Mens versus machine

Hoe (verschillend) presteren bedrijfsjuristen en AI bij het oplossen van een juridische casus?

*Mr. dr. K.L. Maes, prof. dr. S.A. Kruisinga en mr. D.F. Groenevelt**

1 Inleiding: de belofte van AI in de juridische praktijk

Kunstmatige intelligentie (ofwel artificiële intelligentie; hierna: AI) heeft de afgelopen jaren een stormachtige ontwikkeling doorgemaakt. Waar AI-systemen voorheen vooral goed waren in regelgebaseerde taken, kunnen zogeheten *Large Language Models* (hierna: taalmodellen[1]) nu ook overweg met andere taken, waaronder het genereren en begrijpen van complexe (juridische) teksten. AI-technologie wordt dan ook wereldwijd steeds meer toegepast in de juridische sector, van documentanalyse tot jurisprudentieonderzoek.[2] Een gestaag groeiende hoeveelheid onderzoeksresultaten toont aan dat het gebruik van AI tot aanzienlijke efficiëntiewinst en ook kwaliteitsverbetering kan leiden in de juridische praktijk.[3] Toch zijn er ook zorgen, bijvoorbeeld over de betrouwbaarheid van door AI gegenereerde juridische analyses.[4] Bovendien lijkt er sprake te zijn van een gebrek aan kennis: kennis over de mogelijkheden van AI én over de juiste manier om de applicaties te instrueren en gebruiken.[5]

In deze bijdrage worden de resultaten gedeeld van een vergelijkend onderzoek naar de wijze waarop een groep bedrijfsjuristen een casus oplost en hoe verschillende AI-modellen dezelfde casus oplossen. Met dit verslag van een bescheiden, maar

* Mr. dr. K.L. Maes is advocaat bij Van Benthem & Keulen te Utrecht en als onderzoeker en universitair docent verbonden aan de Universiteit Utrecht. Prof. dr. S.A. Kruisinga is als hoogleraar overeenkomstenrecht verbonden aan de Open Universiteit te Heerlen en als Professional Support Lawyer aan Van Benthem & Keulen te Utrecht. Mr. D.F. Groenevelt is voormalig Head of Legal bij ASML en oprichter & CEO van Viridea.ai.

1 Een taalmodel is, kort gezegd, een type kunstmatige intelligentie dat is getraind op grote hoeveelheden tekst om patronen in taal te herkennen en te reproduceren. Het voorspelt welk woord of welke zin waarschijnlijk volgt op basis van de context.

2 American Bar Association, 2024 Legal Technology Survey Report, www.americanbar.org/products/ecd/single/449382161/. Een samenvatting daarvan is te raadplegen via www.lawnext.com/2025/03/aba-tech-survey-finds-growing-adoption-of-ai-in-legal-practice-with-efficiency-gains-as-primary-driver.html.

3 Zo volgt uit het evaluatierapport van de Europese Commissie voor de Efficiëntie van Justitie (p. 161, te raadplegen via www.coe.int/en/web/cepej/special-file) onder meer dat AI kan helpen bij het voorspellen van de (mogelijke) uitkomst van zaken, dat *natural language processing* behulpzaam kan zijn bij juridisch onderzoek ter voorbereiding van een zaak door het samenvatten van documenten en het destilleren van belangrijke informatie, dat generatieve AI kan worden gebruikt om juridische informatie toegankelijk te maken, en dat taalmodellen behulpzaam kunnen zijn bij het juridisch onderzoek dat nodig is voor de ontwikkelingen van de strategie in een bepaalde zaak. Vgl. ook D. Schwartz e.a., AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice (Minnesota Legal Studies Research Paper No. 25-16), 4 maart 2025, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111#, en R.J. Couture, The Impact of Artificial Intelligence on Law Firms' Business Models, Harvard Law School – Center on the Legal Profession, 25 februari 2025, https://clp.law.harvard.edu/knowledge-hub/insights/the-impact-of-artificial-intelligence-on-law-law-firms-business-models.

4 Zie bijv. L. Grutters & K. Haex, Weeg gebruik van AI in de rechtszaal op een goudschaaltje, AA 2025, afl. 2, p. 91.

5 Een grote enquête (onder meer dan 450 respondenten uit alle EU-lidstaten) van de European Legal Tech Association (ELTA) toont aan dat kennis en het vermogen om Gen AI op betrouwbare wijze in te zetten de grootste blokkade vormt voor het gebruik hiervan, zie Generative-AI-Global-Report-2024-ELTA-1 – hier te downloaden: https://elta.org/generative-ai-global-report-2024/. Vgl. ook C.V. Chien & M. Kim, Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap (UC Berkeley Public Law Research Paper), 11 april 2024; een enquête van de Universiteit van Berkeley uit 2024 onder 202 juridisch dienstverleners waaruit naar voren komt dat 90% van de deelnemers geen vertrouwen had in hun vermogen om generatieve AI effectief te gebruiken, zie https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4733061. Overigens wordt in de rechtspraak, zij het zeer voorzichtig, ook geëxperimenteerd met AI, vgl. www.advocatie.nl/nieuws/rechtbank-rotterdam-doet-proef-met-ai-als-schrijfhulp-in-strafvonnis/?utm_source=&utm_medium=email&utm_campaign=0_Advocatie_31032025.

inzichtelijk praktijkonderzoek beogen wij een bijdrage te leveren aan een beter begrip van de huidige mogelijkheden en beperkingen van AI-technologie in een professionele Nederlandse juridische context.

De opbouw van deze bijdrage is als volgt. Allereerst lichten wij de onderzoeksvraag en de door ons gehanteerde onderzoeksmethodiek toe (par. 2). Aansluitend presenteren we de onderzoeksresultaten (par. 3), waaraan we vervolgens een nadere duiding geven (par. 4). We sluiten af met een conclusie en enkele aanbevelingen (par. 5).

