Evaluatiekader voor GenAI-systemen

Een toolkit om GenAI-toepassingen te evalueren van begin tot einde

2025-2027

GPT Academy

gpt-academy.be/evaluatie

UCLL HOGESCHOOL Research & Expertise

Introductie

Waarom een evaluatiekader?
Door snel op de kar te willen springen, durft men soms wel eens voorbij te gaan aan het belang van een diepgaande analyse om te bepalen of een aangekocht systeem daadwerkelijk aan de verwachtingen voldoet en daarmee de investering waard is. Dit kan een tijdsinvestering zijn in het geval van gratis tools, of een geldinvestering bij betalende licenties.

Momenteel evalueren gebruikers deze systemen vaak op basis van een eerste indruk of gevoel – een zogenaamde vibe check. Hoewel dit begrijpelijk is wanneer dit de enige methode voorhanden is, is dit 'gevoel' slechts een oppervlakkige aanpak.

Dit holistisch evaluatiekader helpt niet-technische eindgebruikers in kmo's om de prestaties van GenAI-toepassingen systematisch te beoordelen. Door deze gestructureerde aanpak krijgen bedrijven een genuanceerd beeld van de sterke en zwakke punten van een GenAI-systeem, wat leidt tot beter geïnformeerde beslissingen. Zo hopen we ertoe te leiden dat kmo's hun investeringen in GenAI kunnen optimaliseren en tools selecteren die beter aansluiten bij hun specifieke bedrijfsdoelen en verwachtingen.

Drie fasen
Evaluatie verloopt volgens ons kader in drie fases: (1) een geïsoleerde **pre-evaluatie**, (2) evaluatie **tijdens het gebruik** (bv. in een piloottest), en (3) **post-evaluatie** acties. Tijdens de pre-evaluatie bereid je je vooraf voor om de beste tool of het beste model te kiezen voor hetgeen je voor ogen hebt. Het GenAI-systeem wordt daarna effectief tijdens de ingebruikname geëvalueerd. Op basis van die evaluaties onderneem je achteraf acties om het systeem te verbeteren, om vervolgens de evaluatie iteratief te hernemen. Elke fase wordt gekenmerkt door specifieke vragen en bezorgdheden.

Twee perspectieven
Niet iedereen speelt dezelfde rol in de evaluatie van een GenAI-systeem. Er zijn namelijk bepaalde zaken die specifiek relevant zijn voor **ontwikkelaars** of technische profielen, terwijl andere zaken essentieel zijn voor **niet-technische eindgebruikers** van de GenAI-systemen. Wij bekijken evaluatie vanuit beide perspectieven. Uiteraard ontmoeten die aspecten elkaar en gebeurt de evaluatie door ontwikkelaars in tandem met die van eindgebruikers.

