ecoteers

WHITEPAPER

Verstandig omgaan met AI

Duurzaam en efficiënt

Wat een AI-taak werkelijk kost aan energie, water en CO2, en welke keuzes je als team wel en niet in de hand hebt.

Ecoteers · september 2026 · 27 bronnen, onder meer IEA, CBS, Google, Microsoft en de Europese Commissie

ecoteers
digital impact agency

INHOUD

Wat je in deze whitepaper vindt

1 Waarom dit ertoe doet, ook voor een klein team

2 Wat een AI-taak werkelijk kost aan energie, water en CO2

3 Het grote plaatje: datacenters, het stroomnet en wat de aanbieders zelf rapporteren

4 Waar je als gebruiker wel invloed op hebt

5 Het beslisschema: heb ik hier AI voor nodig?

6 Van losse gewoontes naar beleid

7 Wat de regels zeggen

8 Een eerlijke conclusie, inclusief het rebound-effect

De AI-richtlijn op één pagina

Bronnen

In het kort

- Het verschil tussen zorgvuldig en slordig AI-gebruik is geen kwestie van procenten maar van factoren. Tussen de lichtste en de zwaarste taak in de gemeten literatuur zit ruwweg een factor duizend [19].
- Cijfers over AI en energie lopen ver uiteen omdat partijen verschillende systeemgrenzen hanteren. Google meet voor dezelfde prompt 0,10 Wh binnen een smalle grens en 0,24 Wh binnen een brede grens [4].
- De aanbieders worden per vraag snel zuiniger en zien hun totale voetafdruk tegelijk groeien. Google werd per mediane prompt 33 keer efficiënter en rapporteert over 2025 een uitstoot die 81 procent boven 2019 ligt [3 4].
- Het Internationaal Energieagentschap verwacht dat datacenters in 2030 ongeveer 945 TWh gebruiken, net onder 3 procent van de wereldwijde elektriciteit [1]. In Nederland ging in 2024 al 4,6 procent van de stroom naar datacenters [8].
- Als gebruiker heb je invloed op zeven knoppen: modelkeuze, redeneerstand, promptlengte, regeneraties, batchen en cachen, beeld en video, en leverancier en regio. Ze kosten je geen kwaliteit en meestal ook geen tijd.
- De AI-verordening verplicht aanbieders van modellen voor algemene doeleinden om het bekende of geschatte energieverbruik van hun model te documenteren [10]. Dat maakt inkoopvragen over energie de komende jaren normaler.

HOOFDSTUK 1

Waarom dit ertoe doet, ook voor een klein team

AI zit inmiddels in je tekstverwerker, je advertentieplatform, je CRM en je klantenservice. De vraag of je het gebruikt is voor de meeste teams achterhaald. Wat overblijft is hoe je het gebruikt: waarvoor, hoe vaak, met welk model en met welke afspraken.

Begin met de omvang, zodat je jezelf niet groter of kleiner maakt dan je bent. Het Internationaal Energieagentschap schat dat datacenters in 2024 wereldwijd ongeveer 415 TWh elektriciteit gebruikten, ongeveer 1,5 procent van het totaal, en verwacht een verdubbeling naar ongeveer 945 TWh in 2030 [1]. Datacenters zijn goed voor minder dan een tiende van de groei in de wereldwijde elektriciteitsvraag tussen 2024 en 2030, minder dan elektrificatie, elektrisch rijden en airconditioning [1]. In Nederland is het aandeel groter: het CBS meldt dat datacenters in 2024 ongeveer 5.100 GWh afnamen, 4,6 procent van het nationale elektriciteitsverbruik, tegenover 3,3 procent in 2021 [8]. Ongeveer 45 grote locaties nemen daarvan ruwweg 90 procent voor hun rekening [9].

Jouw prompts kantelen dat plaatje niet. Toch zijn er vier redenen waarom dit ook voor een team van vijf of vijftig mensen de moeite waard is.

Je eigen rekening

Energie en kosten lopen bij AI redelijk parallel. Een zwaar model dat lang nadenkt over een onderwerpregel kost je geld, wachttijd en verbruik tegelijk. De keuzes in dit document verlagen alle drie.

Je verhaal klopt of het klopt niet

Maak je duurzaamheid onderdeel van je positionering, dan is een ongecontroleerde AI-praktijk een zwakke plek. Niet omdat het verbruik enorm is, maar omdat je geen antwoord hebt als een klant of inkoper ernaar vraagt. Een richtlijn die je kunt laten zien, is beter dan een schatting achteraf.