In deze bijdrage gaan we niet in op de hiervoor benoemde gevaren en risico's van het gebruik van AI-systemen. Wel zij hier opgemerkt dat het onderzoek als een 'laagrisicogebruik' van AI-systemen kan worden aangemerkt. Er zijn noch persoonsgegevens, noch (bedrijfs)vertrouwelijke gegevens in de AI-systemen ingevoerd, enkel de casus, de vragen en relevante rechtspraak. Terzijde signaleren wij dat we bij het opstellen van deze bijdrage ook op andere wijze AI hebben gebruikt, te weten als sparringpartner bij het analyseren van de onderzoeksresultaten, bij het visualiseren van resultaten in tabellen en als schrijfhulp (met name voor het zoeken van synoniemen voor bepaalde bewoordingen).[6]

2 Onderzoeksbeschrijving en -methode

2.1 Onderzoeksvraag

In ons onderzoek staat de kernvraag centraal: in hoeverre is AI in staat een juridische casus op te lossen op een niveau dat vergelijkbaar is met dat van menselijke juridische professionals? Daarbij worden de volgende deelvragen beantwoord:

1. Wat is de invloed van verschillende promptstrategieën[7] op de kwaliteit van de uitkomsten bij het oplossen van een juridische casus door AI?
2. Wat is het effect van het aanbieden van additionele juridische informatie zoals jurisprudentie aan de AI-systemen (dus in aanvulling op de informatie die reeds in het taalmodel 'besloten' ligt op basis van hun *pre-training*[8])?
3. Wat zijn de verschillen in prestaties tussen de diverse, voor de Nederlandse jurist eenvoudig toegankelijke (grote) AI-systemen?

Ook hebben we onderzocht in hoeverre de analysemethodologie (de redeneringssystematiek) van AI-systemen verschilt met die van (menselijke) juristen.

2.2 Onderzoeksmethodiek: selectie van casus, bedrijfsjuristen en AI-modellen

Om een zo objectief mogelijke vergelijking mogelijk te maken, is bij de beantwoording van de voornoemde onderzoeksvragen gekozen voor de volgende onderzoeksopzet. Allereerst is er gebruik gemaakt van reeds bestaande juridische casus met bijbehorende vragen uit het curriculum van de Stichting Beroepsopleiding Bedrijfsjuristen (SBB) over het algemeen verbintenissenrecht. Denk hierbij aan thema's als verzuim en ontbinding. De bedrijfsjuristen van de twee meest recente 'jaarlagen' van de SBB zijn benaderd met de vraag of de (reeds in het kader van hun opleiding gegeven) antwoorden op de bij deze casus behorende vragen voor dit onderzoek gebruikt mochten worden. 26 bedrijfsjuristen stemden daarmee in. De antwoorden vormen een goede doorsnede van de manier waarop juridische professionals een casus analyseren. Daarbij is het van belang om op te merken dat de bedrijfsjuristen die aan deze opleiding deelnemen meerdere casus moeten oplossen. Voor ons onderzoek selecteerden we twee lastige casus, dat wil zeggen casus die door de bedrijfsjuristen als lastig werden ervaren en waarop dus gemiddeld geen hoog cijfer behaald werd. Een dergelijke casus gaf ons de mogelijkheid om de AI-systemen stevig 'aan de tand te voelen', en tevens te bezien of de AI-systemen wellicht in staat zouden zijn hogere cijfers te halen dan de bedrijfsjuristen.

Vervolgens selecteerden wij een aantal van de grootste en meest geavanceerde taalmodellen, voor zover deze modellen momenteel beschikbaar zijn voor de gemiddelde Nederlandse jurist (dat wil zeggen middels een eenvoudig via het internet te gebruiken chatbot, en tegen op zijn hoogst een tarief van ongeveer € 20 per maand). Het gaat hier om de volgende modellen:

1. *OpenAI GPT-4o* en *OpenAI o1*: deze modellen vertegenwoordigen ten tijde van het schrijven van deze bijdrage de meest recente taalmodellen van OpenAI, bekend van de applicatie ChatGPT. Het o1-model is een zogenaamd 'redeneringsmodel' (*reasoning model*);
2. *Claude 3.5 Sonnet Pro*: ontwikkeld door het Amerikaanse bedrijf Anthropic. Op ChatGPT na is Claude momenteel de meest populaire AI-chatbot, bekend om haar goede scores op benchmarks ten aanzien van 'creatief schrijven';
3. *Gemini 1.5 Pro*: Googles geavanceerde taalmodel, dat langzaam wordt ingevoerd als AI-motor achter talloze standaard services van Google;
4. *DeepSeek*: een in China ontwikkeld model dat recent de markt heeft bestormd en bekendstaat om zijn diepgaande 'redenerings'-capaciteiten;
5. *Le Chat* (Mistral): een chatbot gebaseerd op het in Frankrijk ontwikkelde taalmodel Mistral, een van de weinige taalmodellen van Europese bodem die zich kunnen meten met de Amerikaanse taalmodellen.

6 De hierbij gebruikte tool is Claude Sonnet Pro 3.5 van Anthropic https://claude.ai. Wij hebben ons voor het overige gehouden aan de 'Gebruiksrichtlijnen voor het inzetten van AI bij auteurswerkzaamheden' (2024) van Boom.

7 Prompten is het geven van een invoer (prompt) aan een taalmodel. De prompt fungeert als instructie of vraag en stuurt het gedrag van het model aan, waardoor het model relevante en contextuele antwoorden produceert.

8 *Pre-training* is het initiële leerproces waarbij een taalmodel wordt getraind op een grote, algemene dataset (corpus) van tekst, met als doel het leren van statistische patronen, grammatica en semantiek in deze dataset. Deze fase legt de basis voor de kennis van het model voordat het wordt bijgesteld voor specifieke toepassingen (*fine-tuning*).

 doi: 10.5553/MvV/157457672025035004001

We hebben deze taalmodellen vervolgens geïnstrueerd ofwel 'geprompt'[9] om de bij de casus behorende vragen te beantwoorden. Daarbij hebben we verschillende promptstrategieën gehanteerd om te onderzoeken hoe de prompts, en het toevoegen van bepaalde additionele informatie (context), de kwaliteit van de antwoorden van de taalmodellen beïnvloeden. Zie daarover nader paragraaf 4.

