Universiteit van Amsterdam

Bachelorscriptie Informatiekunde – scriptie verdediging

Automatische classificatie van citatiefuncties in Nederlandse rechtspraak met Large Language Models

Thijs Beers 15232492

Universiteit van Amsterdam, Faculteit der Natuurwetenschappen, Wiskunde en Informatica (FNWI)

Begeleider: Dr. M.J. (Maarten) Marx

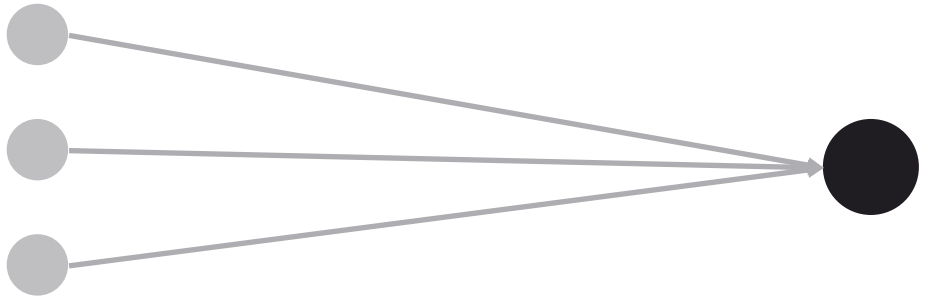
Tweede beoordelaar: Dr. D.P. (David) Graus

3 juli 2026

Probleem en onderzoeksvraag

Veel geciteerd betekent niet automatisch gezaghebbend

Standaard netwerk



Alle verwijzingen tellen gelijk mee. Veel citaties lijken automatisch belangrijk.

Getypeerd netwerk

Steunend, onderscheidend of procedureel: de functie blijft zichtbaar.

Hoofdvraag

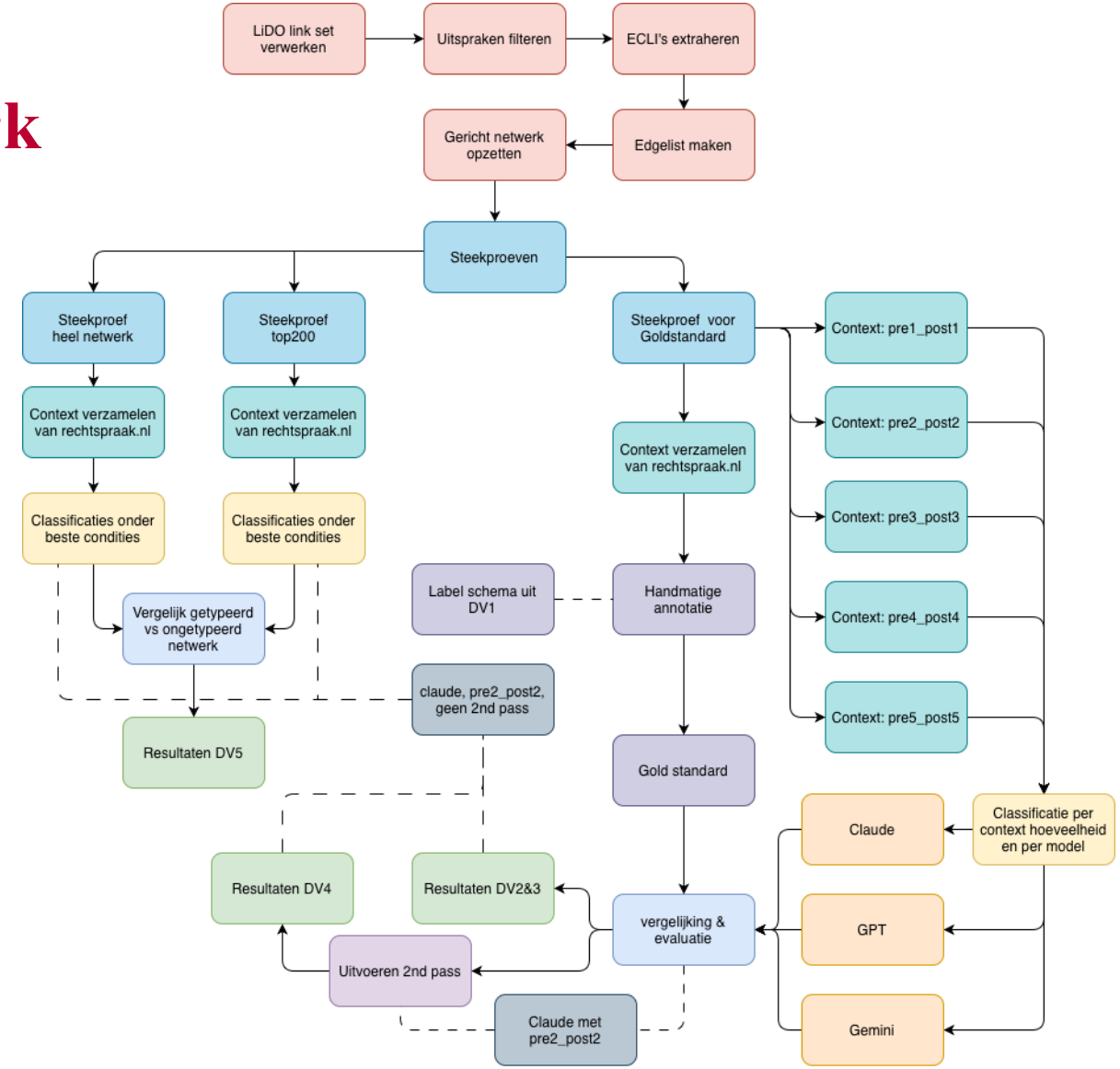
"In hoeverre en onder welke condities kunnen Large Language Models citaties tussen Nederlandse rechterlijke uitspraken classificeren naar hun juridische functie, en wat voegt deze classificatie toe aan de analyse van Nederlandse citatienetwerken?"