De vraag komt sowieso naar je toe

De Europese AI-verordening verplicht aanbieders van modellen voor algemene doeleinden om onder meer de gebruikte rekenkracht en het bekende of geschatte energieverbruik van hun model vast te leggen in technische documentatie [10]. Het Rijk liet in januari 2025 de duurzaamheidskosten van generatief AI-gebruik binnen de overheid doorrekenen en kwam voor drie scenario's uit op 14,6, 27,2 of 53,3 GWh per jaar na drie jaar [16]. Grote inkopers zijn dus al met deze vraag bezig.

Het gaat niet alleen over stroom

De Autoriteit Persoonsgegevens meldde in augustus 2024 meerdere datalekken doordat medewerkers persoonsgegevens in een AI-chatbot plakten, waaronder medische gegevens van patiënten bij een huisartsenpraktijk en een bestand met klantadressen bij een telecombedrijf [17]. De toezichthouder adviseert heldere afspraken te maken over wat medewerkers wel en niet mogen invoeren [17]. Dezelfde richtlijn die je verbruik regelt, regelt ook dit.

Dit document is geen oproep om minder te vragen. Het is een handleiding om per taak een betere keuze te maken.

HOOFDSTUK 2

Wat een AI-taak werkelijk kost aan energie, water en CO2

Er is geen enkel getal dat het antwoord is. Wie zegt dat een AI-vraag zoveel wattuur kost, laat weg welke systeemgrens hij hanteert, welk model hij bedoelt en uit welk jaar de meting komt. Die drie dingen verklaren vrijwel alle verschillen tussen publicaties.

Dezelfde prompt, twee getallen, dezelfde meter

Google publiceerde in 2025 een technisch document over de meting van Gemini in productie. Binnen een brede meetgrens die de actieve chip, de host-CPU en het geheugen, de inactieve reservecapaciteit en de datacenteroverhead meetelt, komt de mediane tekstprompt uit op 0,24 Wh, 0,03 gram CO2-equivalent en 0,26 milliliter water [4]. Binnen de smallere grens die in veel benchmarks gebruikelijk is, alleen de actieve chip en alleen de zuinigste datacenters, komt precies dezelfde prompt uit op 0,10 Wh en 0,12 milliliter [4]. De brede meting is opgebouwd uit 0,14 Wh actieve AI-chips, 0,06 Wh CPU en geheugen, 0,02 Wh inactieve machines en 0,02 Wh overhead [4]. Wie alleen de chip telt, komt er dus ruim twee keer te gunstig uit.

Meetjaar telt net zo zwaar

Dezelfde meting laat zien dat de mediane Gemini-prompt in mei 2024 nog op 7,92 Wh zat en een jaar later op 0,24 Wh, een factor 33 in energie en een factor 44 in CO2 [3 4]. Veelgeciteerde cijfers uit 2023 en 2024 zijn dus niet verzonnen, ze zijn verouderd. De veelgehoorde claim dat een AI-vraag tien keer zoveel stroom kost als een zoekopdracht, ook terug te vinden bij het VN-milieuprogramma, stamt uit die oudere generatie schattingen [15].

Verbruik per taaksoort

Let in deze tabel op de kolom met de meetgrens: die maakt de getallen pas vergelijkbaar.

Taaksoort	Energie per keer	Meetgrens en bron
Tekstclassificatie, sortering, labelen	ongeveer 0,002 Wh	Gemiddelde over 88 modellen, alleen GPU [19]
Tekstgeneratie, open modellen	ongeveer 0,042 Wh	Gemiddelde over 88 modellen, alleen GPU [19]
Korte tekstvraag, moderne productiestack	0,10 tot 0,24 Wh	Smalle tegenover brede meetgrens, mei 2025 [4]
Zelfde vraag, twaalf maanden eerder	7,92 Wh	Brede meetgrens, mei 2024 [4]
Schatting voor een typische GPT-4o-vraag	ongeveer 0,3 Wh	Modelschatting, 500 uitvoertokens [21]
Vraag met ongeveer 10.000 tokens context	ongeveer 2,5 Wh	Modelschatting [21]
Vraag met ongeveer 100.000 tokens context	ongeveer 40 Wh	Modelschatting [21]
Redeneertaak, redeneerstand aan	7,6 tot 12,8 Wh	Gemeten over 1.000 vragen, alleen GPU [18]
Afbeelding van 1024 bij 1024 pixels	0,63 tot 2,9 Wh	Enkel model incl. koeling [22] en gemiddelde over modellen [19]
Video van vijf seconden	ongeveer 944 Wh	Gemeten op een videomodel [22]

De onderste rij is de belangrijkste. Eén gegenereerde video van vijf seconden staat ruwweg gelijk aan vierduizend korte tekstvragen. Tussen de lichtste en de zwaarste taak in het onderzoek van Luccioni en collega's zit een factor van ongeveer 1.450, gemeten over 88 modellen, 10 taken en 30 datasets [19].

