NUMMER 32, 23 MAART 2026

GENTSE ECONOMISCHE INZICHTEN

UNIVERSITEIT GENT

IS AI BETROUWBAAR VOOR HET BEANTWOORDEN VAN PENSIOENVRAGEN?

Kris Boudt, Arno De Block, Feliciaan De Palmenaer,

Andries Fluit, Yanick Inghels

Vakgroep Economie, Universiteit Gent

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

KERNINZICHTEN

- **Artificiële intelligentie wordt steeds vaker ingezet voor kennisintensief advies**, maar in gereguleerde domeinen waar fouten financiële of juridische gevolgen hebben, blijft de betrouwbaarheid van generieke taalmodellen een open vraag. Dit Inzicht onderzoekt die vraag voor het Belgische pensioendomein.
- We evalueerden de prestaties van **zes toonaangevende generieke taalmodellen** op meer dan 170 meerkeuzevragen over het Belgische pensioenstelsel, opgesteld en gevalideerd samen met onafhankelijke pensioenexperts. We onderscheiden daarbij publieksvragen en technischere expertvragen.
- **Generieke taalmodellen presteren behoorlijk op algemene pensioenvragen**, maar schieten tekort bij expertvragen waar uitzonderingsregels, fiscale nuances of precieze berekeningswijzen doorslaggevend zijn. Het best presterende generieke model behaalt 92 procent overall, maar bij expertvragen daalt de nauwkeurigheid tot 68 à 88 procent afhankelijk van het model.
- Een **domeinspecifieke aanpak**, waarbij een model en dataset speciaal ontwikkeld is voor de probleemstelling, behaalt 96 procent overall en 92 procent op expertvragen, met een gemiddelde responstijd van 16 seconden tegenover 28 seconden voor het best presterende generieke model. Dat verschil wordt niet verklaard door een krachtiger taalmodel, maar door de informatieomgeving waarin het opereert: een gecureerde kennisbasis die systematisch wordt geactualiseerd bij wetgevingswijzigingen, een gericht retrievalsysteem dat enkel gevalideerde bronnen doorzoekt, en deterministische rekenmodules voor numerieke parameters.
- De resultaten suggereren dat een **gecureerde, domeinspecifieke architectuur de nauwkeurigheidskloof kan dichten** in gereguleerde kennisdomeinen. Of dat ook geldt voor fiscaliteit, arbeidsrecht of socialezekerheidsrecht, vergt afzonderlijke empirische toetsing.

INLEIDING[1]

Europese banken en verzekeraars investeren steeds nadrukkelijker in AI-gestuurde klantinteractie. Volgens de European Banking Authority gebruikt circa 40 procent van de EU-banken inmiddels generatieve AI, voornamelijk in klantenservice en interne procesoptimalisatie (European Banking Authority, 2024). Maar naarmate de vereiste nauwkeurigheid toeneemt, worden de beperkingen zichtbaar. Generieke taalmodellen zijn getraind op brede data en geven voorrang aan taalkundige plausibiliteit boven feitelijke correctheid. Ze genereren antwoorden die overtuigend klinken maar feitelijk onjuist kunnen zijn. Ze werken met trainingsdata die maanden achterloopt op wetgevingswijzigingen, en ze opereren zonder controleerbaar brongebruik. Voor alledaags gebruik is dat zelden problematisch. In gereguleerde adviesomgevingen, waar fouten financiële, juridische en reputationele gevolgen hebben, is het dat wel.

Het Belgische pensioenadvies illustreert deze spanning bij uitstek.

Kunnen we artificiële intelligentie wel vertrouwen wanneer het gaat om vragen over ons pensioen? De boutade gaat dat als je deze vraag stelt aan twee economen je drie antwoorden krijgt. Daarom hebben we de vraag getest op een ruime set van 170 pensioenvragen, opgesteld en gevalideerd door domeinexperts. We richten ons hierbij uitsluitend op feitelijke pensioenvragen: vragen die betrekking hebben op regelgeving, voorwaarden, berekeningswijzen, fiscale behandeling en uitbetalingsmodaliteiten binnen het Belgische meerpijlersysteem. Vragen die individuele dossiertoegang vereisen, gedragsmatig advies of psychologische begeleiding vallen buiten het bestek van dit onderzoek. Die afbakening is methodologisch gemotiveerd: feitelijke vragen laten eenduidige scoring toe en maken het mogelijk om de nauwkeurigheid van AI-systemen objectief te meten.

Generieke taalmodellen schieten tekort bij complexe pensioenvragen waarvoor uitzonderingsregels, fiscale nuances of precieze berekeningswijzen doorslaggevend zijn.

We vinden dat **generieke taalmodellen** behoorlijk antwoorden op algemene pensioenvragen, maar tekort schieten bij expertvragen waar uitzonderingsregels, fiscale nuances of precieze berekeningswijzen doorslaggevend zijn. Het best presterende generieke model behaalt 92 procent overall, maar bij expertvragen daalt de nauwkeurigheid tot 68 à 88 procent afhankelijk van het model.

Een **domeinspecifieke aanpak**, waarbij een model en dataset speciaal ontwikkeld is voor de probleemstelling, behaalt 96 procent overall en 92 procent op expertvragen, met een gemiddelde responstijd van 16 seconden tegenover 28 seconden voor het best presterende generieke model. Dat verschil

[1] Dit Gents Economisch Inzicht kon genieten van waardevolle inzichten aangereikt door Bart Chiau, Koen Inghelbrecht, Gerrit Van Daele, Rudi Vander Vennet (Vakgroep Economie) en Steven Vanduffel. Wij danken ook Zoë Imhof, het IOF (F2023/IOF-ConcepTT/074), het FWO (F2023/IOF-ConcepTT/074, 3179L0120), en het UGent en VUB Technology Transfer Office voor de steun om dit onderzoek mogelijk te maken.

wordt niet verklaard door een krachtiger taalmodel, maar door de informatieomgeving waarin het opereert: een gecureerde kennisbasis die systematisch wordt geactualiseerd bij wetgevingswijzigingen, een gericht retrievalsysteem dat enkel gevalideerde bronnen doorzoekt, en deterministische rekenmodules voor numerieke parameters.

