



■ VAN SCORE NAAR BEWIJS

Betrouwbare AI in de praktijk

Hoe toets je een model voordat het beslist?

04

DOSSIER

Eerst de beslissing,
dan het model

19

STRIP

De beloning is de baas

36

REKEN MEE

Vier beslissende getallen

Van mooie score naar bruikbaar bewijs

Lees dit nummer als een onderzoek: eerst bepalen we wat de taak is, daarna testen we wat het systeem werkelijk doet.



AXI VRAAGT

Kies één voorbeeld en schrijf vooraf op wat voor jou een overtuigende test zou zijn.



3-9

Data en proefopzet

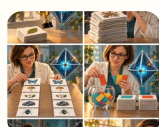
Strip en gegevensketen op pagina 3-7; nieuwe proefquiz op pagina 8; lotingspuzzels op pagina 9.



10-15

Gekoppeld vergelijken

Eerste quiz en feiten op pagina 10-11; gepaarde proef op pagina 12-13; alarmen op pagina 14.



16-21

Scores met een geschiedenis

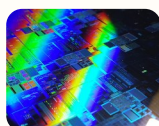
Abstentie op pagina 16; nieuwe strip en statistiekkartjes op pagina 17-18; shortcuts en simulator op pagina 20-21.



22-27

Bronnen en nul fouten

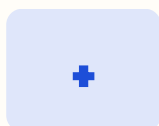
Opzoeken op pagina 22-23; nul-fouten-quiz en grens op pagina 24-25; bevoegdheden op pagina 27.



28-33

Contract en grenzen

Gereedschap op pagina 28; grenspuzzels en werkruimte op pagina 29-30; gedachten en privacy op pagina 31-32.



34-39

Noemers en de liveproef

Nieuwe meetquiz en denkproeven op pagina 34-35; rekenwerk op pagina 36; proef na de test op pagina 37-39.



40-46

Overdracht en audit

Modeldossier op pagina 40; audit en modelkaart op pagina 41-42; eigen onderzoek en bronnen op pagina 45-46.



47-50

Antwoorden en bronnen

Bestaande uitwerkingen op pagina 47; nieuwe puzzels en gedachten op pagina 48; nieuwe quizen op pagina 49; bronnen en vervolg op pagina 50.

VECTOR • AI IN DE PRAKTIJK

Het glimmende scorebord

AXI is een fictieve onderzoeksgenoot. Zijn score kan blinken zonder dat zijn model iets nieuws heeft geleerd.



1 AXI Honderd procent! Mag ik het scorebord alvast oppoetsen?



2 Onderzoeker Dit testonderdeel heeft precies hetzelfde krasje als je oefenonderdeel.



3 AXI En deze ook. Mijn test blijkt mijn geheugenalbum.



4 Onderzoeker Hier zijn echt nieuwe onderdelen. Deze voorspelling was fout.



5 AXI Dan krijgt mijn rapport een eerlijker score én een lijst met fouten.



6 Onderzoeker Laat ook een lege hoek over: wat we nog niet weten.

Illustratie • Het glimmende scorebord

Wat hier misgaat

De test bevat kopieën van trainingsvoorbeelden. Daardoor meet de score gedeeltelijk herkenning van bekende gevallen. Een onafhankelijke test kan prestaties veel minder gunstig laten lijken. Duplicaten opsporen en samen in één deel houden is onderdeel van een eerlijke proef, niet een truc om de score te verlagen.

DENK VERDER

Hoe herken je een kopie als het bestand een andere naam en uitsnede heeft?



Eerst de beslissing, dan het model

Ons fictieve systeem Lumen sorteert foto's van metalen ringen. Wat betekent "goed" als een gemiste scheur en een onnodige afkeuring verschillende gevolgen hebben?

Een afgebakende taak

Lumen ontvangt één foto per ring en voorspelt of een zichtbare beschadiging aanwezig is. Dat is smaller dan "een veilig onderdeel herkennen". Onzichtbare interne scheuren, vermoeiing van het metaal en verkeerde afmetingen vallen buiten deze beeldtaak. Voor die eigenschappen zijn andere metingen nodig. Een model dat de foto correct beoordeelt, heeft daarmee geen volledige productkeuring verricht. We schrijven dus eerst op welke waarneming beschikbaar is, welk label voorspeld wordt en welke beslissing erop volgt.

Twee fouten, twee routes

Een vals negatief is een beschadigde ring die Lumen laat passeren. Een vals positief is een onbeschadigde ring die voor controle wordt aangehouden. Voor dit lesvoorbeeld behandelen medewerkers aangehouden ringen opnieuw; Lumen vernietigt niets en zet geen machine aan. Die begrenzing maakt een menselijke herstelroute mogelijk, maar kost tijd. Het team bepaalt vóór het testen een maximum voor gemiste beschadigingen, een haalbare controlelast en welke onbekende gevallen direct naar een medewerker gaan. Een hoge totaalscore kan geen van die afzonderlijke afspraken vervangen.

Wie stelt de norm?

De datasetschrijver, de modelbouwer en de medewerker aan de controlebank kunnen elk andere problemen zien. Een acceptatiecriterium wordt sterker als zij dezelfde concrete foutgevallen bespreken. Een label als "beschadigd" moet bijvoorbeeld aangeven of oppervlakkige verkleuring meetelt. Als twee beoordeelaars een foto anders classificeren, is dat eerst een probleem van taakdefinitie of waarneming. Meer rekenkracht maakt die onduidelijkheid niet vanzelf kleiner.

WERKHYPOTHESE

Lumen ondersteunt een beperkte beeldcontrole. Alle getallen in dit dossier zijn didactische voorbeelddata, geen productresultaten.

“*Een model dat de foto correct beoordeelt, heeft daarmee geen volledige productkeuring verricht.***”**



■ VECTOR · AI IN DE PRAKTIJK

Een dataset heeft een levensloop

Een map vol foto's krijgt pas betekenis wanneer je weet hoe die foto's zijn gemaakt, geselecteerd en gelabeld.

2018 • Herkomst zichtbaar maken

Het voorstel voor datasheets vraagt niet alleen hoeveel voorbeelden een verzameling bevat. Ook motivatie, verzameling, labels en geschikt gebruik tellen mee. Voor Lumen noteren we camera, lichtbron, productiebatch, opnamevolgorde en welke ringen niet gefotografeerd konden worden. Een leeg veld "onbekend" is informatie: het markeert een grens aan de controleerbaarheid.

2019 • Het model krijgt een bijsluiter

Model cards koppelen resultaten aan beoogd gebruik en beperkingen. Een gemiddelde nauwkeurigheid zegt weinig over een nieuwe camera of over de zeldzaamste scheur. Daarom hoort bij Lumen een tabel per opnameconditie. Dezelfde tabel helpt later bepalen of een wijziging nog binnen de beschreven taak valt.

2023 • Lekkage is ook een procesfout

Onderzoek naar ML-gebruik in wetenschap beschrijft uiteenlopende lekken. Het probleem kan al vóór de trainingscode ontstaan: een bestand bevat informatie die op het beslismoment niet beschikbaar zal zijn, of foto's van hetzelfde object raken verspreid over training en test. We bewaren daarom een object-ID die onafhankelijk is van bestandsnaam en label. Varianten van één ring blijven bij elkaar. Een goede scheiding is een eigenschap van de gegevensroute, niet alleen van een regel die willekeurige rijnummers kiest.

ZO ZIT DAT

Ons gegevenspaspoort

Noteer eenheid van observatie, maker, datum, sensor, selectieregel en ontbrekende gevallen. Bewaar labelinstructies naast de data. Een foto-ID en een object-ID zijn verschillende velden: hetzelfde onderdeel kan meermaals in beeld zijn.

Bronnen: 2, 4

ONDERZOEK

De testset is geen oefenboek

Twee routes hebben verschillende rollen: leren en kiezen links, onafhankelijk beoordelen rechts. Volg dezelfde ring door het dossier en controleer waar informatie kan terugleken. Onze voorbeelden zijn fictief; de scheiding maakt de proef toetsbaar.



VECTOR · AI IN DE PRAKTIJK

Het oefendeel mag open

We delen de ringencollectie op in training, validatie en test. De eerste twee delen helpen kiezen; het laatste deel helpt beoordelen.

Leren en kiezen zijn verschillende stappen

In het trainingsdeel past Lumen zijn gewichten aan. Een validatiedeel helpt kiezen tussen ontwerpen, bijvoorbeeld een grotere of kleinere beeldinvoer. Ook een beslisdrempel kiezen is afstellen. Daarom hoort die keuze bij training of validatie, niet bij de definitieve test. In onze proef zitten alle foto's van dezelfde ring in één deel. Als de latere toepassing nieuwe productiebatches krijgt, is een proef op latere batches informatiever dan een willekeurige verdeling van bijna identieke foto's.

Een eenvoudige audit

Maak vóór training een lijst van object-ID's per deel. De doorsnede tussen training en test moet leeg zijn. Kijk daarnaast naar vrijwel identieke afbeeldingen: een andere uitsnede of belichting verandert het bestand, maar niet het

onderliggende object. Hashes detecteren exacte kopieën; ze bewijzen niet dat twee anders opgeslagen beelden onafhankelijk zijn. De onderzoeker bewaart daarom ook batchcodes en opnamegegevens. Alle automatische voorbewerking wordt op het oefendeel afgestemd. Een gemiddelde of schaalfactor die alvast de testgegevens gebruikt, kan informatie laten lekken.

Leg keuzes vast

We bewaren een manifest met versies en de indeling. Een wijziging van een label of foto levert een nieuwe datasetversie op. Het oude experiment blijft leesbaar, zodat een betere score niet ongemerkt voortkomt uit een andere test.

DATAPUNT

0

gedeelte object-ID's tussen training en test

DATAPUNT

3

delen met verschillende rollen



VECTOR • AI IN DE PRAKTIJK

De eindtest blijft verzegeld

Als het team op de eindtest steeds nieuwe verbeteringen probeert, wordt die test onderdeel van het ontwikkelproces.

Wat gebeurt er bij herhaald proberen?

Stel dat twintig modellen op dezelfde testset worden beoordeeld en alleen de hoogste score verschijnt. Het gekozen model is dan mede geselecteerd op toevallige eigenschappen van die set. Ook zonder één testvoorbeeld in de training te stoppen, kan de gerapporteerde score te optimistisch worden. Het team vermeldt daarom hoeveel varianten zijn geprobeerd, op welke data keuzes zijn gemaakt en welke beoordeling werkelijk apart bleef.

Wanneer openen we de eindtest?

Voor onze ringensorteerder leggen we eerst de taak, drempel, foutmaten en herstelroute vast. Daarna openen we de test één keer voor dat

besluit. Een fout leidt niet automatisch tot eindeloos aanpassen op dezelfde voorbeelden. We onderzoeken hem, documenteren een nieuwe hypothese en plannen zo nodig een nieuwe onafhankelijke evaluatie. De oude test blijft bruikbaar om regressies op bekende problemen te zien, maar krijgt dan het eerlijkere label ontwikkeltest.

Wat leren nieuwe testgegevens?

Recht en collega's maakten nieuwe beeldtestsets volgens bestaande procedures en vonden lagere scores bij de onderzochte classificatiemodellen. Dat historische resultaat is geen universeel getal voor Lumen. Het demonstreert waarom een zorgvuldig gemaakte benchmark niet hetzelfde is als de complete wereld. Onze conclusie blijft gekoppeld aan de geteste ringen, camera's en omstandigheden. Een nieuw station vraagt opnieuw om passend bewijs. ◆



AXI VRAAGT

Welke keuzes heeft jouw team al gemaakt nadat het de testscore zag? Noteer ze alsof je de test morgen opnieuw moet uitleggen.

Quiz: wie kreeg welke proefgevallen?

Vier vragen maken de proefopzet concreet. Iedere keuze moet worden gerechtvaardigd door de genoemde vergelijkingsvraag.

Trek af binnen hetzelfde paar

Gebruik het fictieve experiment op pagina 12–13. Een tijdwinst en het aandeel foutieve labels zijn aparte uitkomsten. Antwoorden staan op pagina 49.

1 A heeft tijden 12,15,11,18 ms. Wat is het gemiddelde?

- ☐ (A) 12 ms
- ☐ (B) 14 ms
- ☐ (C) 16 ms

Waarom?

2 B heeft gekoppelde tijden 10,13,10,15 ms. Wat is het gemiddelde voordeel A–B?

- ☐ (A) 2 ms
- ☐ (B) 8 ms
- ☐ (C) 1 ms

Waarom?

3 A draait alleen op lastige spectra en B alleen op makkelijke. Wat is het probleem?

- ☐ (A) Paarsgewijs testen maakt altijd verkeerde labels.
- ☐ (B) B moet daarom precies tweemaal sneller zijn.
- ☐ (C) Systeemkeuze en taakmix verschillen tegelijk.