Water en CO2: hier lopen de grenzen het verst uiteen

Bron en grens	CO2 per antwoord	Water per antwoord
Google, alleen de serving-stack op de datacenterlocatie [4]	0,03 gCO2e	0,26 mL
Mistral met Carbone 4 en ADEME, hele levenscyclus [7]	1,14 gCO2e	45 mL

Dat lijkt bijna een factor 40 in CO2 en ruim 170 in water, maar het is geen meningsverschil over feiten. Google rekent met het water dat door de koeling van het datacenter gaat, en met een marktgebaseerde emissiefactor die de inkoop van schone stroom verrekent [4]. Mistral maakte met

Carbone 4 en het Franse milieuagentschap ADEME een volledige levenscyclusanalyse volgens de Frugal AI-methode van AFNOR, het GHG Protocol Product Standard en ISO 14040 en 14044, en telt daar het water en de uitstoot bij op van het opwekken van de stroom en het maken van de hardware [7]. Beide getallen kloppen binnen hun eigen grens. Wie er een noemt zonder die grens erbij, zegt weinig.

Trainen tegenover gebruiken

Trainen is een grote eenmalige post. Mistral rapporteert voor het trainen en draaien van Mistral Large 2 over achttien maanden 20,4 kiloton CO2-equivalent, 281.000 kubieke meter water en 660 kilogram antimoon-equivalent aan grondstofuitputting [7]. Afgezet tegen het gebruik in die periode is dat 1,14 gram CO2-equivalent per antwoord van 400 tokens [7]. Op die training heb je als gebruiker geen invloed meer, die is al betaald. Op hoe vaak en hoe zwaar je het model aanroept wel. Daar zit je hele handelingsruimte.

HOOFDSTUK 3

Het grote plaatje: datacenters, het stroomnet en wat de aanbieders zelf rapporteren

Het per-prompt-verhaal wordt pas eerlijk als je er het macroplaatje naast legt. Dat plaatje is niet geruststellend en ook niet apocalyptisch.

Wat de IEA verwacht

De groei in Energy and AI zit vooral bij versnelde servers, de machines waarop AI draait: die groeien in het basisscenario met ongeveer 30 procent per jaar en zijn goed voor bijna de helft van de netto toename tot 2030 [1]. Regionaal verschilt het beeld sterk: de Verenigde Staten komen er ongeveer 240 TWh bij, Europa ruim 45 TWh [1].

In april 2026 meldde de IEA dat de wereldwijde datacenterstroomvraag in 2025 met 17 procent groeide, in lijn met de eerdere projectie, en dat het verbruik van specifiek op AI gerichte datacenters met 50 procent steeg [2]. De investeringen van vijf grote technologiebedrijven liepen in 2025 op tot meer dan 400 miljard dollar en zouden in 2026 nog eens 75 procent hoger uitkomen [2]. Tegelijk lopen die plannen tegen fysieke grenzen aan: levertijden voor gasturbines en transformatoren, en wachttijden voor netaansluitingen en vergunningen [2].

Wat Google en Microsoft zelf rapporteren

Dit is het eerlijkste deel van het verhaal, want het komt uit hun eigen jaarrapporten.

Indicator	Google, boekjaar 2025	Microsoft, boekjaar 2025
Totale uitstoot	ongeveer 14,5 miljoen ton CO2e, 18 procent boven 2024 en 81 procent boven het basisjaar 2019 [3]	21,1 miljoen ton CO2e, tegenover 13,1 miljoen ton in basisjaar 2020 [5]
Uitstoot eigen operatie	2,9 miljoen ton CO2e, 240 procent boven 2019, maar 2 procent lager dan vorig jaar [3]	scope 1 en 2 samen 2,88 miljoen ton CO2e, tegenover 0,57 miljoen ton in 2020 [5]
Elektriciteitsverbruik	43,6 TWh totaal, waarvan 42,4 TWh in datacenters, tegenover 30,6 TWh een jaar eerder [3]	37,0 TWh, tegenover 10,8 TWh in 2020 [5]
Water	zoetwaterverbruik 37 procent hoger dan in 2024 [3]	waterverbruik 8.170 megaliter, tegenover 3.990 in 2020 [5]
Efficiëntie datacenter	gemiddelde PUE van 1,09 over de hele vloot [3]	niet in deze vorm gerapporteerd

Lees die tabel naast hoofdstuk 2 en je ziet de kern van het probleem. Dezelfde bedrijven die per prompt tientallen keren zuiniger werden, zien hun totale voetafdruk fors stijgen, omdat het volume harder groeit dan de efficiëntie. Microsoft schrijft zelf dat de toename wordt gedreven door de uitbreiding van de datacenterinfrastructuur, en dat scope 2 is opgelopen van bijna 2 procent naar 13 procent van de totale uitstoot [6].