In dit *Gents Economisch Inzicht* geven we eerst een overzicht van de gebruikte testset van pensioenvragen. Vervolgens bespreken en evalueren we in sectie 3 bestaande AI-oplossingen voor het beantwoorden van dergelijke vragen. In sectie 4 introduceren en evalueren we een domeinspecifieke AI assistent die we Bikon 1.0 noemen. We besluiten met een bespreking van de implicaties voor adviesproductiviteit en de bredere inzet van AI in gereguleerde domeinen.

OVERZICHT FEITELIJKE PENSIOENVRAGEN

Pensioeninformatie bestrijkt een breed spectrum: van vragen over het wettelijk pensioen, zoals pensioenbedrag, vervroegd pensioen, loopbaanvoorwaarden, over het aanvullend pensioen via werkgever of beroepsactiviteit en de fiscale behandeling daarvan, tot individueel pensioensparen in de derde pijler. In dit onderzoek beperken we ons echter tot de feitelijke dimensie van pensioeninformatie: vragen waarvoor een objectief correct antwoord kan worden geformuleerd op basis van de geldende regelgeving. Vragen die gedragsmatig advies of psychologische begeleiding vereisen, vallen buiten het bestek van deze studie. Deze afbakening is methodologisch gemotiveerd: feitelijke vragen laten eenduidige scoring toe en maken het mogelijk om de nauwkeurigheid van AI-systemen objectief te meten.

Pensioenvragen kunnen vervolgens worden onderverdeeld in algemene en persoonlijke vragen. Persoonlijke vragen vereisen individuele informatie, zoals iemands exacte loopbaan, pensioenbedragen of pensioenrechten. In deze benchmark beschouwen we enkel algemene vragen: vragen die beantwoord kunnen worden zonder nood aan toegang tot een persoonlijk dossier. Binnen die algemene vragen maken we vervolgens nog een onderscheid tussen publieksvragen en meer technische expertvragen. Om de prestaties van AI-modellen te evalueren, hebben we op basis van deze indeling een specifieke vragenset samengesteld rond het Belgische pensioensysteem.

Een eerste categorie bestaat uit publieksvragen: herkenbare en praktische vragen die aansluiten bij wat burgers in de praktijk willen weten over hun pensioen. Voorbeelden zijn vragen zoals “Wat is het minimumpensioen voor een alleenstaande?” of “Telt de periode waarin ik mijn kind opvoedde en niet werkte mee voor de berekening van mijn vroegst mogelijke pensioendatum?”. Dit soort vragen staat dicht bij de leefwereld van veel mensen en vertrekt vaak vanuit een concrete situatie of bezorgdheid. Voor deze categorie hebben we ons bij de samenstelling van de evaluatieset gebaseerd op publieksgerichte informatiebronnen, zoals FAQ’s en communicatie van de Federale Pensioendienst, Wikifin, de FSMA en het RSVZ. Daarnaast werden ook vragen en invalshoeken uit mediabronnen meegenomen, zoals pensioenrubrieken in kranten als De Tijd en HLN.

Daarnaast onderscheiden we expertvragen. Dat zijn technischere vragen waarin uitzonderingsregels, bijzondere statuten, fiscaliteit, berekeningsregels of productkennis centraal staan. Voorbeelden uit de evaluatieset zijn vragen zoals “Op welke voordelige loopbaanbreuk heb ik recht als kleuterjuf voor de berekening van mijn pensioenbedrag?” of “Wat is correct mbt de berekening van het wettelijk pensioen van een beroepsjournalist?”. Dergelijke vragen zijn complexer omdat het juiste antwoord vaak

afhangt van een specifiek statuut, een afwijkende regel of een nuance in de berekeningswijze. Voor dit type vragen hebben we beroep gedaan op de expertise van drie professoren financiële economie en verzekeringen, namelijk Professor Bart Chiau (UGent), Professor Gerrit Van Daele (UGent) en Professor Steven Vanduffel (VUB).

Net de combinatie van beide types vragen is belangrijk. Een evaluatieset met alleen algemene vragen zou te weinig zeggen over de prestaties in complexe of technische gevallen. Omgekeerd zou een set met alleen expertvragen minder goed aansluiten bij de realiteit van eindgebruikers. Door beide types te combineren, weerspiegelt de set zowel de vragen die mensen echt stellen als de technische complexiteit van het pensioendomein. Verdere voorbeeldvragen kunnen worden gevonden in Annex 1.

BESTAANDE AI OPLOSSINGEN

Generatieve AI biedt steeds meer potentieel om feitelijke pensioenvragen te beantwoorden, zowel voor burgers als voor professionele adviseurs. De ontwikkeling van deze modellen wordt vandaag gedomineerd door een klein aantal technologiebedrijven: OpenAI met ChatGPT, Google DeepMind met Gemini en Anthropic met Claude. Binnen deze modellen ontstaat bovendien een belangrijk onderscheid. Klassieke modellen genereren vrijwel onmiddellijk een antwoord, terwijl een nieuwe generatie zogenaamde reasoning-modellen eerst intern redeneert en meerdere stappen doorloopt voordat een antwoord wordt geformuleerd.