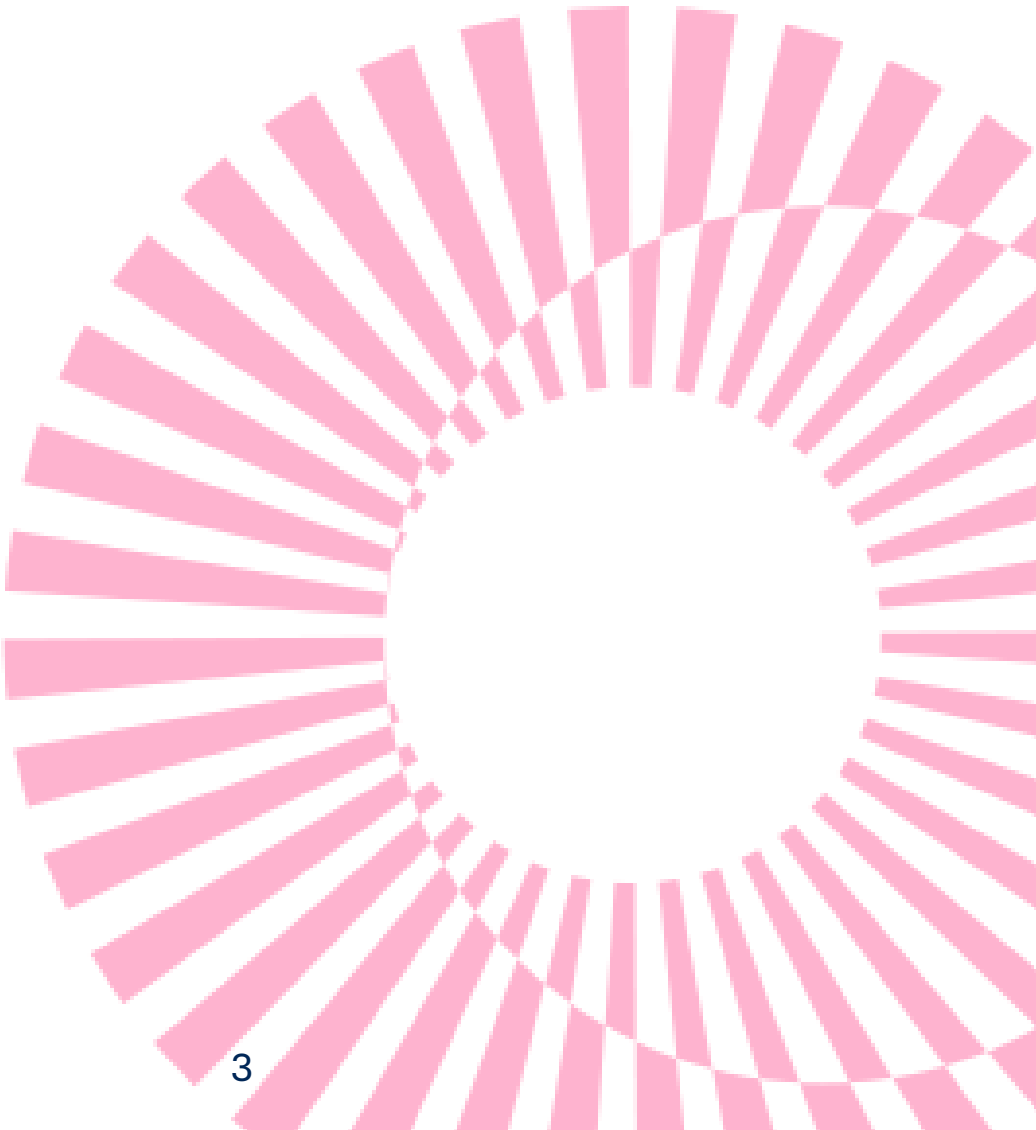
Wanneer eindgebruikers echter een systeem aankopen waarover ze geen echt zeggenschap hebben (bv. Microsoft Copilot, ChatGPT…), valt het luik voor de ontwikkelaars weg en moet er op vlak van evaluatie enkel gekeken worden naar het luik voor de eindgebruiker.

Introductie

Enkele opmerkingen
GenAI-systemen hebben specifieke eigenschappen, maar uiteindelijk blijven het nog altijd softwareproducten die een probleem van een eindgebruiker oplossen. Veel evaluatietechnieken die traditioneel al worden gebruikt om (onder andere) software te beoordelen (A/B-testing, klassieke tevredenheidsenquêtes, focusgroepen, User Acceptance Testing...) blijven dus erg relevant en nemen we op in ons evaluatiekader (zie p. 10).

Strategische overwegingen zoals het identificeren van het onderliggende bedrijfsprobleem, het bepalen of GenAI de meest geschikte oplossingsrichting is, en het afwegen van alternatieve technologieën vallen buiten de scope van dit kader.

Hoe concreet aan de slag?
We bieden een aantal tools, checklists en methodieken aan. Overloop zeker en vast de aanbevelingen die we doen voor elk van de drie fases vanuit jouw eigen perspectief (ontwikkelaar of gebruiker). Het kan ook interessant zijn om die van de andere zijde te bekijken, om een beter zicht te krijgen op hoe evaluatie voor hen in zijn werk gaat.



Vooraf - Eindgebruiker

Voor je een GenAI-toepassing inzet of aankoopt, bekijk je best welke oplossing het best past bij jouw noden. Onderstaande vragen stel je jezelf best op voorhand als eindgebruiker.