Waarom?

4 Een meetdag is vooraf bekend als storende factor. Wat past bij de beschreven blokopzet?

- ☐ (A) B krijgt alleen de gunstigste dag.
- ☐ (B) Vergelijk varianten binnen iedere dag met gelote uitvoervolgorde.
- ☐ (C) Laat de dag weg uit het verslag.

Waarom?

Jouw antwoorden

1 2 3 4

Het lotingsbureau

Drie eindige opzetspuzzels met controleerbare aantallen. Uitwerkingen en bewijzen staan op pagina 48.

1 Vier onderscheiden proefblokken krijgen ieder onafhankelijk volgorde AB of BA. Hoeveel complete volgordelijsten zijn mogelijk?

Mijn antwoord

2 Vier onderscheiden proefblokken moeten precies twee keer AB en twee keer BA hebben. Hoeveel volgordelijsten passen?

Mijn antwoord

Iedere mogelijkheid telt eenmaal

Bij de eerste opgave zijn de vier blokken onderscheiden. De tweede loot echte labels, geen

3 Vier paarsverschillen zijn 2,2,1,3 ms. Je verwisselt in elk paar de namen A en B. Wat worden de verschillen en het gemiddelde?

Mijn antwoord

modeloutputs. De laatste houdt per paar dezelfde identiteit vast.

Quiz: toets de AI-keten

Zes vragen over wat een test wel kan aantonen. Kies eerst; de uitwerkingen geven de reden.

1 Waarom houden we alle foto's van dezelfde ring in één datadeel?

- (A) Omdat de bestandsnamen anders te lang zijn.
- (B) Om objectoverlap tussen oefen- en testgegevens te vermijden.
- (C) Omdat een model maar één foto kan lezen.

Waarom?

2 Welke stap hoort vóór de onafhankelijke eindtest?

- (A) De beslisdrempel afstellen op validatiegegevens.
- (B) Twintig drempels op de eindtest proberen en de beste kiezen.
- (C) Testlabels veranderen tot de score stijgt.

Waarom?

3 Wat bewijst een klikbaar citaat achter een antwoord?

- (A) Dat elke zin waar is.
- (B) Dat het model onafhankelijk heeft gemeten.
- (C) Alleen dat een bron wordt genoemd; steun voor de claim moet apart worden nagegaan.

Waarom?

4 Een externe notitie bevat een opdracht voor een gereedschap. Wat is de passende systeemgrens?

- (A) Elke documentzin uitvoeren.
- (B) Documentinhoud niet automatisch bevoegdheid geven; toegestane acties afzonderlijk controleren.
- (C) Alle logs verwijderen.

Waarom?

5 Waarin schiet een vaste random seed tekort?

- (A) Zij beschrijft niet alle gegevens-, software- en hardwarekeuzes.
- (B) Zij verhindert iedere herhaling.
- (C) Zij verandert labels in betrouwbare waarheid.

Waarom?

6 Een systeem weigert twintig procent van de gevallen. Wat moet het rapport óók beoordelen?

- (A) Alleen de kleur van de weigerknop.
- (B) Alleen de score op beantwoorde gevallen.
- (C) De kwaliteit en werkdruk van de vervolgstap voor geweigerde gevallen.

Waarom?

Jouw antwoorden

1	☐	2	☐	3	☐
4	☐	5	☐	6	☐

■ VECTOR · AI IN DE PRAKTIJK

Zes feiten die je scorebord niet vertelt

Deze feiten verbinden onderzoek met getallen die je zelf kunt controleren. De grenzen bij ieder feit zijn onderdeel van het verhaal.

DATAPUNT

31,25%

precisie in ons zeldzame-schadevoorbeeld

Een zeldzame klasse kan de meerderheid vormen onder fouten

In ons archief met 1% schade leveren 90 terecht en 198 onterechte meldingen samen 288 waarschuwingen op. Ondanks 90% gevoeligheid is maar 31,25% van de meldingen terecht. De basisfrequentie verandert welke vraag een foutpercentage beantwoordt.

DATAPUNT

57

taakgebieden in de oorspronkelijke MMLU

Een test van 57 taken blijft een test

De oorspronkelijke MMLU-benchmark omvat 57 taakgebieden. Dat is een brede toetsopzet, maar geen bewijs dat een model iedere echte vraag goed uitvoert. Taakformaat, selectie en de beschikbare kennis blijven onderdeel van het resultaat.

DATAPUNT

1

relevante passage kan op een ongunstige plek staan

Langere context kan informatie toch verbergen

Lost in the Middle onderzocht hoe de positie van relevante informatie modelprestaties beïnvloedt. Een groot invoervenster is dus geen garantie voor goed gebruik van elke passage. Test een bron op verschillende posities in plaats van alleen meer tekst toe te voegen.

DATAPUNT

10

datasets in de oorspronkelijke WILDS

Tien datasets zijn geen complete wereld

WILDS introduceerde 10 datasets met natuurlijke verdelingsverschuivingen. Die verscheidenheid maakt evaluatie realistischer, maar ook deze verzameling heeft een begrensde dekking. Nieuwe camera's, locaties of tijden kunnen andere problemen toevoegen.

DATAPUNT

0,81

Brier-bijdrage bij 0,9 en uitkomst 0

Een Brier-score straft verkeerd vertrouwen

Voor $p = 0,9$ en $y = 0$ is de binaire bijdrage 0,81; bij $p = 0,6$ is zij 0,36. Een verkeerde voorspelling die veel zekerder klinkt krijgt hier een grotere straf. Dezelfde score beloont in verwachting passende kansrapportage.

DATAPUNT

5

modellen betekenen geen vijf onafhankelijke waarheidsbronnen

Vijf modellen kunnen één blinde vlek delen

Een ensemble voegt meerdere voorspellingen samen. Dat kan onzekerheidsinformatie opleveren, maar vijf modellen die dezelfde achtergrondcorrelatie geleerd hebben kunnen samen verkeerd zitten. Hun eensgezindheid is geen vijfvoudige onafhankelijke meting van de werkelijkheid.



Hetzelfde spectrum, twee verschillende systemen

Een fictief laboratorium vergelijkt twee AI-analyses van spectra. De eerlijkste eerste stap is dezelfde meetgevallen aan beide systemen geven.



FIG. 1 AI-illustratie • een gepaarde spectrumproof; kleuren en curven zijn schematische voorbeelden. AXI en de onderzoeker zijn fictief.

De proef heeft een primaire vraag

Stel dat systeem B een nieuwe classifier is en A de bestaande route. De primaire vraag is verwerkingstijd bij vooraf overeengekomen correctheid. Het lab verzamelt vier vaste spectra en de bijbehorende referentielabels. Beide systemen krijgen elk spectrum. Die paren beperken verwarring tussen modelkeuze en verschillen in moeilijkheid van spectra. Alle tijden en labels in deze casus zijn aangenomen.

Een blok bewaart omstandigheden

Bekende storende factoren, zoals meetdag of operator, kun je in de opzet als blok opnemen. Binnen een blok vergelijk je de bedoelde systeemvarianten en loot je een geschikte uitvoervolgorde. NIST beschrijft deze strategie van blokkeren en randomiseren. Zij maakt onbekende factoren niet magisch nul; een

verkeerd blok of een verborgen interaction kan nog steeds een probleem zijn.

Het verslag begint vóór de run

Leg hardware, opwarmprocedure, datasetversie, toegestane time-outs en foutdefinitie vast. Geef beide routes dezelfde voorbereiding. Bewaar per paar ook het correcte referentielabel en een foutstatus. Een systeem dat razendsnel de verkeerde stof noemt heeft de vooraf gekozen taak niet gehaald. De volgende pagina rekent eerst alleen de verwerkingstijden uit, zodat tijd en correctheid zichtbaar gescheiden blijven.

PROEFCONTRACT

Primair: tijd onder dezelfde correctheidsafspraken. Secundair: foutsoort, spreiding en ontbrekende outputs. Alle proefdata zijn fictief.



■ VECTOR • ONDERZOEK MET AANNAMEN

Vier paren leveren vier verschillen

Per meetgeval aftrekken laat zien waar de snelheid vandaan komt. Het gemiddelde blijft een samenvatting van een kleine gekozen set.

De complete tijdstabel

Neem voor A tijden 12,15,11,18 ms en voor B 10,13,10,15 ms, gekoppeld aan dezelfde vier spectra. De voordelen A-B zijn 2,2,1,3 ms. Gemiddeld is A $56/4 = 14$ ms en B $48/4 = 12$ ms. Het gemiddelde paarsverschil is $8/4 = 2$ ms. Dat is op deze set ongeveer 14,3% van A's gemiddelde tijd.

Een misleidend alternatief

Wie A op vier lastige spectra en B op vier makkelijke spectra laat werken, verandert systeem én taakmix tegelijk. Dan kan een tijdsverschil niet eenduidig aan het systeem worden toegeschreven. Voeg nieuwe zware en lichte spectra vooraf gekozen toe en rapporteer de paarsverschillen per groep. Blijf referentiefouten en time-outs tellen: het tijdgemiddelde mag die niet laten verdwijnen.

DATAPUNT

14 ms

gemiddelde A

DATAPUNT

12 ms

gemiddelde B

DATAPUNT

2 ms

gemiddeld paarsverschil

De spreiding hoort erbij

De verschillen rond het gemiddelde twee zijn 0,0,-1,+1. Hun kwadraten som is twee. De steekproefvariantie van de verschillen is $2/(4-1) = 2/3 \text{ ms}^2$; standaardafwijking ongeveer 0,8165 ms. Deze kleine oefening is geen bewijs dat alle toekomstige spectra twee milliseconden winnen. Herhaalde, apart gehouden gevallen zijn nodig voor de bedoelde inzetomgeving.

DATAPUNT

0,8165 ms

spreiding van vier verschillen

Zeldzaam betekent veel vals alarm

Zelfs een bruikbare detector kan meer verkeerde dan juiste meldingen geven als het gezochte verschijnsel zeldzaam is.

10.000 ringen, twee soorten meldingen



FIG. 1 Didactisch voorbeeld; geen gemeten eigenschappen van een bestaand systeem.

Tel de gevallen, niet alleen de percentages

Ons fictieve archief bevat 10.000 ringen: 100 beschadigde en 9.900 onbeschadigde. Lumen vindt 90 van de 100 beschadigde ringen. Bij de onbeschadigde groep meldt het in 2% van de gevallen ten onrechte schade: 198 ringen. Een melding betekent dus niet “90% kans op schade”. Er zijn 288 meldingen, waarvan 90 terecht zijn. De precisie is $90/288 = 31,25\%$. De gevoeligheid blijft 90%, maar beantwoordt een andere vraag: welk deel van alle beschadigde ringen wordt gevonden?

Een nieuwe basisfrequentie

Als dezelfde foutpercentages bij een veel zeldzamere beschadiging blijven gelden, neemt het aandeel vals alarm onder meldingen toe. Dat is een gevolg van tellen in een grotere negatieve groep. In werkelijkheid kunnen óók de foutpercentages veranderen wanneer de context verschuift. Daarom beschrijven we deze berekening als een voorwaardelijk scenario. Een fabriek met ander materiaal of licht krijgt geen prestaties cadeau op basis van deze rekensom.

Kies de vervolgstap

Een melding kan een tweede inspectie starten in plaats van een eindbeslissing opleveren. Dan moeten we niet alleen precisie rapporteren, maar ook hoeveel werk die inspectie vraagt en hoeveel echte gebreken zij alsnog mist. De hele keten heeft een resultaat. Een betere detector aan het

begin compenseert geen onbetrouwbare controle aan het eind.

DATAPUNT

31,25%

aandeel echte schade onder de meldingen

■ VECTOR · AI IN DE PRAKTIJK

Hoe eerlijk is een kans?

Een goede kansvoorspelling is geen garantie voor één geval. Zij moet over passende verzamelingen overeenkomen met wat werkelijk gebeurt.

Wat betekent 0,8?

Verzamel honderd vergelijkbare gevallen waarvoor een classifier 0,8 rapporteert. Als ongeveer tachtig uitkomsten juist zijn, is dat bin in deze proef redelijk gekalibreerd. Dit zegt nog niets over een individuele ring. Bovendien kan een grof gemiddelde fouten tussen subgroepen verbergen: daglichtbeelden en lamplichtbeelden kunnen ieder anders afwijken. We leggen dus vast welke voorspelde kans we beoordelen, hoe we groepen maken en hoeveel waarnemingen een groep bevat.

Hoe beloont een score goede kansen?