Beperkingen

De prestaties van grote taalmodellen zijn de voorbije jaren duidelijk verbeterd, onder meer in fiscale, financiële, juridische en medische toepassingen (Nie et al., 2024; Siino et al., 2025; Maity and Saikia, 2025). Toch betekent dit niet dat zulke modellen automatisch goed presteren in een specifieke nationale context zoals het Belgische pensioenstelsel. Pensioenregels zijn sterk domeinspecifiek, bevatten veel uitzonderingen en evolueren frequent. Bovendien hebben taalmodellen een zogenaamde knowledge cutoff: hun interne kennis stopt op een bepaald moment in de tijd, waardoor recente wetswijzigingen of nieuwe interpretaties mogelijk niet in het model zijn opgenomen (Gao et al., 2023). Dat vergroot het risico op hallucinaties, waarbij een model een antwoord formuleert dat overtuigend klinkt maar feitelijk onjuist is (Huang et al., 2025).

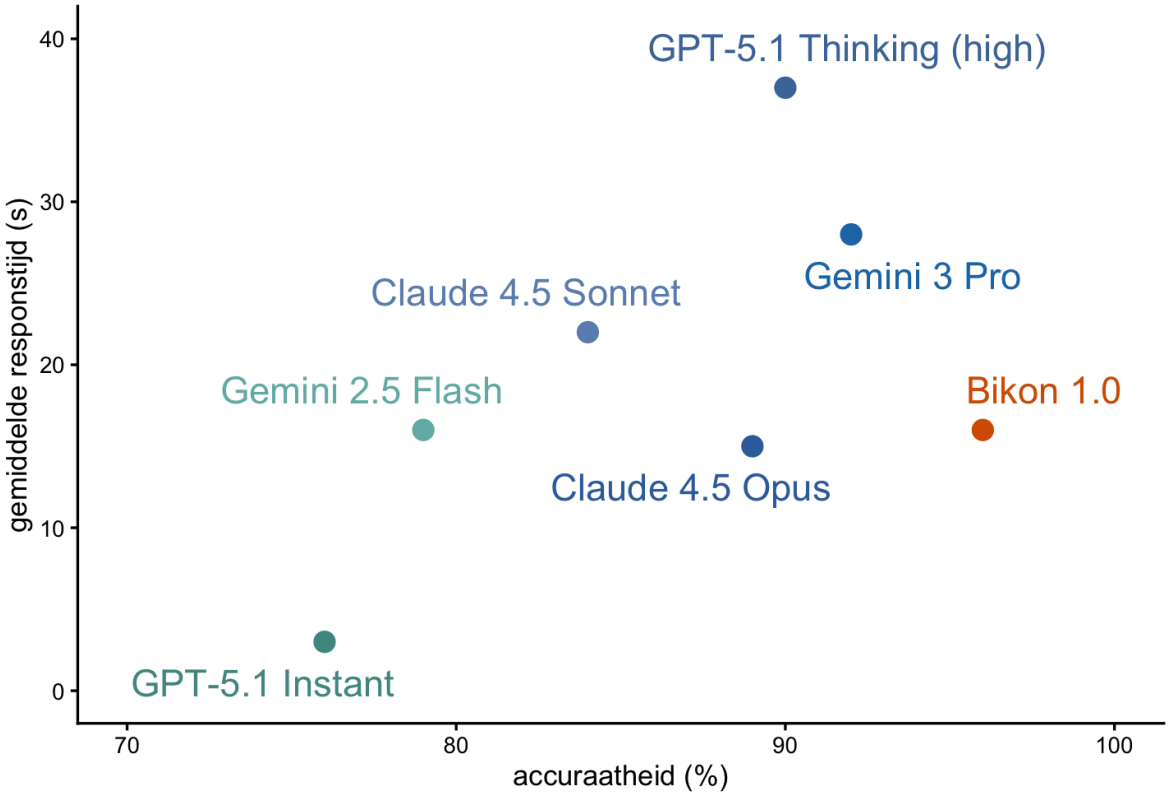
Om verouderde kennis op te vangen, laten AI-bedrijven hun modellen steeds vaker informatie ophalen via externe zoeksystemen. Reasoning-modellen gaan daarin het verst: zij splitsen vragen op in deelvragen en voeren meerdere zoekopdrachten uit voordat ze een antwoord formuleren (Li et al., 2025). Maar deze zoeksystemen zijn algemene infrastructuur, geen domeinspecifieke zoekmachines voor Belgische pensioenvragen. Er is weinig controle over welke bronnen worden opgehaald, hoe die worden gewogen en of cruciale pensioeninformatie effectief wordt teruggevonden. In de praktijk kan een model informatie ophalen uit fora, blogs of verouderde nieuwsartikels, terwijl officiële documentatie van bijvoorbeeld de Federale Pensioendienst niet noodzakelijk prioriteit krijgt (Barnett et al., 2024). Dat verhoogt het risico dat antwoorden gebaseerd zijn op informatie die onvolledig, verouderd of onvoldoende betrouwbaar is.

Het redeneervermogen van reasoning-modellen kent eveneens beperkingen. De redenering is algemeen ontworpen en niet afgestemd op Belgische pensioenregels. Er is bovendien weinig controle over

de redeneringsketen die het model doorloopt. Een model kan een vraag verkeerd opsplitsen, bijvoorbeeld door eerst algemene pensioenregels te bespreken terwijl de kern van de vraag net een specifieke uitzondering of statuut betreft. Het model redeneert dan coherent, maar vertrekt van een verkeerde interpretatie, wat uiteindelijk tot een fout antwoord kan leiden.

