken, at alle afledte systemomkostninger er internaliseret, eller at andre kunder nødvendigvis betaler dem.

Forsyningstilsynets Analyse 11 fra marts 2026 beskriver det danske tarifprincip på både distributions- og transmissionsniveau. Tarifferne er blevet mere differentierede mellem kundegrupper, geografi og tidsperioder, blandt andet for bedre at afspejle de omkostninger, som forskellige kundetyper giver anledning til, og for at udnytte nettet mere

effektivt. En bedre udnyttelse af eksisterende kapacitet kan reducere behovet for nye og tidskrævende netinvesteringer.

*Kilde: Forsyningstilsynet (2026), Analyse 11 - Overblik over elsektorens tariffer. Kilden beskriver den aktuelle danske tarifstruktur og dens omkostningslogik; den giver ikke en generel datacenterregning.*

### MW ved tilslutningen og TWh i driften er to forskellige størrelser

Store datacentre planlægges ofte i MW, mens deres faktiske elforbrug senere opgøres i TWh. Den forskel er afgørende for infrastrukturen. Elnettet skal kunne stille den aftalte kapacitet til rådighed, selv når kunden endnu ikke bruger den fuldt ud.

Energistyrelsens Analyseforudsætninger til Energinet 2025 gør netop denne forskel synlig. AF25 indfører en kapacitetsudnyttelsesfaktor på 60 procent i første driftsår, stigende til 80 procent i det 15. driftsår. Den betalte netadgangskapacitet skal samtidig kunne være til rådighed, selv om den realiserede belastning er lavere. 60 og 80 procent er planlægningsantagelser i AF25, ikke universelle udnyttelsesgrader for datacentre.

Som dansk planlægningsreference ligger AF25 omkring 1.000 MW datacenterkapacitet og cirka 6,0-6,1 TWh elforbrug i 2030. Hvis 1.000 MW blev udnyttet konstant hele året, ville energimængden være 8,76 TWh. Et årsforbrug på 6,0 TWh svarer matematisk til en gennemsnitlig belastning på cirka 685 MW. Beregningen viser alene forskellen mellem reserveret/installeret kapacitet og realiseret energi; den beskriver ikke et bestemt datacenters driftsprofil.

*Kilder: Energistyrelsen (2025), Analyseforudsætninger til Energinet 2025 - Datacentre og sammenfatningsnotat. AF25 er et planlægningsgrundlag og skal versionsmærkes, fordi projektpipeline og metode ændres fra år til år.*

### Mere forbrug kan både udløse investeringer og sprede faste omkostninger

Tariffer har en tæller og en nævner. I tælleren ligger de omkostninger, som skal dækkes. I nævneren ligger den mængde eller kapacitet, som betalingen fordeles over. En stor ny belastning kan påvirke begge dele.

Energinets Tarifkatalog 2026 giver et konkret dansk eksempel på nævnereffekten. Energinet forventede i tarifgrundlaget, at det samlede elforbrug steg fra 40,3 TWh i forecast 2025 til 42,9 TWh i forecast 2026. Den forventede stigning på 6,4 procent, blandt andet fra store datacentre, Power-to-X og transport, reducerede isoleret tariffen med 0,4 øre/kWh. Energinet understregede samtidig, at der endnu ikke var registreret et nævneværdigt forbrug fra de nye sektorer, og at tidspunkt og omfang var usikkert.

Det er derfor forkert at bruge 0,4 øre/kWh som "datacenterrabat". Datacentre var kun én af flere forventede drivere, og det samlede tarifniveau afhænger også af udviklingen i selve omkostningerne. Eksemplet dokumenterer en mekanisme: større realiseret forbrug kan sprede visse omkostninger over flere kWh.

*Kilde: Energinet (2025), Energinets Tarifkatalog 2026, især afsnittet om forventet elforbrug. Energinet skelner selv mellem tarifforudsætninger og de langsigtede analyseforudsætninger for investeringsplanlægning.*

## Et illustrativt regneeksempel: tæller og nævner kan trække i hver sin retning

Antag et fiktivt tarifsystem med 5,0 mia. kr. i årlige omkostninger og 40 TWh betalingsgrundlag. Den simple energitarif ville være 12,5 øre/kWh. Dernæst kommer 3 TWh nyt forbrug. Vi viser to forskellige forløb. Alle tal er konstruerede og har ingen relation til en konkret dansk tarif.

**Tabel 15.2 – Illustrativ betydning af både omkostning og betalingsgrundlag**

| Scenario | Årlig omkostning | Betalingsgrundlag | Illustrativ tarif |
|---|---|---|---|
| Udgangspunkt | 5,0 mia. kr. | 40 TWh | 12,50 øre/kWh |
| 3 TWh ekstra, ingen ny årlig omkostning | 5,0 mia. kr. | 43 TWh | 11,63 øre/kWh |
| 3 TWh ekstra og 0,6 mia. kr. ekstra omkostning | 5,6 mia. kr. | 43 TWh | 13,02 øre/kWh |

*Illustrativ beregning: tarif = årlig omkostning / betalingsgrundlag. Eksemplet isolerer mekanikken og udelader virkelige tarifdesigns med effektbetaling, abonnement, tidsdifferentiering, tab, systemydelser og andre poster.*

Det første forløb giver lavere enhedstarif, fordi nævneren vokser. Det andet giver højere enhedstarif, fordi den ekstra omkostning vokser mere end betalingsgrundlaget kan opveje. Begge resultater er matematisk mulige. Det faktiske svar kræver den konkrete investerings- og tarifmodel.

### Usikker projektpipeline skaber en særlig risiko

Et elsystem kan blive bedt om at reservere eller bygge kapacitet til projekter, som senere bliver forsinket, nedskaleret eller aldrig når den forventede belastning. Den risiko er vigtig, fordi netaktiver typisk har langt længere levetid end en enkelt AI-hardwaregeneration eller et datacenterprojekt.

AF25 viser, hvor hurtigt planlægningsbilledet kan ændre sig. Sammenlignet med AF24 faldt den danske 2030-antagelse for datacentres elforbrug fra 17,1 TWh til cirka 6,0 TWh, mens kapacitetsantagelsen faldt fra 1.950 MW til 1.000 MW. Det svarer til omtrent 65 procent lavere TWh og 49 procent lavere MW i én årlig planlægningsrevision. Forskellen mellem de to årgange viser, hvor følsomt infrastrukturområdet er over for projektmodning, realiseringssandsynlighed, indfasning og kapacitetsudnyttelse.

AF25 sandsynlighedsvægter derfor projekter efter modenhed: 100 procent for allerede tilsluttede projekter eller projekter med tilslutnings-/modningsaftale, 40 procent for projekter i screening og 25 procent for mindre sikre interessetilkendegivelser. Vægtene er planlægningsantagelser, ikke empiriske garantier for det enkelte projekt.

*Kilde: Energistyrelsen, AF24 og AF25. Tallene skal altid ledsages af forecast-vintage.*

### Stranded-asset-risiko kan flyttes, men ikke antages væk

Når en stor kunde udløser en investering, bliver kontraktens varighed, minimumsbetalinger, sikkerhedsstillelse og exitvilkår en del af risikofordelingen. Hvis kunden senere bruger mindre kapacitet end forventet eller forlader området, afgør reglerne, hvor stor en del af den

resterende omkostning der bliver hos kunden, netselskabet eller den bredere kundekreds.

Det danske kapitel behøver ikke importere amerikanske tarifprocenter for at forstå mekanismen. Som designkontrast viser nyere regulatoriske ordninger i Virginia og Missouri, at minimum demand charges, flerårige commitments, collateral og exitbetalinger kan bruges til at placere en større del af risikoen hos den store kunde. Disse ordninger er amerikanske og kan ikke bruges som direkte proxy for europæiske tariffer; de viser blot, at cost shifting ikke er en naturgiven størrelse, men påvirkes af kontrakt- og tarifdesign.

*Kilder: Virginia State Corporation Commission og Missouri Public Service Commission. USA-eksemplerne anvendes kun som tarifdesign-kontrast.*

### Irland viser, at tilslutning også kan være et systemvilkår

Irland er et særligt europæisk eksempel, fordi datacentrene allerede udgjorde 22 procent af den nationale el-efterspørgsel i 2024. Den irske regulator CRU vedtog i december 2025 en ny connection policy, som knytter konkrete krav til nye datacentre.

Nye tilslutninger skal blandt andet levere generation og/eller lagring lokalt eller onsite svarende til den ønskede maksimale importkapacitet. Kapaciteten skal deltage i engrosmarkedet og understøtte systemtilstrækkeligheden. Politikken kræver desuden, at mindst 80 procent af det årlige forbrug matches med yderligere vedvarende projekter i Irland, med en seksårig indfasningsperiode, og at netoperatørerne vurderer den konkrete lokations netbegrænsninger.

Irland dokumenterer dermed, at store nye loads kan mødes med særskilte connection conditions, når systemet er presset. Modellen er irsk og kan ikke overføres direkte til Danmark eller resten af EU. Den er relevant som europæisk eksempel på, at nettilslutning kan være mere end en engangsbetaling for en fysisk ledning.

*Kilde: Commission for Regulation of Utilities, Decision on New Electricity Connection Policy for Data Centres, 12. december 2025. CRU oplyser, at datacentre voksede fra 5 procent af national el-efterspørgsel i 2015 til 22 procent i 2024.*

### Energi, effekt, kapacitet og utilisation er forskellige regnskaber

Et årligt energiforbrug i TWh fortæller, hvor meget elektricitet der bruges over tid. Det fortæller ikke alene, hvor høj effekt belastningen kræver på et bestemt tidspunkt, hvor meget ekstra kapacitet der skal stå til rådighed, eller hvor stor en del af den installerede kapacitet der faktisk udnyttes. De fire størrelser kan bevæge sig forskelligt og udløse forskellige økonomiske omkostninger.

Et peer-reviewed studie i Energy and AI fra september 2026 modellerer netop denne mekanisme for training-orienterede AI-datacentre under usikker skalering. I et IEEE 30-bus-testnet steg den planlagte AI-datacenterkapacitet fra 122,4 MW til 180,6 MW, når planlægningen blev mere konservativ, mens batterilagringen steg fra 103,4 MWh til 301,5 MWh. Den modellerede samlede omkostning steg samtidig fra cirka 2,45 til 4,08 mia. CNY.

Tallene er simulationsresultater og må ikke beskrives som observerede omkostninger ved et konkret datacenter. Værdien ligger i mekanismen: høj forsyningssikkerhed og reservekapacitet kan kræve mere fysisk infrastruktur, end det gennemsnitlige energiforbrug alene antyder.

Fleksibilitet kan trække i den modsatte retning. Microsoft Research-preprintet PowerSlider viser i en konkret LLM-serving-opsætning, at dele af inference-loaden kan drosles under power caps med mindre ydelsestab end de sammenlignede metoder. Det er ikke dokumentation for, at AI-datacentre generelt kan fungere som fleksibel netressource, men det viser, hvorfor lavere effekt i en periode ikke må forveksles med lavere samlet energi over hele opgaven.

*Kilder: Power-infrastructure expansion planning for training-oriented AI data centers under large-model scaling uncertainty, Energy and AI 25 (2026); Li, Yueying et al. (2026): PowerSlider: Exploiting Phase Asymmetry for LLM Serving under Demand Response, Microsoft Research, arXiv preprint.*

### Samfundsregnskabet kræver flere kolonner

En analyse af, om et datacenter er "dyrt" eller "billigt" for et lokalsamfund, bliver hurtigt politisk, hvis forskellige regnskaber blandes sammen. Privat investering, offentlige skattefordele, byggearbejdspladser, permanente jobs, lokale skatteindtægter, netinvesteringer, tariffvirkninger, vand, areal, støj og eventuelle fleksibilitetsydelser er separate poster.

En kommune kan få skatteindtægter og aktivitet samtidig med, at netsystemet får nye investeringsbehov. En netinvestering kan senere blive nyttig for flere kunder samtidig med, at den første kunde udløste behovet. En stor stabil belastning kan øge nogle systemomkostninger og samtidig fordele andre faste omkostninger over et større betalingsgrundlag.

Det er netop derfor kapitlet fastholder et fler-kolonne-regnskab frem for ét enkelt netto-tal. En samlet monetarisering kan være relevant i en egentlig samfundsøkonomisk analyse, men så skal værdisætning, counterfactual og fordelingsvirkninger være synlige.

*Kilde/ramme: Egen syntese med Energistyrelsens samfundsøkonomiske analyseprincipper som generel metodebaggrund.*

**Påstanden under lup**

**Påstand:** Datacentre får de øvrige elkunder til at betale for deres netudbygning.

**Vurdering:** Ikke dokumenteret som en generel påstand.

**Kort svar:** En stor datacenterbelastning kan udløse fælles netinvesteringer, men betalingsfordelingen afhænger af tilslutningsbetaling, tarifdesign, kontraktvilkår, realiseret forbrug og aktivernes senere anvendelse. Større realiseret load kan samtidig sprede visse faste omkostninger over flere kWh.

**Hvad siger evidensen?** Dansk tarifregulering søger at afspejle de omkostninger, som kundegrupper giver anledning til. Energinets 2026-tarifgrundlag viser en konkret nævnereffekt, mens AF25 viser stor usikkerhed om, hvor meget datacenterkapacitet der faktisk bliver realiseret. Irlands connection policy viser, at yderligere systemkrav kan lægges direkte på nye datacenterprojekter.

**Det afgørende forbehold:** Spørgsmålet kan kun besvares for en konkret jurisdiktion og investering. Hvem der udløser, finansierer, betaler og senere får nytte af infrastrukturen skal dokumenteres hver for sig.

## Hvad bør stå ved en påstand om netomkostninger?

Når en rapport, avisartikel eller politisk analyse hævder, at AI eller datacentre kræver en bestemt netinvestering, bør læseren kunne finde mindst seks oplysninger tæt på tallet.

