Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

SAMENVATTING

Zoekmachines en chatbots zijn centraal komen te staan in de manier waarop mensen informatie raadplegen en taken uitvoeren. Met de opkomst van *large language models* (LLM's) zijn informatiesystemen verschoven van puur retrieval-gebaseerde systemen naar tekstgeneratie en *retrieval-augmented generation* (RAG). Hoewel deze ontwikkelingen nieuwe kansen bieden, brengen ze ook uitdagingen met zich mee, zoals verouderde kennis, hallucinaties, bias en fouten. Het waarborgen dat dergelijke systemen robuust, onbevooroordeeld en betrouwbaar zijn, vereist systematische evaluatie over een breed scala aan taken en contexten.

In dit proefschrift hebben we onderzocht hoe retrieval- en generatiemodellen zich gedragen in realistische informatiezoekscenario's, met bijzondere aandacht voor robuustheid en onbevooroordeeldheid als essentiële aspecten bij het bouwen van betrouwbare systemen. Het onderzoek is gestructureerd rond vier centrale uitdagingen:

Generaliseerbaarheid van *ranking models* in lexicaal rijke zoekproblemen. We evalueren *contextualized lexical ranking models* in *query-by-example* (QBE) *retrieval* als voorbeeld van een lexicaal rijk zoekprobleem. Onze resultaten tonen aan dat deze modellen, hoewel effectief in ad hoc *retrieval*, minder goed presteren in QBE *rertieval*, waar BM25 competitief blijft. Het interpoleren van *contextualized lexical ranking* modellen met BM25 leidt echter tot verbeterde *ranking*, wat wijst op complementaire kwaliteiten van relevantiesignalen uit zowel gecontextualiseerde lexicale modellen als traditionele lexicale modellen.

Robuustheid van gebruikerstevredenheidsschatting. We hebben benchmarks voor gebruikerstevredenheidsschatting in taakgerichte dialoogsysteemevaluatie uitgebreid door "omgekeerde" onbevredigende dialogen te genereren met LLM's. Deze generatie is bedoeld om de verhouding tussen bevredigende en onbevredigende voorbeelden in de klassenverdeling van tevredenheidslabels te balanceren. Met de verrijkte testcollecties hebben we ontdekt dat fijn-afgestelde modellen zoals BERT en ASAP goed presteren onder de oorspronkelijke, onevenwichtige klassenverdelingen, maar dat hun prestaties geleidelijk afnemen naarmate het aandeel onbevredigende dialogen toenam. In tegenstelling hiermee blijken *few-shot* in-context LLM's robuuster en gevoeliger voor distributieveranderingen.

Meten van maatschappelijke bias in een geordende documentenlijst. We identificeren een beperking van de veelgebruikte NFaiRR-metriek, die bias in documenten onafhankelijk behandelt en tegengestelde bias in verschillende documenten elkaar niet laat opheffen. Om dit aan te pakken, hebben we een nieuwe fairness-metriek voorgesteld: TExFAIR, die (1) term-gebaseerde associaties gebruikt om documenten via representatieve termen te koppelen aan maatschappelijke

groepen, en (2) een rank-biased discountfactor toepast die de invloed reduceert van niet-representatieve documenten, namelijk documenten zonder groep-representatieve termen.

Attributiegevoeligheid en bias in RAG. We hebben twee evaluatiemetrieken ontwikkeld, *Counterfactually-estimated Attribution Bias* (CAB) en *Counterfactually-estimated Attribution Sensitivity* (CAS), om te kwantificeren hoe *retrieval-augmented* LLM's reageren op auteursmetadata in hun brondocumenten. Met behulp van deze metrieken hebben we ontdekt dat het opnemen van metadata over of een document door een mens of door een AI is geschreven, het attributiegedrag significant beïnvloedt, waarbij modellen consequent de voorkeur geven aan door mensen geschreven bronnen. Dit onthult een systematische bias die eerdere aannames over de voorkeur van LLM's voor AI-gegenereerde *content* tegenspreekt. Bovendien benadrukken onze bevindingen een kritieke kwetsbaarheid in het attributiegedrag van LLM's, aangezien dergelijke metadata-gevoeligheid kan worden misbruikt om output te manipuleren, wat belangrijke zorgen oproept voor de betrouwbaarheid van RAG-systemen.

Ons onderzoek kent beperkingen op basis waarvan nieuwe onderzoeksrichtingen kunnen worden geïnitieerd. In QBE *retrieval* hebben we onze analyse gericht op gecontextualiseerde lexicale modellen. *Dense* en hybride *retrieval* models moeten nog systematisch worden onderzocht in lexicaal rijke condities. In tevredenheidsschatting genereerden we uitsluitend omgekeerde dialogen voor evaluatie en hebben we hun impact op training niet onderzocht; het uitbreiden naar dialoog-niveau tevredenheidsschatting en gebruikerstevredenheid in andere systeemtypes (bijvoorbeeld aanbevelingssystemen) vormt een waardevolle volgende stap. In fairness-evaluatie gaat de afhankelijkheid van term-gebaseerde groepsproxies voorbij aan meer semantische en gebruiker-gecentreerde perspectieven op fairness; toekomstig werk moet evaluatiekaders ontwikkelen die beter aansluiten bij menselijke oordelen. Voor attributie in RAG was onze studie beperkt tot auteursmetadata; het uitbreiden van deze methodologie naar andere metadata (bijv. geslacht, etniciteit, bron) zou interessant zijn. Onze studie richtte zich op het blootleggen en analyseren van attributiebias, niet op het mitigeren ervan; toekomstig onderzoek zou strategieën moeten ontwikkelen om deze bias te verminderen en zo de betrouwbaarheid van RAG-systemen te vergroten.

Tot slot benadrukken we een bredere onderzoeksrichting: nu LLM's steeds vaker worden ingezet als *agentic systems* die meer zelfstandig beslissingen kunnen nemen, wordt robuuste evaluatie nog crucialer. Ons gebruik van omgekeerd redeneren (systematisch verkennen van "wat-als"-scenario's) biedt een basis voor het ontwerpen van evaluatieopzetten die de betrouwbaarheid, fairness en robuustheid van dergelijke systemen waarborgen.

Al met al draagt dit proefschrift bij aan een beter begrip van hoe *retrieval*- en generatiemodellen presteren onder realistische en structureel uitdagende condities, terwijl het tegelijk beperkingen en toekomstperspectieven schetst die de ontwikkeling van meer robuuste, onbevooroordeelde en betrouwbare zoekmachines en chatbots kunnen sturen.