Performantie van generieke taalmodellen op pensioenvragen

In Figuur 1 rapporteren we de accuraatheid en de gemiddelde responstijd over alle pensioenvragen in de testset van zowel non-reasoning als reasoning-modellen. Daarbij beperken we ons tot modellen van OpenAI, Google en Anthropic. Concreet testen we voor de non-reasoning modellen Gemini 2.5 Flash en GPT-5.1 Instant en voor de reasoning-modellen Gemini 3 Pro, GPT-5.1 Thinking (high), Claude 4.5 Opus en Claude 4.5 Sonnet. De tests werden uitgevoerd via de API van de respectievelijke modellen op 28 november 2025, telkens met de op dat moment meest recente versies. Omdat responstijden kunnen variëren naargelang het testmoment en de belasting van de onderliggende infrastructuur, moeten de snelheidsresultaten als indicatief worden gelezen.



Figuur 1 *Prestaties van AI-modellen op pensioenvragen: accuraatheid en gemiddelde responstijd*

Notes: de analyse is uitgevoerd op 28 november 2025. GPT-5.1 Instant en GPT-5.1 Thinking (high) verwijzen naar chatgpt-5.1-chat-latest en gpt-5.1-2025-11-13 (met 'reasoning effort' op 'high') respectievelijk in de API van OpenAI.

De resultaten bevestigen dat reasoning-modellen beter presteren op pensioenvragen dan non-reasoning modellen. De non-reasoning modellen Gemini 2.5 Flash en GPT-5.1 Instant blijven duidelijk achter op accuraatheid. Onder de reasoning-modellen behaalt Gemini 3 Pro de hoogste accuraatheid, wat overeenkomt met het feit dat dit model toegang heeft tot de zoekmachine van Google. De keerzijde

is snelheid: hogere accuraatheid gaat doorgaans samen met langere responstijden. Vooral GPT-5.1 Thinking (high) en Gemini 3 Pro zijn merkbaar trager, wat erop wijst dat diepere reasoning en zoekstappen hun prijs hebben in verwerkingstijd. Langere responstijden verhogen niet alleen de kans dat gebruikers afhaken, maar ook de operationele kosten bij grootschalige inzet.

Bouwblokken voor een domeinspecifieke aanpak

De resultaten tonen dat moderne taalmodellen al relatief goed kunnen presteren op pensioenvragen, vooral wanneer we kijken naar de reasoning modellen. Toch blijven er belangrijke beperkingen bestaan. De accuraatheid is niet perfect, modellen beschikken over een knowledge cutoff en de gebruikte informatiebronnen zijn niet altijd transparant of controleerbaar. Voor sectoren waarin advies een juridische of financiële impact kan hebben, vormt dit een belangrijke belemmering. Zonder duidelijke controle over de gebruikte informatie en redeneringsstappen zijn de modellen in deze sectoren momenteel eerder een risico dan een meerwaarde.

In de literatuur zijn verschillende technieken ontwikkeld om taalmodellen domeinspecifieker en betrouwbaarder te maken. Via prompting kan het gedrag van een model worden gestuurd door expliciete instructies en voorbeelden mee te geven (Brown et al., 2020; Wei et al., 2022). Reasoning workflows splitsen het antwoordproces op in gecontroleerde tussenstappen die worden vastgelegd door de architect van het systeem, in tegenstelling tot reasoning-modellen die zelf bepalen welke stappen ze doorlopen (Zhou et al., 2022; Gao et al., 2023). Externe tools geven het model toegang tot zoekmachines, databases of rekenmodules voor taken waarvoor interne kennis niet volstaat (Qin et al., 2024). Retrieval-augmented generation (RAG) laat het model antwoorden op basis van expliciet opgezochte informatie uit een externe databank in plaats van op basis van interne kennis (Lewis et al., 2020; Gao et al., 2023). De kwaliteit van zo'n systeem hangt sterk af van de onderliggende databank: wanneer die verouderde of onbetrouwbare informatie bevat, zal het model die correct samenvatten maar toch een fout antwoord geven (Yoran et al., 2023; Edge et al., 2024). Bikon 1.0 combineert deze technieken in één geïntegreerde architectuur voor feitelijke pensioenvragen.

BIKON 1.0: EEN DOMEINSPECIFIEKE AI ASSISTENT VOOR FEITELIJKE PENSIOENVRAGEN

Algemene taalmodellen zijn doorgaans niet afgestemd op één specifiek domein, houden onvoldoende rekening met nationale institutionele context en werken vaak met weinig transparantie over welke bronnen precies worden gebruikt om antwoorden te genereren. Voor een domein dat sterk bepaald wordt door nationale regelgeving en institutionele details, kan dit de betrouwbaarheid en consistentie van antwoorden beperken.

Zo kan een domeinspecifieke aanpak betere resultaten opleveren. Wanneer een systeem uitsluitend steunt op gecontroleerde bronnen binnen één beleidsdomein en één institutionele context, zou de kwaliteit en precisie van de antwoorden moeten toenemen. Om deze hypothese te onderzoeken ontwikkelden we **Bikon 1.0**, een domeinspecifieke AI assistent voor feitelijke pensioenvragen in België.

Data

Veel van de zwaktes van direct werken met generieke taalmodellen zijn verbonden aan de data die ze beschikbaar hebben, en die onvoldoende specifiek is voor het generen van een antwoord. Het eerste bouwblok dat we gebruiken in onze architectuur is retrieval-augmented generation (RAG), waarbij een taalmodel antwoorden genereert op basis van expliciet opgezochte informatie uit een databank. De kwaliteit en kracht van dit systeem wordt daarom in sterke mate bepaald door deze onderliggende databank. Het opbouwen ervan verloopt in drie stappen: (1) selectie van relevante bronnen en teksten, (2) omzetting van ruwe tekst naar gestructureerde informatie en (3) het structureren en doorzoekbaar maken van de documenten op een slimme manier.