**Tabel 15.3 – Minimumsoplysninger ved påstande om infrastruktur- og netomkostninger**

| Oplysning | Spørgsmål |
|---|---|
| Projekt og belastning | Hvilket konkret load eller hvilken projektpipeline ligger bag? |
| Kapacitet vs. energi | Er tallet MW til rådighed, gennemsnitlig MW eller årlige TWh? |
| Counterfactual | Ville investeringen være nødvendig uden det nye datacenterload? |
| Betalingsmekanisme | Hvilke dele betales via tilslutning, effekt, energi, abonnement eller andre vilkår? |
| Forecast-vintage | Hvilket års planlægningsforudsætninger anvendes, og hvor moden er projektpipen? |
| Rest- og nytteværdi | Kan aktivet bruges af andre kunder eller systemet, hvis load ændres eller forsvinder? |

*Kilde/ramme: Egen syntese af de danske og internationale large-load-kilder omtalt i kapitlet.*

## Infrastrukturregningen er hverken automatisk privat eller automatisk fælles

Det centrale resultat i dette kapitel er derfor metodisk. Et stort AI-relateret elforbrug kan skabe behov for ny infrastruktur, men omkostningen kan ikke placeres korrekt ved at se på én kabelregning eller én tarif alene. Samme investering kan indeholde kundespecifikke aktiver, fælles netaktiver, fremtidig systemværdi og risiko for uudnyttet kapacitet.

Den politisk neutrale analyse begynder med at vise pengestrømmen og risikoen. Først derefter giver det mening at diskutere, om fordelingen er rimelig, effektiv eller ønskelig.

Med dette kapitel er den direkte og indirekte økonomiske regning kortlagt fra virksomhedens abonnement og API-forbrug til menneskelig arbejdstid og den fælles infrastruktur. I næste del vender vi perspektivet. AI har også potentiale til at undgå omkostninger, reducere ressourceforbrug og skabe ny værdi. De gevinster skal vurderes med samme disciplin som belastningerne.

### Kilder nævnt i kapitlet


• Forsyningstilsynet (2026): Analyse 11 - Overblik over elsektorens tariffer. 25. marts 2026.

• Energinet (2025): Energinets Tarifkatalog 2026. Officiel tarifpublikation med forecast for omkostninger og forbrugsgrundlag.

• Energistyrelsen (2025): Analyseforudsætninger til Energinet 2025 - Datacentre samt sammenfatningsnotat og datasæt.

• Commission for Regulation of Utilities, Ireland (2025): Decision on New Electricity Connection Policy for Data Centres. 12. december 2025.

• Virginia State Corporation Commission og Missouri Public Service Commission: nyere large-load-tarifmekanismer, anvendt som ikke-europæisk designkontrast.

• Energy and AI (2026): Power-infrastructure expansion planning for training-oriented AI data centers under large-model scaling uncertainty, vol. 25.

• Li, Yueying et al. / Microsoft Research (2026): PowerSlider: Exploiting Phase Asymmetry for LLM Serving under Demand Response, arXiv preprint.

Del 5

# Det AI kan spare

Del 5 undersøger den anden side af regnskabet. AI kan skabe dokumenterbare gevinster og undgåede omkostninger, men de skal måles mod en tydelig baseline og efterfølgende korrigeres for ændringer i aktivitet og rebound.

## Kapitel 16

# Gevinster og undgåede omkostninger

En bog om AI's omkostninger bliver skæv, hvis den kun tæller det, teknologien bruger. AI kan også reducere arbejdstid, fejl, energiforbrug, ventetid, spild eller behovet for andre ressourcer. De gevinster skal med i regnskabet, når de kan dokumenteres.

Metodekravene er de samme for positive påstande som for påstande om strøm, vand og $CO_2$. En besparelse eksisterer kun i forhold til noget, der ellers ville være sket. Derfor begynder gevinstregnskabet med en baseline og et kontrafaktisk scenarie.

Hvis en medarbejder bruger AI til at skrive et notat på 30 minutter i stedet for 50, kan forskellen være relevant. Hvis kvaliteten samtidig ændres, hvis notatet ellers ikke ville være skrevet, eller hvis den frigivne tid bruges til flere opgaver, er '20 minutter sparet' ikke længere hele resultatet. Den samme logik gælder for energi, materialer og emissioner.

### En gevinst kræver en tydelig baseline

Den kontrafaktiske baseline er beskrivelsen af, hvad der realistisk ville være sket uden AI. Den kan være den nuværende arbejdsproces, et eksisterende regelsystem, traditionel analytics, manuel kontrol eller en anden teknologisk løsning. Det relevante sammenligningspunkt er ikke nødvendigvis 'ingen teknologi'.

GHG Protocol bruger samme grundprincip i project accounting: effekten af et projekt vurderes ved at sammenligne projektforløbet med et kontrafaktisk baselineforløb uden interventionen. Den metode er særlig vigtig ved påstande om undgåede emissioner, fordi et positivt klimaresultat ikke kan udledes af AI-løsningens eget lave energiforbrug alene.

Bogens metode anvender derfor en fast regel: den undgåede ressource skal navngives, den nye AI-belastning skal med, og kvaliteten eller funktionen skal være sammenlignelig mellem de to scenarier.

*Kilde: GHG Protocol, Inventory and Project Accounting: A Comparative Review (2023), suppleret med bogens egen scenariesyntese.*

**Tabel 16.1 – Forskellige typer gevinst må ikke tælles som det samme**

| Gevinsttype | Hvad der kan forbedres | Hvad der skal sammenlignes | Typisk risiko for fejltolkning |
|---|---|---|---|
| Tids-/ kapacitetsgevinst | Kortere arbejdstid eller højere throughput | Samme accepterede output med og uden AI | Sparet tid behandles automatisk som kontant lønbesparelse. |
| Kvalitetsgevinst | Færre fejl, bedre svar, højere first-pass-kvalitet | Samme opgave og uafhængigt kvalitetskriterium | Højere kvalitet tælles oven i tidsgevinsten, selv om de to mål overlapper. |
| Undgået direkte omkostning | Mindre køb af ekstern service, overtid, rework eller anden udgift | Den konkrete udgift, som faktisk bortfalder | En teoretisk tidsværdi behandles som en realiseret cash saving. |
| Undgået fysisk ressource | Energi, brændstof, materialer, transport, spild eller kapacitet | Samme service, produktion og sikkerhedsniveau | Modelpotentiale omtales som målt nettobesparelse. |
| Undgået risiko/tab | Færre fejl, hændelser, nedetid eller forsinkelser | Observeret eller modelleret risikoforskel med samme eksponering | Et muligt undgået tab behandles som sikker årlig gevinst. |
| Ny aktivitet / mulighed | En aktivitet bliver mulig, som ellers ikke ville være udført | Værdi og ekstra ressourcebrug rapporteres separat | Ny aktivitet kaldes “besparelse”, selv om der ikke findes en undgået baseline. |

*Kilde/ramme: Egen syntese af kilderne og beregningerne i dette kapitel.*

## Produktivitet er et resultatmål, ikke en universel faktor

Der findes nu solide eksperimenter, hvor generativ AI har forbedret produktivitet på konkrete opgaver. Resultaterne spænder samtidig så bredt, at de ikke kan omsættes til én generel procent for 'AI-produktivitet'.

Noy og Zhang fandt i et randomiseret forsøg med 453 professionelle deltagere, at ChatGPT reducerede tiden på afgrænsede skriveopgaver med 40 procent og løftede den bedømte kvalitet med 18 procent. Brynjolfsson, Li og Raymond fandt i en virkelig kundeserviceorganisation med 5.172 agenter, at en GPT-baseret assistent øgede antallet af løste kundesager pr. time med omkring 15 procent i gennemsnit, med langt større gevinst hos mindre erfarne medarbejdere end hos de mest erfarne.

Dell'Acqua og kolleger viste i et randomiseret forsøg med 758 management consultants, at GPT-4 inden for modellens kapabilitetsgrænse gav 12,2 procent flere gennemførte delopgaver og 25,1 procent højere hastighed samt bedre kvalitet. På en opgave uden for frontieren var AI-brugerne derimod 19 procent mindre tilbøjelige til at nå det korrekte resultat.

Det modsatte resultat findes også. METR målte i 2025 erfarne open-source-udviklere på egne, velkendte repositories og fandt, at de tilladte AI-værktøjer gjorde opgaveløsningen 19 procent langsommere i den konkrete setting, selv om deltagerne både før og efter forsøget mente, at AI havde gjort dem hurtigere.

De fire studier modsiger ikke hinanden. De måler forskellige opgaver, brugere, modeller, kvalitetskrav og arbejdssammenhænge. De dokumenterer tilsammen, at gevinstens størrelse er lokal til workflowet.

*Kilder: Noy & Zhang (Science, 2023); Brynjolfsson, Li & Raymond (Quarterly Journal of Economics, 2025); Dell'Acqua et al. (Organization Science, 2026); Becker et al./METR (2025).*

**Tabel 16.2 – Fire produktivitetsstudier, fire forskellige resultater**

| Studie / setting | Primært mål | Observeret effekt | Vigtig begrænsning |
|---|---|---|---|
| Noy & Zhang, 2023 - professionelle skriveopgaver | Tid + bedømt kvalitet | 40 % kortere tid; 18 % højere kvalitet | Afgrænsede, incentiverede skriveopgaver - ikke et helt job. |
| Brynjolfsson et al., 2025 - kundeservice | Løste sager pr. time | Ca. 15 % højere i gennemsnit | Én virksomhed og rolle; stor forskel efter erfaring/skill. |

| Studie / setting | Primært mål | Observeret effekt | Vigtig begrænsning |
|---|---|---|---|
| Dell'Acqua et al., 2026 - management consulting | Tasks, hastighed og kvalitet | +12,2 % tasks; +25,1 % hastighed inden frontier | Effekten vendte på en opgave uden for AI-frontieren. |
| METR, 2025 - erfarne open-source-udviklere | Faktisk completion time | 19 % længere tid med de testede AI-værktøjer | Snæver high-context setting og early-2025 tools. |

*Kilde: Noy & Zhang (2023); Brynjolfsson, Li & Raymond (2025); Dell'Acqua et al. (2026); Becker et al./METR (2025). Tallene er studiespecifikke og må ikke anvendes som universelle business-case-faktorer.*

## En opgavegevinst bliver ikke automatisk en virksomheds- eller samfundsgevinst

Der er flere trin mellem at løse en enkelt opgave hurtigere og at ændre virksomhedens samlede omkostninger, produktion eller beskæftigelse. Det danske studie af Humlum og Vestergaard er nyttigt netop her. Forskerne kombinerer store adoptionsundersøgelser med administrative arbejdsmarkedsdata og finder omfattende brug af AI-chatbots og nye AI-relaterede opgaver, men ingen målbar effekt på registrerede arbejdstimer eller indtjening på arbejdsplads- og medarbejderniveau i den observerede periode; i den seneste revision kan effekter over 2 procent afvises.

Resultatet peger på, at gevinster på opgaveniveau kan blive omsat til ændret arbejdsfordeling, højere kvalitet, mindre arbejde uden for normal tid, nye opgaver eller andre former for værdi, før de bliver synlige som færre løntimer eller højere samlet output.

For denne bog er konsekvensen enkel: opgaveeffekt, procesgevinst, virksomhedsøkonomi og samfundsøkonomi skal stå i hver sin kolonne. Man må ikke hoppe direkte fra den første til den sidste.

*Kilde: Humlum & Vestergaard, Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI, NBER Working Paper 33777, revision marts 2026.*

## AI kan også reducere fysiske ressourcer

Gevinstregnskabet stopper ikke ved kontorarbejde. AI og avanceret analytics kan bruges til styring af bygninger, industrielle processer, transport og elsystemer. Her kan den relevante gevinst være mindre

energiforbrug, lavere brændselsforbrug, mindre spild, færre tab eller udskudt kapacitetsbehov.

IEA's Energy and AI analyserer sådanne anvendelser som et selvstændigt spor. Rapporten bruger blandt andet optimeret HVAC-styring som et eksempel, hvor omkring 10 procent energibesparelse kan være muligt, og peger på tilsvarende muligheder i transport og industri. IEA's Widespread Adoption Case modellerer potentielle emissionsreduktioner på omkring 1.400 Mt $CO_2$ i 2035 fra eksisterende AI-anvendelser på tværs af sektorer.

Det store tal er et scenarieresultat, ikke en observeret global besparelse. IEA knytter selv resultatet til udbredt adoption og til barrierer som data, kompetencer, digital infrastruktur, regulering og sikkerhed. En politisk eller økonomisk analyse bør derfor ikke skrive, at AI 'vil spare 1,4 Gt $CO_2$'. Den præcise formulering er, at IEA i et bestemt udbredelsesscenarie modellerer dette potentiale for 2035.

*Kilde: International Energy Agency, Energy and AI (2025), afsnittet "AI and climate change". IEA-resultatet er et Widespread Adoption-scenarie og må ikke behandles som målt eller garanteret global effekt.*

### Potentiale, realiseret bruttoeffekt og nettoeffekt er tre forskellige tal

En teknisk model kan vise, at en proces kan bruge mindre energi med AI. Først når løsningen implementeres, ved vi hvor meget af potentialet der realiseres. Derefter skal AI-systemets egne ekstra ressourcer, integration og drift trækkes fra for at finde nettoeffekten inden eventuel rebound.

Denne rækkefølge er værd at gøre synlig, fordi den forhindrer, at en teknisk procent bliver til et ubetinget bæredygtighedsløfte.

**Tabel 16.3 – Illustrativt eksempel: fra teknisk potentiale til nettoeffekt**

| Led | Illustrativ beregning | Resultat |
|---|---|---|
| Baseline | Proces uden AI | 1.000 MWh/år |
| Teknisk potentiale | 12 % lavere ressourcebehov | 120 MWh potentiel besparelse |
| Realiseret bruttoeffekt | 75 % af potentialet opnås i drift | 90 MWh realiseret bruttobesparelse |
| AI-/styringsoverhead | Ekstra compute, sensors og drift | 5 MWh/år |
| Netto før rebound | 90 - 5 MWh | 85 MWh/år = 8,5 % |

*Illustrativ beregning. Alle tal er konstruerede for at vise metoden. De er ikke et benchmark for bygninger, industri eller andre AI-anvendelser. Kapitel 17 behandler aktivitetseffekt og rebound.*

## Ny aktivitet kan være værdifuld uden at være en besparelse

Nogle af de vigtigste AI-anvendelser har ingen meningsfuld 'undgået' aktivitet. En virksomhed kan analysere flere kundesager, oversætte materiale til sprog, den tidligere ikke dækkede, eller gennemføre en kontrol, der før blev fravalgt af tidsmæssige grunde. I sådanne tilfælde gør AI en ny aktivitet mulig.