Het prompten werd in dit onderzoek uitgevoerd door mr. Douwe Groenevelt, ten tijde van het schrijven van deze bijdrage Head of Legal bij ASML, tevens een van de auteurs van deze bijdrage, die over ruime ervaring beschikt ten aanzien van het inzetten van AI in de juridische praktijk.

2.3 Beoordeling van de (geanonimiseerde) antwoorden door experts

De door de bedrijfsjuristen gegeven antwoorden zijn daarna als volgt met de door AI gegenereerde antwoorden vergeleken. Prof. dr. Sonja Kruisinga (hoogleraar overeenkomstenrecht aan de Open Universiteit en Professional Support Lawyer bij Van Benthem & Keulen) en mr. dr. Kirsten Maes (advocaat bij Van Benthem & Keulen en docent en onderzoeker aan de Universiteit Utrecht) – tevens twee van de auteurs van deze bijdrage – hebben beiden, afzonderlijk van elkaar, alle antwoorden nagekeken en met een cijfer beoordeeld. Daarbij was voor hen niet bekend van wie of wat de na te kijken antwoorden afkomstig waren. Vervolgens hebben zij de beoordeling met elkaar besproken en daar waar die uiteenliep, hebben zij gezamenlijk besproken wat de beoordeling zou moeten zijn, zodat uiteindelijk de beoordeling op hetzelfde cijfer uitkwam.

Voor een consistente en objectieve beoordeling van alle antwoorden werd een vooraf opgesteld en gestandaardiseerd beoordelingskader gebruikt. Dit kader was gebaseerd op de criteria die in het juridisch onderwijs worden gehanteerd voor het beoordelen van casusanalyses. Per casus konden maximaal tien punten worden behaald. Het beoordelingskader omvatte onder meer de volgende elementen:

1. identificatie en toepassing van relevante wetsartikelen;
2. juiste verwijzing naar en interpretatie van relevante jurisprudentie;
3. volledigheid van de juridische analyse;
4. logische structuur en onderbouwing van de redenering.

Daarbij was voor de beoordelaars in de eerste fase van het onderzoek niet zichtbaar welke antwoorden waren vervaardigd door de AI-systemen en welke door de (menselijke) bedrijfsjuristen. Om dat te bereiken zijn de antwoorden van de taalmodellen in dezelfde '*formatting*' gegoten als de antwoorden van de bedrijfsjuristen en bovendien in willekeurige volgorde tussen deze uitwerkingen gevoegd. In totaal werden 44 uitgewerkte antwoorden beoordeeld in de eerste fase: 26 van bedrijfsjuristen en 18 van AI-modellen (de antwoorden van zes verschillende modellen die op drie verschillende manieren waren geprompt). In de tweede fase werden aanvullende tests uitgevoerd met geoptimaliseerde prompts om de maximale potentie van de AI-modellen te verkennen, zoals hierna nader wordt toegelicht.

3 De belangrijkste onderzoeksresultaten

De belangrijkste resultaten (van beide fases van het onderzoek) worden in tabel 1, 2 en 3 gepresenteerd en in paragraaf 4 van deze bijdrage nader geduid en toegelicht.

9 Zie noot 7.

Tabel 1 De hoogst behaalde cijfers uitgelicht

Beste Cijfers*

Type	Casus 1	Casus 2
Bedrijfsjuristen	**8,0**	**10,0**
AI Modellen		
↳*One-shot*	*4,0*	*7,0*
↳*Met (initiële) prompt, zonder context*	*8,0*	*7,0*
↳*Met (initiële) prompt, met context*	*9,0*	*8,0*
↳*Verbeterde prompt, met context*	*10,0*	*10,0*

* Dus: de hoogste cijfers voor casus 1 en voor casus 2 die zijn behaald door (a) de (voor die casus) best scorende bedrijfsjurist en (b) het (voor die casus) best scorende AI-model.

Tabel 2 De gemiddeld behaalde cijfers uitgelicht

Gemiddelde Cijfers*

Type	Casus 1	Casus 2	Gemiddeld
Bedrijfsjuristen	**4,8**	**5,8**	**5,3**
AI Modellen			
↳*One-shot*	*2,3*	*4,5*	*3,4*
↳*Met (initiële) prompt, zonder context*	*3,7*	*4,2*	*3,9*
↳*Met (initiële) prompt, met context*	*7,7*	*4,5*	*6,1*
↳*Verbeterde prompt, met context*	*9,4*	*5,7*	*7,5*

* Dus: de gemiddelde cijfers voor casus 1, voor casus 2 en over beide casus gemiddeld van (a) de 26 bedrijfsjuristen en (b) de zes verschillende geteste AI-modellen (op vier verschillende manieren geprompt; zie verder tabel 3).