Bronnen en tekstselectie

De kern van de databank bestaat uitsluitend uit regulatoire en institutionele bronnen over het Belgische pensioenstelsel. Websites van publieke instellingen worden op regelmatige tijdstippen automatisch geraadpleegd, waaronder onder meer de Federale Pensioendienst en het Rijksinstituut voor de Sociale Verzekeringen der Zelfstandigen.

Daarnaast wordt elke bron systematisch geclassificeerd volgens drie dimensies. Ten eerste het type vragen dat ermee kan worden beantwoord. Ten tweede het inhoudelijke domein, bijvoorbeeld het wettelijk pensioen, aanvullende pensioenen of het zelfstandigenstatuut. Ten derde het maturiteitsniveau van de bron, waarbij een onderscheid wordt gemaakt tussen primaire institutionele informatie en secundaire, geïnterpreteerde bronnen.

Binnen elke bron wordt bovendien een verdere selectie gemaakt van de meest relevante pagina's en teksten. Op die manier wordt de databank gecontroleerd opgebouwd uit teksten die inhoudelijk relevant zijn voor pensioenvragen binnen de Belgische context.

Van ruwe tekst naar gestructureerde tekst

Na het ophalen doorloopt elke tekst een verwerkingspipeline die gericht is op het structureren van de tekst en verrijken van de informatie over de tekst (*metadata)*, en dit in de volgende stappen:

Eerst worden webpagina's verwerkt (*geparsed)* en genormaliseerd. De inhoud wordt omgezet naar een uniforme tekststructuur zodat documenten in een gelijkaardige vorm beschikbaar zijn voor verdere verwerking. De terminologie en inhoud van de bron blijven daarbij ongewijzigd.

Vervolgens wordt de tekst verrijkt met contextuele metadata, zoals doelgroep, toepassingsgebied, geldigheidsdatum en thematische labels. Daarna worden teksten opgesplitst in betekenisvolle fragmenten via zogenoemde context-aware chunking. Elk tekstfragment bevat expliciete contextinformatie, zoals bron, datum en toepassingsgebied, zodat het model de informatie binnen een duidelijk normatief kader kan interpreteren.

Ten slotte wordt versiebeheer toegepast: historische versies van teksten blijven beschikbaar, terwijl enkel de meest recente en geldige versie actief wordt gebruikt in de retrievallaag.

Empirisch blijkt dat de leesbaarheid en structurele helderheid van bronnen een directe impact hebben op de kwaliteit van de gegenereerde antwoorden. Bronnen die duidelijk gestructureerd en interpreteerbaar zijn, leveren doorgaans betere resultaten dan ruwe teksten.

Structureren en doorzoekbaar maken

De gestructureerde teksten worden vervolgens geïndexeerd zodat ze efficiënt kunnen worden opgezocht binnen het RAG-systeem. Hiervoor wordt een combinatie gebruikt van semantische embeddings, BM25-zoektechniek en metadatafiltering.

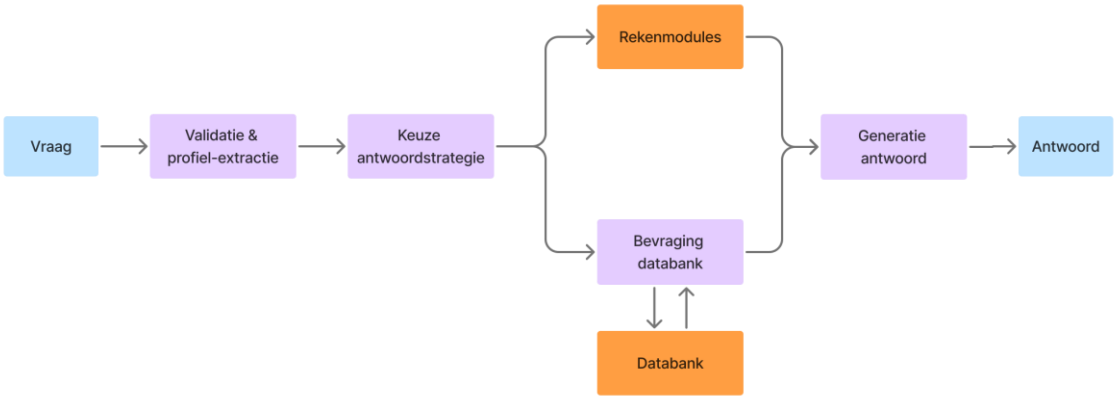
Semantische embeddings zetten teksten om in numerieke vectoren die de inhoudelijke betekenis van een tekst representeren. Hierdoor kan het systeem documenten terugvinden die inhoudelijk gelijkaardig zijn aan een vraag, ook wanneer andere woorden worden gebruikt.

BM25 is een klassieke zoekmethode uit de informatieretrieval die teksten rangschikt op basis van het voorkomen van specifieke zoektermen (Boudt et al., 2025). Deze methode is bijzonder effectief wanneer de vraag expliciete termen bevat die ook in de bron voorkomen.

Door beide methoden te combineren kan het systeem zowel semantisch gelijkaardige teksten vinden als documenten waarin de exacte terminologie voorkomt. Metadatafiltering maakt het daarnaast mogelijk om de zoekruimte verder te beperken tot relevante bronnen of domeinen. Hierdoor kan het systeem bij elke vraag snel de meest relevante informatie uit de databank ophalen en gebruiken voor het genereren van een antwoord.

Architectuur

Voortbouwend op de gestructureerde en doorzoekbare databank, lichten we hiertoe hoe het systeem inkomende vragen verwerkt tot een antwoord. Deze architectuur is specifiek ontwikkeld om de afweging tussen operationele kosten (zoals rekenkracht en tokengebruik), antwoordsnelheid en betrouwbaarheid te optimaliseren. Het antwoordsysteem volgt geen vaste, lineaire pijplijn, maar is ontworpen als een dynamische beslissingsstructuur. Dit betekent dat het systeem per vraag autonoom kiest tussen snelle, lichte modellen en zwaardere, meer contextuele analyses met meerdere modellen.