Waar de stroom vandaan komt, telt zwaarder dan hoe zuinig je prompt is

Twee factoren buiten het model bepalen een groot deel van de uitkomst. De eerste is de overhead van het datacenter, uitgedrukt in PUE. De Uptime Institute-enquête van 2026 komt over 644 respondenten op een gemiddelde jaarlijkse PUE van 1,52, een cijfer dat al jaren nauwelijks beweegt, en 23 procent zit onder 1,3 [23]. Google rapporteert 1,09 over de eigen vloot [3]. Tussen een gemiddelde serverruimte en een moderne hyperscale-locatie zit dus ruwweg 40 procent extra stroom bovenop hetzelfde rekenwerk.

De tweede is de stroommix. Volgens Electricity Maps had het Nederlandse net in 2025 een gemiddelde intensiteit van ongeveer 262 gram CO2-equivalent per kilowattuur [24], terwijl Frankrijk op ongeveer 32 gram uitkwam bij een aandeel koolstofvrije stroom van 96,3 procent [25]. Hetzelfde rekenwerk op een andere plek scheelt dus ruwweg een factor acht in CO2. Dat is een grotere hefboom dan alles wat je in je prompt kunt doen, en hij ligt bij je leverancier.

HOOFDSTUK 4

Waar je als gebruiker wel invloed op hebt

Zeven knoppen, op volgorde van effect. Geen ervan kost je kwaliteit, en de meeste leveren ook tijd op.

Modelkeuze

Voor dezelfde taak lopen modellen ver uiteen: in de metingen van Luccioni en collega's over 88 modellen scheelt het tussen taaksoorten ruwweg een factor 1.450, en binnen dezelfde taaksoort nog eens ordes van grootte tussen modellen [19]. Laat samenvatten, herschrijven, vertalen en sorteren dus door het lichtste model dat de kwaliteit haalt, en bewaar het zwaarste model voor werk waar je echt een redenering van wilt zien.

De redeneerstand

Dit is de grootste losse knop. Hugging Face meet dat modellen met een redeneerstand gemiddeld ongeveer 30 keer meer energie gebruiken dan modellen zonder [18]. Zet je bij hetzelfde model de stand aan, dan zijn de verschillen extremer: van 49,53 naar 7.626,53 Wh per 1.000 vragen bij een model van 70 miljard parameters, en van 18,35 naar 12.791,22 Wh bij een model van 3 miljard [18]. De oorzaak is dat redeneermodellen 300 tot 800 keer meer tokens produceren [18].

Een peer-reviewed studie in Frontiers in Communication komt lager uit: over 14 modellen en 1.000 MMLU-vragen leverden de redeneerstanden een 4 tot 6 keer hogere uitstoot op dan de tekstmodus, waarbij de redeneervarianten gemiddeld 543,5 denk-tokens per meerkeuzevraag produceerden tegenover 37,7 antwoordtokens [20]. Het verschil zit in de selectie: Hugging Face meet over een brede ranglijst inclusief kleine modellen die enorm doorredeneren, de Frontiers-studie meet een vaste benchmark. Neem 4 tot 30 keer als werkbare bandbreedte en behandel de redeneerstand als een bewuste keuze, niet als standaardinstelling.

Promptlengte en outputlengte

Lange invoer telt kwadratisch mee, omdat het model elk tokenpaar tegen elkaar afweegt. Epoch AI schat dat een vraag met ongeveer 10.000 tokens context op ongeveer 2,5 Wh uitkomt en met ongeveer 100.000 tokens op ongeveer 40 Wh, tegenover ongeveer 0,3 Wh voor een gewone vraag [21]. Een heel contract meeplakken in plaats van de twee relevante artikelen is dus geen kleine slordigheid. Aan de uitvoerkant is de outputlengte de directe kostendrijver [20]. Zeg in je prompt hoeveel woorden je wilt en in welke vorm.

Regeneraties

Elke druk op opnieuw genereren is een volledige nieuwe aanroep, inclusief het opnieuw verwerken van alle context. Vijf keer regenereren en dan de derde kiezen, kost vijf keer zoveel als één keer goed herformuleren. Spreek af dat je na een teleurstellend antwoord niet regenereert, maar

opschrijft wat er mis was.

Batchen en cachen

Hier sla je iets over in plaats van iets te beperken. Bij prompt-caching hergebruikt het model een al berekende vaste voorkant van je prompt: systeeminstructies, merkrichtlijnen, productdata. Anthropic rekent voor gecachte invoertokens 0,1 keer de normale invoerprijs, tegenover 1,25 keer voor het wegschrijven van de cache [26]. Prijs is geen energie, maar de rekenstap die je overslaat is echt. Bij batchverwerking laat je werk dat niet nu af hoeft in een rustiger venster draaien: OpenAI rekent daarvoor 50 procent van het normale tarief met een venster van 24 uur [27]. Vierhonderd productteksten hoeven zelden vandaag klaar te zijn.