Tabel 3 De verschillende modellen, getest met de verschillende prompttechnieken[10]

	Model	Prompttechniek	Cijfer casus 1	Cijfer casus 2	Gemiddeld cijfer
1.	OpenAI o1	One-shot, *geen* context	4	4	4
2.	OpenAI o1	Met initiële prompt, *geen* context	3	5	4
3.	OpenAI o1	Met initiële prompt, *met* context	5	4	4,5

10 Alleen het model OpenAI o1 hebben wij i.v.m. een technisch probleem op het moment van testen niet getest met de verbeterde prompt.

 doi: 10.5553/MvV/157457672025035004001

Tabel 3 (Vervolg)

	Model	Prompttechniek	Cijfer casus 1	Cijfer casus 2	Gemiddeld cijfer
4.	OpenAI GPT-4o	One-shot, *geen* context	3	5	4
5.	OpenAI GPT-4o	Met initiële prompt, *geen* context	4	4	4
6.	OpenAI GPT-4o	Met initiële prompt, *met* context	8	5	6,5
7.	OpenAI GPT-4o	Verbeterde prompt, *met* context	9	2	5,5
8.	Claude 3.5 Sonnet Pro	One-shot, *geen* context	2	5	3,5
9.	Claude 3.5 Sonnet Pro	Met initiële prompt, *geen* context	4	5	4,5
10.	Claude 3.5 Sonnet Pro	Met initiële prompt, *met* context	8	4	6
11.	Claude 3.5 Sonnet Pro	Verbeterde prompt, *met* context	10	10	10
12.	Gemini 1.5 Pro	One-shot, *geen* context	2	6	4
13.	Gemini 1.5 Pro	Met initiële prompt, *geen* context	3	4	3,5
14.	Gemini 1.5 Pro	Met initiële prompt, *met* context	8	4	6
15.	Gemini 1.5 Pro	Verbeterde prompt, *met* context	9	7,5	7,75
16.	DeepSeek R1	One-shot, *geen* context	3	7	5
17.	DeepSeek R1	Met initiële prompt, *geen* context	8	7	7,5
18.	DeepSeek R1	Met initiële prompt, *met* context	9	8	8,5
19.	DeepSeek R1	Verbeterde prompt, *met* context	10	6	8
20.	Mistral	One-shot, *geen* context	0	0	0
21.	Mistral	Met initiële prompt, *geen* context	0	0	0
22.	Mistral	Met initiële prompt, *met* context	8	2	5
23.	Mistral	Verbeterde prompt, *met* context	9	3	6

4 Duiding belangrijkste onderzoeksresultaten

4.1 Verschillende promptstrategieën

Een van de meest significante bevindingen betrof de cruciale impact van verschillende promptstrategieën op de kwaliteit van de AI-antwoorden. De resultaten tonen een duidelijke progressie in prestaties naarmate de prompts verbeterden.[11] Dat laat zich als volgt toelichten.

We testten vier verschillende promptstrategieën:

1. *one-shot*: het model kreeg als 'input' alleen de casus en de bijbehorende vragen, zonder verdere instructies of context. Deze manier van prompten wordt ook wel '*one shot*' of '*zero shot*' genoemd, omdat het een eerste, snelle 'voor de vuist weg'-poging betreft;
2. *verbeterde prompt, maar zonder context*: het model kreeg meer specifieke instructies over hoe de juridische analyse moest worden uitgevoerd, maar kreeg geen aanvullende juridische informatie zoals met name (mogelijk) relevante jurisprudentie. De prompt die hierbij is gebruikt, is onder 'Bevinding 2' in paragraaf 4.2 integraal opgenomen;
3. *verbeterde prompt, met context*: het model kreeg naast de meer specifieke instructies ten aanzien van de analyse eveneens aanvullende relevante jurisprudentie (meer concreet: een viertal arresten dat ook in de syllabus van de bedrijfsjuristen was opgenomen);
4. *verder geoptimaliseerde prompt, met context*: op basis van onze eerste bevindingen na het doorlopen van voormelde eerste drie promptstrategieën hebben we onze prompt aangepast om te corrigeren voor geconstateerde lacunes in de redeneringsmethodiek van de AI-modellen tegen de context van het beoordelingskader. Deze verbeterde

11 Voor de duidelijkheid zij opgemerkt dat de *volgorde* van het toepassen van deze strategieën geen invloed heeft gehad op de individuele uitkomsten. De AI-systemen 'leerden' niet van voorgaande stappen (wij hebben in dat verband elke strategie ook in een aparte 'chat' getest, waarbij geldt dat de historie van de ene chat geen invloed heeft op een andere chat binnen hetzelfde AI-systeem).

prompt is integraal opgenomen onder 'Bevinding 4' in paragraaf 4.2.

4.2 De belangrijkste bevindingen

Bevinding 1: one-shot = bad shot

Een eerste bevinding is dat de meeste van de AI-modellen met het enkele invoeren van de casus met vragen een onvoldoende scoren voor de beantwoording. Dit verbaast niet. De genoemde AI-systemen kunnen weliswaar een zeer divers palet aan generieke gebruikersvragen beantwoorden, maar zijn niet specifiek voorbereid op de gewenste manier van beantwoorden van nichevragen zoals juridische casusvragen (qua redeneringsmethodiek of beoordelingskader). Het is een bekend fenomeen dat taalmodellen suboptimale antwoorden geven wanneer zij 'koud' worden geprompt op expertisedomeinen.[12] Een meer uitgebreide prompt waarin de context van de vraag wordt geschetst, en waarin meer instructies worden gegeven ten aanzien van het gewenste beantwoordingsformat, eventuele specifiek te overwegen factoren of toonzetting, werkt daarom veelal beter (zie ook de volgende bevinding). Alleen Gemini en DeepSeek wisten op casus 2 een voldoende te scoren. Gemini scoorde een 6,0 en DeepSeek scoorde zelfs een 7,0. Op dit verrassende resultaat van DeepSeek gaan wij nader in onder 'Bevinding 5'.

Bevinding 2: goed prompten is belangrijk, maar biedt geen garantie op succes

Na de one-shotpoging hebben we een meer uitgebreide prompt gebruikt conform een populair doch simpel *prompting framework*: COSTAR. Dit acroniem staat voor Context (schets de achtergrond van de vraag, het 'waarom'), Objective (stel het doel van de taak duidelijk; wat verwacht je qua output?), Style, Tone (omschrijf de schrijfstijl en toonzetting van het antwoord), Audience (omschrijf voor wie de output bedoeld is) en Response (omschrijf in welke format het antwoord moet worden gegoten, bijvoorbeeld in tekst, een puntsgewijze opsomming, of een tabel).[13] De prompt die we gebruikten, was:

> 'Jij bent een ervaren bedrijfsjurist die deelneemt aan een cursus verbintenissenrecht. Hieronder volgt een casus, met een bijbehorende vraag. Beantwoord deze vraag naar Nederlands recht. Identificeer in dat verband eerst de relevante wetsartikelen en relevante rechtspraak. Onderzoek de casus vervolgens nauwkeurig tegen de context van deze relevante wetsartikelen en rechtspraak. Geef een genuanceerd en weloverwogen antwoord, in een gepaste formele stijl, maar beperk je antwoord tot 300 woorden.'