DV1: Label schema, DV2: Overeenstemming, DV3: Context grootte, DV4: 2e pass en DV5: getypeerde netwerk analyse

Methode

Van linkset tot getypeerd netwerk

- LiDO-linkset verwerken tot een citatienetwerk.
- *Gold standard* handmatig annoteren.
- Drie LLM's evalueren op de *gold standard*.
- Invloed van contextgrootte onderzoeken.
- Two-pass verificatie evalueren.
- Best presterende model toepassen op het netwerk.
- Getypeerd en standaard netwerk vergelijken.



Deelvraag 1 & 2

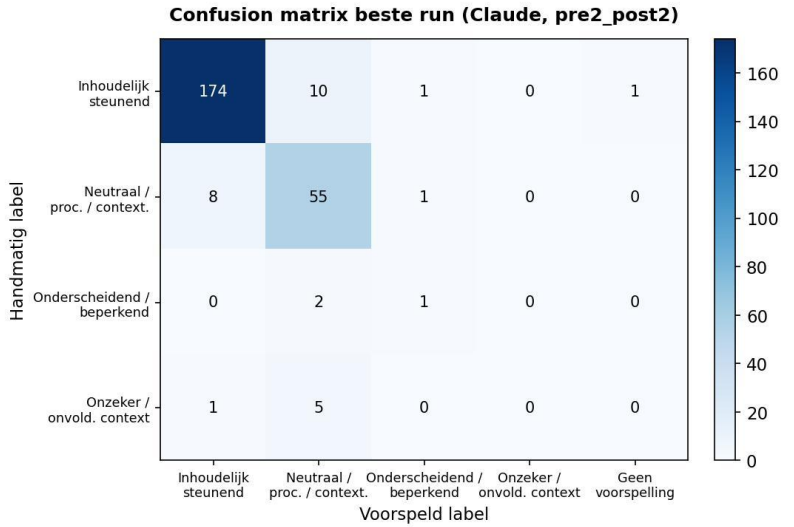
Labelschema en classificatieprestaties

Vijf citatiefuncties: steunend, neutral, onderscheidend, kritisch en onzeker

Context	Model	Accuracy	Macro-F1
pre2_post2	Claude	0,89	0,52
pre4_post4	Gemini	0,87	0,49
pre2_post2	Gemini	0,87	0,51
pre3_post3	Claude	0,86	0,43
pre4_post4	Claude	0,86	0,43