Den nye aktivitet kan have stor økonomisk eller samfundsmæssig værdi. Ressourceregnskabet ser alligevel anderledes ud, fordi hele den nye aktivitet er inkrementel i forhold til baseline. Der findes ikke en tidligere proces, hvis energi eller arbejdstid automatisk kan trækkes fra.

Det er derfor mere præcist at rapportere værdien og den ekstra ressourcebelastning separat end at forsøge at presse den nye aktivitet ind i et besparelsesregnskab.

**Påstanden under lup**

**Påstand:** AI vil spare mere $CO_2$ i resten af økonomien, end datacentrene udleder.

**Vurdering:** Ikke dokumenteret som en generel fremtidig konklusion.

**Kort svar:** Der findes troværdige use cases og scenarier, hvor AI kan reducere emissioner langt uden for datacenteret. Nettoresultatet afhænger af adoption, baseline, faktisk realiserede besparelser, AI-systemets egen belastning og efterfølgende ændringer i aktivitet.

**Hvad siger evidensen?** IEA modellerer i sit Widespread Adoption Case omkring 1.400 Mt $CO_2$ i potentielle reduktioner i 2035 fra eksisterende AI-anvendelser og sammenligner dette med datacenteremissioner i sine scenarier. Resultatet viser et muligt størrelsesforhold under bestemte antagelser - ikke en målt global klimabalance for AI.

**Det afgørende forbehold:** Undgåede emissioner er et kontrafaktisk impact-regnskab og skal holdes adskilt fra virksomheders Scope 1-3-inventar. GHG Protocol kræver, at avoided emissions rapporteres separat og ikke trækkes fra virksomhedens emissionsinventar.

*Kilder: IEA, Energy and AI (2025); GHG Protocol, Inventory and Project Accounting: A Comparative Review (2023) og Scope 3 FAQ om avoided emissions.*

### Den samme gevinst må kun tælles én gang

Gevinstregnskaber bliver let for optimistiske, når flere KPI'er beskriver den samme underliggende forbedring. Hvis en medarbejder bliver 20 procent hurtigere og derfor håndterer 25 procent flere ens sager i timen, er tidsgevinsten og throughputgevinsten ikke nødvendigvis to uafhængige værdier. Hvis virksomheden samtidig værdisætter de frigivne minutter som lønbesparelse og de ekstra sager som omsætning, kan den samme kapacitet blive monetiseret to gange.

Kvalitet kan skabe en selvstændig gevinst, når færre fejl faktisk reducerer rework, reklamationer eller risiko. Her bør den økonomiske værdi knyttes til den dokumenterbare konsekvens i stedet for at lægge en vilkårlig procent oven på produktivitetsgevinsten.

En robust business case viser derfor både de fysiske eller operationelle KPI'er og den økonomiske konvertering. Den gør også klart, hvilke gevinster der overlapper.

**Tabel 16.4 – Minimumsoplysninger ved en AI-gevinstpåstand**

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Baseline | Hvad ville realistisk være sket uden AI? |
| Functional unit | Er output, kvalitet, service og sikkerhed sammenlignelige? |
| Gevinstmål | Måles tid, kvalitet, throughput, energi, omkostning, risiko eller noget andet? |
| AI-merbelastning | Er compute, tools, integration, review og drift medregnet? |
| Realiseringsgrad | Er tallet teknisk potentiale, observeret bruttoeffekt eller nettoeffekt? |
| Dobbeltregning | Beskriver flere KPI'er den samme underliggende gevinst? |
| Tid og geografi | Gælder resultatet den aktuelle model, proces, lokation og periode? |
| Scale/rebound | Ændrer lavere pris eller hurtigere proces den samlede aktivitet? |

*Kilde/ramme: Egen syntese af kilderne i kapitlet. Rebound og samlet aktivitet behandles særskilt i kapitel 17.*

## Gevinsten er en del af omkostningsregnskabet

AI's omkostninger kan ikke vurderes seriøst ved kun at tælle datacenterets ressourcer. Hvis en AI-anvendelse dokumenterbart reducerer en større ressource eller omkostning andetsteds, hører den undgåede belastning med i beslutningsgrundlaget.

Det samme krav gælder begge veje: negative effekter må ikke overdrives ved at ignorere gevinster, og positive effekter må ikke overdrives ved at bruge modelpotentiale som realiseret besparelse.

Den mest robuste formulering er derfor sjældent 'AI sparer 20 procent'. Den beskriver i stedet hvilken proces der blev sammenlignet, hvilket resultat der blev opnået, hvilke ekstra AI-ressourcer der blev lagt til, og om effekten er målt, modelleret eller fremskrevet.

Der mangler endnu ét led. Når en opgave bliver billigere, hurtigere eller mere energieffektiv, kan vi begynde at udføre flere af dem. Den ekstra aktivitet kan reducere, fjerne eller i nogle tilfælde overgå den oprindelige besparelse. Kapitel 17 undersøger denne rebound-effekt.

### Kilder nævnt i kapitlet


• Noy, Shakked & Whitney Zhang (2023): Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654), 187-192.

• Brynjolfsson, Erik; Danielle Li & Lindsey Raymond (2025): Generative AI at Work. The Quarterly Journal of Economics 140(2), 889-942.

• Dell'Acqua, Fabrizio et al. (2026): Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science 37(2), 403-423.

• Becker, Joel et al. / METR (2025): Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.

• Humlum, Anders & Emilie Vestergaard (2025; rev. 2026): Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI. NBER Working Paper 33777.

• International Energy Agency (2025): Energy and AI - AI and climate change.

• GHG Protocol (2023): Inventory and Project Accounting: A Comparative Review; samt guidance/FAQ om avoided emissions.


## Kapitel 17

# Rebound og de indirekte effekter

Kapitel 16 viste, at AI kan skabe dokumenterbare gevinster: mindre arbejdstid, højere kvalitet, lavere energiforbrug i en proces eller en anden undgået omkostning. En effektiviseringsgevinst er imidlertid kun første led i regnestykket. Når en opgave bliver billigere, hurtigere eller mindre ressourcekrævende, kan organisationer og brugere begynde at udføre flere af den. Nye anvendelser kan opstå, flere brugere kan få adgang, og den frigjorte tid eller de frigjorte penge kan blive brugt andre steder.

Det er her rebound-effekten bliver relevant. Den beskriver den del af en forventet ressourcebesparelse, som bliver modvirket af ændret aktivitet efter en effektivisering. Rebound kan være lille, stor eller i nogle tilfælde større end den oprindelige besparelse. Størrelsen kan ikke antages på forhånd; den skal måles eller modelleres for det konkrete system.

For AI er problemstillingen særlig vigtig, fordi både pris, ydeevne og tilgængelighed ændrer sig hurtigt. En inference kan blive billigere og mindre energikrævende, samtidig med at antallet af inferenser, brugere, modeller og anvendelser vokser. Derfor skal enhedsforbrug og samlet aktivitet vises i hver sin kolonne.

*Kilder/ramme: Freire-González, Font Vivanco & Puig-Ventosa (2021); Luccioni, Strubell & Crawford (FAccT 2025).*

### Rebound begynder efter effektivitetsgevinsten

Antag, at en bestemt AI-opgave kan udføres med 50 procent mindre energi efter en software- eller hardwareforbedring. Hvis antallet af opgaver er uændret, falder det samlede energiforbrug til den opgave tilsvarende. Hvis opgavevolumen samtidig fordobles, forsvinder hele den forventede energibesparelse. Vokser aktiviteten endnu mere, kan det samlede forbrug ende over udgangspunktet.

Rebound er dermed ikke en påstand om, at effektivisering er nytteløs. Effektiviseringen kan være både teknisk reel og økonomisk værdifuld. Rebound beskriver, hvordan adfærd og aktivitet påvirker den besparelse, som ellers ville være opnået ved uændret aktivitet.

Den rigtige reference er derfor et kontrafaktisk forløb: Hvad ville det samlede ressourceforbrug have været uden effektivitetsforbedringen, under sammenlignelige betingelser? Det spørgsmål er mere krævende end blot at sammenligne årets samlede elforbrug med sidste års.

### Rebound og Jevons er ikke det samme

Begreberne bruges ofte som synonymer, men det gør diskussionen unødigt polariseret. Rebound kan dække enhver situation, hvor en del af

den potentielle effektiviseringsgevinst bliver spist op af mere aktivitet. Hvis 30 procent af den forventede besparelse forsvinder, er der rebound, mens den samlede ressourceanvendelse stadig ligger under baseline.

Jevons-paradokset eller backfire er den stærkere situation, hvor rebound overstiger 100 procent. Den øgede aktivitet er da så stor, at det samlede ressourceforbrug efter effektiviseringen bliver højere end i det relevante udgangspunkt. Den generelle rebound-litteratur bruger netop denne skelnen, og bogen fastholder den for at undgå at kalde enhver vækst i AI-forbrug for "Jevons".

For AI findes der plausible mekanismer for begge dele, men der findes endnu ikke en robust empirisk faktor, som viser, at generativ AI som helhed har en bestemt rebound-procent. Luccioni, Strubell og Crawford kortlægger i deres FAccT-studie fra 2025 en række mekanismer og fremhæver behovet for at analysere både tekniske, økonomiske og adfærdsmæssige effekter. Studiet er en tværfaglig syntese, ikke en måling af én samlet global AI-rebound.

*Kilder: Freire-González, Font Vivanco & Puig-Ventosa (2021), Energy Research & Social Science; Luccioni, Strubell & Crawford (2025), FAccT, DOI 10.1145/3715275.3732007.*

### Flere forskellige mekanismer kan øge aktiviteten

Rebound omkring AI behøver ikke opstå gennem én enkel prisreaktion. En billigere model kan blive brugt oftere. En hurtigere workflow kan gøre det attraktivt at behandle sager, som tidligere blev fravalgt. Et bedre system kan skabe helt nye produkter. AI kan også påvirke efterspørgslen uden for datacenteret gennem anbefalinger, automatisering og ændret forbrugeradfærd.

**Tabel 17.1 – Fire mekanismer, der kan ændre aktivitet efter en AI-effektivisering**

| Mekanisme | Hvad ændres? | AI-eksempel | Hvad den ikke dokumenterer alene |
|---|---|---|---|
| Direkte/service-level rebound | Den samme tjeneste bliver billigere eller lettere at bruge. | Lavere pris eller kortere svartid fører til flere requests. | At samlet energiforbrug nødvendigvis stiger. |
| Skala og nye use cases | Flere brugere, opgaver eller funktioner bliver økonomisk mulige. | AI bruges på dokumenter, billeder eller sager, der tidligere ikke blev behandlet. | At væksten skyldes effektivisering frem for bedre capabilities eller adoption. |
| Indirekte/economy-wide | Sparet tid eller penge anvendes i andre aktiviteter. | Frigjort arbejdstid bruges til mere produktion eller andre digitale tjenester. | At den nye aktivitet har samme ressourceprofil som den gamle. |
| Adfærd og induktion | AI ændrer valg og efterspørgsel. | Anbefalingssystemer eller chatbots påvirker køb, mobilitet eller serviceforbrug. | At effekten altid er negativ; AI kan også understøtte mere bæredygtige valg. |

*Kilde/ramme: Komprimeret syntese af Luccioni, Strubell & Crawford (2025) og de øvrige rebound-kilder i kapitlet. Taxonomien er et kort over mekanismer, ikke målte rebound-procenter.*

## Enhedsforbrug og samlet forbrug skal stå side om side

Den enkleste måde at holde regnskabet klart på er at vise både ressourceforbruget pr. funktionel enhed og antallet af funktionelle enheder. For en afgrænset AI-tjeneste kan den første tilnærmelse skrives som: samlet ressourceforbrug = ressource pr. accepteret opgave × antal opgaver. I et fuldt systemregnskab kommer faste server-, netværks-, køle- og øvrige overheadposter oveni.

Følgende eksempel er konstrueret for at vise matematikken. Vi antager, at en softwareforbedring reducerer energien pr. opgave fra 1,00 Wh til 0,25 Wh, altså 75 procent. Udgangspunktet er 100 millioner opgaver.

**Tabel 17.2 – Illustrativt eksempel: 75 procent lavere energi pr. opgave kan give forskellige totaler**

| Scenario | Energi pr. opgave | Antal opgaver | Samlet energi og ændring mod baseline |
|---|---|---|---|
| Baseline | 1,00 Wh | 100 mio. | 100 MWh |
| Samme aktivitet efter effektivisering | 0,25 Wh | 100 mio. | 25 MWh (-75 %) |
| Aktivitet fordobles | 0,25 Wh | 200 mio. | 50 MWh (-50 %) |
| Aktivitet firedobles | 0,25 Wh | 400 mio. | 100 MWh (0 %) |
| Aktivitet seksdobles | 0,25 Wh | 600 mio. | 150 MWh (+50 %) |

*Illustrativ beregning. Alle tal er konstruerede for at vise sammenhængen mellem enhedsforbrug og aktivitet; de beskriver ingen konkret model eller tjeneste.*

I dette eksempel er firedobling af aktiviteten break-even for energien. Under den grænse er der fortsat en nettoenergibesparelse i forhold til baseline; over den bliver det samlede forbrug højere end før. Det afgørende er, at rebound ikke kan læses ud af effektiviseringsprocenten alene.

## Effektivitet kan virke, selv om det samlede forbrug vokser

Det modsatte fejlslutning forekommer også: Hvis datacentrenes samlede elforbrug vokser, konkluderes det nogle gange, at effektiviseringerne ikke virker. En sådan konklusion kræver et kontrafaktisk scenario.

IEA giver et nyttigt eksempel. Organisationen estimerer datacentrenes globale elforbrug til omkring 415 TWh i 2024. I Energy and AI opstilles et High Efficiency Case, hvor efterspørgslen efter digitale tjenester og AI holdes på samme niveau som i Base Case, men software, hardware og infrastruktur antages at blive mere energieffektive. High Efficiency Case når omkring 970 TWh i 2035 og ligger mere end 15 procent under Base Case.