Beeld en video

Hier zit de grootste enkele beslissing. Een afbeelding kost tussen 0,63 en 2,9 Wh, afhankelijk van model en meetgrens [19][22]. Een video van vijf seconden kost ongeveer 944 Wh [22]. Vraag daarom altijd eerst of een bestaande foto, een screenshot of een eenvoudige grafiek het ook doet. Meestal is dat inhoudelijk ook de betere keuze: echt beeld van je eigen werk is geloofwaardiger dan gegenereerd beeld.

Hergebruik, leverancier en regio

De goedkoopste aanroep is de aanroep die je niet doet omdat een collega het antwoord al had. Leg goede prompts en bruikbare output vast op een vindbare plek. En stel je leverancier twee vragen: in welke regio draait dit werk, en welke PUE rapporteert de locatie. Het verschil tussen 1,52 en 1,09 aan overhead [3][23] en tussen 262 en 32 gram CO2 per kilowattuur [24][25] is groter dan alles wat je in het promptvenster kunt doen.

HOOFDSTUK 5

Het beslisschema: heb ik hier AI voor nodig?

Een voornemen om zuiniger te prompten houdt niemand vol. Een schema van zes stappen dat je in tien seconden doorloopt wel.

1. **Weet ik het antwoord zelf, of staat het in onze eigen documenten?** Zoek het dan op. Een model raadplegen om te bevestigen wat je al weet, kost tijd en introduceert een risico op een onjuist antwoord dat je vervolgens moet controleren.
2. **Is het een feit, een som, een filter of een opzoekactie?** Gebruik dan een zoekmachine, een spreadsheet of een filter in je CRM. Daar is de uitkomst reproduceerbaar en controleerbaar. Classificeren en filteren kost in de metingen ongeveer 0,002 Wh per keer, honderd keer minder dan een generatieve aanroep [19].
3. **Gaat het om taal, structuur of varianten?** Dan is een taalmodel op zijn plek. Kies het lichtste model dat de kwaliteit haalt, geef alleen de context die nodig is en zeg hoe lang het antwoord mag zijn.
4. **Moet er echt geredeneerd, gepland of tegengesproken worden?** Pas dan zet je de redeneerstand aan. Dat geldt voor analyses met meerdere stappen, voor het doorrekenen van scenario's en voor het vinden van fouten in een redenering. Niet voor het herschrijven van een alinea.
5. **Wil ik beeld of video?** Vraag eerst of bestaand beeld volstaat. Zo nee, spreek af dat een tweede persoon meekijkt voordat er video gegenereerd wordt. Dit is de enige categorie waar één klik echt telt [22].
6. **Kan dit wachten tot vanavond of tot morgen?** Zo ja, zet het in een batch in plaats van in je chatvenster [27]. Dat scheelt de helft van het tarief en het haalt het werk uit de spits.

De eerste twee stappen leveren de grootste besparing op en tegelijk het beste antwoord. Dat is geen toeval: een model inzetten voor werk waar het niet voor bedoeld is, is meestal zowel duurder als slechter.

HOOFDSTUK 6

Van losse gewoontes naar beleid

Individueel gedrag verwatert binnen een maand. Wat blijft, is wat je opschrijft en wie het bijhoudt. Een AI-richtlijn hoeft niet langer te zijn dan twee pagina's.

Wat je vastlegt

Zes onderwerpen zijn genoeg, en ze overlappen grotendeels met wat je om privacyredenen toch al regelt.

- **Modelkeuze per taaksoort.** Welk model is het uitgangspunt voor schrijven, samenvatten, analyseren en code, en bij welk signaal schaal je op naar iets zwaarders.
- **De redeneerstand.** Standaard uit, met een korte opsomming van gevallen waarin hij aan mag.
- **Beeld en video.** Een reden die iemand kan navertellen, en voor video een tweede paar ogen vooraf.
- **Gegevens.** Welke categorieën nooit in een extern model gaan, en wat je doet als het toch gebeurt. De Autoriteit Persoonsgegevens beschouwt dat in veel gevallen als een meldplichtig datalek [17].
- **Kwaliteit.** Dat een model geen cijfers, citaten of klantresultaten invult, dat elk getal een bron heeft, en wie eindredacteur is.
- **Inkoop.** Welke vragen je aan een leverancier stelt over regio, PUE en gerapporteerd energieverbruik.