Vijf hoogste accuracy-runs, met bijbehorende macro-F1

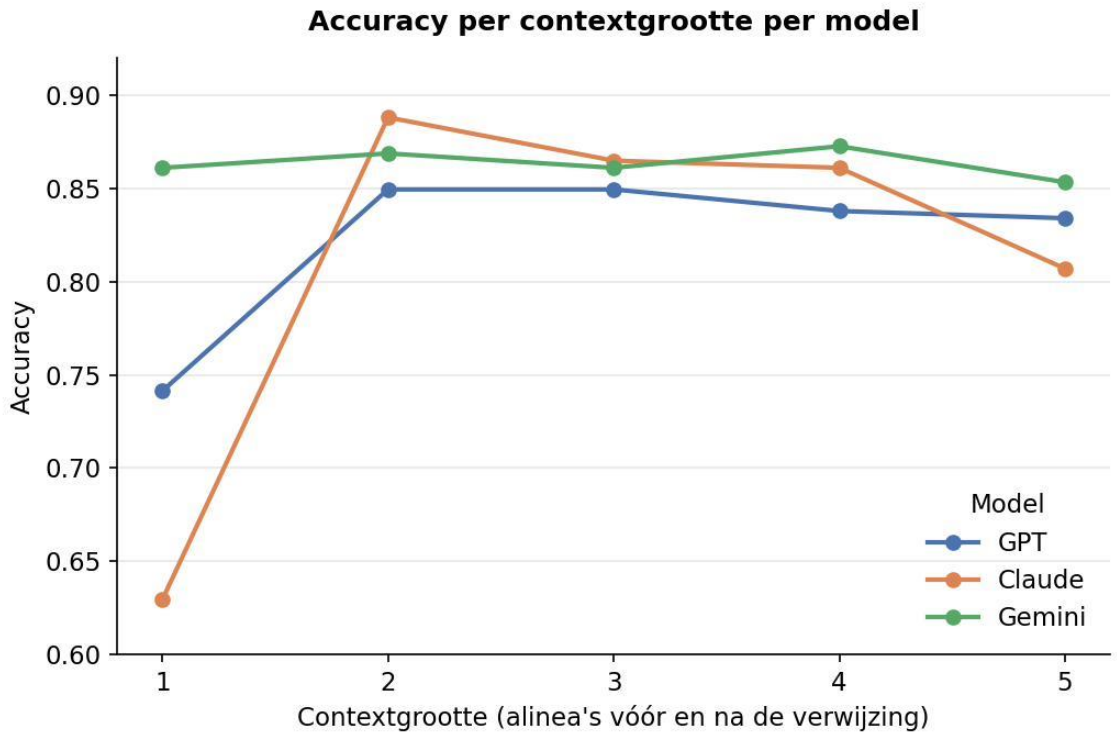
Accuracy tussen 0,86 en 0,89, maar macro-F1 rond 0,52: de twee grote klassen worden goed herkend, de minderheidsklassen opgeslokt.



Confusion matrix van de beste run (Claude, pre2_post2)

Deelvraag 3

Hoeveel context heeft het model nodig?



Gemiddelde accuracy per contextgrootte

Accuracy van 0,74 naar 0,87

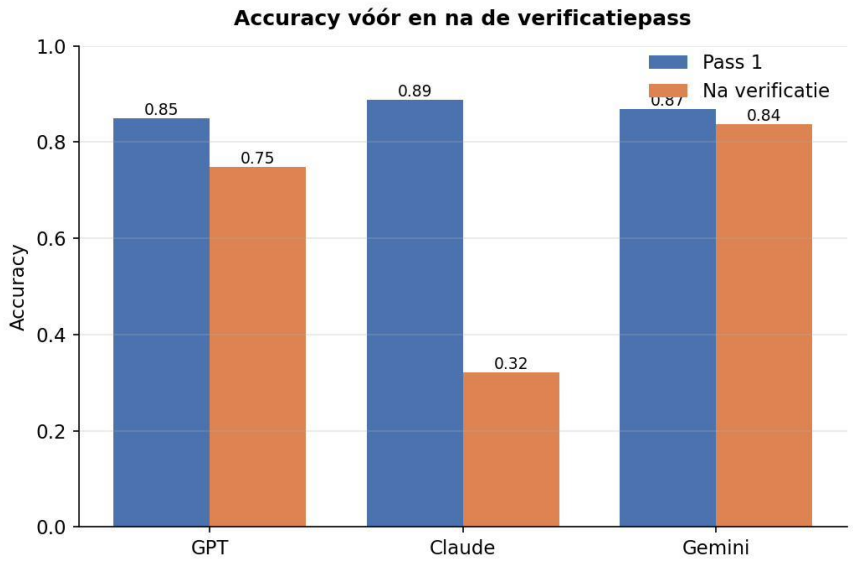
van één naar twee alinea's context

Twee alinea's vóór en na de citatie is optimaal.