Det samlede forbrug i High Efficiency Case er stadig mere end dobbelt så højt som 2024-niveauet. Effektiviseringen har alligevel en stor effekt i modellen, fordi det relevante sammenligningspunkt er det højere forbrug, der ville være opstået med samme serviceefterspørgsel og

svagere effektivisering. IEA-scenariet er derfor et godt eksempel på forskellen mellem absolut udvikling og undgået forbrug.

High Efficiency Case er ikke en måling af AI-rebound. Tværtimod holder IEA serviceefterspørgslen sammenlignelig netop for at isolere effekten af stærkere effektivisering. Eksemplet bruges her til at vise, hvorfor vækst i totalforbruget hverken beviser, at effektivisering er mislykket, eller at Jevons-paradokset er indtruffet.

*Kilde: International Energy Agency, Energy and AI (2025), "Energy demand from AI". IEA: ca. 415 TWh datacenterforbrug i 2024; High Efficiency Case ca. 970 TWh i 2035 og mere end 15 % lavere end Base Case ved samme efterspørgsel efter digitale tjenester og AI.*

### Prisfald og adoption viser mekanismen - ikke dens størrelse

AI-markedet har flere af de forhold, der kan gøre rebound relevant. Stanford AI Index 2025 rapporterede, at inferenceprisen ved omtrent GPT-3.5-benchmarkniveau faldt mere end 280 gange fra november 2022 til oktober 2024. Samtidig viser nyere AI Index-data hurtig udbredelse af generativ AI blandt både borgere og organisationer.

De to observationer kan ikke kædes sammen til en kausal rebound-procent. Prisfald kan gøre mere brug økonomisk mulig, mens adoption samtidig påvirkes af nye capabilities, bedre brugerflader, større udbud, organisatoriske investeringer og andre forhold. Adoption er heller ikke det samme som requests, tokens eller energiforbrug.

Evidensen er derfor stærk nok til at dokumentere en relevant mekanisme og en kraftig skalering, men ikke til at skrive, at faldende pris har øget AI-energiforbruget med en bestemt procent. Denne sondring er central for bogens neutrale behandling af rebound.

*Kilde: Stanford Institute for Human-Centered AI, AI Index 2025 og 2026.*

### Den samlede AI-rebound er endnu ikke robust kvantificeret

Tianqi Xiao og kolleger modellerede i Nature Sustainability i 2025 miljøpåvirkningen fra AI-servere i USA frem mod 2030 og lod forskellige udviklingsbaner skelne mellem højere hardwareeffektivitet og højere

applikationsvækst. Studiet viser metodisk, hvorfor efficiency og activity bør modelleres separat.

Forfatterne fremhæver samtidig rebound som en mulig modvirkning af effektivitetsgevinster, men skriver, at den komplekse rebound-proces ikke kunne simuleres robust med tilstrækkeligt friske data. Det er en vigtig negativ konklusion: forskningen giver grundlag for at tage mekanismen alvorligt, men ikke for at sætte en universel procent på den.

Når nye studier forsøger at levere ét samlet reboundtal for AI, bør læseren derfor kontrollere systemgrænse, geografi, periode, funktionel enhed, prisrespons, modeludvikling og hvilke former for aktivitet der tælles med. Et makrotal uden disse oplysninger risikerer at skjule mere, end det forklarer.

*Kilde: Xiao et al. (2025), Nature Sustainability, "Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA". USA-modellen anvendes som metodisk evidens, ikke som direkte europæisk proxy.*

### Rebound kan ligge uden for datacenteret

Kapitel 16 viste, at AI kan bruges til at reducere ressourceforbrug i andre sektorer. Den samme systemgrænse skal bruges, når indirekte effekter vurderes. Hvis AI gør en transporttjeneste billigere, kan mobiliteten stige. Hvis AI gør markedsføring og individualiserede anbefalinger mere effektive, kan forbruget af varer eller tjenester ændre sig. Hvis automation frigør arbejdstid, kan den bruges til mere produktion.

IEA bruger autonome køretøjer som et eksempel på denne mekanisme: højere effektivitet kan reducere energi pr. transportydelse, mens lavere omkostninger og større tilgængelighed samtidig kan øge personmobiliteten eller flytte brugere fra kollektiv transport. Transporteksemplet er ikke en måling af rebound for generativ AI; det viser, at AI's nettoeffekt kan opstå langt fra selve datacenteret.

European Environment Agency fremhævede i 2026 et tilsvarende europæisk styringsproblem omkring forbrug. AI kan hjælpe borgere,

virksomheder og offentlige indkøbere til mere bæredygtige valg, men personalisering, anbefalingssystemer og andre AI-funktioner kan også forstærke ressourceintensive forbrugsmønstre. EEA vurderer samtidig, at evidensen om de nuværende og historiske effekter er begrænset. Det er derfor mere præcist at beskrive disse som potentielle mekanismer og styringsspørgsmål end som målte nettogevinster eller nettotab.

*Kilder: IEA, Energy and AI (2025), system-/transporteksempel; European Environment Agency, Briefing 09/2026, Artificial intelligence and sustainable consumption in Europe. EEA understreger begrænset evidens og betingede effekter.*

### Rebound er ikke automatisk spild

Et større ressourceforbrug efter en effektivisering kan være miljømæssigt vigtigt uden at være økonomisk eller samfundsmæssigt værdiløst. En billigere AI-tjeneste kan gøre oversættelse, tilgængelighed, analyse eller rådgivning mulig for mennesker og organisationer, der ellers ikke ville have fået den service. Den ekstra aktivitet har en ressourceomkostning, men kan samtidig skabe reel værdi.

Derfor bør miljøregnskabet og værdiregnskabet stå ved siden af hinanden. Hvis en løsning bruger 20 procent mindre energi pr. accepteret resultat, men virksomheden producerer dobbelt så mange nyttige resultater, skal begge oplysninger vises. Den samlede energi kan stige, mens produktiviteten eller den leverede service også stiger.

Denne opdeling forhindrer to modsatrettede overdrivelser. Rebound bør ikke bruges som bevis for, at effektivisering er meningsløs. Omvendt bør ny økonomisk eller social værdi ikke bruges til at skjule, at det samlede fysiske ressourceforbrug er vokset.

*Kilde/ramme: Luccioni, Strubell & Crawford (2025) samt bogens egen scenariesyntese.*

**Påstanden under lup**

**Påstand:** Når AI bliver mere energieffektiv, falder AI's samlede energiforbrug.

**Vurdering:** Ikke dokumenteret som en automatisk sammenhæng.

**Kort svar:** Lavere energi pr. sammenlignelig opgave er en reel effektiviseringsgevinst. Det samlede energiforbrug afhænger samtidig af antallet af opgaver, brugere, modeller, capabilities, datacenterarkitektur og anden aktivitet i systemet.

**Hvad siger evidensen?** IEA viser i sit High Efficiency Case, at stærkere effektivisering kan reducere det kontrafaktiske datacenterforbrug markant, selv mens den absolutte efterspørgsel fortsat vokser. AI-litteraturen identificerer samtidig flere plausible reboundmekanismer, men den samlede rebound for generativ AI er endnu ikke robust kvantificeret.

**Det afgørende forbehold:** Vækst i samlet datacenter- eller AI-energiforbrug er forenelig med rebound, men er ikke i sig selv bevis for Jevons-paradokset. En kausal vurdering kræver en baseline og en forklaring på, hvor meget af aktivitetsvæksten der skyldes effektivisering, og hvor meget der skyldes nye capabilities, flere brugere, større workloads eller andre forhold.

*Kilder: IEA, Energy and AI (2025); Luccioni, Strubell & Crawford (2025); Xiao et al. (2025).*

## Hvad bør måles, når rebound undersøges?

Et rebound-regnskab bør kunne efterprøves fra enhed til total. Det kræver flere oplysninger end en graf over samlet datacenterforbrug. Særligt vigtigt er det at fastholde samme funktionelle enhed over tid, fordi bedre modeller kan levere en anden kvalitet eller løse opgaver, som ældre modeller slet ikke kunne løse.

**Tabel 17.3 – Minimumsoplysninger ved en påstand om AI-rebound**

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Baseline | Hvad ville det samlede ressourceforbrug have været uden den konkrete effektiviseringsændring? |
| Funktionel enhed | Sammenlignes samme nyttige output, kvalitet og sikkerhed? |

| Oplysning | Spørgsmål læseren bør kunne besvare |
|---|---|
| Enhedsforbrug | Hvor meget energi, vand, tid eller penge bruges pr. opgave/resultat før og efter? |
| Aktivitetsvolumen | Hvor mange opgaver, brugere eller outputs køres før og efter? |
| Årsag til vækst | Skyldes mere aktivitet lavere pris/energi, bedre capabilities, nye use cases, markedsvækst eller en kombination? |
| Systemgrænse | Tælles kun inference, eller også træning, netværk, storage, facility-overhead og relevante indirekte effekter? |
| Geografi og tid | Hvor og hvornår opstår ressourceforbruget, og er resultaterne overførbare? |
| Værdi/output | Hvilken ekstra service eller produktion følger med den øgede aktivitet? |

*Kilde/ramme: Egen syntese af rebound-kilderne samt kapitlerne 3, 16 og 17.*

## Fra rebound til pris pr. accepteret resultat

Rebound-analysen understreger, hvorfor samlede etiketter som miljøgevinst eller miljøbelastning bliver for grove. Enhedsforbruget kan falde, totalforbruget kan stige, og den leverede værdi kan samtidig vokse. Alle tre udsagn kan være sande på samme tid.

For beslutningstagere er den vigtigste praktiske regel derfor todelt: mål først omkostningen og ressourceforbruget pr. sammenligneligt nyttigt resultat, og mål derefter hvor mange resultater organisationen faktisk producerer. Effektivitetsmålet uden aktivitetsmålet er ufuldstændigt; totalmålet uden en funktionel enhed fortæller heller ikke, om teknologien er blevet bedre eller dårligere.

Det leder direkte til bogens sidste metodiske syntese. Prisen pr. token eller pr. API-kald fortæller kun om én teknisk enhed. I næste kapitel samler vi modelpris, menneskelig reviewtid, fejl, retries, tools og fallback i et mere beslutningsrelevant mål: omkostningen pr. accepteret resultat.

### Kilder nævnt i kapitlet

• Luccioni, Alexandra Sasha; Emma Strubell & Kate Crawford (2025): From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. FAccT '25, DOI 10.1145/3715275.3732007.

• Freire-González, Jordi; Ignacio Font Vivanco & Lluc Puig-Ventosa (2021): Policies and systemic alternatives to avoid the rebound effect. Energy Research & Social Science 72, 101732.

• Xiao, Tianqi et al. (2025): Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. Nature Sustainability.

• International Energy Agency (2025): Energy and AI - Energy demand from AI samt relevante systemeksempler om indirekte effekter.

• Stanford Institute for Human-Centered AI (2025; 2026): AI Index Report - Research and Development samt Economy/Overview.

• European Environment Agency (2026): Artificial intelligence and sustainable consumption in Europe. EEA Briefing 09/2026.

## Del 6

# Hvad koster det nyttige resultat?

Del 6 samler bogens metode omkring det resultat, der faktisk kan bruges. En lav enhedspris er først beslutningsrelevant, når kvalitet, retries, menneskelig kontrol, fallback og den samlede systemgrænse er gjort synlige.

## Kapitel 18

# Fra pris pr. token til pris pr. accepteret resultat

En AI-model kan være billig pr. million tokens og dyr i den arbejdsgang, virksomheden faktisk skal have til at fungere. Omvendt kan en model med en højere teknisk enhedspris vise sig billigere, hvis den oftere leverer et resultat, der kan bruges, uden ekstra forsøg, omfattende rettelser eller menneskelig overtagelse.

Det er derfor bogens økonomiske analyse ender et andet sted end ved pris pr. token. Den relevante beslutningsenhed er det resultat, som opfylder den kvalitet, sikkerhed og funktion, organisationen på forhånd har krævet. I dette kapitel kalder vi det et accepteret resultat.

Prisen pr. accepteret resultat samler flere af de spor, vi allerede har gennemgået. Kapitel 13 viste den direkte leverandør- og platformpris. Kapitel 14 viste menneskelig review- og korrektionsarbejde. Kapitel 16 og 17 viste, at gevinster, aktivitet og rebound kræver en tydelig baseline. Nu samles disse elementer i én workflow-måling, der kan bruges til at sammenligne alternative modeller og processer på samme kvalitetsniveau.

*Metoderamme: Boominathan et al. (2026) giver uafhængig benchmark-evidens for outcome-normaliseret API-omkostning; OpenAI (2026) bruges som leverandørens enterprise-ramme for fully loaded cost pr. successful task. Bogens samlede workflowmodel er en egen syntese af de dokumenterede omkostningsled.*

### Et accepteret resultat skal defineres før prisen måles

Et accepteret resultat er ikke det samme som et genereret svar. Det er et output eller en afsluttet opgave, som opfylder en på forhånd fastlagt kvalitetsgrænse. For en kundeservicefunktion kan det være en sag, der er løst korrekt og dokumenteret efter virksomhedens regler. For software

kan det være en ændring, der består relevante tests og review. For en juridisk eller økonomisk arbejdsgang kan det være et udkast, der opfylder de faglige og procesmæssige krav, organisationen har sat.

Kvalitetsgrænsen skal være den samme, når to modeller eller workflows sammenlignes. Hvis den ene løsning accepterer flere fejl eller kræver mindre dokumentation, er en lavere pris pr. output ikke en reel økonomisk forbedring på samme funktionelle enhed.

Det samme gælder sikkerhed og ansvarlig eskalation. Et system, der stopper og sender en vanskelig sag til et menneske på det rigtige tidspunkt, kan være mere værdifuldt end et system, der altid afleverer et svar. En korrekt eskalation kan derfor være en del af det designede workflow og ikke nødvendigvis en fejl. Omkostningen ved den menneskelige overtagelse skal stadig med i regnestykket.