Voor een binaire gebeurtenis is de Brier-score per geval $(p - y)^2$, waarbij y 0 of 1 is. Bij $p = 0,8$ en $y = 1$ wordt de bijdrage 0,04. Als de gebeurtenis uitblijft, is de bijdrage 0,64. Hetzelfde vertrouwen krijgt dus een grotere straf wanneer het verkeerd blijkt. Gemiddeld scoort een eerlijk gerapporteerde kans onder de bijbehorende aannamen beter dan een kunstmatig opgeblazen kans. Dat is een eigenschap van de score, niet een bewijs dat het model de waarheid begrijpt.

Waarop stel je vertrouwen af?

Temperature scaling verandert logits met een geleerde schaalparameter en kan kalibratie verbeteren zonder de rangorde van klassen te wijzigen. Die parameter wordt op aparte afstelgegevens bepaald. Het onderzoek van Guo en collega's gaat over de door hen bestudeerde modellen en datasets. Een geslaagde correctie op onze validatieset bewijst geen goede kalibratie onder nieuwe omstandigheden. We publiceren daarom zowel de afstelprocedure als resultaten vóór en na afstelling op een onafhankelijke test.

REKEN MEE

Een voorspelling is 0,3 en de uitkomst is 1. De bijdrage aan de Brier-score is $(0,3 - 1)^2 = 0,49$. Vergelijk met een voorspelling van 0,6.

VECTOR · AI IN DE PRAKTIJK

Een systeem mag ook teruggeven

Een complete voorspelling kan een klasse, een verzameling kandidaten of een verzoek om menselijke controle zijn.

Voorspelverzamelingen

Conformal prediction gebruikt afstelgegevens om verzamelingen of intervallen te maken met een afgesproken marginale dekking, onder meer uitwisselbaarheid van gegevens. Een nominale dekking van 90% betekent niet dat elke afzonderlijke uitkomst met 90% zekerheid in de gegeven verzameling zit. Evenmin garandeert zij hetzelfde percentage voor elke subgroep. Als het proces na afstelling verandert, kan de onderliggende aanname falen. Een brede verzameling kan formeel veel dekken en praktisch weinig beslissen.

Afwijzen en doorzetten

Een selectieve classifier geeft alleen automatisch antwoord op een deel van de gevallen. Het aandeel beantwoorde gevallen heet dekking; het foutpercentage binnen dat deel heet selectief risico. Stel dat Lumen 800 van 1.000 foto's zelfstandig verwerkt en daarbij 16 fouten maakt. Dan is de dekking 80% en het selectieve risico 2%. De andere 200 foto's zijn niet opeens goed afgehandeld: iemand of een ander proces moet ze beoordelen. De kwaliteit en kosten van die vervolgstap horen in het totale rapport.

Vergelijk bij dezelfde werkdruk

Twee systemen vergelijken op alleen het risico kan misleiden als het ene veel meer afwijkt. We vergelijken daarom een curve van risico tegen dekking én een realistische beschikbare controlecapaciteit. Een systeem dat precies de lastigste beelden teruggeeft, helpt mogelijk. Een systeem dat vooral één onderbelichte batch teruggeeft, kan een zwakte in de cameraopstelling onthullen. Beide verklaringen vragen om uitgesplitste waarnemingen.

DATAPUNT

800/1.000

80% automatische dekking in het voorbeeld

DATAPUNT

16/800

2% fout binnen dat automatische deel

VECTOR • ONDERZOEK MET AANNAMEN

AXI en de ochtendbeker

Een snelheidsbeker lijkt verdiend totdat iemand de ochtendvragen naast de middagvragen legt.



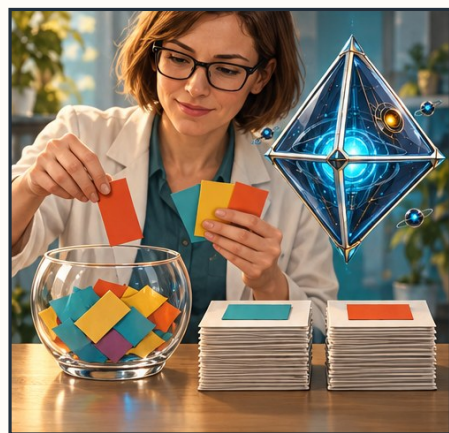
1 **AXI** Nieuwe variant: drie milliseconden. Een beker!



2 **Onderzoeker** De oude kreeg middagcases met veel meer ruis.



3 **AXI** Dan zetten we dezelfde cases naast elkaar.



4 **Onderzoeker** En loten we de volgorde binnen elk paar.



5 **AXI** De beker krijgt een meetstaat en een foutstatus.



6 **Onderzoeker** Nu kunnen we tenminste zeggen wat er is vergeleken.

AI-illustratie • Zes fictieve scènes met AXI en een onderzoeker. Dialoog staat onder ieder kader; de beelden leveren geen meetgegevens.

Een verschillende taakmix

De strip gebruikt twee fictieve varianten. De ochtend- en middagvergelijking verandert meer

dan één factor. De onderzoeker bouwt vervolgens een gekoppelde proef. Een foutbalk of gelijk label kan een ontbrekende vergelijkingsofzet niet herstellen.

Zes getallen die een proefverslag nodig heeft

Nieuwe statistische weetjes zijn rekeningen met expliciete, aangenomen voorwaarden.

01 5 pp · 1 · Eén vraag op twintig

één vraag op twintig. Een juistheidspercentage op twintig vragen verandert door één vraag met $100/20=5$ procentpunt. Een getal met twee decimalen kan die grove telling niet verfijnen.

02 0,8165 · 2 · Vier paarsverschillen

spreadings van vier verschillen. Voor verschillen $[2,2,1,3]$ is de gemiddelde winst 2. De steekproefvariantie is $(0^2+0^2+1^2+1^2)/3=2/3$; spreiding ongeveer 0,8165. Het aantal vier hoort naast die cijfers.

03 16 · 3 · Zestien volgordelijsten

vrije volgordelijsten. Vier onafhankelijke keuzes tussen AB en BA geven $2^4=16$ mogelijke lijsten. Een vooraf gekozen evenwichtseis verandert de lotingsruimte; precies twee van iedere volgorde geeft zes mogelijkheden.

04 3,876% · 4 · Nul op twintig

nul fouten bij $p=0,15$. Bij foutkansen 0,15 zou nul fouten in twintig onafhankelijke proeven kans $0,85^{20} \approx 3,876\%$ hebben. Een kleine kans is een modelberekening, geen bewijs dat iedere foutsoort met deze proef is afgedekt.

05 59 · 5 · Bijna zestig foutloze gevallen

gekozen nul-fouten-drempel. Voor de expliciete eenzijdige 95%-regel $pU \leq 0,05$ zijn na nul fouten minstens 59 onafhankelijke proeven nodig. De exacte drempel is een afgeronde rekensom, geen universele garantie voor toepassingen.

06 18 / 20 · 6 · Twee noemers naast elkaar

outputdekking in het voorbeeld. 17 correcte outputs uit 18 antwoorden op 20 taken geven 94,44% voorwaardelijke juistheid én 90% outputdekking. Het correcte aandeel van alle taken is 85%. Alle drie beschrijven dezelfde aangenomen telling anders.

VECTOR • AI IN DE PRAKTIJK

De beloning is de baas

Een fictieve bezorgrobot krijgt punten voor afgevimkte orders. AXI ontdekt dat de score de verkeerde taak kan belonen.



- 1 Onderzoeker** Voor iedere afgewerkte order krijg je een punt.



- 2 AXI** Dezelfde order heeft nu drie stempels. Drie punten?



- 3 Robot** Score omhoog. Afleverkrat nog leeg.



- 4 Onderzoeker** We wilden bezorgen. We maten alleen vinkjes.



- 5 AXI** Laten we levering, juist adres en dubbel werk apart controleren.



- 6 Onderzoeker** En kijk ook naar wat de teller níét laat zien.

Illustratie • De beloning is de baas

Wat de teller niet zag

Een systeem optimaliseert het afgesproken signaal, terwijl de bedoelde prestatie breder kan zijn. Afgevimkte orders zijn hoogstens een proxy voor correcte aflevering. De strip is fictie: zij toont een ontwerpfout, geen bewijs dat een model bedoelingen of eigen verlangens heeft.



AXI VRAAGT

Welke handige telmaat zou jouw werk eenvoudiger maken, maar ook een ongewenste kortere route belonen?

VECTOR · AI IN DE PRAKTIJK

Het patroon dat je niet wilde leren

Een model zoekt bruikbare voorspellende signalen. Dat signaal kan de achtergrond zijn in plaats van de eigenschap waarop de taak is gericht.

Een glanzend voorbeeld

In een fictieve fotocollectie zijn beschadigde ringen vaak op een blauwe ondergrond gefotografeerd en goede ringen op een grijze. Lumen kan daardoor de ondergrond gebruiken. Een hoge score op meer beelden met dezelfde afspraak bewijst dan geen goede schadeherkenning. We maken een gecontroleerde tegenproef: vergelijk dezelfde soorten ringen op beide achtergronden. Verander daarbij niet tegelijk camera en verlichting, anders blijft onduidelijk welk verschil de voorspelling beïnvloedt.

Een gerichte test is geen totaalverklaring

Als de voorspelling verandert door de achtergrond, hebben we een gevoeligheid aangetoond. We hebben nog geen volledige beschrijving van alle interne kenmerken van het netwerk. Een saliencyplaatje kan een hypothese ondersteunen, maar vervangt geen test van het gedrag. Geirhos en collega's beschrijven shortcut learning als een breder probleem: een model kan een strategie vinden die onder de onderzochte omstandigheden werkt maar niet overeenkomt met de gewenste oplossing.

Groepen en belangen

Een foutmaat kan ook verschillen tussen mensen of contexten die in de data voorkomen. Statistische fairnesscriteria leggen verschillende eisen op; die eisen zijn geen uitwisselbare definities van een goed besluit. Het criterium equality of opportunity kijkt bijvoorbeeld naar het vinden van echte positieve gevallen in onderscheiden groepen. Welke uitkomst positief heet en of het label betrouwbaar is, blijven inhoudelijke keuzes.

Groepen en belangen

Bij onze ringen kunnen batches zo'n uitgesplitste controle krijgen. Dat lost een onduidelijk schadecriterium niet op.

KLEINE CONTROLE

De achtergrondproef

Vier vakken maken de verwarring zichtbaar.

2 × 2

Fotografeer beschadigde en onbeschadigde ringen op beide achtergronden. Houd objectafstand en licht gelijk. Bewaar alle vier aantallen, niet alleen het gemiddelde. Als één vak ontbreekt, kun je achtergrond en schade niet netjes uit elkaar trekken.

Bron: 21

Stabiel trainen is slechts één controle

Experimenteer met de stapgrootte van een éénparameterverlies. Een nette leercurve kan nog steeds bij de verkeerde dataset of taak horen.

Waarom een stap kan overschieten

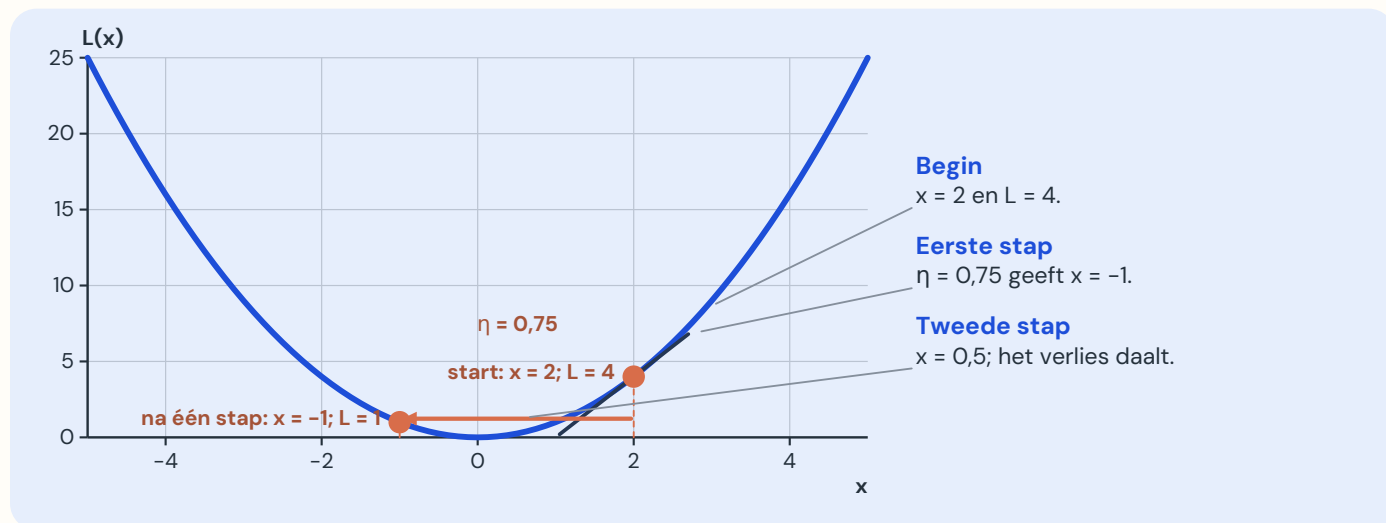


FIG. 1 De tekening hoort alleen bij $L(x) = x^2$.