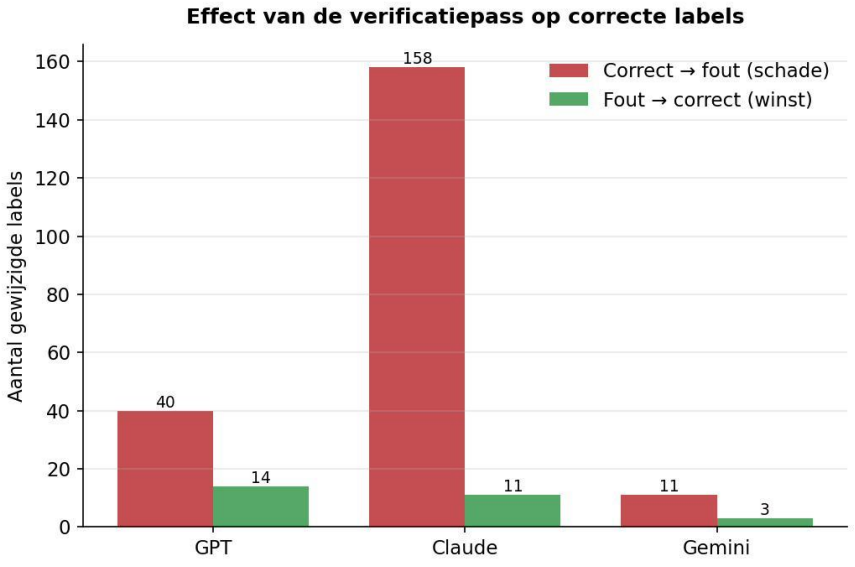
Meer context verbetert niets en voegt bij de grootste vensters eerder ruis toe.

Deelvraag 4

Helpt een tweede beoordelingsronde?



Accuracy vóór (pass 1) en na verificatie



Wijzigingen: schade versus winst per model

Verslechtert

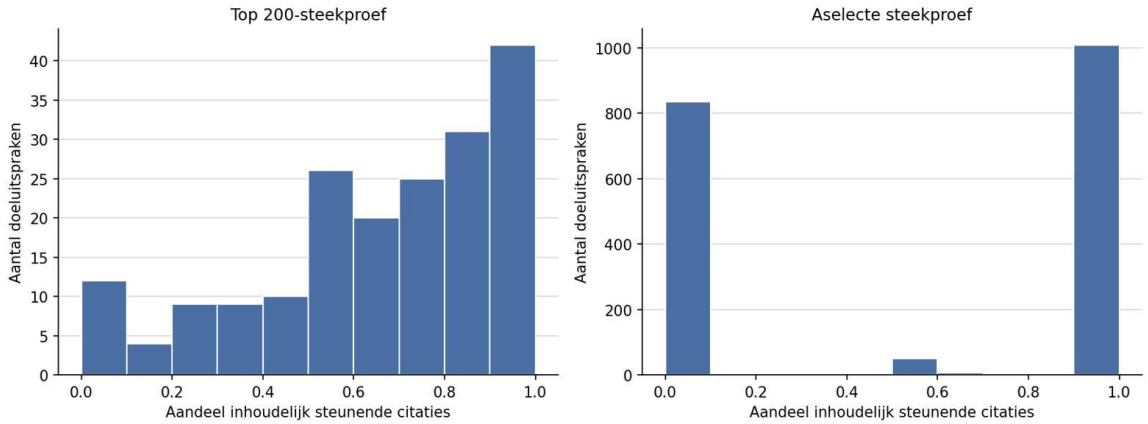
bij alle drie de modellen

68% van Claude's eigen labels werden verworpen en vaak waren dit correcte labels.

Zonder externe input duwt de verifier voorspellingen systematisch weg van de dominante klasse.

Deelvraag 5

Wat verandert er in het netwerk?



Aandeel inhoudelijk steunende citaties per uitspraak

Het aandeel steunende citaties is ongelijk verdeeld, dus de rangorde verschuift zodra alleen steun meetelt.

Uitspraak	Totaal	Steun	Rang verandering
HR:2016:252	119	78	1 naar 1
HR:2022:853	80	25	2 naar 7
RVS:2020:1560	65	24	3 naar 9
CRVB:2009:BH1009	46	40	5 naar 2
RBMNE:2023:4482	45	0	6 naar 178
CRVB:2015:4920	33	29	10 naar 4

Rangverschuiving (totale en steunende in-degree)

Plek 6 met 45 citaties, maar geen enkele steunende citatie zakt naar plek 178. Dus een hoge in-degree wijst niet direct naar juridisch gezag.