Figuur 2 *Architectuur en stappen van Bikon 1.0 antwoordsysteem*

Dit proces verloopt in zes samenhangende stappen, geschetst in *Figuur 2*:

1. **Validatie en profiel-extractie.** Elke vraag doorloopt eerst een initiële controlelaag. Hierbij valideert het systeem of de vraag binnen het domein van het pensioenstelsel valt, en of het mag antwoorden via *guard rails*. Tegelijkertijd extraheert het systeem specifieke gebruikerskenmerken uit de tekst, zoals leeftijd, beroepsstatuut, loopbaanfase en pensioenpijler via *tools*. Door deze informatie gestructureerd op te slaan, kan het systeem de zoekruimte in de databank direct inperken tot uitsluitend de relevante bronnen.

2. **Keuze van de antwoordstrategie.** Afhankelijk van de geanalyseerde complexiteit van de vraag, kiest het systeem de meest efficiënte oplossingsstrategie. Wanneer de vraag eenvoudig is en alle parameters beschikbaar zijn, kiest het voor een directe beantwoording of een gerichte herformulering. Bij complexe vragen schakelt het systeem een gespecialiseerd model in dat de hoofdvraag opsplitst in behapbare, doelgerichte deelvragen waarover dan specifieke teksten gevonden kunnen worden. Daarnaast kijkt het systeem ook nog welke berekeningen moeten gebeuren. We doen dit nog hoofdzakelijk door gebruik te maken van verschillende *prompting* technieken zoals *few-shot prompting*.

3. **Informatie ophalen en herwaarderen (retrieval en reranking).** Om de juiste teksten aan het taalmodel mee te geven, gebruikt de architectuur een tweetrapsmethode. Eerst selecteert het systeem teksten en specifieke tekstfragmenten in de gemaakte databank. Omdat deze initiële zoekopdracht soms net niet precies genoeg is, volgt een kwalitatieve correctie (*reranking*). Hierbij beoordeelt een compact, efficiënt model de gevonden kandidaten opnieuw om de semantische precisie te verhogen en de meest relevante fragmenten bovenaan te plaatsen. Als we meer tekst meegeven hebben we ook het risico dat de taalmodellen meer vergeten wat er exact meegegeven is.

4. **Deterministische rekenmodules** Omdat taalmodellen van nature minder geschikt zijn voor wiskundige bewerkingen, bevat het systeem specifieke, voorgeprogrammeerde rekenmodules. Voor parameters die wiskundig vastliggen zoals de wettelijke pensioenleeftijd voert het systeem exacte berekeningen uit in plaats van deze louter tekstueel te interpreteren. Deze harde numerieke data worden vervolgens aan de context toegevoegd. Dit verhoogt de consistentie van de output aanzienlijk, zeker wanneer er gebruik wordt gemaakt van kleinere en snellere taalmodellen.

5. **Generatie** In de laatste fase combineert het systeem de geselecteerde tekstfragmenten en de uitgevoerde berekeningen om een gestructureerd antwoord te genereren. Een cruciaal principe hierbij is dat het model niet vrij mag redeneren over het Belgische pensioenstelsel; het construeert antwoorden uitsluitend binnen de expliciet afgebakende grenzen van de aangeleverde teksten en data. Deze strikte benadering reduceert de kans op fictieve antwoorden (*hallucinaties*) en garandeert een hoge mate van reproduceerbaarheid.

RESULTATEN

We hebben al gezien dat sterke generieke taalmodellen pensioenvragen behoorlijk goed kunnen beantwoorden, maar dat dit vaak gepaard gaat met een duidelijke trade-off tussen accuraatheid en snelheid. Binnen die groep kwam Claude 4.5 Opus naar voren als het meest gebalanceerde model. De relevante vraag is dan hoe een domeinspecifieke oplossing zoals Bikon 1.0 presteert op zowel herkenbare publieksvragen, zoals *"Wat is het minimumpensioen voor een alleenstaande?"*, als meer technische expertvragen, zoals *"Wordt het wettelijk pensioen van universiteitsprofessoren op dezelfde manier berekend voor een hoogleraar als voor een docent?"*.

Zoals Tabel 1 laat zien, behaalt Bikon 1.0 een accuraatheid van 95% over alle vragen heen, tegenover 92% voor het best presterende generieke model. Op algemene vragen loopt dat op tot 100%, waar het beste generieke model 96% haalt. Het verschil blijft ook zichtbaar bij de expertvragen, waar Bikon 1.0 92% behaalt, tegenover 88% voor het sterkste generieke model.