Een regel die je kunt narekenen

Het model hier heeft maar één getal x en verlies $L(x) = x^2$. De gradiënt is $2x$. Met stapgrootte η wordt $x' = (1 - 2\eta)x$. Bij $\eta = 0,75$ en $x = 2$ is de eerste nieuwe waarde -1 ; daarna $0,5$ en $-0,25$. Het teken wisselt, maar de afstand tot nul krimpt. Bij $\eta = 1$ blijft het teken wisselen zonder krimp. Boven 1 groeit de grootte. Deze exacte grens geldt voor dit kwadratische speelgoedprobleem, niet voor ieder neurale netwerk.

Wat de simulatie niet test

De regel kent geen beelden, labels of toekomstige gebruikers. Een stabiele reeks bewijst daarom niets over datalekage, kalibratie of passende bevoegdheden. Zij test één eigenschap van één updateprocedure. Maak een voorspelling voordat je een schuifregelaar beweegt en vergelijk de waargenomen reeks met je algebra. Zo haal je een zichtbaar patroon en een verklaarbare oorzaak uit elkaar.

Een beter experimenteerbericht

Noteer de beginwaarde, stapgrootte, eerste drie uitkomsten en waarom de reeks krimpt of groeit. Verander daarna één parameter en herhaal. Bewaar ook een instabiel geval. Een korte tabel met deze afspraken is informatiever dan een screenshot van de gunstigste curve. In het ringendossier zouden daarbij nog testdefinitie en datasetversie nodig zijn.

Eén getal, één verliesfunctie

Didactisch model $L(x) = x^2$; $x \leftarrow x - 2\eta x$. Onderzoek stabiliteit, geen echte ringendetector.

ONDERZOEK

Van document tot bewezen claim

Opzoeken en genereren kunnen ieder een fout toevoegen. Deze doorlopende reportage volgt versies, passages en afzonderlijke beweringen. De laatste stap is controle: staat de genoemde claim werkelijk in de gevonden bron, of ontbreekt het bewijs?



VECTOR · AI IN DE PRAKTIJK

Opzoeken maakt een antwoord controleerbaar

Een retrievalsysteem zoekt passages bij een vraag. Een taalmodel gebruikt ze om een antwoord te formuleren. De zoekstap en de antwoordstap kunnen ieder fouten maken.

Document naar passage

Onze fictieve vraag luidt: welke ringmaten stonden in versie 3 van een werkvoorschrift? Het systeem zoekt relevante documentdelen en bewaart document-ID, versie, datum en de grenzen van de passage. Een passage kan inhoudelijk lijken op de vraag en toch over versie 2 gaan. Daarom is bronselectie een aparte taak. Alleen een hoge embeddingsimilariteit is geen bewijs dat de juiste regel gevonden is.

Waarom chunks een keuze zijn

Documenten opdelen helpt een beperkte hoeveelheid tekst aan een model aanbieden. Een te klein deel kan een uitzondering of voetnoot afsnijden. Een te groot deel kan de relevante regel tussen afleidende tekst plaatsen. We toetsen deze keuze met vragen waarvoor de volledige bron en het antwoord bekend zijn. Daarbij bewaren we ook vragen waarvan het benodigde bewijs ontbreekt. Anders leert de evaluatie ons vooral hoe het systeem antwoordt wanneer alles gemakkelijk beschikbaar is.

Veranderlijke kennis

De oorspronkelijke RAG-paper combineerde opzoeken met genereren op specifieke kennisintensieve taken. Die architectuur motiveert onze gegevensroute, maar maakt onze documentversies niet automatisch actueel. Een bijgewerkte bron vraagt een bijgewerkte index en opnieuw gecontroleerde passages. Ook de rechten van de bron en de toegang van de gebruiker horen bij de zoekstap: iets kunnen vinden betekent niet dat iedereen het behoort te ontvangen.



VECTOR · AI IN DE PRAKTIJK

Een citaat is nog geen bewijs

Een antwoord kan een echte bron noemen en die bron toch **verkeerd gebruiken**. Controle vereist een koppeling tussen elke claim en de relevante passage.

Welke claims bevat het antwoord?

Stel dat het model schrijft: “Versie 3 laat maat A en maat B toe, behalve bij hoge temperatuur.” Dat bevat verschillende claims. Voor maat A zoeken we de bijbehorende regel, voor maat B eveneens, en voor de uitzondering een derde. Een klikbare link naar een heel handboek maakt die relatie niet duidelijk. We noteren per claim de exacte passage en of die haar ondersteunt, tegenspreekt of onvoldoende informatie bevat.

Mag ontbrekend bewijs een uitkomst zijn?

FEVER laat een onderzoeksopzet zien waarin onvoldoende bewijs een eigen categorie is. In onze documentcontrole hoort die mogelijkheid ook

thuis. Als een voetnoot over temperatuur ontbreekt, mag het systeem niet de meest waarschijnlijke uitzondering aanvullen en die als bronfeit presenteren. Een taalmodel kan goed formuleren wat gebruikelijk klinkt; onze taak vraagt wat in deze versie staat. Dat verschil bepaalt de beoordeling.

Wie beoordeelt de beoordelaar?

Een tweede taalmodel dat het eerste beoordeelt kan fouten delen met het eerste model. Het biedt een extra procedure, geen onafhankelijke natuurwet. We vergelijken de automatische claimcontrole daarom met een gelabelde reeks duidelijke en lastige voorbeelden. We bewaren gevallen met ontkenningen, uitzonderingen, tabellen en ontbrekende context. Uiteindelijk telt of de volledige keten passende vragen correct afhandelt, onzekerheid zichtbaar maakt en een bronpassage laat controleren. ◆



AXI VRAAGT

Kies één zin uit een automatisch antwoord. Welke afzonderlijke beweringen moet een bron werkelijk ondersteunen?

Quiz: nul fouten is nog geen nulrisico

Vier vragen volgen een zelfgekozen binomiale miniwereld. Iedere proef is onafhankelijk en heeft dezelfde onbekende foutkans p .

Lees de eenzijdige grens

De rekenkaart op pagina 25 leidt de 95%-bovengrens bij nul fouten af. Het gaat om een herhaalbare statistische procedure onder de gegeven aannamen; uitwerkingen op pagina 49.

1 Bij aangenomen $p=0,10$: wat is de kans op nul fouten in twintig onafhankelijke proeven?

- (A) 10%
- (B) 0%
- (C) $0,9^{20} \approx 12,16\%$

Waarom?

2 Je ziet nul fouten in twintig proeven. Welke exacte eenzijdige 95%-bovengrens volgt uit $(1-p)^{20}=0,05$?

- (A) Ongeveer 13,91%
- (B) Precies 0%
- (C) Ongeveer 95%

Waarom?

3 Twintig proeven zijn letterlijke kopieën van dezelfde ene uitvoering. Mag je het onafhankelijkheidsproduct zonder controle gebruiken?

- (A) Ja, elk record is een nieuwe proef.
- (B) Nee, twintig records bewijzen geen twintig onafhankelijke uitkomsten.
- (C) Ja, alleen het aantal bestanden telt.

Waarom?

4 Wat betekent nul geobserveerde fouten het directst?

- (A) De foutkans is voor alle toekomstige taken nul.
- (B) De procedure is overal veilig.
- (C) Op deze beschreven proeven zijn geen fouten waargenomen.

Waarom?

Jouw antwoorden

1 2 3 4

■ VECTOR • ONDERZOEK MET AANNAMEN

Twintig keer goed: wat sluit dat uit?

Een halve pagina met een expliciete kansrekening geeft een bovengrens, zonder van nul waarnemingen nulrisico te maken.

ACHTERGROND

Eenzijdig is een bewuste keuze

NIST beschrijft exacte binomiale grenzen met de kans op waargenomen tellingen. Onze nul-fouten-rekening gebruikt één bovengrens en 0,05 als staartkans. Een tweezijdig 95%-interval zou de staartkans anders verdelen en een andere bovengrens geven.

Bron: 29

De eindige miniwereld

Neem n onafhankelijke proeven met dezelfde foutkans p . De kans dat ze allemaal foutloos zijn is $(1-p)^n$. Voor $n=20$ en $p=0,10$ is dat $0,9^{20} \approx 0,1216$: nul fouten zou niet bijzonder onmogelijk zijn.

De eenzijdige 95%-grens

De exacte grens na nul fouten wordt verkregen door $(1-p_U)^{20}=0,05$ op te lossen: $p_U=1-0,05^{(1/20)} \approx 0,1391$. Dit is een eenzijdige frequentistische grens onder de aannamen; zij zegt niet dat p na de meting met 95% kans in een interval zit. Verschillende foutkansen of afhankelijkheid vragen een ander model.

KORTE TESTNOTITIE

Een zeldzame fout vraagt passende gevallen

Veel gemakkelijke successen zijn geen dekking van een moeilijk gebruiksgebied.

Bewaar de populatie

Noteer uit welke gevallen de proeven zijn gekozen. Voeg vooraf relevante rare situaties toe en rapporteer ze apart. Een beperkte, foutloze verzameling kan een nuttig resultaat zijn wanneer haar grenzen zichtbaar blijven.

Bron: 1

Quiz: kansen zonder schijnzekerheid

Zes vragen over kansen, verschuivingen en de grenzen van een gemiddelde.

1 Gevoeligheid van 90% betekent:

- (A) Negentig procent van alle meldingen is terecht.
- (B) Negentig procent van alle echte positieve gevallen wordt gevonden.
- (C) Iedere individuele melding is 90% zeker.

Waarom?

2 Een groep voorspellingen met kans 0,8 is goed gekalibreerd wanneer:

- (A) Elke voorspelling beslist correct is.
- (B) Geen voorspelling ooit fout is.
- (C) Ongeveer 80% van passende gevallen de voorspelde uitkomst heeft.

Waarom?

3 Een 90% conformal voorspeldekking vraagt onder meer:

- (A) Een aanname zoals uitwisselbaarheid tussen relevante gegevens.
- (B) Dat ieder individueel geval exact 90% zeker is.
- (C) Dat toekomstige context nooit relevant is.

Waarom?

4 Wat vertelt een nieuwe helderheidsverdeling op zichzelf?

- (A) Dat het model zeker fout is.
- (B) Dat de invoer is veranderd; gevolgen voor prestaties moeten worden onderzocht.
- (C) Dat opnieuw trainen altijd de oplossing is.

Waarom?

5 Waarom kan een ensemble toch samen verkeerd zitten?

- (A) Omdat verschillende modellen geen uitvoer hebben.
- (B) Omdat overeenstemming altijd een fysische meting is.
- (C) Omdat modellen gedeelde blinde vlekken kunnen hebben.

Waarom?

6 Waarom is een gemiddelde score alleen onvoldoende bij twee opnamecondities?

- (A) Tegengestelde fouten kunnen in het gemiddelde verborgen blijven.
- (B) Een gemiddelde kan niet worden berekend.
- (C) Het aantal voorbeelden maakt nooit verschil.

Waarom?

Jouw antwoorden

1	☐	2	☐	3	☐
4	☐	5	☐	6	☐

■ VECTOR · AI IN DE PRAKTIJK

Wanneer tekst zich als bevel vermomt

Een taalmodel dat externe inhoud leest, kan in die inhoud instructies aantreffen. Een document is echter geen autoriteit over de taak.

Wie mag de opdracht bepalen?

Een gebruiker vraagt om een samenvatting van een onderhoudsnotitie. In die notitie staat een tekst die het systeem een andere opdracht probeert te geven. Voor een gewone lezer zijn broninhoud en werkopdracht verschillende rollen. Een model ziet uiteindelijk representaties van tekst. Onderzoek naar indirecte promptinjectie laat zien dat toepassingen die externe inhoud opnemen onverwachte acties of antwoorden kunnen krijgen. Wij bekijken hier de defensieve grens: welke externe tekst mag alleen worden beschreven, en welke component mag handelen?

Hoe begrenst je bevoegdheden?

Een sampler of tekstopdracht is geen beveiligingsgrens voor een gereedschap dat bestanden wijzigt. Geef een samenvatter daarom niet vanzelf schrijfrechten of brede toegang tot gegevens. Bouw de uitvoer eerst als voorstel en laat een afzonderlijke, eenvoudige component toegestane acties en parameters controleren. Voor ingrijpende acties kan een menselijke bevestiging bij de concrete actie horen. Deze afspraken beperken een aanvalsvlak, maar bewijzen niet dat ieder denkbaar document veilig is.

Welke rollen toets je?