Conclusie & Discussie

LLM’s kunnen citaties bruikbaar classificeren, en dit heeft invloed op het netwerk beeld

Belangrijkste conclusies

Classificatie

Betrouwbaar voor veelvoorkomende functies, zwak voor de zeldzame door scheve data. Twee alinea’s context is optimaal.

Zelfcontrole

Een tweede ronde in deze opzet voegt niks toe en verslechtert de classificatie.

Netwerk

Typeren maakt zichtbaar dat een hoge in-degree niet altijd juridische autoriteit betekent.

Vervolgonderzoek

- Grotere, evenwichtiger dataset zodat ook zeldzame functies betrouwbaar te beoordelen zijn.
- Annotatie door meerdere beoordelaars om de labels te valideren.

Beperkingen

- Slechts 259 geannoteerde citaties, sterk scheef verdeeld.
- Een annotator, geen interbeoordelaarsbetrouwbaarheid.
- Een two-pass-opzet getest.

Verantwoording AI-gebruik

Waarvoor AI is ingezet

Voor mijn scriptie heb ik gebruik gemaakt van verschillende modellen van Claude en Open AI.

Schrijven

AI ondersteunde mijn schrijfproces door het helpen met het omschrijven van alineas naar beter lopende alineas.

Ook heft AI geholpen met het maken van keuzes over waar in mijn scriptie ik bepaalde argumenten moet maken.

Programmeren

AI heft geholpen met het schrijven van de code. Ik biedt bijvoorbeeld de structuur / een beetje pseudo code aan aan de AI die het daarna implementeerd. Hierna verifier ik de werking en zo ga ik stap voor stap verder.

Literatuur

AI heeft geholpen met het samenvatten van artikelen zodat ik makkelijker een selectie kan maken van belangrijke en relevante papers.

Voorbeeldprompts

Voorbeeld prompt

Onderzoeksvraag aanscherpen

"Ik vind mijn onderzoeks vraag wel goed, maar denk dat hij beter word door iets toe te voegen over welke condities er nodig zijn voor de classificatie, wat vind jij hiervan?"

Voorbeeld prompt

Classificatie concurrent maken

"Ik heb de code die we tot nu toe hebben geschreven in dit bericht geplakt. Bestudeer deze goed, en geef een paar richtingen om de code zo te maken dat we meerdere API calls tegelijk kunnen sturen per provider zodat de classificatie wat sneller gaat."

Voorbeeld prompt

Tekst herformuleren

Het antwoord op deelvraag 3 is dat vooral de overgang van weinig (1 alinea) naar matige context (2 alineas) goed helpt (accuracy gaat van 0.74 naar 0.87) . Daarna is er weinig verbetering en gaat het zelfs slechter bij teveel context. Schrijf dit om naar een alinea voor in mijn resultaten.

De hoeveelheid context beïnvloedt het classificatievermogen vooral bij de overgang van zeer weinig naar matige context. Het uitbreiden van één naar twee alinea's vóór en na de verwijzing verhoogt de gemiddelde accuracy duidelijk, van 0,74 naar 0,87. Daarboven treedt geen verdere verbetering meer op en neemt de prestatie bij de grootste context licht af. De beste balans tussen nauwkeurigheid en efficiëntie wordt geboden bij twee alinea's vóór en na de verwijzing.

Resultaat in scriptie

5.3.4 Antwoord op deelvraag 3

De hoeveelheid context beïnvloedt het classificatievermogen vooral bij de overgang van zeer weinig naar matige context. Het uitbreiden van één naar twee alinea's vóór en na de verwijzing verhoogt de gemiddelde accuracy duidelijk, van 0,74 naar 0,87. Daarboven treedt geen verbetering meer op en neemt de prestatie bij de grootste context licht af. De beste balans tussen nauwkeurigheid en efficiëntie wordt geboden door pre2_post2.

Referenties

Fowler, J. H., & Jeon, S. (2008). The authority of supreme court precedent. Social Networks, 30(1), 16-30. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378873307000378 doi:https://doi.org/10.1016/j.socnet.2007.05.00

van Kuppevelt, D., van Dijck, G., & Schaper, M. (2020). Purposes and challenges of legal citation network analysis on case law. In R. Whalen (Ed.), Computational legal studies (pp. 265–292). United Kingdom: Edward Elgar Publishing. doi: 10.4337/9781788977456.00017

Whalen, R. (2016). Legal networks: The promises and challenges of legal network analysis. Michigan state law review, 2016, 539. Retrieved from https://api.semanticscholar.org/CorpusID:32690001 Wu, Z., Zeng, Q., Zhang, Z., Tan, Z., Shen, C., & J