Allereerst verdiep je jezelf best in het brede landschap aan AI-toepassingen. Ontdek zo welke aansluiten bij jouw noden. Kijk eerst wat de grote, populaire (chat)tools bieden (zoals OpenAI ChatGPT, Microsoft Copilot, Claude Anthropic...), maar kijk zeker ook naar andere meer specifieke tools. Die eerste categorie zijn vaak “jack-of-all-trades, but master of none”. Dit wil zeggen dat ze goed zijn in het uitvoeren van de meeste taken, maar nergens in gespecialiseerd zijn.

In de laatste categorie kan je tools vinden die specifiek gemaakt zijn voor een bepaald doel, bv. een transcriptietool, een tool om e-mails te optimaliseren, etc. Een hulp bij het vinden van de verschillende AI-tools zijn websites zoals **There’s An AI For That**.

Soms heb je echter een eigen AI-tool op maat van je bedrijf nodig, of dat nu is omwille van dataveiligheid of omdat de specifieke toepassing die jij zoekt nu eenmaal nog niet op de markt is.



Let op: Deze checklist houdt geen rekening met de vraag of GenAI überhaupt de oplossing voor je probleem is.

Hoe zit het bij jou?

- [ ] Wat is het grotere plaatje? Wat wil je bereiken? Je kan **dit bord** dupliceren en invullen.
- [ ] Bestaat de toepassing al, of moet deze ontwikkeld worden?
- [ ] Is er opleiding nodig voor je er mee aan de slag kan gaan?
- [ ] Kan de toepassing bv. integreren met je bedrijfsspecifieke systemen?
- [ ] Is het mogelijk om de toepassing op je bedrijfsdata te laten werken?
- [ ] Is de toepassing EU AI Act- en GDPR-compliant?
- [ ] Is het mogelijk om de output van de tool te exporteren naar verschillende formaten?
- [ ] Wat is de kostprijs? Is er een maandelijkse of jaarlijkse betaling?
- [ ] Kan je de toepassing op je eigen systemen draaien of ben je verplicht dit te doen op de servers van het bedrijf?

Vooraf - Eindgebruiker

Toolmatchmaking flowchart



* Tegenwoordig bieden bedrijfslicenties ook vaak al een bepaald niveau van bescherming. Check dus welke afspraken je maakt met de provider en hoe veilig je gegevens zijn!

Tijdens - Eindgebruiker

Hoe ga je tijdens de pilot meten of iets goed werkt?

Het bepalen en ranken van criteria
In het vorige deel hebben we de ambitie bepaald. Je kent het grotere plaatje en weet wat je wil bereiken door de GenAI-toepassing in te zetten. Dit kan je al invullen in het **Miro-bord** (eerst dupliceren) dat we hebben opgesteld om je te helpen tot je evaluatiecriteria te komen.

Dit hoofddoel kan opgedeeld worden in '**jobs to be done**', de deeltaken of subdoelen die de GenAI-toepassing moet ondersteunen. Er kunnen meerdere jobs to be done onder één ambitie vallen. Een job to be done kan gedefinieerd worden als: 'Wanneer X zich voordoet (situatie), wil ik Y (motivatie), zodat ik Z (verwachte uitkomst)'.

Het is niet altijd eenvoudig om in te schatten wanneer jobs to be done succesvol zijn uitgevoerd. Vaak worden **harde succescriteria** gebruikt (bv. tijdswinst, meer klanten helpen per dag, minder fouten in je teksten), maar ook **zachte criteria** spelen een rol (bv. productiviteitsperceptie, algemene tevredenheid, appreciatie). Er kunnen meerdere taakspecifieke succescriteria onder één job to be done vallen.

Het is belangrijk om na te denken hoe je deze laatste succescriteria **SMART** (Specifiek, Meetbaar, Acceptabel en Realistisch) kunt opstellen, zeker in het geval van zachte criteria waarbij dit minder voor de hand ligt.

Niet alle criteria zijn van even groot belang. Het kan helpen om ze te ranken in 'Must have', 'Should have' en 'Nice to have' categorieën. Wanneer er niet voldaan wordt aan de 'Must have' criteria, is een GenAI-toepassing niet geslaagd. Denk hier dus goed over na. Wanneer al je criteria 'Must have' zijn of wanneer je erg veel 'Should have' criteria hebt, kan de aandacht naar de echt allerbelangrijkste criteria verloren gaan.