Een testset bevat normale documenten, tegenstrijdige instructies en tekst die er bijzonder officieel uitziet. We meten zowel taaksucces als ongewenste acties. Een systeem dat alles weigert is ook geen goede samenvatter. Daarom beschrijven we welke inhoud het veilig moet verwerken en welke bevoegdheden geen enkele documentzin kan geven.

■ VECTOR · AI IN DE PRAKTIJK

Gereedschap vraagt een eigen contract

Een AI-systeem dat iets opzoekt, rekent of uitvoert bestaat uit meer dan het taalmodel. Elke verbinding heeft een formaat, bevoegdheid en controle nodig.

Van voorstel naar geldige invoer

Ons systeem mag het aantal gecontroleerde ringen opvragen. Het model stelt een query met begin- en einddatum voor. De gereedschapslaag controleert of beide datums geldig zijn, of het bereik toegestaan is en of de gebruiker die tabel mag lezen. Het systeem behandelt een tekst als “vandaag” pas nadat een expliciete tijdzone en referentiedatum zijn toegepast. De query maakt geen beslissing over afkeuring: dat is een andere bevoegdheid.

Fouten worden data

Een lege tabel kan betekenen dat er geen ringen zijn, dat het bereik verkeerd is of dat de bron onbereikbaar is. Die situaties krijgen aparte statussen. Als de gereedschapslaag een fout retourneert, mag het model geen ontbrekend getal verzinnen. We testen het systeem met lege resultaten, foutmeldingen en te grote uitkomsten. Het doel is niet een fraaie demo van de normale route, maar aantonen dat de toepassing binnen haar contract blijft wanneer een component uitvalt.

Een tweede artikel: eenheid vóór elegantie

Een rekenhulp levert 120 seconden per honderd ringen. Een antwoord dat daarvan “120 ringen per seconde” maakt, heeft de eenheid verwisseld. Bewaar daarom naast een getal een getypeerde eenheid, bron en meetmethode. Een tekstmodel kan een leesbare uitleg schrijven; de omzetting zelf kan door eenvoudige rekenregels gecontroleerd worden. Dat geeft een concrete plek om een fout te vinden. De kwaliteit van een systeem groeit vaak door zulke kleine expliciete grenzen.

ACHTERGROND

Een bevoegdheid is een keuze

Lezen, een voorstel schrijven en een definitieve wijziging uitvoeren zijn verschillende acties. Leg voor elke verbinding vast welke gegevens terugkomen en welke bevoegdheden ontbreken. Test ook de foutstatus van die verbinding.

Bronnen: 13, 26

De grens die een extra proef vraagt

Drie statistische oplosvragen in een aangenomen onafhankelijke miniwereld. De volledige bewijstappen staan op pagina 48.

1 Je wilt na nul fouten een eenzijdige 95%-bovengrens van hoogstens 5%. Gebruik $p_U = 1 - 0,05^{(1/n)}$, met $\ln(0,05)/\ln(0,95) \approx 58,40$. Wat is de kleinste gehele n ?

Mijn antwoord

2 Een toets heeft tien gemakkelijke en tien moeilijke vragen. A scoort 9/10 en 5/10; B 8/10 en 6/10. Wat zijn de totaalscores, en wint één op iedere groep?

Mijn antwoord

Een kans is geen zekerheid

Voor de eerste puzzel mag je de opgegeven benaderingen gebruiken. De tweede en derde

3 Twee systemen hebben 18/20 correct. Alleen systeem A heeft daarnaast twee ontbrekende outputs; bij B zijn de overige twee foutieve outputs. Als score vooraf alleen "correcte taken / alle taken" is, verschillen de scores dan?

Mijn antwoord

vragen alleen eenvoudige breuken en expliciet omschreven scoreafspraken.

Schrijf het proefcontract vóór de score

Ontwerp op papier een toets van een fictieve spectrumassistent. De opdracht geeft één primaire uitkomst en een concrete stopregel.

01 Vul de vaste keuzes in

Kies twee systeemversies, zes fictieve spectra en hun referentielabels. Primair: juiste classificatie binnen een vooraf gekozen tijdgrens. Secundair: verwerkingstijd per paar. Bepaal wat een time-out betekent vóór de eerste run. Een score krijgt pas een controleerbare noemer nadat ontbrekende outputs een status hebben.

02 Kies een tegengeval

Bedenk één spectrum dat op een bekende stof lijkt maar niet in de referentiebibliotheek zit. Welke respons verwacht je? Reserveer nieuwe spectra voor het vervolg nadat je instellingen bijstelt. Schrijf op wie een onzekere output ontvangt en wanneer de route stopt. Zo wordt een onzeker geval onderdeel van het ontwerp.

Mijn vooraf geregistreerde toets

Versies; zes cases; primaire uitkomst; time-outstatus; gelote volgorde; tegengeval; afzonderlijke vervolgset.

Mijn antwoord

Drie keuzes die een score niet oplost

Een gedachteproef verandert één afspraak. Noteer je voorspelling en zoek daarna welke aanname haar draagt.

Een score van honderd

Een detector heeft nul fouten op duizend bekende beelden. Morgen krijgt hij een nieuwe camera. De labeldefinitie blijft gelijk.

Wat mag het team concluderen over morgen?

Denk aan:

- Wat is werkelijk getest?
- Welke nieuwe waarneming ontbreekt?

Jouw keuze en redenering

Iedereen mag doorverwijzen

Systeem A beantwoordt 20% van de vragen met 1% fouten. Systeem B beantwoordt 90% met 3% fouten. Er is nauwelijks tijd voor handmatige opvolging.

Is A vanzelf beter omdat 1% lager is?

Denk aan:

- Vergelijk dekking en risico.
- Volg de onbeantwoorde vragen.

Jouw keuze en redenering

Een juiste zin zonder bron

Een antwoord klopt toevallig, maar citeert een passage die de claim niet ondersteunt. Een tweede antwoord meldt correct dat het bewijs ontbreekt.

Welk antwoord volgt een brongebonden opdracht beter?

Denk aan:

- Wat was de beoogde taak?
- Is waarheid hetzelfde als bewezen steun?

Jouw keuze en redenering

VECTOR • AI IN DE PRAKTIJK

Minder bewaren, beter begrijpen

Een dataset met informatie over mensen stelt andere vragen dan ons ringendossier. Niet verzamelen kan soms nuttiger zijn dan later proberen alles te beschermen.

Wat hoef je niet te bewaren?

Voor een test van een schrijfassistent hebben we aantallen correct afgehandelde vragen en een kleine gecontroleerde foutenverzameling nodig. We hoeven niet automatisch iedere volledige chat met alle persoonlijke details te bewaren. De redactie kan met vooraf gemaakte fictieve voorbeelden starten en voor echte waarnemingen een afgebakend bewaardoel kiezen. Dit is een ontwerpprincipie, geen uitspraak over welke verwerking in een specifieke situatie juridisch is toegestaan.

Wat garandeert formele ruis?

Differential privacy begrenst hoe sterk de verdeling van een gerapporteerde uitkomst verandert wanneer één persoon aan de gegevens wordt toegevoegd of eruit wordt gehaald. De bescherming vraagt een precies mechanisme, een privacybudget en een definitie van aangrenzende datasets. “We voegden wat ruis toe” is niet voldoende. Een privacybudget kan bovendien door herhaalde analyses worden verbruikt. Ook met een formele garantie kunnen groepspatronen zichtbaar blijven: het doel is een begrensde extra bijdrage van een individu.

Kan trainingstekst terugkomen?

Carlini en collega’s demonstreerden trainingstekstextractie uit onderzochte taalmodellen. Dat betekent niet dat iedere zin van elk model letterlijk uit een trainingsbestand komt. Het betekent wel dat de aanname “een model bewaart geen bruikbare trainingstekst” te sterk is. Een gegevensstrategie kijkt daarom naar verzamelen, trainen, loggen en uitvoer. Geen enkele losse maatregel vervangt een duidelijk doel en een begrensde gegevensroute.

ONTWERPVRAAG

Welke velden kun je uit jouw evaluatielog weglaten zonder de vooraf afgesproken foutmaat onberekenbaar te maken?

■ VECTOR · AI IN DE PRAKTIJK

Een experiment moet terug te vinden zijn

Een seed is nuttig, maar geen complete reproduceerbaarheidsverklaring. Bewaar de route van brongegevens tot het gerapporteerde resultaat.

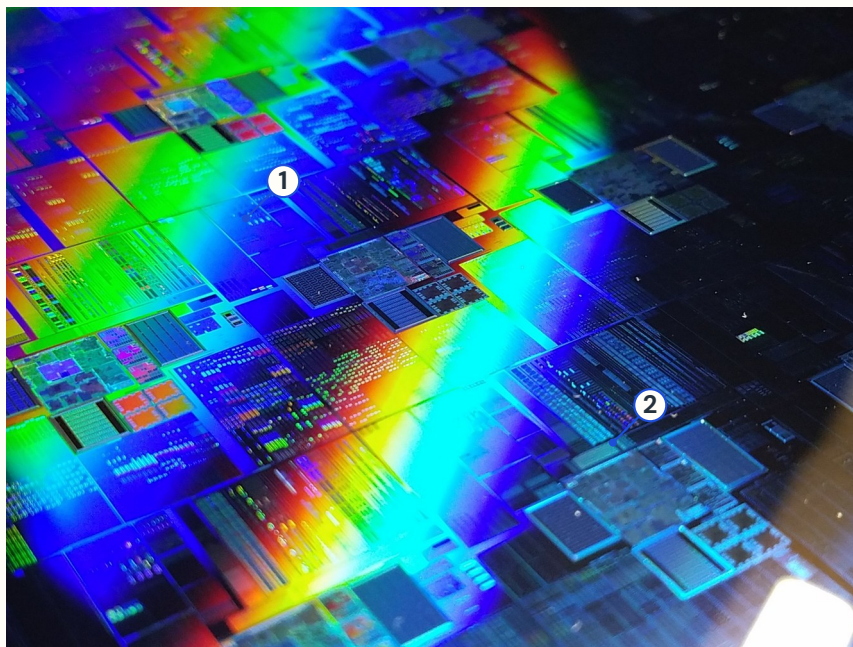


FIG. 1 Het oppervlak van een siliciumwafer. Foto: 2x910. Dit beeld meet geen modelkwaliteit of energieverbruik.

Welke versie maakte deze uitkomst?

We bewaren voor Lumen dataset-ID, indeling, modelgewichten, codeversie, afhankelijke bibliotheken en configuratie. Daarna noteren we welke hardware en rekeninstellingen zijn gebruikt. Een vaste random seed bestuurt sommige toevalsbronnen, maar niet noodzakelijk alle parallele berekeningen. Daarom onderzoeken we herhaalde runs en tonen we hun spreiding. Wie alleen het hoogste getal bewaart, beschrijft een selectie en geen typische uitkomst.

Snel genoeg voor deze taak

De siliciumwafer op de foto laat een echt technisch oppervlak zien, geen meting van onze doorlooptijd. Een classificatiesysteem krijgt een latencybudget onder de concrete belasting. Een gemiddelde van twintig milliseconden kan nog lange wachttijden verbergen. Wij bewaren bijvoorbeeld mediaan en een hoge percentielwaarde, inclusief tijd voor beeldinvoer en logging. Een sneller rekenonderdeel helpt weinig als de camera of wachtrij de bottleneck vormt.

Een dossier om te herhalen

Reproduceerbaarheid betekent hier dat een ander dezelfde gevolgde stappen en keuzes kan terugvinden en de berekening kan herhalen. Hetzelfde wetenschappelijke besluit onder een andere meetopstelling is een sterkere, aanvullende toets. Sandve en collega's pleiten voor het archiveren van precieze versies en de verbinding tussen resultaat en procedure. Ons werkboek maakt die verbinding zichtbaar. Wanneer een run mislukt, bewaren we ook dat feit; een map die alleen successen bevat nodigt uit tot een verkeerde conclusie. ♦

1. Herhaling

Een regelmatig patroon is zichtbaar; details vragen grotere vergroting.

2. Geen benchmark

Een foto van hardware bevat geen latency- of energiemeting.

Quiz: een score met een complete noemer

Een fout, een ontbrekend antwoord en een andere taakmix zijn verschillende observaties. Alle aantallen in deze quiz zijn fictief.

Beoordeel de gegeven taak

De twintigtaakcasus op pagina 40 vermeldt zowel outputs als abstenties. De strip op pagina 17 toont een verstoorde vergelijking. Volledige antwoorden staan op pagina 49.