Tabel 1 *Accuraatheid van Bikon 1.0 en generieke taalmodellen op pensioenvragen*

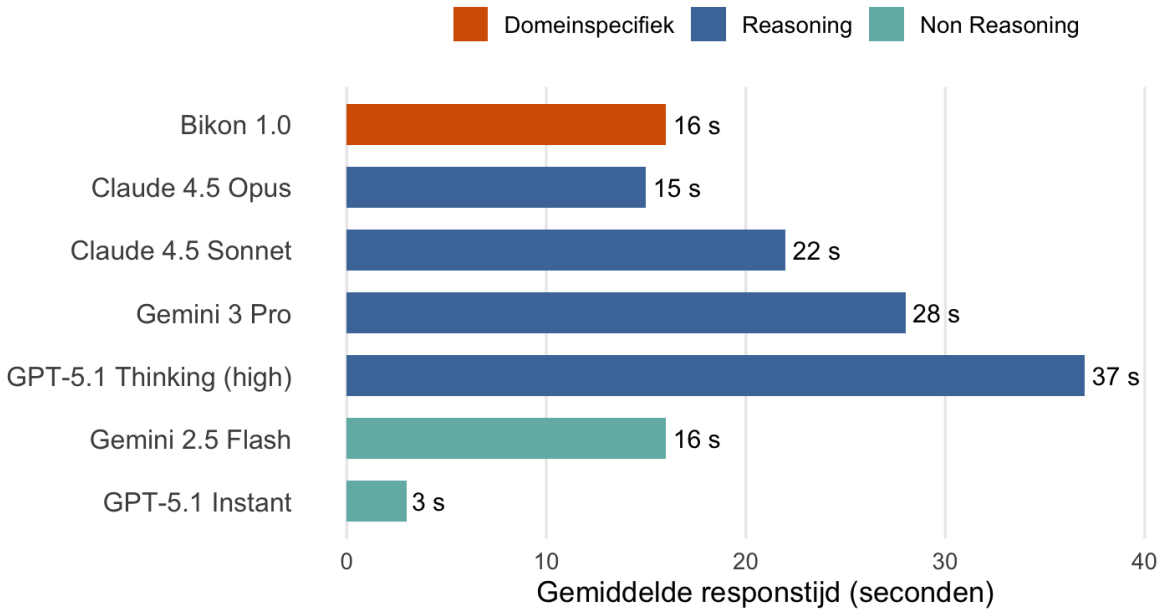
Model	Alle vragen	Algemene vragen	Expert vragen
Bikon 1.0	**96%**	**100%**	**92%**
Reasoning			
Gemini 3 Pro	92%	95%	88%
GPT-5.1 Thinking (high)	90%	93%	87%
Claude 4.5 Opus	89%	91%	87%
Claude 4.5 Sonnet	84%	91%	76%
Non Reasoning			
Gemini 2.5 Flash	79%	88%	69%
GPT-5.1 Instant	76%	84%	68%

Notes: de analyse is uitgevoerd op 28 november 2025. GPT-5.1 Instant en GPT-5.1 Thinking (high) verwijzen naar chatgpt-5.1-chat-latest en gpt-5.1-2025-11-13 (met 'reasoning effort' op 'high') respectievelijk in de API van OpenAI.

Bikon 1.0 presteert dus niet alleen sterk op toegankelijke, publieksgerichte vragen, maar ook op de technischere vragen waar statuten, uitzonderingsregels of berekeningswijzen doorslaggevend zijn. Tegelijk toont Figuur 3 dat de gemiddelde responstijd met 16 seconden praktisch bruikbaar blijft en dicht aansluit bij modellen zoals Claude 4.5 Opus (15 seconden), terwijl sommige andere sterke modellen duidelijk trager zijn, zoals Gemini 3 Pro (28 seconden) en GPT-5.1 Thinking (37 seconden).

Dat suggereert dat de meerwaarde van een domeinspecifieke pensioenassistent zoals Bikon 1.0 niet alleen zit in extra inhoudelijke juistheid, maar ook in het vermogen om die kwaliteit te leveren zonder de sterke snelheidskost die sommige generieke topmodellen kennen. Waar generieke modellen vaak

een keuze lijken af te dwingen tussen snelheid en accuraatheid, schuift een domeinspecifieke AI-oplossing die grens duidelijk op.



Figuur 3 *Gemiddelde responstijd van Bikon 1.0 en generieke taalmodellen op pensioenvragen*

CONCLUSIE

De centrale vraag van dit Inzicht luidt: kunnen we artificiële intelligentie vertrouwen om pensioenvragen te beantwoorden? De resultaten van onze benchmark suggereren dat het antwoord genuanceerd is. Generieke AI-modellen kunnen een groot deel van de pensioenvragen correct beantwoorden, maar hun betrouwbaarheid blijft beperkt zodra het antwoord afhangt van uitzonderingsregels, fiscale nuances of precieze parameters uit de regelgeving. In zulke gevallen blijken fouten systematisch voor te komen.

In onze testset van meer dan 170 meerkeuzevragen over het Belgische pensioenstelsel behaalt het best presterende generieke model een accuraatheid van 92 procent. Wanneer we focussen op de technischere expertvragen, daalt die nauwkeurigheid tot 68 à 88 procent, afhankelijk van het model. Dat verschil is relevant, omdat net deze vragen vaak bepalend zijn in adviescontexten waar kleine interpretatieverschillen grote gevolgen kunnen hebben.

Een domeinspecifieke aanpak behaalt daarentegen 96 procent accuraatheid over alle vragen en 92 procent op expertvragen, en doet dat bovendien met een gemiddelde responstijd van 16 seconden, tegenover 28 seconden voor het best presterende generieke model. De resultaten suggereren dus dat AI wel degelijk betrouwbaar kan worden ingezet voor pensioenvragen, maar dat de betrouwbaarheid sterk afhangt van de architectuur waarin het taalmodel wordt ingebed.

Het verschil wordt niet uitsluitend verklaard door het taalmodel zelf. Generieke modellen, ook wanneer ze agentic werken en via een zoekmachine actuele informatie ophalen, doorzoeken het open internet zonder controle over bronselectie, actualiteit of domeinrelevantie. De domeinspecifieke aanpak verschilt op drie punten. Ten eerste werkt ze met een afgebakende kennisbasis waarvan de bronnen expliciet geselecteerd en systematisch geactualiseerd worden. Ten tweede gebruikt ze een retrievalsysteem dat de zoekruimte beperkt tot relevante en gevalideerde documenten. Ten derde bevat ze deterministische rekenmodules voor numerieke parameters zoals pensioenleeftijd en loopbaanvoorwaarden. Het taalmodel zelf blijft in essentie generiek; het is de informatieomgeving waarin het opereert die het verschil verklaart.