Zoals hiervoor al werd beschreven, volgt uit diverse onderzoeken dat de antwoorden normaliter significant beter worden wanneer je een dergelijk prompt framework hanteert.[14] De verschillen tussen de one-shotmethode en de methode met uitgebreide 'COSTAR'-prompt in dit onderzoek resulteerden slechts in een bescheiden verbetering (het gemiddelde cijfer steeg van een 3,4 naar een 3,9 over beide casus). Bij sommige modellen (GPT-4o en Gemini, casus 2) resulteerde het gebruik van de verbeterde prompt echter in een slechter cijfer.

Een goede verklaring daarvoor lijkt te zijn dat we met de genoemde prompt de AI-modellen weliswaar instrueerden de juiste redeneringsmethodiek te volgen, maar dat een cruciaal onderdeel voor het uitvoeren van die methodiek – de relevante jurisprudentie – simpelweg ontbrak in het parate geheugen van de taalmodellen (zie de volgende bevinding). Hierdoor kon de casus nog steeds niet optimaal worden beantwoord.

Bevinding 3: (jurisprudentiële) context is alles

Onze resultaten bevestigen dat het aanvullen van de 'standaardkennis' van de grote taalmodellen met specifieke relevante juridische context een beslissende factor is in hun vermogen om een juridische casus correct te analyseren. Althans, dat geldt in elk geval voor zover het gaat om Nederlandse jurisprudentie op een specifiek rechtsgebied zoals het verbintenissenrecht.

De meeste AI-modellen hadden weinig moeite de relevante wetsartikelen te identificeren, hoewel de modellen soms ook bepalingen noemden die niet relevant waren voor de te beantwoorden vragen (zie ook 'Bevinding 6' hierna). Kennelijk is deze informatie, ook als het gaat om Nederlandse wetgeving, afdoende ingebed in de '(pre-)training' van deze modellen. Geen van de modellen slaagde er echter in om de voor deze casus specifiek relevante jurisprudentie te identificeren zonder dat we deze expliciet aan het 'kortetermijngeheugen' van de AI-systemen hadden toegevoegd.[15] Na deze toevoeging verbeterde de kwaliteit van al hun analyses aanzienlijk, hetgeen ook leidde tot een significante verbetering van hun scores. Zo steeg het gemiddelde cijfer van de AI-systemen over beide casus van een 3,9 (met initiële prompt, maar zonder context) naar een 6,1 (mét context), een significant hoger gemiddelde dan de bedrijfsjuristen (5,3). Dit is te verklaren door de substantiële stijging van het gemiddelde cijfer van de AI-systemen voor casus 1 van een 3,7 (met initiële prompt, maar zonder context)

12 Zie voor een breed overzicht aan studies die aantonen hoe verschillende prompttechnieken (prompt engineering) significante verbeteringen opleveren ten opzichte van *one-shot* of *zero-shot prompting*: www.prompthub.us/blog/prompt-engineering-with-reasoning-models.

13 Zie https://portkey.ai/blog/what-is-costar-prompt-engineering voor een uitleg van (de voordelen van) het COSTAR framework.

14 Zie noot 7. Terzijde merken wij op dat aan de bedrijfsjuristen uitsluitend is gevraagd de bij de casus behorende vragen te beantwoorden; voor hen was daarnaast een syllabus beschikbaar waarin zowel relevante rechtspraak als literatuur was opgenomen.

15 Relevant om te vermelden hierbij is dat we de tools hebben 'gevoed' met de relevante jurisprudentie, terwijl de bedrijfsjuristen een veel langere lijst met rechtspraak hadden in de syllabus. In hoeverre de tools in staat zijn om uit een grotere hoeveelheid jurisprudentie de relevante uitspraken te vinden, hebben we dus niet onderzocht. Wel viel uit de antwoorden op te maken of de tools de rechtspraak – al dan niet – op correcte wijze konden toepassen op de casus.

 doi: 10.5553/MvV/157457672025035004001

naar een 7,7 (mét context), een significant hoger gemiddelde dan de bedrijfsjuristen voor die casus (4,8).

Op casus 2 scoorden de AI-modellen dan weer vrijwel even goed mét context (4,5) in vergelijking met de variant zónder context (4,2) en scoorde de one-shotmethode (4,5) zelfs even hoog als de variant mét initiële prompt en context, maar daarvoor is de verklaring waarschijnlijk gelegen in het feit dat jurisprudentie voor die casus niet relevant was voor de beantwoording van de bijbehorende vragen.

Het onderzoek bevestigt in elk geval het voor de Nederlandse rechtspraktijk belangrijke inzicht dat de grootste, meest geavanceerde en meest populaire AI-modellen vooralsnog niet beschikken over een brede kennis van de Nederlandse jurisprudentie. De modellen hebben wel een goed begrip van de meer bekende, klassieke arresten,[16] maar daarbuiten wordt de kennis al gauw meer beperkt. Voor onderzoeksvragen waarbij zulke jurisprudentie van belang is, is het dan ook zaak om de betreffende jurisprudentie toe te voegen aan de AI-systemen, hetzij door handmatig een (relevante) selectie toe te voegen, hetzij door een automatische koppeling tot stand te brengen met jurisprudentiedatabases.[17]

Tot slot zij opgemerkt dat AI-systemen een veel groter en breder begrip van Amerikaanse jurisprudentie hebben, omdat daarvan een zeer veel groter corpus aan data bestaat, die bovendien openbaar toegankelijk is en dus goed vertegenwoordigd in de (pre-)trainingssets van de modellen. De hoeveelheid gepubliceerde Nederlandse jurisprudentie is aanmerkelijk kleiner in vergelijking met de Amerikaanse en dat betekent dat – kort gezegd – aan Nederlandse uitspraken een veel kleiner gewicht toekomt in de algoritmische (statistische) modellen van de AI-systemen. Dat leidt ertoe dat alleen de bekendste uitspraken op een adequate manier door de betreffende chatbots kunnen worden gevonden c.q. 'gereproduceerd'.