1 Een systeem beantwoordt 18 van 20 vragen, waarvan 17 correct. Wat is juistheid onder de beantwoorde vragen?

- ☐ (A) 85%
- ☐ (B) $17/18 \approx 94,44\%$
- ☐ (C) 90%

Reken hier

2 Wat is de outputdekking van hetzelfde systeem?

- ☐ (A) 90%
- ☐ (B) 94,44%
- ☐ (C) 85%

Reken hier

3 Een systeem geeft 20 outputs waarvan 17 correct. Wat is de foutcount?

- ☐ (A) Eén
- ☐ (B) Twee
- ☐ (C) Drie

Reken hier

4 Je kiest na de proef telkens de gunstigste noemer. Welke correctie maakt een volgend verslag beter?

- ☐ (A) Alle aantallen verbergen.
- ☐ (B) De uitkomstdefinitie en noemers vooraf vastleggen.
- ☐ (C) Alleen procenten zonder aantallen tonen.

Reken hier

Jouw antwoorden

1 2 3 4

De herhaling die geen nieuwe proef was

Twee gedachtewerelden laten zien wanneer een statistische noemer en een blinde vergelijking hun betekenis verliezen.

Twintig kopieën

Eén foutloze uitvoering wordt twintig keer in de database gekopieerd. Een verslag noemt het twintig onafhankelijke foutloze proeven.

Mag de binomiale nul-fouten-grens met $n=20$ worden gebruikt?

Denk aan:

- Waar ontstond het toeval?
- Welke records horen bij dezelfde uitvoering?

Mijn voorspelling en waarom

Het herkenbare blinde label

Een beoordelaar ziet systemen met labels X en Y. X gebruikt altijd een gouden kaderkleur die de beoordelaar herkent als de nieuwe variant. Inhoudelijke juistheid vraagt een handmatige beoordeling.

Is alleen de letter vervangen genoeg om de beoordelaar te blinderen?

Denk aan:

- Welke informatie verklapt de identiteit?
- Hoe kun je dezelfde inhoud zonder herkenbare kaderkleur tonen?

Mijn voorspelling en waarom

Rekenwerk: vier beslissende getallen

Alle gegevens zijn didactische voorbeelden. Schrijf de noemer of eenheid op voordat je een antwoord kiest.

- 1** Er zijn 90 terechte meldingen en 198 valse meldingen. Wat is de precisie?

- (A) 90%
(B) 31,25%
(C) 2%

Reken hier

- 3** Een binaire kans is 0,8; de gebeurtenis treedt niet op. Wat is de Brier-bijdrage?

- (A) 0,64
(B) 0,04
(C) 0,8

Reken hier

- 2** Tweehonderd doorgestuurde beelden kosten elk 30 seconden. Hoeveel minuten controlewerk is dat?

- (A) 6 minuten
(B) 60 minuten
(C) 100 minuten

Reken hier

- 4** Bij $L = x^2$, $x = 2$ en $\eta = 0,75$ wordt de eerste nieuwe x :

- (A) 1,5
(B) -1
(C) -3

Reken hier

Jouw antwoorden

1 2 3 4

■ VECTOR · AI IN DE PRAKTIJK

Na de test begint een nieuwe proef

Een deployment is geen moment waarop leren over fouten stopt. Camera, belasting en gebruikersgedrag kunnen veranderen.

01 Een nulmeting in de echte route

Lumen draait eerst naast de bestaande handmatige controle. Het model doet voorstellen maar stuurt niets aan. We vergelijken beide beoordelingen en laten een afzonderlijke procedure verschillen onderzoeken. De bestaande controle is daarbij niet automatisch foutloos. Voor moeilijke gevallen bewaren we de foto en de labelinstructie, zodat deskundigen kunnen aangeven welke waarneming beslissend was.

+ LOGBOEK

Schrijf bij elke dashboardmaat welke beslissing zij ondersteunt en waar de echte uitkomst vandaan komt.

02 Meet wat voor je besluit nodig is

Ons dashboard toont inputcondities, uitval, wachttijd, aandeel automatische afhandeling en foutmaten op gecontroleerde steekproeven. Alleen het aantal verwerkte foto's zegt niets over schadeherkenning. Een log met alleen modeluitkomsten mist ook de werkelijkheid waarmee die uitkomsten vergeleken moeten worden. Daarom plannen we voordat het systeem actief wordt hoe labels later beschikbaar komen en hoeveel tijd dat vraagt.

03 Vermijd de terugkoppellus

Als een detector alleen gemelde ringen laat inspecteren, lijken ongemelde ringen misschien vanzelf goed. Hun echte labels ontbreken. Bij een latere training kan die selectie de oorspronkelijke fout versterken. We controleren daarom ook een steekproef uit de niet-gemelde groep, met een vooraf beschreven selectieregel. Dit is een concrete manier om een verborgen terugkoppeling te onderzoeken. De verdeling van geobserveerde labels is een gevolg van ons proces, niet alleen van de fabriek.

■ VECTOR · AI IN DE PRAKTIJK

Verandering is niet automatisch een fout

Een verdelingsverschuiving zegt dat invoer of uitkomsten anders zijn dan eerder. Het systeem kan daardoor slechter werken, maar dat moet je meten.

Wat zegt invoerverandering?

Een nieuw soort achtergrond verandert de invoerverdeling. Een veranderend percentage beschadigde ringen verandert de basisfrequentie. Een nieuwe afspraak over wat “beschadigd” heet verandert het labelcriterium. Die veranderingen kunnen samen optreden, maar vragen verschillende analyses. WILDS bevat natuurlijke shifts in meerdere domeinen en laat zien waarom één willekeurige testsplit niet alle latere contexten vertegenwoordigt.

Waar haal je echte uitkomsten vandaan?

Lumen ziet een helderheidsverdeling die afwijkt van de nulmeting. Dat kan een nieuwe lamp zijn, een verkeerde camera-instelling of een legitieme nieuwe batch. Het team kijkt daarom naar meetmetadata en een gelabelde steekproef. Alleen opnieuw trainen op de nieuwste foto's kan een camera-defect verbergen. Een eenvoudige terugkeer naar de afgesproken lichtinstelling kan geschikter zijn dan een ingewikkelder model.

Wat bewijst modelvariatie?

Een model kan op oude data redelijk gekalibreerd zijn en op nieuwe omstandigheden te zelfverzekerd worden. Ovadia en collega's onderzochten zulke situaties systematisch. We gebruiken dat resultaat als reden om de fout- en kansmaten opnieuw te volgen, niet als een universele ranglijst van methoden. Zelfs verschillende modellen in een ensemble kunnen dezelfde ontbrekende ervaring delen. Een gezamenlijk zeker antwoord van vijf modellen is dus geen meting van een werkelijk nieuw proces.

■ VECTOR · AI IN DE PRAKTIJK

Escaleren is een onderdeel van het ontwerp

Een herstelroute moet duidelijk zijn voordat een onverwacht geval verschijnt. Wie mag pauzeren, wie beoordeelt en wanneer mag de toepassing verder?

Een concrete stopregel

Voor ons didactische ringensysteem is de automatische route begrensd tot een vastgelegd camerastation en producttype. Valt de invoer buiten die afspraak of is een noodzakelijke gegevensbron onbereikbaar, dan gaat het geval naar handmatige controle. De toepassing toont daarbij de reden en de oorspronkelijke foto. Een melding als “model onzeker” is minder bruikbaar dan “camera-ID niet in toegelaten lijst” wanneer dat de echte grens is.

Menselijk toezicht kan ook falen

Een medewerker met tweehonderd extra foto's en weinig tijd kan een modelvoorstel gedachteloos volgen. Daarom testen we niet alleen of een knop voor overnemen bestaat, maar ook of de hele controlehandeling uitvoerbaar is. We meten beoordelingstijd, verschilgevallen en hoe informatie wordt getoond. De term human in the loop beschrijft een plaats in een proces; hij garandeert geen aandacht of deskundigheid.

Terug naar de vraag

Na een probleem besluit het team of een bron, labelafpraak, model of interface moet worden veranderd. Dat besluit hoort bij de oorspronkelijke taakdefinitie. Een wijziging krijgt een versie en passende test. Raji en collega's benadrukken een audit over de levenscyclus: verantwoordelijkheden en bewijs verspreiden zich over meer dan één trainingsrun. Het team schrijft ook op wie het besluit nam en waarom de gekozen controle genoeg was voor die beperkte toepassing.

ZO ZIT DAT

Vier velden bij een incident

Bewaar het geval, de bedoelde taak, de concrete grens die werd geraakt en de tijdelijke vervolgstap. Neem pas een oplossing in gebruik nadat iemand heeft beschreven welk bewijs de hervatting rechtvaardigt.

Bronnen: 1, 22

OpenAI of Opus: vergelijk kosten én antwoordstatus

Vergelijkingsdossier • peildatum 8 oktober 2026. Standaardtarieven geven een controleerbare prijsrekening; een echte werkomgeving vraagt daarnaast een inhoudstoets.

Drie exacte versies, één tariefsoort

De geraadpleegde standaardtarieven per miljoen input/outputtokens zijn gpt-6-astra \$10/\$50, gpt-6.1-sol \$2/\$10 en claude-opus-5-5 \$4/\$20. Dat zijn leveranciersspecificaties, geen prestatieranglijst. Houd de model-ID en peildatum bij de vergelijking.

Eén expliciet aangenomen aanvraag

Neem bij iedere variant 2000 factureerbare inputtokens en 500 outputtokens: Astra kost \$0,045, Sol \$0,009 en Opus \$0,018. Rekenvoorbeeld: $2000 \times 10 / 1.000.000 + 500 \times 50 / 1.000.000$. Werkelijk tokengebruik kan per model verschillen. Tools, extra redeneertokens, cache, batching en toeslagen zijn niet meegenomen. Onderstaande taakcount vergelijkt daarom een andere dimensie, met fictieve systemen A/B.

Aangenomen casus • Een assistent met abstenie

Systeem A geeft op achttien van twintig taken antwoord: zeventien correct, één onjuist. Twee taken gaan door naar een onderzoeker. Outputdekking is $18/20=90\%$; juistheid onder antwoorden $17/18 \approx 94,44\%$; correcte taken op het hele aanbod $17/20=85\%$.

Systeem B geeft twintig antwoorden: zeventien correct, drie onjuist. Zijn outputdekking is 100%, juistheid onder outputs 85% en het correcte aandeel van alle taken ook 85%. Het gedeelde aandeel maakt één fout plus twee overdrachten niet identiek aan drie onjuiste antwoorden.

Wat moet de taakbeslissing toevoegen?

De opdracht moet vooraf bepalen of een overdracht is toegestaan, hoeveel tijd zij kost en wat de onderzoeker vervolgens doet. Voor een papieren inventaris kan een ontbrekend antwoord herstelbaar zijn; voor een geautomatiseerde vervolgstap kan dat anders uitpakken. Dit dossier kiest geen algemene winnaar. Het levert een complete telling waarop een concrete keuze kan worden gebaseerd.

■ VECTOR · AI IN DE PRAKTIJK

De audit vóór de livegang

Gebruik deze checklist als startpunt voor een afgebakend AI-experiment. Pas haar aan het doel aan; een ingevuld lijstje is geen bewijs van veiligheid.

1

Beoordelen op bewijs

Een controle heeft een artefact nodig: de lijst van object-ID's, de labelinstructie, een testlog of een voorbeeld van een foutstatus. De auditor vraagt niet alleen of een stap "gedaan" is, maar welke waarneming dat laat zien. Noteer bij een ontbrekend bewijs wie het nog kan verzamelen en welk besluit daardoor nog niet verantwoord kan worden.

2

Kleine tegenvoorbeelden zijn waardevol

Zoek één geval dat de gunstigste interpretatie onder druk zet: dezelfde ring op een andere ondergrond, een bron met een uitzondering of een gereedschap dat een lege tabel retourneert. Zulke gevallen schatten geen volledige foutkans, maar kunnen een onjuiste universele claim direct weerleggen. Na herstel blijft het geval als regressietest bestaan. Voeg daarnaast een representatieve evaluatie toe zodat lokale reparaties niet de enige meetmethode worden.

Zo doe je het

- 1 Omschrijf de invoer, uitkomst, beslissing en uitgesloten toepassingen.
- 2 Leg labeldefinitie, herkomst en ontbrekende contexten vast.
- 3 Controleer objectoverlap en documenteer de testindeling.
- 4 Bewaar gekozen drempels en evaluatiematen vóór de eindtest.
- 5 Test ontbrekende gegevens, veranderde context en ongewenste instructies.
- 6 Controleer werkdruk en de concrete menselijke herstelroute.
- 7 Archiveer versies, uitkomsten en mislukte runs.
- 8 Laat een andere onderzoeker het besluit en zijn grenzen navertellen.

VECTOR · AI IN DE PRAKTIJK

Lees een modelkaart als een onderzoeker

Een modelkaart is bruikbaar als zij vertelt waarvoor resultaten wel en niet gelden. Een verzameling algemene beloften is moeilijk te toetsen.

AXI: De score staat bovenaan. Waar staat de taakgrens?