## Fra teknisk enhedspris til beslutningsrelevant pris

**Tabel 18.1 – Fire forskellige prisniveauer, der ikke bør forveksles**

| Måleenhed | Hvad den fortæller | Hvad den typisk udelader | God til |
|---|---|---|---|
| Pris pr. token | Leverandørens tekniske pris for input/output eller beslægtede token-enheder. | Success rate, retries, tools, review, correction og fallback. | Billing og optimering af modelkald. |
| Pris pr. kald eller forsøg | Hvad ét konkret model-/tool-forsøg koster. | Om forsøget ender i et brugbart resultat. | Usage-kontrol og teknisk drift. |
| API-pris pr. korrekt løst benchmarkopgave | Modelomkostningen normaliseret efter korrekt løsning i en defineret test. | Virksomhedens review, integration, downstream-risiko og lokale proceskrav. | Uafhængig model-/workflowbenchmarking. |
| Fully loaded pris pr. accepteret resultat | Alle relevante workflowomkostninger divideret med resultater, der når den fastlagte quality bar. | Værdi eller gevinst pr. resultat, medmindre den beregnes separat. | Virksomhedsbeslutninger og sammenligning af reelle workflows. |

*Kilde/ramme: Boominathan et al. (2026); OpenAI, A scorecard for the AI age (17. juli 2026). De fire niveauer er bogens analytiske opdeling.*

## Den fulde formel

Den enkleste generelle form kan skrives som nedenfor. Hvilke poster der faktisk indgår, afhænger af systemgrænsen og workflowet, men nævneren skal altid være de resultater, der har nået den aftalte kvalitetsgrænse.

**Pris pr. accepteret resultat =**
**(model/API + tools + retries + human review + correction/rework +**
**escalation/fallback + relevante driftsomkostninger)**
**/ antal accepterede resultater**

Integrations- og platformomkostninger kan være faste eller periodiske frem for variable pr. opgave. De bør derfor fordeles efter en dokumenteret allokeringsregel, hvis de skal med i en enhedspris. En business case kan med fordel vise både den variable workflowpris og en fully loaded pris inklusive allokerede faste omkostninger. Dermed bliver det tydeligt, hvad der ændrer sig ved større volumen, og hvad der ligger fast.

Værdien af resultatet hører til et separat regnskab. Et accepteret resultat kan koste 20 kroner og skabe meget forskellig økonomisk værdi afhængigt af opgaven. Cost per accepted outcome gør omkostningssiden sammenlignelig; den dokumenterer ikke i sig selv ROI.

## First-pass acceptance og eventual success er to forskellige KPI'er

En løsning, der afleverer et brugbart resultat i første forsøg, har en anden økonomisk profil end en løsning, der når samme slutresultat efter to eller tre forsøg. Den endelige succesrate kan være identisk, mens compute, ventetid, menneskelig opmærksomhed og tool-forbrug er forskellige.

Boominathan og kolleger publicerede i 2026 et peer-reviewed benchmark af syv store sprogmodeller på 814 Python-opgaver inden for data science. Hver opgave kunne få op til tre forsøg. Gemini 2.5 Pro

havde i dette benchmark den højeste samlede succesrate på 81,3 procent og en Pass@1 på 62,2 procent. Retries løftede den samlede succes på tværs af modellerne med omtrent 19-24 procentpoint, og feedback-guidede retries genvandt omtrent 20-31 procent af de oprindelige fejl.

Tallene beskriver dette konkrete codingbenchmark og må ikke bruges som forventede retry-rater i andre workflows. Den overførbare pointe er selve måleskellet: first-pass reliability, retry recovery og eventual success er forskellige egenskaber med forskellige økonomiske konsekvenser.

*Kilde: Boominathan et al. (2026), Empirical Software Engineering 31, article 186, DOI 10.1007/s10664-026-10927-y. Benchmark: 814 data-science codingopgaver, syv LLM'er, op til tre forsøg.*

## Billigst pr. løst benchmarkopgave kan være en anden model end bedst på accuracy

Det samme studie gør prisproblemet konkret. Gemini 2.5 Pro havde den højeste samlede succes i benchmarken, men median API-pris pr. korrekt løst problem blev opgjort til 0,10740 USD. Qwen3-Coder lå på 0,00034 USD pr. løst problem i samme benchmark, en forskel på en faktor 316. Samtidig var der ingen enkelt model, der dominerede alle dimensioner som accuracy, consistency, recovery og cost.

Forskellen er interessant, fordi den viser, at pris pr. token, pris pr. løst problem og kvalitet ikke nødvendigvis giver samme rangordning. Benchmarken er dog stadig en kontrolleret codingtest. Den inkluderer ikke virksomhedens menneskelige review, integration, latency-værdi, compliancekrav eller downstream-konsekvenser. Den kan derfor validere outcome-normalisering uden at fungere som en færdig enterprise-TCO.

*Kilde: Boominathan et al. (2026). Bruges som uafhængig empirisk illustration af cost/accuracy/reliability-trade-offs; ikke som prisbenchmark for andre opgavetyper eller som virksomhedens samlede TCO.*

## Et illustrativt workflow: den dyrere model kan ende billigere

Følgende regneeksempel er konstrueret for at vise metoden. Tallene er ikke markedspriser, benchmarkresultater eller en anbefaling af bestemte

modeller. Begge modeller behandles på samme 1.000 opgaver og skal nå den samme kvalitetsgrænse. Alle output får ét minuts menneskeligt review. Retries sker automatisk i eksemplet. Sager, der fortsat ikke kan afsluttes, går til et manuelt fallback på otte minutter.

**Tabel 18.2 – Illustrativ sammenligning af to AI-workflows ved samme quality bar**

| Parameter | Model A - lav enhedspris | Model B - højere enhedspris |
|---|---|---|
| Opgaver | 1.000 | 1.000 |
| Model/tool-pris pr. forsøg | 0,20 kr. | 0,80 kr. |
| Accepteret i første forsøg | 650 | 850 |
| Opgaver med ét automatisk retry | 350 | 150 |
| Accepteret efter retry | 200 | 100 |
| Manuel fallback | 150 | 50 |
| Menneskeligt review | 1 min. pr. opgave | 1 min. pr. opgave |
| Manuel fallback-tid | 8 min. pr. fallback | 8 min. pr. fallback |
| Illustrativ loaded time cost | 400 kr./time | 400 kr./time |

*Illustrativt metodeeksempel. Alle tal er konstruerede. Review- og fallbacktider er ikke standarder; de skal måles lokalt for den konkrete opgave, risiko og kvalitetsgrænse.*

Model A bruger 1.350 model-/tool-forsøg og får dermed en direkte model-/tool-omkostning på 270 kr. Model B bruger 1.150 forsøg og koster 920 kr. på samme tekniske post. Ser man kun på leverandørregningen, er Model A klart billigst.

Review af de 1.000 opgaver koster i begge scenarier 6.666,67 kr. ved den illustrative timeomkostning. Model A har 150 manuelle fallbacks, svarende til 1.200 minutter eller 8.000 kr. Model B har 50 fallbacks, svarende til 400 minutter eller 2.666,67 kr.

**Tabel 18.3 – Illustrativ fully loaded cost**

| Omkostningspost | Model A | Model B |
|---|---|---|
| Model/tools inkl. retry | 270,00 kr. | 920,00 kr. |
| Menneskeligt review | 6.666,67 kr. | 6.666,67 kr. |
| Manuel fallback | 8.000,00 kr. | 2.666,67 kr. |
| Samlet workflowomkostning | 14.936,67 kr. | 10.253,34 kr. |
| Accepterede/slutafsluttede resultater | 1.000 | 1.000 |
| Pris pr. accepteret resultat | 14,94 kr. | 10,25 kr. |

*Illustrativ beregning. I dette konstruerede eksempel er Model B's model/tool-post 650 kr. højere, mens den samlede pris pr. accepteret resultat er ca. 31 % lavere, fordi færre sager kræver manuelt fallback.*

Eksemplet viser, hvorfor den billigste tekniske komponent ikke nødvendigvis giver det billigste workflow. Det modsatte kan også forekomme. Hvis den dyrere model kun giver en lille kvalitetsforbedring, eller hvis menneskelig review og fallback er næsten identisk, kan dens højere enhedspris dominere. Resultatet skal derfor måles lokalt frem for at blive antaget ud fra modelklasse eller listepris.

## Review, correction og escalation skal måles hver for sig

Menneskelig kontrol er ikke én homogen aktivitet. Et hurtigt faktatjek, en sproglig rettelse, en faglig genberegning og en fuld menneskelig overtagelse har vidt forskellige omkostninger. Derfor bør en virksomhed registrere både hvor ofte de opstår og hvor lang tid de faktisk tager.

De peer-reviewede studier, vi brugte i kapitel 14, viser netop, hvorfor standardminutter er problematiske. I én specialiseret assessment-opgave tog ekspertrevision af AI-genereret materiale længere tid end revision af menneskeskrevet materiale. I et randomiseret farmaceutstudie medførte uhensigtsmæssige AI-råd mere kognitiv verificeringsaktivitet. Ingen af studierne giver en generel minutkoefficient, der kan flyttes til almindeligt kontorarbejde.

Den praktiske konsekvens er enkel: workflowet skal instrumenteres. First-pass acceptance, correction rate, escalation rate og faktisk tidsforbrug bør stå sammen. En høj acceptance rate er kun positiv, hvis quality bar er tilstrækkelig streng; ellers kan den blot skjule fejl, der slipper videre.

*Kilder: Kowal (2025); JMIR pharmacy RCT (2025).*

## Fejl, der slipper igennem, kræver et risikoregnskab

Et accepted-outcome-regnskab kan udvides, når fejl har væsentlige downstream-konsekvenser. Her er det utilstrækkeligt kun at måle reviewtid. En billig prompt kan blive dyr, hvis et forkert svar accepteres

og udløser juridisk håndtering, kundefejl, remediation eller andre hændelser.

Bogens analyse bruger derfor en enkel risikoramme: relevant fejlhyppighed gange sandsynligheden for, at fejlen slipper gennem kontrollen, gange den gennemsnitlige konsekvens, plus den løbende kontrol- og korrektionsomkostning. Hver komponent skal måles eller scenarieberegnes use-case-specifikt. Der findes ingen universel hallucinationsrate, escape-rate eller standardpris på en fejl.

**Forventet risikoomkostning =**
**fejlhyppighed × sandsynlighed for escape × konsekvens + løbende kontrolomkostning**

*Metoderamme: NIST AI 600-1. Formlen er en generisk risikoramme; ingen af komponenterne har en universel standardværdi.*

## Hvad skal virksomheden logge?

Cost per accepted outcome kræver ikke et avanceret forskningsprojekt, men det kræver en mere præcis telemetry end antal prompts og samlet spend. Målingen bør følge opgaven fra første forsøg til enten et accepteret resultat eller et dokumenteret stop.

**Tabel 18.4 – Minimumsdata for cost per accepted outcome**

| Målepunkt | Hvad registreres? | Hvorfor det er nødvendigt |
|---|---|---|
| Attempted tasks | Antal opgaver, der går ind i workflowet. | Giver nævner for success-, correction- og escalation-rater. |
| First-pass accepted | Resultater der opfylder quality bar uden nyt forsøg eller rettelse. | Måler reliability og skjult arbejdsbelastning. |
| Retries | Antal ekstra model-/tool-forsøg og årsag. | Gør retry-cost og stop-regel synlig. |
| Corrections | Andel, severity og faktisk correction-tid. | Skelner små rettelser fra tungt rework. |
| Escalations/fallback | Antal menneskelige overtagelser og completion-tid. | Synliggør restarbejdet og korrekt sikker eskalation. |
| Model/tool spend | Faktisk faktureret eller målt usage pr. workflow. | Knytter teknisk spend til opgaven. |

| Målepunkt | Hvad registreres? | Hvorfor det er nødvendigt |
|---|---|---|
| Human minutes | Review, correction, escalation og evt. integrationstid. | Gør den skjulte arbejdspris målbar. |
| Final accepted outcomes | Resultater der faktisk når den definerede quality bar. | Er nævneren i outcome-cost. |
| Quality/risk bar | Versioneret definition af acceptkrav. | Sikrer at sammenligninger over tid er funktionelt konsistente. |

*Kilde/ramme: Egen økonomisk syntese. Kategorierne er generelle; tærskler og tidsværdier fastlægges lokalt.*

## Mål ikke licensaktivitet som om det var resultatværdi

Licensdata kan være nyttige, men assigned seats, ever-use, månedligt aktive brugere, ugentlig brug og accepterede resultater er forskellige mål. En medarbejder kan bruge en licens sjældent til en meget værdifuld opgave. En anden kan være aktiv hver dag uden at skabe resultater, der kan forsvare omkostningen.

Den økonomiske kæde bør derfor gå fra adgang til relevant brug og videre til accepterede resultater. Licensprisen kan fordeles på de workflows, der faktisk anvender værktøjet, men værdien bør ikke udledes af aktivitetsprocenten alene. De randomiserede workplace-studier omtalt i kapitlet viser netop, at usage-definition og måleperiode kan ændre det rapporterede procenttal markant.

*Kilder: Dillon et al. (2025; 2026).*

## Hold omkostning og værdi adskilt

Når en virksomhed har beregnet 12 kroner pr. accepteret resultat, mangler stadig spørgsmålet om, hvad resultatet er værd. En sparet arbejdstime kan blive til ekstra output, mindre overarbejde, bedre kvalitet, mere kapacitet eller i nogle tilfælde en faktisk reduceret løn- eller contractor-omkostning. De samme frigjorte minutter må ikke samtidig bogføres som både payroll-besparelse og ekstra output, medmindre begge effekter er realiseret og kan dokumenteres separat.

Det er derfor nyttigt at føre to kolonner: omkostning pr. accepteret resultat og dokumenteret værdi pr. accepteret resultat. Forskellen kan bruges i en lokal business case, mens miljøregnskabet fra bogens tidligere dele fortsat står ved siden af. Ingen enkelt kroneværdi bør skjule elektricitet, vand, materialer eller andre fysiske påvirkninger, når de er beslutningsrelevante.

*Kilde/ramme: Kapitel 14 og 16 samt de underliggende produktivitets- og arbejdsstudier.*

**Påstanden under lup**

**Påstand:** Den billigste AI-model pr. token er også den billigste model for virksomheden.

**Vurdering:** Misvisende som generel påstand.

**Kort svar:** Tokenprisen er kun én del af den samlede workflowøkonomi. Success rate, first-pass reliability, retries, tools, menneskelig review, correction og fallback kan ændre rangordningen mellem modeller.

**Hvad siger evidensen?** Boominathan et al. (2026) viser i et peer-reviewed codingbenchmark, at modelrangering ændrer sig, når pris normaliseres pr. korrekt løst problem, og at accuracy, reliability og cost giver forskellige trade-offs. Enterprise-rammen for cost per successful task udvider derefter tælleren med menneskelig tid, review, retries og rework.