Wie eigenaar is

Eén persoon, bij naam, met een kwartaalmoment. Bij een klein team is dat vaak degene die ook de abonnementen beheert, want die ziet de factuur. Bij een groter team leg je het bij de persoon die contentkwaliteit bewaakt, omdat de meeste afspraken daarover gaan. Belangrijker dan de functie is dat het niet bij iedereen ligt.

Het onderzoek dat Universiteit Utrecht voor het ministerie van Binnenlandse Zaken deed, komt tot vergelijkbare aanbevelingen: aandacht voor generatieve AI in je strategie voor duurzame digitalisering, duurzaamheid meewegen in de keuze voor een model, en de duurzaamheidskosten monitoren zodat ze een rol kunnen spelen bij inkoop [16]. Het rapport gebruikt daarvoor de term digital sobriety: bewustzijn over de impact als uitgangspunt, in plaats van een verbod [16].

Meten zonder schijnprecisie

Hier gaat het vaakst mis. De verleiding is groot om een dashboard te bouwen dat het CO2-verbruik van je marketingteam in grammen laat zien. Dat getal is niet te onderbouwen. De OESO constateerde al in 2022 dat gegevens over de milieu-impact van AI niet breed beschikbaar zijn in gestandaardiseerde en gevalideerde vorm, en dat dit vooral geldt voor waterverbruik en volledige

levenscyclusimpact [14]. Het VN-milieuprogramma beveelt aan om eerst gestandaardiseerde meetprocedures en openbaarmakingsregels op te zetten [15]. Zolang die er niet zijn, meet je je eigen praktijk met eenvoudiger maatstaven. Vier die wel werken:

1. **Je factuur per maand, uitgesplitst naar model.** De enige harde cijferreeks die je zelf hebt. Een piek is bijna altijd een piek in verbruik.
2. **Het aandeel aanroepen met de redeneerstand aan.** De knop met het grootste effect, en goed te volgen als percentage.
3. **Het aantal gegenereerde video's per kwartaal.** Klein getal, grote weging.
4. **Het aandeel hergebruikte output.** Grof te schatten, maar het stuurt het juiste gedrag.

Zeg er intern en extern bij dat dit indicatoren zijn en geen voetafdruk. Dat is eerlijker dan een getal met twee decimalen dat op zand rust.

HOOFDSTUK 7

Wat de regels zeggen

Dit hoofdstuk beschrijft de hoofdlijn en is geen juridisch advies. Laat de toepassing op jouw situatie toetsen.

De AI-verordening en transparantie over energie

De Europese AI-verordening, verordening (EU) 2024/1689, legt verplichtingen op aan aanbieders van AI-modellen voor algemene doeleinden. Volgens de Europese Commissie gelden die sinds 2 augustus 2025 [12]. Eén ervan raakt dit onderwerp direct: de technische documentatie die een aanbieder moet opstellen, bevat volgens bijlage XI onder meer de rekenkracht die is gebruikt om het model te trainen en de trainingstijd, plus het bekende of geschatte energieverbruik van het model [10]. Is dat energieverbruik onbekend, dan mag het worden gebaseerd op informatie over de gebruikte rekenkracht [10]. Voor de praktische invulling publiceerde de Commissie een gedragscode met een modelformulier waarmee aanbieders die informatie kunnen vastleggen [13]. Dat de wetgever een schatting toestaat, zegt iets over de staat van de meetpraktijk: precieze cijfers zijn nog niet voor elk model beschikbaar.

Wat dat voor jou betekent

Als gebruiker van een AI-assistent ben je doorgaans geen aanbieder van een model voor algemene doeleinden. De verplichting ligt bij de partij die het model op de markt brengt. Maar het effect komt wel jouw kant op: als aanbieders deze gegevens moeten vastleggen, wordt het normaler om er in een inkoopgesprek naar te vragen. Zet die vraag alvast in je standaardlijst, naast wat je al vraagt over verwerking van persoonsgegevens en datacenterregio.

Vrijwillige gedragscodes over duurzaamheid

Artikel 95 bepaalt daarnaast dat lidstaten gedragscodes faciliteren die onder meer gaan over het beoordelen en beperken van de impact van AI-systemen op ecologische duurzaamheid, waaronder energie-efficiënt programmeren en technieken voor efficiënt ontwerp, efficiënte training en efficiënt gebruik van AI [11]. Dat is geen verplichting, maar het laat zien welke richting de wetgever op wijst en het is een bruikbaar haakje om je eigen richtlijn aan te verankeren.

Privacy staat hier los van en is urgenter

De AVG geldt onverkort. De Autoriteit Persoonsgegevens kreeg meerdere datalekmeldingen doordat medewerkers persoonsgegevens deelden met een AI-chatbot, vaak op eigen initiatief en tegen de afspraken in [17]. De toezichthouder wijst erop dat de meeste aanbieders alles opslaan wat wordt ingevoerd, en adviseert duidelijk te maken welke gegevens medewerkers wel en niet mogen invoeren en zo nodig met de leverancier af te spreken dat ingevoerde gegevens niet worden bewaard [17]. Voor de meeste mkb-teams is dit het onderdeel met de grootste kans op directe schade.