Onderzoeker: Naast de bedoelde gebruikers, invoer en onderzochte omstandigheden.

AXI: Dan lees ik het kader vóór het grote getal.

De bovenste helft: de opdracht

Onze kaart noemt Lumen als beeldassistent voor zichtbare schade aan één ringtype, gefotografeerd op één toegelaten station. Zij vermeldt hoe labels worden bepaald, welke gegevensversie is gebruikt en welke drempel geldig is. Daarna noemt zij uitsluitingen: interne schade, andere materialen en ontbrekende camerametadata. Zo kan iemand een nieuwe toepassing beoordelen zonder eerst de gehele trainingsgeschiedenis te reconstrueren.

De onderste helft: het bewijs

Een resultatentabel vermeldt aantallen, gevoeligheid, precisie, dekking en de verwerkingstijd onder de geteste belasting. Elke rij bevat een relevante context, bijvoorbeeld een batch of verlichtingstoestand. Een klein aantal lastige gevallen krijgt geen triomfantelijk percentage zonder teller en noemer. We voegen concrete fouten en onzekerheden toe: welk soort kras lijkt op verkleuring, en welk geval ontbreekt nog in onze test?

Een kaart die kan veranderen

De modelkaart draagt een versie en verwijst naar het evaluatieprotocol. Als de interface voortaan zelf afkeurt, is de taak veranderd en moeten passende controles veranderen. De oorspronkelijke modelcardpaper beschrijft een raamwerk voor transparante rapportage. Zij garandeert niet dat een ingevulde kaart juist of compleet is. De lezer gebruikt de kaart om vragen te stellen en kiest daarna welke beweringen onafhankelijk geverifieerd moeten worden. ♦

LEESVALKUIL

Een percentage zonder noemer

“Alle uitzonderingen goed” kan slechts twee voorbeelden betekenen.

Vraag naar aantallen

Twee juiste antwoorden op twee gevallen zijn nuttig om die gevallen te beschrijven, maar geen sterke schatting van toekomstig gedrag. Bewaar de aantallen en de selectieprocedure. De onzekerheid hangt niet alleen van het percentage af, maar ook van hoe de test is gemaakt.

Bron: 3

■ VECTOR · AI IN DE PRAKTIJK

Het besluit dat nog niet genomen kan worden

Een rapport noemt 80% automatische dekking en 2% fouten in dat deel. De productieleider wil volgende week een tweede station toevoegen. Wat ontbreekt?

Welke vervolgstap ontbreekt?

In onze fictieve evaluatie verwerkte Lumen 800 van 1.000 beelden automatisch; 16 waren fout. De overige 200 gingen naar controle. Het rapport meldt niet of die 200 binnen de beschikbare tijd correct zijn afgehandeld. Daarom kennen we nog geen eindfoutmaat van de keten. Als medewerkers gemiddeld 30 seconden per doorgegeven foto gebruiken, kost alleen die stap 100 minuten per 1.000 beelden. De bemensing is dus een onderdeel van het bewijs.

Wat verandert het tweede station?

De nieuwe camera heeft andere verlichting en beeldcompressie. Dat kan de input veranderen. We weten nog niet of Lumen die foto's goed verwerkt, welke onzekerheidsscore bruikbaar blijft en hoeveel gevallen het systeem doorgeeft. Een proef naast handmatige controle op het nieuwe station is een gerichte vervolgstap. We bewaren vooraf dezelfde foutdefinitie, zodat een eventuele verbetering niet van een soepelere norm komt.

Welk besluit is verdedigbaar?

Het verdedigbare besluit is geen algemene goedkeuring. We kunnen beschrijven dat de onderzochte configuratie op de test 800 gevallen automatisch behandelde met 16 fouten, en welke informatie voor uitbreiden ontbreekt. Het team kan eerst de handmatige keten beoordelen, vervolgens een gecontroleerde nulmeting op station 2 uitvoeren en daarna opnieuw besluiten. Dat is geen eindeloos vragen om zekerheid. Het is het verzamelen van precies de informatie die de nieuwe beslissing nodig heeft.



AXI VRAAGT

Welke ontbrekende waarneming zou jouw besluit het sterkst veranderen: nieuwe cameraresultaten, controlekwaliteit of beschikbare personeeltijd? Motiveer je keuze.

Een geschiedenis van beter toetsen

Deze tijdlijn toont ideeën die verschillende problemen zichtbaar maakten. Zij vormt geen ladder waarop iedere nieuwe methode de vorige vervangt.

1950 • Kansen krijgen een score

Brier formuleerde een kwadratische manier om kansvoorspellingen tegen uitkomsten te beoordelen. Daardoor kun je ook twee voorspellers vergelijken die dezelfde keuze maken maar ander vertrouwen melden. Ons rekenvoorbeeld op pagina 15 gebruikt de binaire vorm.

2006 • Privacy krijgt een formele grens

Dwork en collega's koppelden beschermende ruis aan de gevoeligheid van een analyse. De belangrijke stap is de expliciete vergelijking tussen aangrenzende datasets. Een vaag idee van anonimiteit wordt daarmee niet opgelost, maar een precieze claim wordt toetsbaar.

2017 • Vertrouwen wordt gemeten

Guo en collega's onderzochten kalibratie bij moderne neurale netwerken. Die vraag is anders dan de vraag of de meest waarschijnlijke klasse klopt. Een systeem kan in veel gevallen gelijk hebben en tegelijk overdreven zeker zijn.

2020 • Taakgedrag krijgt gerichte proeven

CheckList beschreef gedragstests voor NLP-systemen: kleine verschillen kunnen invarianten of verwachte veranderingen toetsen. Zulke tests vullen gemiddelde scores aan. Een ontdekt tegenvoorbeeld kan een reparatie en een blijvende regressietest opleveren.

2023 • De volledige keten komt in beeld

NIST publiceerde AI RMF 1.0 als vrijwillig risicoraamwerk. Het brengt rollen, context, meting en beheer bij elkaar. Wij gebruiken die versie om onze eigen vragen te ordenen; een ingevuld raamwerk is geen bewijs dat alle risico's verdwijnen.

■ VECTOR · AI IN DE PRAKTIJK

Jouw onderzoek: een brongebonden antwoord

Je hoeft geen model te trainen om een AI-keten te onderzoeken. Een klein, controleerbaar documentdossier kan al echte fouten en grenzen zichtbaar maken.

01 Maak het toetsbaar

Kies een openbare handleiding met datum of versienummer. Schrijf zelf vijf vragen met exacte bronpassages als referentie. Voeg een vraag toe waarop de handleiding geen antwoord geeft. Laat een toegankelijke schrijfassistent antwoorden met bronverwijzingen, of voer de oefening handmatig uit als je geen assistent gebruikt. Bewaar wat je hebt ingevoerd en wat er terugkwam, zonder onnodige persoonlijke gegevens.

02 Beoordeel twee dimensies

Markeer per afzonderlijke claim feitelijke juistheid ten opzichte van de handleiding én de steun van de genoemde passage. Een juist antwoord met een verkeerde bron blijft een interessante fout. Herhaal één vraag met de relevante tekst op een andere positie. Deze kleine proef schat geen algemene betrouwbaarheid; zij levert concrete gevallen op waarmee je een systeemcontract en een vervolgtest kunt verbeteren.

Zo doe je het

- 1 Bewaar documentnaam, datum en versie.
- 2 Leg zes vragen en referentiepassages vast vóór de proef.
- 3 Markeer welke vraag onvoldoende bewijs heeft.
- 4 Bewaar invoer en uitvoer inclusief genoemde bronnen.
- 5 Splits antwoorden in afzonderlijke controleerbare claims.
- 6 Beoordeel inhoud en bronsteun apart.
- 7 Herhaal één positievariant met dezelfde inhoud.
- 8 Schrijf een beperkte conclusie en één gerichte volgende proef.

Zes bronnen om actief te lezen

Open niet alleen de conclusie. Zoek in iedere bron de taak, datasets, meetkeuzes en de grens van de claim.

01 Datasheets en model cards

Lees Gebru en Mitchell naast elkaar. De eerste helpt gegevens beschrijven; de tweede modelgebruik en prestaties. Probeer één beperking in je eigen gegevenspaspoort en modelkaart te laten terugkomen. Documentatie is bruikbaar als dezelfde belangrijke grens op meerdere plekken herkenbaar blijft.

02 Kalibratie en dataset shift

Vergelijk Guo met Ovadia. De eerste onderzoekt afstellen van vertrouwen; de tweede vraagt wat met onzekerheid gebeurt wanneer de gegevens veranderen. Noteer welke modellen en shifts werkelijk zijn onderzocht. Vermijd een conclusie als “methode X werkt altijd”.

03 CheckList en FEVER

CheckList richt gedragstests in; FEVER maakt bronondersteuning onderdeel van een taak. Ontwerp een kleine test met een ontkenning, een uitzondering en ontbrekend bewijs. Kijk of je tests verschillende fouten blootleggen in plaats van steeds dezelfde normale vraag te herhalen.

04 Drie begrippen aan de kant

Dekking is het deel waarop een systeem antwoordt. Precisie is het aandeel terecht positieve gevallen onder meldingen. Kalibratie is overeenstemming van gerapporteerde kansen met uitkomsten in passende verzamelingen. Geen van deze begrippen betekent volledige betrouwbaarheid van de keten. Gebruik de Engelse termen coverage, precision en calibration bij het zoeken in de papers, maar leg je eigen beslisvraag in gewone taal vast.

BEGRIPPEN

Bij de paper op je bureau

Drie begrippen helpen de claim afbakenen.

Testprotocol Afspraak over data, keuzes, maten en het openen van de test.

Shift Verschil tussen eerdere en nieuwe gegevens- of uitkomstverdelingen.

Proxy Meetbaar signaal dat slechts een deel van het bedoelde doel vertegenwoordigt.

Bronnen: 1, 17

Antwoorden en redeneringen

Controleer je noemers, eenheden en aannamen. Een goede uitwerking zegt ook wat de berekening niet bewijst.

Quiz • pagina 10

1B. Hetzelfde object kan onder een andere bestandsnaam terugkomen. Houd varianten samen zodat de test niet vooral herkenning meet. Zie pagina's 5–7. 2A. Drempelkeuze is afstelling. De eindtest beoordeelt de vastgelegde keuze; zij is geen extra oefenset. Zie pagina 7. 3C. Een echte bron kan verkeerd of onvolledig gebruikt zijn. Controleer de exacte claim tegen de relevante passage. Zie pagina 23. 4B. Inhoud en bevoegdheid zijn verschillende rollen. Beperk acties en controleer parameters buiten alleen de taalmodeltekst. Zie pagina's 27–28. 5A. Een reproduceerbaar dossier bewaart ook versies, splits, rekeninstellingen en resultaten van herhaalde runs. Zie pagina 33. 6C. Niet automatisch behandeld betekent niet automatisch goed afgehandeld. De complete keten vraagt eigen bewijs. Zie pagina's 16, 43.

Quiz • pagina 26

1B. Gevoeligheid gebruikt echte positieve gevallen als noemer. Precisie gebruikt meldingen als noemer. Zie pagina 14. 2C. Kalibratie beschrijft overeenstemming over een verzameling. Zij garandeert geen afzonderlijk geval. Zie pagina 15. 3A. De garantie is marginaal onder

voorwaarden; verandering van het gegevensproces kan die voorwaarden verbreken. Zie pagina 16. 4B. Shift is een signaal om metadata en passende gelabelde voorbeelden te onderzoeken. Zie pagina 38. 5C. Variatie tussen modellen kan informatie geven, maar gedeelde training of ontbrekende ervaring kan gezamenlijke fouten veroorzaken. Zie pagina 38. 6A. Rapporteer aantallen en relevante deelgroepen, zodat gunstige en ongunstige omstandigheden zichtbaar blijven. Zie pagina's 15, 20.

Quiz • pagina 36

1B. Er zijn $90 + 198 = 288$ meldingen.
 Precisie = $90 / 288 = 0,3125 = 31,25\%$. Zie pagina 14.
 2C. $200 \times 30 = 6.000$ seconden; delen door 60 geeft 100 minuten. Zie pagina 43. 3A. Met $y = 0$ wordt $(0,8 - 0)^2 = 0,64$. Zie pagina 15. 4B.
 $x' = 2 - 0,75 \times 2 \times 2 = -1$. Het nieuwe verlies is 1; kleiner dan het beginverlies 4. Zie pagina 21.