Bevinding 4: promptoptimalisatie leidt tot perfectie

Na onze initiële testronde hebben we, met de inzichten uit de eerste resultaten en met kennis van de juiste antwoorden, een verder geoptimaliseerde promptstrategie ontwikkeld. Deze geoptimaliseerde prompts combineerden de volgende factoren:

1. zeer specifieke instructies over de te volgen juridische methodologie, inclusief het expliciet overwegen van toepasbaarheid van wetsartikelen;
2. verzoeken om reflectie; en
3. instructie om nuance te betrachten bij de eindconclusies.

De volledige geoptimaliseerde prompt luidde als volgt:

> 'Jij bent een ervaren bedrijfsjurist die deelneemt aan een cursus verbintenissenrecht. Hieronder volgt een casus, met een bijbehorende vraag. Beantwoord deze vraag naar Nederlands recht. Identificeer in dat verband de relevante wetsartikelen en relevante rechtspraak, waarbij je met name kunt putten uit de 3 uitspraken die ik je hierbij geef (zie aangehechte bestanden).
>
> Analyseer de casus vervolgens nauwkeurig tegen de context van deze relevante wetsartikelen en rechtspraak, waarbij je als volgt te werk moet gaan:
>
> 1. loop eerst de relevante wetsartikelen af, identificeer expliciet elk van de verschillende artikelleden / sub-artikelen en concludeer voor ieder van deze afzonderlijk of deze wel of niet van toepassing zijn;
> 2. geef dan aan tot welke initiële conclusie die analyse leidt;
> 3. analyseer vervolgens de relevante uitspraken, en onderzoek in hoeverre deze uitspraken de initiële conclusie kunnen nuanceren of veranderen;
> 4. geef ten slotte je eindconclusie, waarbij je duidelijk rekenschap geeft van eventuele verschillende mogelijke interpretaties en aangeeft welke interpretatie naar jouw smaak het best verdedigbaar lijkt.'

Deze aanpak leidde tot een doorbraak in de prestaties van de meeste AI-modellen, met spectaculaire resultaten tot gevolg:

1. Claude 3.5 Sonnet Pro behaalde een score van 10,0 bij zowel casus 1 als casus 2. Dat is derhalve een significante verbetering ten opzichte van de scores met de initiële prompt (met context), te weten een 8,0 bij casus 1 en een 4,0 bij casus 2. Claude behaalde daarmee als enige een perfecte score voor beide casus in het hele onderzoek.
2. GPT-4o verbeterde ten opzichte van de initiële prompt met context van een 8,0 naar een 9,0 bij casus 1, al scoorde deze slechts een 2,0 bij casus 2.
3. Gemini 1.5 Pro verbeterde ten opzichte van de initiële prompt met context van een 8,0 naar een 9,0 bij casus 1 en van een 4,0 naar een 7,5 bij casus 2.
4. DeepSeek verbeterde ten opzichte van de initiële prompt met context van een 9,0 naar een 10,0 bij casus 1, maar ging terug van een 8,0 naar een 6,0 bij casus 2.
5. Mistral verbeterde ten opzichte van de initiële prompt met context van een 8,0 naar een 9,0 bij casus 1 en van een 2,0 naar een 3,0 bij casus 2.

16 Als korte test hebben wij geverifieerd dat de modellen gedegen kennis hadden van arresten als Lindenbaum/Cohen, Kelderluik, Guldemond/Noordwijkerhout, Quint/Te Poel, Berg en Dalse Watertoren, Hofland/Hennis, Pos/Van den Bosch, Haviltex, Jetblast, Nooteboom/Bouwman, het Harmonisatiewet-arrest, het DES-arrest en het Urgenda-arrest. Hiertoe hebben we de modellen geprompt om een 'top 10' van meest belangrijke Nederlandse arresten te produceren.

17 Een dergelijke koppeling kan op verschillende technische manieren worden bewerkstelligd, bijv. middels een zogeheten *application programming interface* (API) tussen het taalmodel en de betreffende juridische database. De aanbieder van de database dient daar echter vaak wel toestemming voor te verlenen. Een andere manier zou zijn om een (aanvullend) Nederlands 'specialistisch' juridisch model te trainen met grote hoeveelheden Nederlandse jurisprudentie (hetgeen in de ons omringende landen wel gebeurt), maar het nadeel is dat een dergelijk model dan nog steeds niet zal beschikken over de meest actuele of 'live' data. Het valt buiten het bestek van deze bijdrage en onze onderzoeksvraag om hierover verder uit te weiden.

De voornaamste reden van het succes van deze prompt is gelegen in de manier waarop de prompt aansluiting met het beoordelingskader afdwingt. Waar AI-modellen van nature geneigd zijn om 'high-level' te redeneren en dan snel naar de juiste conclusie te springen, dwingt de verbeterde prompt deze tot het volgen van de stapsgewijze methodiek die in het juridisch onderwijs wordt verwacht, waarin bijvoorbeeld de verschillende onderdelen van een wetsartikel worden afgelopen en beoordeeld op toepasselijkheid. Dit verklaart waarom de modellen zonder deze specifieke instructies vaak punten misten, ondanks dat ze materieel soms wel tot juiste conclusies kwamen. Bovendien zijn AI-systemen geneigd stappen over te slaan die er niet toe doen in hun perceptie, terwijl het beoordelingskader juist vereist dat de relevante wetsartikelen expliciet worden behandeld, en waarbij dient te worden aangegeven welk onderdeel van de bepaling van toepassing is. Door dit specifiek voor te schrijven corrigeren we een inherente 'bias' in de AI-redeneermethode.