**Det afgørende forbehold:** Benchmarkresultaterne kan ikke overføres som enterprise-TCO eller som en universel modelrangering. Virksomheden skal sammenligne samme opgave, samme quality bar og sin egen faktiske workflowtelemetry.

*Kilder: Boominathan et al. (2026); OpenAI, A scorecard for the AI age (2026).*

## Fra enhedspris til bogens samlede konklusion

Pris pr. accepteret resultat løser et centralt måleproblem, men den gør ikke alle AI-omkostninger til ét universelt tal. En workflowpris i kroner fortæller ikke i sig selv, hvor meget elektricitet der bruges, hvilken emissionsfaktor der gælder, hvilket vand der trækkes fra en lokal ressource, eller hvilke materialer der ligger i infrastrukturen. De regnskaber skal fortsat stå med deres egne enheder og systemgrænser.

Metoden gør derimod sammenligningen skarpere. Hvis to AI-løsninger leverer den samme funktion og kvalitet, kan økonomi, energi, vand og andre relevante omkostninger normaliseres pr. accepteret resultat og derefter skaleres med den faktiske aktivitet. På den måde bindes bogens første kapitler om systemgrænser sammen med kapitlerne om drift, økonomi, gevinster og rebound.

Det efterlader ét sidste spørgsmål: Hvad kan vi med rimelig sikkerhed konkludere efter hele denne kæde af evidens, og hvor bør vi fortsat sige, at data er utilstrækkelige? Det er opgaven for bogens afsluttende kapitel.

### Kilder nævnt i kapitlet


• Boominathan, Santhosh Anitha et al. (2026): Empirical benchmarking of large language models for data science coding: a multidimensional evaluation. Empirical Software Engineering 31, article 186. DOI 10.1007/s10664-026-10927-y.

• OpenAI (2026): A scorecard for the AI age, 17. juli 2026. Leverandørens enterprise-metoderamme for successful-task cost og dependability.

• OpenAI (2026): How to manage AI investments in the agentic era, 14. juli 2026. Leverandørvejledning om outcome ROI og usage/spend visibility.

• Kowal, Martin (2025): Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items. International Journal of Selection and Assessment.

• Journal of Medical Internet Research (2025): Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Pharmacists: Randomized Controlled Trial.

• NIST (2024): Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1.


## Kapitel 19

# Hvad kan vi så konkludere?

Efter råstoffer, chips, datacentre, elektricitet, vand, priser, menneskelig arbejdstid, infrastruktur, gevinster og rebound står én konklusion klart:

AI har ikke én omkostning, som kan udtrykkes med ét tal uden samtidig at forklare, hvad tallet dækker.

Et tal kan være korrekt inden for sin egen systemgrænse og stadig blive misvisende, når det bruges til et andet spørgsmål. En watt-time pr. forespørgsel beskriver ikke et datacenters samlede energiforbrug. Et nationalt gennemsnit i gram $CO_2e$ pr. kWh beskriver ikke en teknologis livscyklusudledning. Et vandinput beskriver ikke automatisk vandforbrug i betydningen consumption. En afskrivningsperiode fortæller ikke, hvornår en accelerator fysisk går i stykker. En API-pris fortæller ikke, hvad et accepteret resultat koster virksomheden.

Bogens vigtigste resultat er derfor en metode til at stille det rigtige spørgsmål før tallet vælges. Når funktion, systemgrænse, geografi, tidspunkt, aktivitet, kvalitet og regnskabsmetode er synlige, bliver meget af den tilsyneladende uenighed om AI-omkostninger lettere at forstå.

*Metoderamme: ITU-T L.1801 (2026); NIST AI 600-1; EEA- og UNECE-kilderne anvendt i kapitel 9-10. Kapitel 19 er en syntese af bogens dokumenterede fund og datagab.*

### To skel beskytter mod to almindelige effektivitetsfejl

Det første skel går mellem intensitet og absolut forbrug. Wh pr. token, PUE, WUE, $CO_2$ pr. beregning eller embodied carbon pr. GPU kan alle forbedres, samtidig med at systemets samlede belastning vokser, hvis der produceres eller anvendes langt flere enheder. Et effektivitetsmål skal derfor stå ved siden af det relevante absolutte forbrug, når spørgsmålet handler om samlet belastning.

Det andet skel går mellem energi, effekt, kapacitet og utilisation. Energi beskriver forbruget over tid. Effekt beskriver belastningen på et givent tidspunkt. Kapacitet beskriver den compute-, net-, storage- eller forsyningskapacitet, der er bygget eller reserveret. Utilisation beskriver, hvor stor en del af kapaciteten der faktisk bruges. To systemer med samme årlige TWh kan derfor have meget forskellige krav til net, transformere, backup, lagring og reservekapacitet.

De to skel gør regnskabet mere præcist uden at kræve ét nyt universelt AI-tal. De viser i stedet, hvilke oplysninger der skal stå ved siden af tallet, før en effektivisering eller en systemomkostning kan fortolkes.

*Kilde/ramme: NEXTDC FY26; Energy and AI 25 (2026) om power-infrastructure planning; PowerSlider (2026); samt kapitel 11 og 17. NEXTDC er en konkret australsk driftscase, mens Energy and AI-studiet er simulationsbaseret.*

## Hvad evidensen kan bære

Nogle konklusioner er robuste nok til at fungere som faste holdepunkt for en beslutningstager. De giver stadig ikke ét universelt AI-tal, men de afgrænser, hvilke typer udsagn der kan forsvares, og hvilke forbehold der bør følge med.

**Tabel 19.1 – Konklusioner, som evidensen kan bære**

| Område | Konklusion | Det afgørende forbehold |
|---|---|---|
| Systemgrænse | AI-omkostninger ændrer sig med den valgte systemgrænse. | Sammenligninger kræver samme funktionelle enhed og samme boundary. |
| Hardware | Fremstilling af acceleratorer, HBM, servere og støtteinfrastruktur har et reelt materialemæssigt og klimamæssigt aftryk. | Der findes ikke én robust universel embodied-faktor for alle moderne AI-systemer. |
| Levetid | Regnskabsmæssig useful life, fysisk levetid, økonomisk anvendelighed og faktisk retirement er forskellige størrelser. | CoreWeave-casen viser, at ældre A100-kapacitet kan have økonomisk værdi langt efter lanceringen, men giver ikke en industriwide retirement-fordeling. |
| Træning/inference | Både træning og inference kan være væsentlige livscyklusposter. | Der findes ingen universel procentfordeling mellem de to faser. |
| Elektricitet | Samme kWh kan give meget forskellige $CO_2$-resultater afhængigt af geografi, tidspunkt og metode. | Nationalt gennemsnit, marginal faktor og teknologisk LCA er forskellige målelag. |
| Vand | Vandpåvirkning afhænger af mængde, kilde, withdrawal, return flow, consumption og lokal/sæsonmæssig stress. | En liter er ikke miljømæssigt ens alle steder, og input må ikke automatisk kaldes consumption. |
| Placering | Cloud, edge og lokal AI kan hver være fordelagtige i bestemte workloads. | Sammenligningen skal ske ved samme funktion og kvalitet og med relevante geografi- og hardwaregrænser. |
| Økonomi | Abonnement eller API-pris er kun | Review, retries, rework, fallback, |

| | | |
|---|---|---|
| | en del af virksomhedens direkte og indirekte AI-omkostning. | integration og andre platformposter kan ændre rangordningen. |
| Gevinster | AI kan dokumenterbart forbedre tid, kvalitet eller output på bestemte opgaver. | Effekten varierer mellem opgaver og kan ikke omsættes til én universel produktivitetsprocent. |
| Rebound | Lavere ressourceforbrug pr. opgave kan eksistere samtidig med højere samlet ressourceforbrug. | Stigende totalforbrug alene beviser ikke rebound eller Jevons/backfire. |
| Energi og effekt | Årligt energiforbrug, momentan effekt, installeret eller reserveret kapacitet og utilisation beskriver forskellige egenskaber ved AI-infrastrukturen. | Samme TWh kan give forskellige krav til net, storage, backup og reservekapacitet; model- og case-resultater er ikke universelle koefficienter. |

*Kilde/ramme: Syntese af kapitel 3-18 og de bærende kilder, herunder ITU-T L.1801, UNECE, EEA, IEA, EU's datacenter-rapporteringsramme, NIST samt de peer-reviewede hardware-, workflow- og produktivitetsstudier. Version 1.1 inddrager desuden TSMC 2025 Sustainability Report, CoreWeave Q2 2026, NEXTDC FY26, Alexandroff et al. (2026) og Energy and AI 25 (2026).*

## Tre typer spørgsmål kræver tre typer svar

En del af forvirringen i AI-debatten opstår, fordi tre forskellige spørgsmål bliver besvaret med den samme type tal. Et attribution-regnskab spørger, hvor stor en del af en eksisterende belastning der kan henføres til AI. Et konsekvensregnskab spørger, hvad der ændrer sig, hvis AI-brugen øges, flyttes eller effektiviseres. Et beslutningsregnskab spørger, hvilken løsning der leverer et bestemt resultat billigst eller med lavest relevant belastning.

Et nationalt års-gennemsnit kan være rimeligt i et location-based driftsregnskab og samtidig være utilstrækkeligt til at vurdere den marginale effekt af et nyt stort datacenter. En tokenpris kan være korrekt som leverandørpris og samtidig være utilstrækkelig til at vælge mellem to workflows. En lifecycle-faktor for vindkraft kan være relevant i en teknologi-LCA, men den kan ikke erstatte et lands konkrete elmix.

*Metoderamme: Kapitel 3, 9, 15 og 18; ITU-T L.1801; GHG Protocol Scope 2 Guidance; bogens faste skel mellem attribution, consequence og beslutningsanalyse.*

## Geografi og tidspunkt er en del af regnestykket

Bogen har bevidst Europa som primært referencepunkt, når resultatet afhænger af geografi. Det er ikke et spørgsmål om, hvor evidensen er produceret. Global forskning kan være metodisk stærk og empirisk værdifuld. Problemet opstår først, når en amerikansk elintensitet, lokal vandstress, tarifstruktur eller datacenterarkitektur bruges som direkte proxy for europæiske forhold.

Kapitel 9 viste forskellen tydeligt: EEA's 2024-gennemsnit for elproduktion varierede fra 6 g $CO_2$e/kWh i Sverige til 566 g i Polen i de fem lande, vi sammenlignede. Det samme fysiske forbrug gav derfor næsten en faktor 94 mellem de to illustrative location-based resultater. UNECE's teknologispecifikke livscyklusfaktorer var et andet målelag og blev holdt separat.

Det samme princip gælder vand. En kubikmeter i et område med høj sæsonmæssig vandstress kan have en anden lokal betydning end den samme mængde i et vandrigt område. Derfor bør geografi og tidspunkt stå ved siden af tallet, når de påvirker fortolkningen.

*Kilder: European Environment Agency, 2024-data for GHG-intensitet af elproduktion; UNECE, Life Cycle Assessment of Electricity Generation Options (2022); EEA WEI+; kapitel 9-10. Nationale gennemsnit og teknologiske livscyklusfaktorer er forskellige størrelser.*

## Det, vi stadig ikke kan kvantificere robust

Et troværdigt evidensgrundlag skal også kunne sige, hvor data stopper. Flere væsentlige spørgsmål er metodisk forståelige, men mangler stadig de offentlige, sammenlignelige data, der skulle til for at give et generelt tal. Disse mangler er i sig selv vigtige resultater, fordi de sætter grænser for, hvor præcist en politiker, journalist eller virksomhed kan konkludere.

**Tabel 19.2 – Dokumenterede datagab, der fortsat bør stå åbne**

| Datagab | Hvad vi kan sige nu | Hvad der mangler |
|---|---|---|
| Faktisk retirement-age for moderne AI-acceleratorer | Reliability-data, regnskabsmæssige useful-life-intervaller og konkret markedsadfærd findes. | En robust offentlig fordeling af faktisk retirement, driftstid og utilisation for moderne acceleratorflåder. |

| | | |
|---|---|---|
| | CoreWeave har kontraheret A100-kapacitet ind i 2029. | |
| AI-specifikt e-affald i Europa | EU har solide WEEE-statistikker for brede udstyrskategorier. | Officielle data, der isolerer AI-servere eller acceleratorer som egen affaldsstrøm. |
| Universel vandfaktor pr. AI-accelerator | Fab-, cooling- og site-data samt TSMC's producentdata dokumenterer relevante processer og størrelsesordener. | En transparent, repræsentativ AI-specifik allokering fra fab, HBM, packaging og resten af forsyningskæden til én moderne accelerator. |
| Fast training/inference-fordeling | Begge faser kan måles og skal holdes adskilt. Ny case-LCA viser, at hardware, streaming og systemgrænse kan ændre fordelingen markant. | En universel livstidsandel på tværs af modeller, workloads, volumen, hardwarelevetid og funktionel enhed findes ikke. |
| Samlet AI-rebound | Mekanismer og scenarier kan analyseres, og unit/total kan måles separat. | Robust empirisk kvantificering af samlet AI-rebound på tværs af sektorer. |
| Langsigtet arbejdsmarkeds- og produktivitetseffekt | RCT'er og feltstudier dokumenterer konkrete opgaveeffekter. | Stabile langsigtede kausale estimater for organisationer og hele økonomier efter fuld adoption og omorganisering. |

*Kilde/ramme: Egen syntese af bogens kilder. Datagab beholdes åbne, når evidensen er utilstrækkelig til en generel empirisk faktor, selv om selve analysemetoden er afklaret.*

## Manglende data er ikke det samme som nul

Når et område ikke kan kvantificeres robust, bør det hverken sættes til nul eller udfyldes med et belejligt proxy-tal uden tydelig mærkning. Hvis officiel WEEE-statistik ikke kan isolere AI-hardware, kan vi ikke konkludere, at AI-e-affald er nul. Vi kan heller ikke tage hele WEEE-strømmen og kalde den AI-affald. Hvis acceleratorernes faktiske retirement-fordeling ikke er offentlig, kan en regnskabsmæssig afskrivningsperiode bruges som økonomisk parameter, men ikke som fysisk levetidsmåling.