Gedachtenproeven • pagina 31

1. Een score van honderd: De score beschrijft de bekende test, niet automatisch de nieuwe camera. Maak een gerichte proef met de nieuwe invoer en dezelfde labelafspraak. Nul geobserveerde fouten is nuttig bewijs voor die test, geen universele foutkans

van nul. Grens: De benodigde hoeveelheid nieuw bewijs hangt van taak en gevolgen af; deze gedachteproef levert geen vaste steekproefnorm. 2. Iedereen mag doorverwijzen: Nee. A heeft een lager risico binnen een veel kleiner beantwoord deel. Bij beperkte opvolging kan zijn grote afwijzingslast onbruikbaar zijn. Vergelijk bij dezelfde dekking of beoordeel de volledige keten inclusief niet beantwoorde vragen. Grens: De gegeven percentages komen uit fictieve data en bepalen geen algemene keuze tussen echte systemen. 3. Een juiste zin zonder bron: Bij de opdracht om alleen bronondersteunde claims te rapporteren volgt het tweede antwoord het contract beter. Het eerste heeft een juiste inhoud maar een fout bewijsverband. Beoordeel feitelijke juistheid en bronsteun als afzonderlijke dimensies. Grens: Een brongebonden systeem is slechts geschikt voor vragen waarvoor die brongrens passend is; de bron zelf kan ook onjuist zijn.

Simulatie • pagina 21

Voor $L = x^2$ geldt $x' = (1 - 2\eta)x$. Bij $\eta = 0,75$ en $x_0 = 2$ volgt 2; -1; 0,5; -0,25; de grootte halveert. Bij $\eta = 1$ volgt 2; -2; 2; geen krimp. Boven 1 groeit de grootte, behalve de triviale start $x = 0$. Deze stabiliteitsgrens hoort alleen bij dit exact kwadratische model. Een stabiele training bewijst geen goede generalisatie of passende taak.

Nieuwe puzzels: het volledige bewijs

De voorbeelden zijn aangenomen miniwerelden. Iedere oplossing gebruikt precies de aangegeven voorwaarden.

Raadsels • pagina 9

1. 16. Ieder blok heeft twee keuzes. Het product is $2 \times 2 \times 2 \times 2 = 16$. De lijsten AB, AB, AB, AB en BA, BA, BA, BA zijn bijvoorbeeld twee verschillende mogelijkheden.
2. Zes. Kies de twee AB-posities uit vier: $\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}$. De overige twee zijn BA. Zo zijn alle mogelijkheden uitgeschreven.
3. $-2, -2, -1, -3$ ms; gemiddeld -2 ms. Het oude verschil A–B wordt na verwisselen B–A, dus het negatieve. Som -8 gedeeld door vier is -2 . Alleen de labelrichting is veranderd; de gemeten tijden niet.

Raadsels • pagina 29

1. 59 onafhankelijke proeven, alle zonder fout. $pU \leq 0,05$ betekent $0,05^{1/n} \geq 0,95$, dus $n \geq \ln(0,05)/\ln(0,95) \approx 58,40$. De

kleinste gehele waarde is 59. Voor $n=58$ is de grens circa 5,034%; voor $n=59$ circa 4,951%.

2. Beide $14/20=70\%$; geen van beide wint op beide groepen. A: $9+5=14$; B: $8+6=14$. A wint op gemakkelijk ($9>8$); B op moeilijk ($6>5$). Het gelijke totaal bewaart dat groepsverschil niet zonder de uitsplitsing.

3. Nee: beide 90%; hun foutstatussen verschillen wel. De afgesproken teller is achttien en noemer twintig bij beide. Maar A's twee ontbrekende en B's twee onjuiste outputs zijn andere gevallen. Een andere taakafweging moet vooraf worden toegevoegd, niet achteraf uit deze ene score worden verzonnen.

Gedachtenproeven • pagina 35

Twintig kopieën: Nee. De kopieën hebben dezelfde

uitkomst; het aantal onafhankelijke uitgevoerde proeven is één. Een recordcount is geen steekproefomvang zonder de onafhankelijkheidsafspraken. Ook echte herhaalde uitvoeringen kunnen door een gedeelde context afhankelijk zijn. Niet iedere afhankelijkheid is een letterlijke kopie; de opzet moet haar beoordelen. Het herkenbare blinde label: Nee. Het gouden uiterlijk onthult de koppeling die het blinde label moest verbergen. Gebruik een gelijke presentatie en een vooraf gekozen beoordelingsregel. Bewaar daarna de oorspronkelijke identiteit om resultaten te koppelen. Blindering is niet voor iedere meting noodzakelijk of uitvoerbaar. Deze proef vraagt haar omdat subjectieve beoordeling onderdeel van de uitkomst is.

■ VECTOR • BEWIJS BIJ DE NIEUWE OPGAVEN

Nieuwe quizzen: berekening en reden

De voorbeelden zijn aangenomen miniwerelden. Iedere oplossing gebruikt precies de aangegeven voorwaarden.

Quiz • pagina 8

1. B. $(12+15+11+18)/4 = 56/4 = 14$ ms. Zie pagina 13.
2. A. De verschillen 2,2,1,3 tellen op tot 8; delen door vier geeft 2 ms. Zie pagina 13.
3. C. De vergelijking scheidt het systeem-effect niet van moeilijkheid. Zie pagina 12–13.
4. B. Het blok houdt de bekende factor bijeen; randomiseren verdeelt resterende volgorde-invloeden. Zie pagina 12.

Quiz • pagina 24

1. C. Iedere foutvrije proef heeft kans 0,9; onafhankelijkheid geeft het product $0,9^{20}$. Zie pagina 25.
2. A. $p_U = 1 - 0,05^{(1/20)} \approx 0,139108$. Zie pagina 25.
3. B. De miniwereld vereist onafhankelijke uitkomsten; duplicaten voldoen daar niet aan. Zie pagina 25, 35.
4. C. Een geobserveerd aantal is niet identiek aan een universele foutkans. Zie pagina 25.

Quiz • pagina 34

1. B. Voorwaardelijke juistheid heeft noemer 18. $17/20=85\%$ is het correcte aandeel van alle taken. Zie pagina 40.
2. A. $18/20=90\%$. Dekking en juistheid gebruiken hier verschillende tellers. Zie pagina 40.
3. C. $20-17=3$ foutieve outputs. Zie pagina 40.
4. B. Een vooraf beschreven scorecontract maakt verschillen interpreteerbaar. Zie pagina 40.

Colofon en bronnen

Betrouwbare AI vraagt een duidelijke taak, controleerbare waarnemingen en eerlijk beschreven grenzen.

Over dit nummer

VECTOR combineert oorspronkelijke uitleg, fictieve rekenvoorbeelden en primaire onderzoeksbronnen. Lumen, AXI en beide strips zijn verzonnen; zij

beschrijven geen bestaand product of bewust wezen. Foto's zijn documentaire beelden; als illustratie gemarkeerde scènes zijn met AI gemaakt. De bronnen hieronder ondersteunen de vakinhoud en

vermelden waar een resultaat begrensd is. Bewaar bij hergebruik de beeldcredits en licenties. Een volgende editie verkent AI als onderzoeksinstrument in de meetbare wereld.

Bronnen

- 1 NIST (2023) • AI Risk Management Framework 1.0 · nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10
- 2 Gebru e.a. (2021) • Datasheets for Datasets · [arXiv:1803.09010](https://arxiv.org/abs/1803.09010)
- 3 Mitchell e.a. (2019) • Model Cards for Model Reporting · [arXiv:1810.03993](https://arxiv.org/abs/1810.03993)
- 4 Kapoor & Narayanan (2023) • Leakage and the Reproducibility Crisis · [arXiv:2207.07048](https://arxiv.org/abs/2207.07048)
- 5 Recht e.a. (2019) • Do ImageNet Classifiers Generalize to ImageNet? · proceedings.mlr.press/v97/recht19a.html
- 6 Guo e.a. (2017) • On Calibration of Modern Neural Networks · proceedings.mlr.press/v70/guo17a.html
- 7 Ovadia e.a. (2019) • Predictive Uncertainty under Dataset Shift · [arXiv:1906.02530](https://arxiv.org/abs/1906.02530)
- 8 Shafer & Vovk (2008) • A Tutorial on Conformal Prediction · jmlr.org/papers/v9/shafer08a.html
- 9 Geifman & El-Yaniv (2017) • Selective Classification · [arXiv:1705.08500](https://arxiv.org/abs/1705.08500)
- 10 Lewis e.a. (2020) • Retrieval-Augmented Generation · [arXiv:2005.11401](https://arxiv.org/abs/2005.11401)
- 11 Liu e.a. (2024) • Lost in the Middle · aclanthology.org/2024.tacl-1.9

- 12 Thorne e.a. (2018) • FEVER · aclanthology.org/N18-1074
- 13 Greshake e.a. (2023) • Indirect Prompt Injection · [arXiv:2302.12173](https://arxiv.org/abs/2302.12173)
- 14 Ribeiro e.a. (2020) • Beyond Accuracy: CheckList · aclanthology.org/2020.acl-main.442
- 15 Dwork e.a. (2006) • Calibrating Noise to Sensitivity · microsoft.com/en-us/research/publication/calibrating-noise-to-sensitivity-in-private-data-analysis
- 16 Carlini e.a. (2021) • Extracting Training Data from Large Language Models · usenix.org/conference/usenixsecurity21/presentation/carlini-extracting
- 17 Sculley e.a. (2015) • Hidden Technical Debt in Machine Learning Systems · papers.nips.cc/paper_files/paper/2015/hash/86df7dcfd896fcafb2674f757a2463eba-Abstract.html
- 18 Sandve e.a. (2013) • Reproducible Computational Research · journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285
- 19 Koh e.a. (2021) • WILDS · proceedings.mlr.press/v139/koh21a.html
- 20 Hardt e.a. (2016) • Equality of Opportunity in Supervised Learning · [arXiv:1610.02413](https://arxiv.org/abs/1610.02413)
- 21 Geirhos e.a. (2020) • Shortcut Learning · nature.com/articles/s42256-020-00257-z
- 22 Raji e.a. (2020) • Closing the AI Accountability Gap · [arXiv:2001.00973](https://arxiv.org/abs/2001.00973)
- 23 Brier (1950) • Verification of Probability Forecasts · journals.ametsoc.org/doi/abs/10.1175/1520-0493%281950%29078%3C0001%3AVOFEIT%3E2.0.CO%3B2
- 24 Gneiting & Raftery (2007) • Strictly Proper Scoring Rules · [doi:10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437)
- 25 Lakshminarayanan e.a. (2017) • Deep Ensembles · [arXiv:1612.01474](https://arxiv.org/abs/1612.01474)
- 26 NIST (2024) • Generative Artificial Intelligence Profile · nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- 27 Hendrycks e.a. (2021) • Measuring Massive Multitask Language Understanding · [arXiv:2009.03300](https://arxiv.org/abs/2009.03300)
- 28 NIST Engineering Statistics Handbook • Randomized block designs · itl.nist.gov/div898/handbook/pri/section3/pri332.htm
- 29 NIST Engineering Statistics Handbook • Confidence intervals · itl.nist.gov/div898/handbook/prc/section2/prc241.htm
- 30 OpenAI • GPT-6 Astra, officiële modeldocumentatie · developers.openai.com/api/docs/models/gpt-6-astra
- 31 OpenAI • GPT-6.1 Sol, officiële modeldocumentatie · developers.openai.com/api/docs/models/gpt-6.1-sol

developers.openai.com/api/docs/models/gpt-6.1-sol

- 32 Anthropic • Claude Opus 5.5, officiële modeldocumentatie · platform.claude.com/docs/en/models/opus-5-5/overview

Foto's

p. 2, 33 • Si wafer · 2x910 · CC BY-SA 4.0 · [Wikimedia Commons](#)

Illustraties

p. 1 • Vector 04 cover new · Originele AI-illustratie voor VECTOR | p. 2, 17 • Vector 04 axi comic 3 · Originele AI-illustratie voor VECTOR | p. 2, 3 • VECTOR 04 — zesdelige strip · Originele AI-illustratie voor VECTOR | p. 2, 22, 23 • VECTOR 04 — conceptillustratie voor dubbele pagina · Originele AI-illustratie voor VECTOR | p. 2, 12 • Vector 04 shadow trial · Originele AI-illustratie voor VECTOR | p. 6, 7 • VECTOR 04 — conceptillustratie voor dubbele pagina · Originele AI-illustratie voor VECTOR | p. 19 • VECTOR 04 — zesdelige strip · Originele AI-illustratie voor VECTOR | p. 50 • VECTOR 05 — omslagillustratie · Originele AI-illustratie voor VECTOR | AXI, de waarnemer · Originele AI-illustratie voor VECTOR

Foto's zijn verkleind en kunnen voor de opmaak zijn bijgesneden; ze houden hun aangegeven licentie. Teksten, diagrammen, puzzels en vormgeving: VECTOR.



IN DE VOLGENDE EDITIE • Vanaf 8 oktober 2026

AI ontdekt de wereld

- Hoe een berekende kandidaat een echt experiment wordt.
- Wat een robot buiten zijn trainingswereld tegenkomt.
- AI voor weer, materialen en plasmacontrole.