Door expliciet te vragen om reflectie op de initiële conclusie in het licht van jurisprudentie, en rekening te houden met meerdere interpretatiemogelijkheden, voorkomen we dat het model te snel tevreden is met zijn eerste antwoord én dat er een te ongenuanceerd antwoord wordt gegeven. Deze methode van reflectief redeneren leidt tot meer genuanceerde, complete antwoorden die beter aansluiten bij de verwachtingen van de beoordelaars. Juist dit element versterkt de juridische diepgang van de antwoorden aanzienlijk.[18]

Bevinding 5: DeepSeek (de 'AI-sensatie' uit China) maakt indruk, terwijl Mistral (het enige Europese model) enigszins teleurstelt

Mistrals Le Chat scoorde opvallend laag met 0 punten voor beide casus wanneer geen context was toegevoegd. Na het toevoegen van context verbeterden de antwoorden significant, met een 8,0 (initiële prompt) respectievelijk een 9,0 (verbeterde prompt) bij casus 1, maar nog steeds met een teleurstellende 2,0 (initiële prompt) respectievelijk 3,0 (verbeterde prompt) bij casus 2. DeepSeek (R1) presteerde echter zeer goed, met de beste gemiddelde scores over beide casus bij de one-shot-prompt (een 5,0 en bij casus 2 zelfs een 7,0), bij de initiële prompt zonder context (een 7,5), alsook bij de initiële prompt mét context (een 8,5). Ook behaalde DeepSeek met de verbeterde prompt de perfecte score (10) bij casus 1. In dat opzicht valt DeepSeek *overall* en de resultaten van alle prompttechnieken beschouwend als 'winnaar' te bestempelen van ons onderzoek.

Wij hebben geen duidelijke verklaring voor Mistrals teleurstellende prestatie. Het zou kunnen zitten in de focus en trainingsstrategie van het model. Hoewel Mistral wordt gepositioneerd als een Europees alternatief voor de Amerikaanse en Chinese AI-giganten, is het model mogelijk niet veel getraind op specifiek juridische redeneringen en in het bijzonder niet op het Nederlandse recht. Dit suggereert dat de (pre-)trainingsdataset mogelijk minder Nederlandse juridische data (waaronder wetsteksten) bevatte, of dat het model minder geoptimaliseerd is voor de stapsgewijze, systematische redeneermethodiek die juridisch onderwijs vereist.

DeepSeek daarentegen verraste zeer positief met zijn vermogen om op academisch niveau juridische casus te analyseren. De goede prestaties van DeepSeek suggereren dat het model over een robuuste juridische basis beschikt, mogelijk door een trainingsregime waarbij aanzienlijke hoeveelheden Nederlandse juridische data zijn verwerkt. Bovendien lijkt DeepSeek beter in staat om de gestructureerde, stapsgewijze benadering te volgen die we in onze geoptimaliseerde prompt voorschreven, wat duidt op een goede capaciteit voor het opvolgen van complexe procesinstructies.

Een andere mogelijke verklaring voor het contrast tussen deze twee modellen ligt in hun schaalgrootte en trainingsfilosofie. DeepSeek is naar verluidt ontwikkeld met substantiële computationele middelen en een breed trainingsregime met veel academische bronnen, terwijl Mistral zich profileert als een efficiënter, compacter model (hetgeen onder meer tot uiting komt in het gegeven dat Mistral van alle modellen verreweg het snelst – praktisch instantaan – antwoord gaf). Deze resultaten suggereren dat, althans voor complexere juridische redeneertaken, de grotere modellen nog steeds een voorsprong hebben in prestatievermogen, ondanks de vooruitgang in efficiëntie van kleinere modellen.

Bevinding 6: AI-antwoorden doen verrassend menselijk aan, maar zijn ook ietwat wijdlopig

Een belangrijke observatie is bovendien dat veel van de door AI gegenereerde antwoorden niet direct te onderscheiden waren van die van de bedrijfsjuristen (op sommige AI-antwoorden na, die eruit sprongen door een overmatig gebruik van 'kopjes' of een bijzonder uitgebreide inleiding, waarin een breed palet aan mogelijk relevante artikelen genoemd werd). De (blinde) beoordeling bevestigde dat AI-modellen in staat zijn juridische teksten te produceren die, qua structuur en inhoud, sterk lijken op die van menselijke professionals. Dit weerlegt in zekere zin de gedachte dat AI-gegenereerde juridische analyses altijd gemakkelijk te herkennen zouden zijn aan hun kunstmatige karakter.

Opvallend was dus wel dat de AI-modellen vaak meer wetsartikelen citeerden en meer wijdlopige antwoorden gaven dan de bedrijfsjuristen. In sommige gevallen werden wetsartikelen geciteerd die niet relevant waren. De juiste antwoorden moesten vaak wat meer worden 'gezocht' in een langere tekst, die naast

18 Wij merken voor de meer technisch georiënteerde lezer nog op dat het model OpenAI o1 een zogeheten 'redeneringsmodel' is, dat in principe reeds 'geïnstrueerd' is om bepaalde van de in onze verbeterde prompt gebruikte redeneringstechnieken te hanteren. De reden waarom o1 dan toch niet goed scoort, is waarschijnlijk daarin gelegen dat die instructies niet dusdanig specifiek zijn dat zij het model voorschrijven om (in dit voorkomende geval) exact aan te geven welke onderdelen van een bepaling – al dan niet – relevant zijn.

 doi: 10.5553/MvV/157457672025035004001

het specifieke antwoord ook allerlei inleidende, meer algemene overwegingen bevatte. Dit dus in tegenstelling tot de antwoorden van de bedrijfsjuristen, die directer op de specifieke vraag reageerden. Een en ander valt te verklaren doordat de betreffende chatbots geprogrammeerd zijn om behulpzaam te zijn, en het vanuit didactisch oogpunt begrijpelijk is dat voor een antwoord op een specifieke vraag eerst een bredere context wordt geschetst.