Wanneer je een meetbaar criterium opstelt, vergeet dan ook niet om de status quo te meten. Zonder zo'n nulmeting of baseline kan je namelijk niet vergelijken. Je kan hiervoor bv. het framework Performance Measurement Baseline gebruiken.

Achteraf - Eindgebruiker

Welke acties kan je ondernemen om de toepassing beter voor je te doen werken?

Hou je prompts bij
Zo kan je ze opnieuw uitproberen wanneer er door de ontwikkelaar verbeteringen zijn aangebracht.

Verbeter je prompt
- Wees volledig. Geef voldoende context mee.
- Structureer je prompt. Deel hem op in delen en benoem ze.
- Als je je prompt in een gesprek veel hebt moeten bijsturen, vraag dan aan de chatbot hoe je de prompt in het vervolg beter had kunnen verwoorden. Werkt niet altijd, maar kan inspiratie geven.

Zorg voor propere data
Gebruik je een tool die in je data moet zoeken? Bekijk of die gegevens ordelijk, up to date en makkelijk doorzoekbaar zijn.

Blijf bijleren
Sommige tools vereisen misschien meer voorkennis van jezelf of je medewerkers dan je op voorhand voorzien had. Met een training leer je het onderste uit de kan te halen.

Vooraf - Ontwikkelaar

Welke resources gebruik ik? Welk model wordt de basis van mijn tool? Onderstaande vragen stel je jezelf best op voorhand als ontwikkelaar.

De vereisten waaraan je GenAI-toepassing en de resources die je gebruikt om deze te ontwikkelen aan moeten voldoen, kan je vanuit ontwikkelaarsperspectief opdelen in twee categorieën: functionele, inhoudelijke aspecten en non-functionele requirements. Onder de laatste categorie vallen vooral droge, harde cijfers, zoals de grootte van het model en de kostprijs.

De functionele aspecten zijn een stukje desk research. Zo goed als elk model komt met een korte paper, waar de resultaten van evaluatie op verschillende benchmarks wordt weergegeven, en een model card of system card.

Naast deze interne evaluatie, worden modellen ook regelmatig door externen op open source benchmarks en datasets getest. Enkele voorbeelden zijn:

- **Chatbot Arena**: een van de meest populaire benchmarks voor chatbots
- **GLUE**: General Language Understanding Evaluation, bestaande uit 9 assessment onderdelen
- **GPQA**: Graduate-Level Google-Proof Q&A Benchmark
- **EuroEval**: een framework voor benchmarks waar Nederlands ook meer en meer aandacht krijgt

Non-functionele requirements

- [ ] Hoe snel is het?
- [ ] Hoe groot is het model?
- [ ] Is het open source beschikbaar, of komt er een kost bij kijken?
- [ ] Draait het op je GPU of op een externe server?
- [ ] Is alles EU AI Act- en GDPR-compliant?

Functionele aspecten

- [ ] Ik heb de modelcard bekeken.
- [ ] Het model is getest op relevante benchmarks.

Om te weten wat er nodig is, zit je natuurlijk ook samen met de klant of de eindgebruiker waarvoor je de GenAI-toepassing maakt. Hun versie van de checklist (zie p. 4) kan hierbij als leidraad gebruikt worden.

Tijdens - Ontwikkelaar

Hoe ga je tijdens de ontwikkeling meten of iets goed werkt?

Enerzijds is er **intrinsieke evaluatie**, waarmee je bekijkt hoe goed het model an sich is. Anderzijds is er **extrinsieke evaluatie**, waarmee je bekijkt hoe goed het model werkt in het systeem waarin het zit.

We geven een kort algemeen overzicht van intrinsieke metrieken, meer info vind je in:

- Ons **Google Colab notebook** waarin we enkele van deze metrieken verder uitdiepen en demonstreren.
- De bronnen van het opensourceproject HuggingFace, zoals de **demo's per metriek** en de **documentatie over de evaluate-bibliotheek**

Generische metrieken
Denk hierbij aan accuracy, precision, recall, F1-score.