Den samme disciplin gælder fremtidsscenarier. Et scenario kan være nyttigt til at undersøge konsekvenser af bestemte antagelser, men et modelresultat skal fortsat mærkes som modelleret. Det bliver først observeret evidens, når den relevante udvikling faktisk er målt.

## Fra tal til beslutning: hvornår er et AI-omkostningstal brugbart?

Et beslutningsegnet tal behøver ikke være perfekt. Det skal være tilstrækkeligt transparent til, at en anden person kan se, hvilket spørgsmål tallet besvarer, hvilke data der ligger bag, og hvilke antagelser der kan ændre resultatet. For bogens målgruppe er dette en mere praktisk kvalitetsgrænse end et krav om én universel standardberegning.

**Tabel 19.3 – Minimumskrav til et beslutningsegnet AI-omkostningstal**

| Kontrolpunkt | Spørgsmål, der skal kunne besvares |
|---|---|
| Formål | Hvilken beslutning eller påstand skal tallet understøtte? |
| Funktionel enhed | Pr. hvad måles omkostningen - prompt, token, GPU-time, kWh, bruger, sag eller accepteret resultat? |
| Systemgrænse | Hvilke led er med: model, server, datacenter, net, hardwareproduktion, menneskelig tid, end-of-life? |
| Geografi | Hvilket land, net, vandopland, cloudregion eller leverandørsetup gælder data for? |
| Tid | Hvilket år, hvilken periode og eventuelt hvilke timer beskriver tallet? |
| Datastatus | Er tallet målt, leverandørrapporteret, estimeret, modelleret eller et scenario? |
| Regnskabsmetode | Er el f.eks. location-based, market-based, gennemsnitlig eller marginal; er vand withdrawal eller consumption? |
| Kvalitetsgrænse | Hvis løsninger sammenlignes, leverer de samme funktion og quality bar? |
| Aktivitet | Er enhedstallet skaleret med den faktiske eller scenarie-baserede mængde aktivitet? |
| Usikkerhed | Hvilke antagelser eller følsomheder kan ændre konklusionen væsentligt? |
| Kilde | Kan læseren finde den oprindelige kilde og kontrollere, hvad den faktisk målte? |

*Kilde/ramme: Egen syntese af bogens metodekrav fra kapitel 1-18, særligt kapitel 3, 9, 10, 15 og 18.*

## En hurtig kontrol for journalister og beslutningstagere

Når et meget præcist AI-tal optræder i en pressehistorie, rapport eller business case, er den første kontrol derfor ikke, om tallet har mange decimaler. Kontrollen er, om måleenheden og systemgrænsen er synlige. Derefter kommer geografi og tidspunkt, datastatus og

regnskabsmetode. Først til sidst giver det mening at diskutere, om tallet skal skaleres op til en virksomhed, et land eller hele AI-sektoren.

Denne rækkefølge er særlig vigtig ved populære one-liners. 'Et glas vand pr. prompt', '10 gange en Google-søgning', 'AI bruger strøm som et land' og 'en GPU holder tre til fem år' kan alle spores til konkrete målinger, estimater eller fortolkninger. Problemet opstår, når betingelserne falder væk, mens tallet eller formuleringen bliver stående som en universel konstant.

Påstanden under lup

Påstand: Man kan beregne, hvad AI koster, med ét samlet tal.
Vurdering: Misvisende som generel påstand.
Kort svar: AI har flere omkostningsregnskaber. De kan sammenholdes i en konkret beslutningsanalyse, men energi, klima, vand, materialer, infrastruktur, menneskelig tid og økonomi kræver forskellige enheder og metoder.
Hvad siger evidensen? Bogens livscyklus-, datacenter-, el-, vand-, arbejds- og workflowkilder viser gennemgående, at resultater ændrer sig med systemgrænse, geografi, aktivitet og kvalitetskrav. Cost per accepted outcome kan gøre økonomiske workflows mere sammenlignelige, men omsætter ikke automatisk alle miljø- og samfundspåvirkninger til den samme enhed.
Det afgørende forbehold: Et samlet beslutningstal kan godt konstrueres, hvis en organisation vælger en eksplicit monetariserings- eller vægtningsmetode. Resultatet vil da være et normativt beslutningsværktøj med valgte priser og vægte - ikke en naturgiven universel 'AI-omkostning'.

*Kilde/ramme: Syntese af kapitel 1-18; ITU-T L.1801; NIST AI 600-1.*

## Det mest nyttige fælles mål er ofte resultatet

Når to konkrete AI-løsninger skal sammenlignes, er det ofte mest nyttigt at begynde med det resultat, organisationen faktisk ønsker. Kapitel 18 brugte accepteret resultat som økonomisk nævner, fordi tokenpris, retries og menneskelig fallback ellers kan skjule den reelle workflowpris. Den samme tankegang kan bruges på energi og andre fysiske størrelser, når data findes: Wh pr. accepteret resultat, liter pr. accepteret resultat eller kg $CO_2$e pr. accepteret resultat.

Normaliseringen løser dog kun sammenligningsproblemet, hvis quality bar og systemgrænse er ens. En lille lokal model, der løser en simpel klassifikation, kan ikke sammenlignes direkte med en større cloudmodel,

der løser en anden og vanskeligere opgave, blot fordi begge producerer ét output. Resultatet skal være funktionelt sammenligneligt.

*Kilde/ramme: Kapitel 12 og 18; ITU-T L.1801 om funktionel enhed og systemgrænse; Boominathan et al. (2026) som empirisk eksempel på outcome-normaliseret cost.*

### Bogen ender derfor uden ét facit

Det er muligt at dokumentere reelle belastninger fra AI. Der bruges materialer, elektricitet, vand, datacenterkapacitet og menneskelig arbejdstid. Der opstår investeringer i hardware og infrastruktur, og nogle konsekvenser ligger uden for virksomhedens egen faktura.

Det er samtidig muligt at dokumentere reelle gevinster. På bestemte opgaver kan AI reducere arbejdstid, forbedre kvalitet, øge kapacitet eller gøre nye ydelser økonomisk mulige. Effektiviseringer kan mindske ressourceforbrug pr. nyttigt resultat, selv når den samlede aktivitet senere vokser.

Ingen af de to observationer kan stå alene som en samlet dom over AI. Det relevante spørgsmål er, hvilken opgave der udføres, med hvilken teknologi, i hvilken geografi, over hvilken periode, med hvilken kvalitet, og hvad der sker med aktiviteten, når løsningen bliver billigere eller bedre.

Den vigtigste konklusion i denne bog er derfor en arbejdsregel: Bed om systemgrænsen før du accepterer tallet. Bed om geografien før du accepterer $CO_2$- eller vandfortolkningen. Bed om quality bar og fallback før du accepterer den økonomiske besparelse. Bed om baseline før du accepterer gevinsten. Bed om aktivitet før du konkluderer noget om rebound.

Når de oplysninger er til stede, kan AI's omkostninger analyseres seriøst. Når de mangler, bør præcisionen i konklusionen falde tilsvarende.

Det er et mindre spektakulært svar end ét stort tal. Til gengæld er det langt mere anvendeligt for den, der faktisk skal træffe en beslutning.

## Kilder nævnt i kapitlet


• ITU-T (2026): Recommendation L.1801, Guidelines for assessing the environmental impact of artificial intelligence systems.

• NIST (2024): Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1.

• UNECE (2022): Life Cycle Assessment of Electricity Generation Options.

• European Environment Agency: Greenhouse gas emission intensity of electricity generation; Water Exploitation Index Plus (WEI+); relevante 2024-2026 datasæt og briefings.

• International Energy Agency (2025): Energy and AI.

• European Commission / EU datacenter-rapporteringsrammen under energieffektivitetsdirektivet, anvendt i kapitel 10-11.

• GHG Protocol: Scope 2 Guidance, anvendt til skellet mellem location-based og market-based elregnskab.

• Boominathan et al. (2026): Empirical benchmarking of large language models for data science coding: a multidimensional evaluation.

# Påstande undersøgt i bogen

Bogen undersøger gennemgående udbredte AI-påstande i deres oprindelige kontekst. Registeret gør det muligt hurtigt at finde vurderingen og gå tilbage til det kapitel, hvor evidensen og forbeholdene er forklaret.

| Kapitel | Påstand | Vurdering | Side |
|---|---|---|---|
| 2 | En AI-forespørgsel bruger cirka 10 gange så meget strøm som en Google-søgning. | Forældet som generel nutidig påstand. | 29 |
| 3 | AI bruger lige så meget strøm som Japan. | Misvisende som generel påstand. | 37 |
| 4 | AI's råstofproblem ligger primært i GPU'en. | Misvisende. | 50 |
| 5 | Det meste af acceleratorens produktionsaftryk kommer fra selve GPU-chippen. | Misvisende som generel påstand. | 61 |
| 6 | En AI-GPU holder typisk tre til fem år. | Ikke dokumenteret som en generel fysisk levetid. | 73 |
| 7 | Når AI-hardware udskiftes, bliver den til e-affald. | Misvisende som generel påstand. | 82 |
| 8 | Det er modeltræningen, der står for AI's største energi- og klimaaftryk; inference er kun en mindre restpost. | For kategorisk. | 93 |
| 9 | Hvis to datacentre bruger lige mange kWh, har de samme $CO_2$-aftryk fra elektriciteten. | Forkert som generel påstand. | 102 |
| 10 | En AI-prompt bruger et glas eller en halv liter vand. | Misvisende som generel påstand. | 110 |
| 11 | Et datacenter med PUE 1,1 er cirka 91 procent energieffektivt. | Misvisende uden en præcis definition. | 119 |
| 12 | Lokal AI er grønnere end cloud. | Kan være korrekt i konkrete cases, men er ikke en generel regel. | 129 |
| 13 | Virksomhedens AI-omkostning er abonnementet eller API-regningen. | Misvisende som fuld direkte omkostning. | 139 |

| | | | |
|---|---|---|---|
| 14 | Hvis AI sparer en medarbejder én arbejdstime, har virksomheden sparet én times løn. | Misvisende som generel økonomisk konklusion. | 146 |
| 15 | Datacentre får de øvrige elkunder til at betale for deres netudbygning. | Ikke dokumenteret som en generel påstand. | 156 |
| 16 | AI vil spare mere $CO_2$ i resten af økonomien, end datacentrene udleder. | Ikke dokumenteret som en generel fremtidig konklusion. | 166 |
| 17 | Når AI bliver mere energieffektiv, falder AI's samlede energiforbrug. | Ikke dokumenteret som en automatisk sammenhæng. | 176 |
| 18 | Den billigste AI-model pr. token er også den billigste model for virksomheden. | Misvisende som generel påstand. | 188 |
| 19 | Man kan beregne, hvad AI koster, med ét samlet tal. | Misvisende som generel påstand. | 196 |

# Samlet kildeliste

Kilderne er samlet fra bogens 19 kapitler. De skal læses sammen med de løbende henvisninger i teksten, hvor det fremgår, hvilken påstand, beregning eller metode den enkelte kilde understøtter. Hvor en kilde findes i flere udgaver eller versioner, er den relevante udgave angivet i det kapitel, hvor den anvendes.


ITU-T L.1801 (2026): Guidelines for assessing the environmental impact of artificial intelligence systems

IEA (2025): Energy and AI

EEA (2025): Water scarcity conditions in Europe / Water Exploitation Index Plus (WEI+)

Google (2025): Measuring the environmental impact of AI inference

EPRI (2024): Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption

IEA (2026): Key Questions on Energy and AI

Li et al. (2025): Making AI Less "Thirsty"

Lei et al. / Lawrence Berkeley National Laboratory (2025): The water use of data center workloads

IEA (2025): Energy and AI – AI and energy security

IEA (2026): Global Critical Minerals Outlook 2026 – Executive summary og Outlook

USGS (2025/2026): Key Minerals in Data Centers Infographic

USGS (2025): About the 2025 List of Critical Minerals

IEA (2024): Recycling of Critical Minerals

IEA (2023): Sustainable and Responsible Critical Mineral Supply Chains

Amoah, Brown, Simon, Bazilian & Matisek (2026): Mineral demand from AI data centers: Infrastructure intensity, processing bottlenecks, and supply competition, Resources Policy

NIST / CHIPS Program Office (2024): Final Programmatic Environmental Assessment for Modernization and Expansion of Existing Semiconductor Fabrication Facilities

Wang, Qi et al. (2023): Environmental data and facts in the semiconductor manufacturing industry: An unexpected high water and energy consumption situation

NVIDIA / WSP (2025): Product Carbon Footprint Summary for NVIDIA HGX H100

NVIDIA / WSP (2025): Product Carbon Footprint Summary for NVIDIA HGX B200

imec / KU Leuven (2026): Reducing the Environmental Impact of Wet Chemical Processes for Advanced Semiconductor Manufacturing

Ouattara et al. / Cleaner Waste Systems (2026): Towards cleaner semiconductor packaging: Life cycle assessment of material and location decisions on carbon and water scarcity footprints

European Commission (2026): SWD(2026) 504 final – Chips Act 2.0 Impact Assessment Report

Amazon.com, Inc. (2026): Annual Report 2025 – Form 10-K

Microsoft (2025): Microsoft Annual Report 2025

Meta Platforms, Inc. (2026): Annual Report 2025 – Form 10-K

Ostrouchov, George et al. / Oak Ridge National Laboratory (2020): GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability

Wadenstein, Mattias & Wim Vanderbauwhede (2025): Life cycle analysis for emissions of scientific computing centres, European Physical Journal C

Stojkovic, Jovan et al. / Microsoft Research (2026): Rearchitecting the Datacenter Lifecycle for AI

Li, Ruihao et al. / Meta & The University of Texas at Austin (2026): Hardware Lifecycle-Aware Power Planning in Commercial Hyperscale Datacenters, OSDI 2026

Google Sustainability (2026): Our operations – waste reduction and data center hardware reuse

Microsoft (2025–2026): Circular Centers / zero-waste cloud hardware reporting

European Parliament and Council (2012, amended): Directive 2012/19/EU on waste electrical and electronic equipment (WEEE)

Eurostat (2025): Waste statistics – electrical and electronic equipment, data through 2023

Eurostat (2025): WEEE metadata – env_waselee / open-scope six product categories

Google Sustainability (2026): Bridging the Gap: Operationalizing Circularity in Data Centers