Bevinding 7: analyse van succesvolle en minder succesvolle antwoordpatronen

Door de verschillende antwoorden te analyseren konden er enkele patronen worden geïdentificeerd die kenmerkend waren voor succesvolle versus minder succesvolle AI-antwoorden.

Kenmerken van succesvolle AI-antwoorden:

1. *systematische structuur*: de hoogst scorende antwoorden volgden een duidelijke, systematische structuur bij het analyseren van de juridische vraagstukken, vergelijkbaar met de traditionele methodologie die in het juridisch onderwijs wordt gedoceerd;
2. *expliciet redeneren*: succesvolle antwoorden maakten hun redeneerproces expliciet, met duidelijke verwijzingen naar relevante wetsartikelen en jurisprudentie en een 'stap voor stap'-analyse van hoe deze bronnen van toepassing waren op de feiten van de casus;
3. *kritische reflectie*: de beste antwoorden vertoonden tekenen van kritische reflectie en het vermogen om initiële conclusies te heroverwegen en verschillende interpretaties tegen elkaar af te wegen;
4. *balans tussen volledigheid en relevantie*: succesvolle antwoorden vonden een goede balans tussen het systematisch doorlopen van alle relevante juridische aspecten, zonder te verzanden in niet-relevante details.

Kenmerken van minder succesvolle AI-antwoorden:

1. *overgeneralisatie*: minder succesvolle antwoorden hadden de neiging om te algemeen te blijven, zonder voldoende specifieke toepassing van de juridische bepalingen op de casus;
2. *overmatige focus op wetsartikelen*: sommige antwoorden citeerden uitgebreid wetsartikelen, zonder adequaat uit te leggen hoe deze precies van toepassing waren op de casus;
3. *gebrek aan systematiek*: minder succesvolle antwoorden misten vaak een duidelijke, systematische aanpak en benoemden verschillende juridische concepten zonder een coherent verhaal te creëren;
4. *onvoldoende kritische analyse*: de zwakste antwoorden presenteerden conclusies zonder voldoende onderbouwing of overweging van alternatieve interpretaties.

Interessant genoeg vertoonden de hoog scorende AI-antwoorden meer overeenkomsten met de antwoorden van de bedrijfsjuristen dan met die van minder goed presterende AI-modellen. Dit suggereert dat er bepaalde fundamentele kenmerken van effectief juridisch redeneren zijn die zowel door mensen als door geavanceerde AI-systemen kunnen worden toegepast.

5 Conclusie en praktische aanbevelingen

Uit ons bescheiden onderzoek kan worden geconcludeerd dat AI-taalmodellen, mits optimaal geïnstrueerd en van relevante context (met name rechtspraak) voorzien, in staat zijn om Nederlandse juridische casus te analyseren op een niveau dat vergelijkbaar is met of zelfs beter is dan dat van juridische professionals. De belangrijkste bevindingen kunnen als volgt worden samengevat:

1. One-shot prompting levert – althans bij de huidige stand van de techniek – ondermaatse resultaten op voor de juridische analyse van een casus.
2. Goed gestructureerde prompts zijn belangrijk, maar zonder relevante context blijft de verbetering beperkt.
3. Het is essentieel om te weten welk type (juridische) kennis wel en niet voldoende aanwezig is in de datasets van de modellen, opdat je eventueel ontbrekende c.q. niet voldoende vertegenwoordigde kennis als specifieke context kunt toevoegen aan de prompt. Ons onderzoek toont aan dat de geteste AI-modellen in ieder geval gedegen kennis ontberen van Nederlandse jurisprudentie. Deze jurisprudentie moet derhalve worden toegevoegd om een hoogwaardige juridische analyse van een casus te realiseren.
4. Promptoptimalisatie gericht op systematische analyse, reflectie en nuance leidt tot significant betere resultaten.
5. Er bestaan grote verschillen in prestaties tussen de AI-modellen. Daarbij lijken Europese modellen achter te blijven ten opzichte van zowel Amerikaanse als Chinese alternatieven. Het verdient aanbeveling te experimenteren met de verschillende modellen om te bepalen welke het best presteert voor een specifieke juridische taak of behoefte.
6. AI-gegenereerde antwoorden doen inmiddels zeer menselijk aan, maar zijn soms nog wat te algemeen of wijdlopig.

Dit onderzoek heeft aangetoond dat AI-taalmodellen, mits optimaal ingezet en van de relevante context voorzien, in staat zijn om een casus te analyseren op een niveau dat vergelijkbaar is met of zelfs beter is dan dat van juridische professionals. Tegelijkertijd heeft het de cruciale rol belicht die menselijke expertise blijft spelen in het sturen, contextualiseren en verifiëren van deze analyses.

Tot slot: de toekomst van juridisch werk ligt waarschijnlijk niet in een *keuze* tussen mens en machine, maar in een doordachte samenwerking waarbij beide hun unieke sterktes inbrengen. Door te begrijpen hoe AI-systemen juridische problemen benaderen, en hoe we hun prestaties kunnen optimaliseren, kunnen we deze krachtige tools effectief integreren in zowel het juridisch onderwijs als de praktijk, om uiteindelijk betere, toegankelijkere en efficiëntere juridische diensten te leveren. De vraag is wat ons betreft dan ook niet zozeer óf AI een rol zal spelen in juridische analyse, maar hoe (juridische) professionals deze technologie het beste kunnen inzetten om de kwaliteit van juridisch werk te verbeteren.