HOOFDSTUK 8

Een eerlijke conclusie, inclusief het rebound-effect

De verleiding is om te eindigen met opluchting of met schuldgevoel. Geen van beide is terecht.

Het rebound-effect

De OESO beschreef in haar rapport over de AI-voetafdruk dat efficiëntiewinst kan worden tenietgedaan door rebound-effecten, waarbij technologische vooruitgang leidt tot een versnelling van het verbruik in plaats van een daling [14]. Dat is geen theorie meer. Google meet per mediane prompt een factor 33 minder energie in twaalf maanden [4] en rapporteert over 2025 een totale uitstoot van ongeveer 14,5 miljoen ton CO2-equivalent, 81 procent boven het basisjaar 2019 [3]. Microsoft komt uit op 21,1 miljoen ton over boekjaar 2025, tegenover 13,1 miljoen ton in basisjaar 2020, bij een elektriciteitsverbruik dat ruim verdrievoudigde [5].

Per vraag gaat het hard de goede kant op. In totaal gaat het hard de verkeerde kant op, omdat er veel meer vragen gesteld worden. Efficiëntie is dus noodzakelijk maar niet voldoende, en dat is precies het argument om ook naar het volume te kijken, niet alleen naar de zuinigheid per stuk.

Wat we niet beweren

- Dat individueel gedrag het macroprobleem oplost. Datacenters zijn goed voor minder dan een tiende van de groei in de wereldwijde elektriciteitsvraag tot 2030, minder dan elektrisch rijden en airconditioning [1].
- Dat er een betrouwbaar besparingspercentage te beloven valt. De spreiding tussen bronnen is te groot en hangt volledig af van model, taak, meetgrens en regio.
- Dat AI per definitie een probleem is. Dezelfde rapporten beschrijven toepassingen die energie besparen, en de IEA wijst erop dat de helft van de groei in datacentervraag met hernieuwbare bronnen wordt gedekt [2].

Wat wel klopt

Er is een reële post op je eigen rekening, en de keuzes om die post te verkleinen zijn gratis. Het lichtste model dat volstaat. De redeneerstand uit tenzij je hem nodig hebt. Een korte prompt en een korte gevraagde output. Geen video omdat het kan. Niets door een model halen wat een filter of een formule ook doet. En één persoon die dat elk kwartaal nakijkt op de factuur. Dat is geen groen verhaal en geen technisch verhaal, het is gewoon beter werken.

OM TE GEBRUIKEN

De AI-richtlijn op één pagina

- [ ] Leg per taaksoort vast welk model het uitgangspunt is, en bij welk signaal je opschaalt naar een zwaarder model.
- [ ] Zet de redeneerstand standaard uit. Hij gaat alleen aan bij analyses met meerdere stappen, en de aanvrager noemt kort waarom.
- [ ] Vraag in elke prompt om een lengte en een vorm, bijvoorbeeld 120 woorden in drie alinea's.
- [ ] Plak geen hele documenten mee. Knip de passages eruit die ertoe doen en zeg erbij wat je ermee wilt.
- [ ] Regenereer maximaal één keer. Daarna schrijf je op wat er mis was en pas je de prompt aan.
- [ ] Cache vaste context zoals systeeminstructies, merkrichtlijnen en productdata in plaats van ze elke keer opnieuw mee te sturen.
- [ ] Zet werk dat kan wachten in een batch met een venster van 24 uur, niet in je chatvenster.
- [ ] Genereer beeld alleen met een reden die iemand kan navertellen. Voor video kijkt eerst een tweede persoon mee.
- [ ] Gebruik geen model voor wat een zoekopdracht, een filter of een spreadsheetformule ook doet.
- [ ] Bewaar goede prompts en bruikbare output op één vindbare plek, zodat collega's werk hergebruiken in plaats van overdoen.
- [ ] Leg vast welke categorieën gegevens nooit in een extern model gaan, en behandel een fout hierin als mogelijk datalek.
- [ ] Wijs één eigenaar aan en evalueer deze richtlijn elk kwartaal aan de hand van de factuur per model.