Microsoft (2025–2026): Circular Centers / progress towards zero waste

UNITAR & ITU (2024): Global E-waste Monitor 2024

ITU-T (2026): Recommendation L.1801 – Guidelines for assessing the environmental impact of artificial intelligence systems

Chien, Andrew A. et al. (2026): Strategies and design for increasing AI sustainability, Nature Reviews Clean Technology

Applied Energy (2026): Generative AI impact assessment through a life cycle analysis of multiple data center typologies

European Commission, DG CONNECT (2026): Targeted consultation on measuring energy consumption and emissions of AI models and systems

Vartziotis, Tina et al. (2026): From Tokens to Watt-hours: Analytical Energy Estimation for LLM Inference on Modern GPUs (technical preprint; used for method triangulation)

European Environment Agency (EEA): Greenhouse gas emission intensity of electricity generation in Europe – 2024 country data and methodology

UNECE (2022): Life Cycle Assessment of Electricity Generation Options

GHG Protocol: Scope 2 Guidance – location-based and market-based accounting

GHG Protocol (29 July 2026): Scope 2 consultation feedback and corporate standards development update

IEA (2022): Advancing Decarbonisation through Clean Electricity Procurement

Google Cloud (2026): Carbon free energy for Google Cloud regions – 2024 regional data

Brander, Haxhimusa, Liebensteiner & Taschini (2025): From Average to Marginal: Estimating Emission Factors in Europe's Electricity Markets, SSRN working paper

Global Reporting Initiative (GRI): GRI 303 Water and Effluents 2018 samt GRI Standards Glossary 2025

Commission Delegated Regulation (EU) 2024/1364 om fælles EU-ratingordning og rapportering for datacentre

European Environment Agency (EEA): Water Exploitation Index Plus (WEI+) dataset, 2000-2023, publiceret 29. januar 2026

Parlement de Wallonie (26. oktober 2023): myndighedssvar om Crystal Computing/Google Saint-Ghislain, pumpetilladelse, indtag og retur

Google Data Centers: St. Ghislain, Belgium; Hamina, Finland

Li, Pengfei; Yang, Jianyi; Islam, Mohammad A.; Ren, Shaolei: Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Microsoft (2025): Environmental Sustainability Report og datacenter-water disclosures

Noy, Shakked & Whitney Zhang (2023): Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654), 187-192.

Brynjolfsson, Erik; Danielle Li & Lindsey Raymond (2025): Generative AI at Work. The Quarterly Journal of Economics 140(2), 889-942.

Dell'Acqua, Fabrizio et al. (2026): Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science 37(2), 403-423.

Becker, Joel et al. / METR (2025): Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.

Humlum, Anders & Emilie Vestergaard (2025; rev. 2026): Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI. NBER Working Paper 33777.

International Energy Agency (2025): Energy and AI - AI and climate change.

GHG Protocol (2023): Inventory and Project Accounting: A Comparative Review; samt guidance/FAQ om avoided emissions.

Luccioni, Alexandra Sasha; Emma Strubell & Kate Crawford (2025): From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. FAccT '25, DOI 10.1145/3715275.3732007.

Freire-González, Jordi; Ignacio Font Vivanco & Lluc Puig-Ventosa (2021): Policies and systemic alternatives to avoid the rebound effect. Energy Research & Social Science 72, 101732.

Xiao, Tianqi et al. (2025): Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. Nature Sustainability.

International Energy Agency (2025): Energy and AI - Energy demand from AI samt relevante systemeksempler om indirekte effekter.

Stanford Institute for Human-Centered AI (2025; 2026): AI Index Report - Research and Development samt Economy/Overview.

European Environment Agency (2026): Artificial intelligence and sustainable consumption in Europe. EEA Briefing 09/2026.

Boominathan, Santhosh Anitha et al. (2026): Empirical benchmarking of large language models for data science coding: a multidimensional evaluation. Empirical Software Engineering 31, article 186. DOI 10.1007/s10664-026-10927-y.

OpenAI (2026): A scorecard for the AI age, 17. juli 2026. Leverandørens enterprise-metoderamme for successful-task cost og dependability.

OpenAI (2026): How to manage AI investments in the agentic era, 14. juli 2026. Leverandørvejledning om outcome ROI og usage/spend visibility.

Kowal, Martin (2025): Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items. International Journal of Selection and Assessment.

Journal of Medical Internet Research (2025): Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Pharmacists: Randomized Controlled Trial.

NIST (2024): Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1.

ITU-T (2026): Recommendation L.1801, Guidelines for assessing the environmental impact of artificial intelligence systems.

European Environment Agency: Greenhouse gas emission intensity of electricity generation; Water Exploitation Index Plus (WEI+); relevante 2024-2026 datasæt og briefings.

International Energy Agency (2025): Energy and AI.

European Commission / EU datacenter-rapporteringsrammen under energieffektivitetsdirektivet, anvendt i kapitel 10-11.

GHG Protocol: Scope 2 Guidance, anvendt til skellet mellem location-based og market-based elregnskab.

Boominathan et al. (2026): Empirical benchmarking of large language models for data science coding: a multidimensional evaluation.

TSMC (2026): 2025 Sustainability Report.

CoreWeave (2026): Q2 2026 earnings call, 11. august 2026.

Alexandroff, Ronja et al. (2026): Life cycle assessment of a generative artificial intelligence system: the case of a radio application, International Journal of Life Cycle Assessment.

NEXTDC (2026): FY26 Annual Report.

Reuters (28. august 2026): NEXTDC FY26 reporting, herunder WUE-udviklingen.

Energy and AI (2026): Power-infrastructure expansion planning for training-oriented AI data centers under large-model scaling uncertainty, vol. 25.

Li, Yueying et al. / Microsoft Research (2026): PowerSlider: Exploiting Phase Asymmetry for LLM Serving under Demand Response, arXiv preprint.

# Stikordsregister

Stikordsregisteret samler bogens centrale fagbegreber og gør det lettere at finde tilbage til de steder, hvor begrebet forklares eller anvendes i en central analyse.

| | |
|---|---|
| WUE (Water Usage Effectiveness) | 105–106, 113, 117, 119, 121 |

# Ændringslog

Ændringsloggen viser væsentlige faglige ændringer mellem offentlige versioner. Interne manuskript- og produktionsversioner indgår ikke.

| Version | Dato | Væsentlige ændringer |
|---|---|---|
| 1.0 | 2026 | Første offentliggjorte udgave. |
| 1.1 | 5. september 2026 | Opdateret med nye producent- og driftsdata for semiconductor- og datacenterleddet, ny peer-reviewed case om training/inference, ny dokumentation for længere økonomisk anvendelighed af ældre AI-acceleratorer samt et skarpere skel mellem intensitet og absolut forbrug og mellem energi, effekt, kapacitet og utilisation. Evidensstatus for hardwarelevetid og semiconductor-leddet er styrket. |

# Om forfatteren

Bent Karise er forfatter, underviser og praktisk formidler med mange års erfaring i at omsætte komplekst stof til viden, mennesker kan bruge i deres arbejde.

Han underviser til daglig i tekniske sikkerhedssystemer, herunder alarm, adgangskontrol, brand, videoovervågning og integration mellem systemer. Det er områder, hvor præcision, dokumentation, ansvar og praktisk anvendelse er nødvendige vilkår. Den samme tilgang præger hans arbejde med kunstig intelligens og med kildekritik.

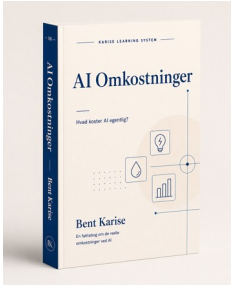
I sit arbejde med AI udvikler Bent Karise undervisningsmaterialer, tekniske manualer, læringsforløb, dokumentation, research og større strukturerede projekter. Karise Learning System er udviklet som en praktisk ramme, der forbinder forståelse med afprøvning, kontrol, dokumentation og anvendelse.

AI Omkostninger udspringer af et andet behov end de øvrige praktiske KLS-bøger. Her er målet at undersøge selve regnestykket bag AI: økonomi, energi, vand, materialer, infrastruktur, menneskelig arbejdstid, gevinster og indirekte effekter. Arbejdet bygger på et løbende evidensgrundlag med peer-reviewede studier, myndighedskilder og teknisk dokumentation.

Ud over undervisnings- og formidlingsarbejdet har Bent Karise bred praktisk erhvervserfaring fra virksomhedsledelse, indkøb, lagerstyring, logistik, produktion, salg, markedsføring, administration, facility management, procesoptimering og personaleledelse.

Læs mere og find den seneste version af bøger og materialer på www.bentkarise.com.

# Karise Learning System

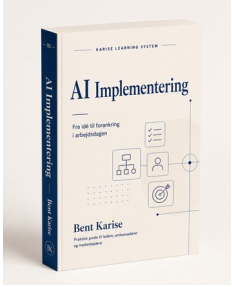



Ét sammenhængende AI-univers
Karise Learning System samler bøger og materialer, der behandler AI fra forskellige vinkler: den daglige brug, implementeringen, sikkerheden, agenterne, ledelsen, omkostningerne og AI i seniorlivet.
Bøgerne kan læses hver for sig. Samlet giver de et bredere billede af, hvordan AI kan forstås, indføres og bruges med tydeligt menneskeligt ansvar.
AI Omkostninger er en del af serien. På de næste sider finder du de øvrige titler.

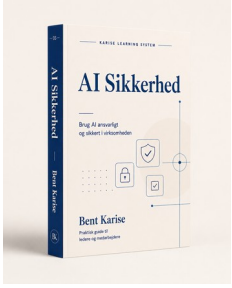
# Flere bøger i Karise Learning System

Bøgernes fokus er forskelligt, men de deler den samme ambition: at gøre AI forståelig og anvendelig i virkelige situationer.


02 · AI i praksis
Fra prompt til automation
Et komplet AI-kursus til virksomheder, ledere og medarbejdere. Bogen fører læseren fra de første prompts og arbejdsmetoder til workflows og automation med fokus på praktisk anvendelse, kontrol og ansvar.

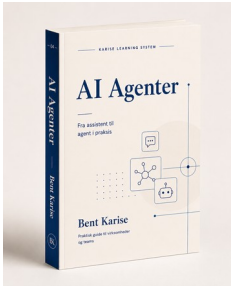



02 · AI Implementering
Fra beslutning til fælles praksis
En praktisk guide til at implementere AI i virksomheden. Bogen fører ledelsen og organisationen fra de første beslutninger og Shadow AI-kortlægning til pilot, kompetencer, måling, drift og forankring.

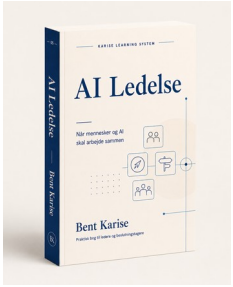



03 · AI Sikkerhed
Sikker AI-brug i virksomheden
Fra daglig brug til robust praksis. Bogen giver medarbejdere og ledere et fælles sprog for data, adgang, prompt injection, menneskelig godkendelse, hændelser og de rammer, der skal være på plads, når AI får mere ansvar.

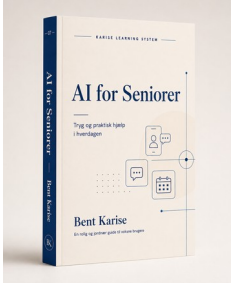
## Fra assistent til ledelse og livskvalitet




04 · AI Agenter
Fra AI-assistent til selvkørende workflow
Forstå, byg og styr agentbaseret arbejde.
Bogen viser vejen fra en almindelig AI-assistent til workflows og agenter, der kan udføre flere trin under klare adgangs-, test- og godkendelsesrammer.




05 · AI Ledelse
Mennesker og teknologi
Når teknologi møder ledelse. Bogen handler om retning, ansvar, kompetencer, tempo, tillid og de ledelsesbeslutninger, der opstår, når AI bliver en reel del af organisationens arbejde.




07 · AI for seniorer
Lettere hverdag. Mere frihed. Mere livskvalitet.
En praktisk AI-guide til dig, der vil have mere ud af hverdagen. Bogen viser AI som en rolig hjælper til kommunikation, praktiske gøremål, indkøb, oplevelser, minder og andre hverdagssituationer.

**Se aktuelle bøger, udgaver og materialer på www.bentkarise.com.**

# Hvad koster AI egentlig?

Det spørgsmål lyder enkelt, men svaret afhænger af, hvad man vælger at tælle med. For nogle handler AI's omkostninger om abonnementer, licenser og API-forbrug. For andre handler det om strøm, datacentre, vand, råstoffer, infrastruktur og klima. I virksomheder ligger en del af regningen også i medarbejdertid, kontrol, implementering og arbejdet med at få et resultat, der faktisk kan bruges.

AI Omkostninger er en faktabog om de reelle omkostninger ved AI. Bent Karise undersøger både den synlige pris og de regninger, der ofte forsvinder i debatten. Bogen følger AI's omkostninger fra råstoffer, chipproduktion og servere til træning, inference, elektricitet, vandforbrug, drift, reboundeffekter og den økonomiske hverdag i praksis.

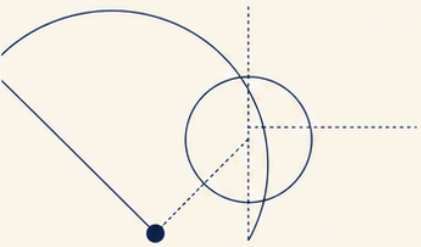
Undervejs tager bogen også fat på en række udbredte påstande: Bruger en AI-forespørgsel ti gange så meget strøm som en søgning? Koster en prompt et glas vand? Hvem betaler for den voksende infrastruktur bag AI? Og hvordan vurderer man egentlig prisen på et nyttigt resultat frem for blot prisen pr. token?

Bogen er skrevet til beslutningstagere, journalister, undervisere, virksomheder og andre læsere, der ønsker et mere præcist, nuanceret og faktabaseret grundlag for at forstå, hvad AI faktisk koster.

---

BK

**Bent Karise** er forfatter, underviser og fagformidler. Han skriver praktiske og oplysende bøger i Karise Learning System med fokus på, hvordan mennesker og virksomheder kan forstå og bruge AI mere bevidst, sikkert og ansvarligt.



bentkarise.com