Die vaststelling heeft implicaties die verder reiken dan pensioenadvies. In elk gereguleerd kennisdomein waar de foutkosten hoog zijn, de regelgeving frequent evolueert en het antwoord afhangt van specifieke gebruikersparameters, botsen generieke modellen op dezelfde structurele beperkingen. Fiscaal advies, waar jaarlijkse parameterverschuivingen en samenloopregels het verschil maken tussen een correcte en een kostelijke aangifte. Arbeidsrecht, waar de interactie tussen cao's, sectorale afspraken en individuele contractvoorwaarden een precisie vereist die brede trainingsdata niet kan garanderen. Socialezekerheidsrecht, waar de overgang tussen statuten gepaard gaat met voorwaarden die per situatie verschillen. De hypothese die uit dit onderzoek volgt, is dat een gecureerde, domeinspecifieke aanpak de nauwkeurigheidskloof ook in deze domeinen kan dichten. Die hypothese vergt echter afzonderlijke empirische toetsing.

Dit onderzoek kent een aantal beperkingen. De evaluatie is gebaseerd op meerkeuzevragen met één correct antwoord, een intussen industriestandaard die eenduidige scoring mogelijk maakt, maar de complexiteit van een adviesgesprek met open vragen, bijvragen en contextafhankelijke nuance niet volledig weerspiegelt. Bovendien is de benchmark een momentopname: zowel generieke modellen als de domeinspecifieke aanpak evolueren, waardoor de prestatieverhoudingen over tijd kunnen verschuiven. Ten slotte maakt dit onderzoek de informatie-input en de brontraceerbaarheid van het antwoord transparant, maar niet de tussenstappen die het taalmodel doorloopt om van bronmateriaal tot een geformuleerd antwoord te komen, het zogenaamde inferentieproces. Die resterende ondoorzichtigheid is een open onderzoeksvraag, zowel voor domeinspecifieke als voor generieke toepassingen.

De auteurs van dit Inzicht zijn betrokken bij de ontwikkeling van de domeinspecifieke aanpak die in dit onderzoek wordt geëvalueerd. Die betrokkenheid heeft de opzet mogelijk gemaakt: toegang tot de architectuur, de kennisbasis en de iteratieve ontwikkeling. De objectiviteit van de evaluatie is geborgd doordat de benchmarkvragen zijn opgesteld en gevalideerd door onafhankelijke pensioenexperten, en de generieke modellen zijn geëvalueerd via hun publiek beschikbare interfaces.

REFERENTIES

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* (pp. 194-199).

Boudt, K., Delmarcelle, O., & Ringoot, P. (2025). A news monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions. *Risk Sciences*, *1*, 100018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

European Banking Authority. (2024). *Special topic – Artificial intelligence*. In *Risk assessment report – November 2024*. https://www.eba.europa.eu/publications-and-media/publications/special-topic-artificial-intelligence

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1), 32.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., ... & Dou, Z. (2025, November). Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 5420-5438).

Maity, S., & Saikia, M. J. (2025). Large language models in healthcare and medical applications: a review. *Bioengineering, 12*(6), 631.

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903.*

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., ... & Sun, M. (2024). Tool learning with foundation models. *ACM Computing Surveys*, *57*(4), 1-40.

Siino, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access, 13*, 18253-18276.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824-24837.

Yoran, O., Wolfson, T., Ram, O., & Berant, J. (2023). Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., ... & Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625.*

Prof. dr. Kris Boudt is professor finance en econometrie aan de vakgroep Economie van de Universiteit Gent en Vrije Universiteit Brussel. Zijn onderzoek focust op de ontwikkeling van data-gedreven methodes voor het beantwoorden van financieel-economische vragen. Contact: Kris.Boudt@UGent.be.

Arno De Block is doctoraatsonderzoeker aan de vakgroep Economie van de Universiteit Gent en Vrije Universiteit Brussel. Zijn onderzoek focust op nowcasting en economische tekstanalyse. Contacteren kan via Arno.DeBlock@UGent.be of arno@bikon.ai.

Feliciaan De Palmenaer is doctoraatsonderzoeker aan de vakgroep Economie van de Universiteit Gent en Vrije Universiteit Brussel. Zijn onderzoek focust op big data en economie. Contacteren kan via Feliciaan.DePalmenaer@UGent.be of feliciaan@bikon.ai.

Andries Fluit is strategisch communicatieconsultant bij akkanto en wetenschappelijk medewerker aan de vakgroep Economie van de Universiteit Gent, waar hij betrokken is bij de ontwikkeling van Bikon als business developer. Contacteren kan via andries@bikon.ai.

Yanick Inghels is doctoraatsonderzoeker aan de vakgroep Economie van de Universiteit Gent. Zijn onderzoek focust op grootschalige tekstanalyse van Europese onderzoeksfinanciering. Contacteren kan via Yanick.Inghels@UGent.be of yanick@bikon.ai.

Gentse Economische Inzichten vormen een forum voor toegankelijk gecommuniceerde onderzoeksbevindingen en beleidsaanbevelingen door vorsers van de Universiteit Gent. De Inzichten vertolken alleen de visie van de auteur(s). Zij kunnen niet doorgaan als de visie van een onderzoeksgroep, de Vakgroep Economie of Universiteit Gent.

Voor meer onderzoek van de Vakgroep Economie verwijzen we naar https://www.ugent.be/eb/economics/en.

Gentse Economische Inzichten wordt mede mogelijk gemaakt door een gift van www.sustinvest.eu.
SUSTINVEST heeft geen inspraak over inhoud en beleidsaanbevelingen.

SUSTINVEST