VERANTWOORDING

Bronnen

1 International Energy Agency. *Energy demand from AI (Energy and AI)* (2025). https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

2 International Energy Agency. *Data centre electricity use surged in 2025, even with tightening bottlenecks driving a scramble for solutions* (2026). https://www.iea.org/news/data-centre-electricity-use-surged-in-2025-even-with-tightening-bottlenecks-driving-a-scramble-for-solutions

3 Google. *2026 Environmental Report* (2026). https://sustainability.google/reports/google-2026-environmental-report/

4 Google. *Measuring the environmental impact of delivering AI at Google Scale (arXiv 2508.15734)* (2025). https://arxiv.org/abs/2508.15734

5 Microsoft. *2026 Environmental Sustainability Report: Data Fact Sheet* (2026). https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2026-Microsoft-Environmental-Data-Fact-Sheet-PDF.pdf

6 Microsoft. *Responsibly building the AI future* (2026). https://blogs.microsoft.com/on-the-issues/2026/07/09/responsibly-building-the-ai-future/

7 Mistral AI, Carbone 4 en ADEME. *Our contribution to a global environmental standard for AI* (2025). https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai

8 Centraal Bureau voor de Statistiek. *Datacenters verbruiken 4,6 procent van de elektriciteit* (2025). https://www.cbs.nl/nl-nl/nieuws/2025/51/datacenters-verbruiken-4-6-procent-van-de-elektriciteit

9 Centraal Bureau voor de Statistiek. *Elektriciteit geleverd aan datacenters, 2017-2024* (2025). https://www.cbs.nl/nl-nl/maatwerk/2025/51/elektriciteit-geleverd-aan-datacenters-2017-2024

10 Europese Unie. *AI-verordening (EU) 2024/1689, bijlage XI: technische documentatie voor modellen voor algemene doeleinden* (2024). https://artificialintelligenceact.eu/annex/11/

11 Europese Unie. *AI-verordening (EU) 2024/1689, artikel 95: gedragscodes voor vrijwillige toepassing* (2024). https://artificialintelligenceact.eu/article/95/

12 Europese Commissie. *General-purpose AI models in the AI Act: questions and answers* (2025). https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers

13 Europese Commissie. *The General-Purpose AI Code of Practice* (2025). https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

14 OESO. *Measuring the environmental impacts of artificial intelligence compute and applications: the AI footprint (Digital Economy Papers 341)* (2022). https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/11/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_3dddded5/7babf571-en.pdf

15 UN Environment Programme. *AI has an environmental problem. Here is what the world can do about that* (2025). https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about

16 Rijksoverheid en Universiteit Utrecht. *Eindrapport Generatieve AI en duurzaamheid* (2025). https://www.rijksoverheid.nl/documenten/rapporten/2025/01/31/eindrapport-generatieve-ai-en-duurzaamheid

17 Autoriteit Persoonsgegevens. *Pas op: gebruik van AI-chatbot kan leiden tot datalekken* (2024). https://www.autoriteitpersoonsgegevens.nl/en/current/caution-use-of-ai-chatbot-may-lead-to-data-breaches

18 Hugging Face. *AI Energy Score v2 (Luccioni en Gamazaychikov)* (2025). https://huggingface.co/blog/sasha/ai-energy-score-v2

19 Luccioni, Jernite en Strubell (ACM FAccT). *Power Hungry Processing: watts driving the cost of AI deployment?* (2024).
https://arxiv.org/abs/2311.16863

20 Dauner en Socher (Frontiers in Communication). *Energy costs of communicating with AI* (2025).
https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1572947/full

21 Epoch AI. *How much energy does ChatGPT use?* (2025).
https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

22 MIT Technology Review. *We did the math on AI's energy footprint* (2025).
https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/

23 Uptime Institute. *Global Data Center Survey 2026* (2026).
https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/UptimeInstitute.GlobalDataCenterSurvey.2026.pdf

24 Electricity Maps. *Grid in Review 2025: the Netherlands* (2026).
https://www.electricitymaps.com/grid-in-review-2025/the-netherlands

25 Electricity Maps. *Grid in Review 2025: France* (2026).
https://www.electricitymaps.com/grid-in-review-2025/france

26 Anthropic. *Prompt caching: prijsverhoudingen voor cache-lees- en schrijftokens (documentatie)* (2026).
https://platform.claude.com/docs/en/docs/build-with-claude/prompt-caching

27 OpenAI. *Batch API (documentatie)* (2026).
https://developers.openai.com/api/docs/guides/batch

Hier helpt Ecoteers mee

Gevonden worden in AI-antwoorden en in Google, AI verstandig inzetten in je marketingteam, en een marketingaanpak met één vast team en één aanspreekpunt. Sinds 2019 voor honderden organisaties.

ecoteers.nl/ai · onze visie en diensten rond AI

ecoteers.nl/cases · wat we voor klanten bereikten

Plan een kennismaking: ecoteers.nl/contact · info@ecoteers.nl · 030 - 227 0475

ONDER MEER VOOR

de Buren · CFP GREEN BUILDINGS · OM DENKEN · Landmarkt · BIJENBAAS