Taakspecifieke metrieken
Hieronder bevinden zich bv. de statistical scorers voor eerder 'old school NLP' taken. Enkele voorbeelden zijn:

- BLEU voor machine translation
- ROUGE voor summarization, met een focus op recall
- METEOR voor machine translation, met meer balans tussen precision en recall

Datasetspecifieke metrieken
Benchmarks met eigen specifieke datasets

Een geschiedenis van de evaluatie van tekstgeneratie
Tekstgeneratie kan je zien als een specifieke taak. Er zijn pogingen gedaan om dit te evalueren met de taakspecifieke **metrieken van andere NLP-taken**, maar dit was niet altijd succesvol. Vaak zijn deze ook te rigide voor een open-ended taak als tekstgeneratie.

Met de '**NLP-as-a-judge**' krijgen we een kleine verbetering: een taalmodel - dat geen LLM is - geeft een score aan een gegenereerde output. Voorbeelden zijn BLEURT, COMET en BERTScore. BERTScore kijkt bv. naar cosine distance tussen embeddings i.p.v. naar een exacte match. Je hebt met deze metrieken echter nog altijd een referentietekst nodig, en ze bleken niet altijd goed overeen te komen met hoe mensen open-ended outputs van generatieve modellen beoordelen.

Hier komt '**LLM-as-a-judge**' kijken, een metriek die erg populair is, zeker ook als extrinsieke evaluatiemethode. Hierbij beoordeelt een LLM de output van een andere LLM, op basis van criteria die jij kiest. Voorbeelden zijn **G-Eval via DeepEval** en GPTScore. DeepEval biedt nog heel wat andere mogelijkheden, bijvoorbeeld specifieke evaluatie van **samenvattingen** of van feitelijk correcte antwoorden (**RAGAS**). Ook **TruLens** kan interessant zijn voor het tracken van AI Agents.



Experimenteer! Voeg LLM-as-a-judge op verschillende stappen in de flow toe.

Tijdens - Ontwikkelaar

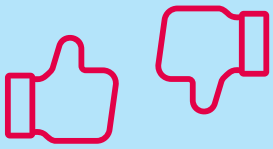
Hoe ga je tijdens de pilot meten of iets goed werkt?

Traditionele evaluatietechnieken uit Software Development
Er zijn verschillende technieken die al een lange geschiedenis hebben. We focussen hier minder op, maar hier zijn een aantal voorbeelden:

- A/B-testing
- User Acceptance Testing
- Tevredenheidsenquêtes
- Focusgroepen



Voeg een like/dislike knop toe voor de gebruiker.



Achteraf - Ontwikkelaar

Welke acties kan je ondernemen om de toepassing te verbeteren?

Ga aan de slag met de feedback
Met Reinforcement Learning from Human Feedback (RLHF) kan je een model afstemmen op de voorkeuren van mensen. In de vorige stap heb je de input van je eindgebruikers verzameld. Je kan deze input gebruiken om het model te verbeteren met RLHF, in het geval dat je toegang hebt tot de gewichten van het model en het kan finetunen. Als dit niet mogelijk is, kan je de feedback gebruiken om het systeem te optimaliseren (bv. door prompts aan te passen, relevantere context mee te geven, etc.)

Verbeter je (systeem)prompt
- Wees volledig. Geef voldoende context mee.
- Structureer je prompt. Deel hem op in delen en benoem ze.
- Als je je prompt in een gesprek veel hebt moeten bijsturen, vraag dan aan de chatbot hoe je de prompt in het vervolg beter had kunnen verwoorden. Werkt niet altijd, maar kan inspiratie geven.

Zorg voor propere data
Gebruik je een tool die in data moet zoeken? Bekijk of die gegevens ordelijk, up to date en makkelijk doorzoekbaar zijn.

Colofon

Met veel dank aan de partners in het projectconsortium, Hogeschool UCLL en alle deelnemers en bedrijven die input gaven.

Vanwege van het team van GPT Academy
Heike Pauli en Michael Bauwens

gpt-academy.be/evaluatie

