A Reproducibility Study of LLMs for Query Generation

Moritz Staudinger, Wojciech Kusa, Florina Piroi, Aldo Lipani, Allan Hanbury
moritz.staudinger@tuwien.ac.at

Query Generation

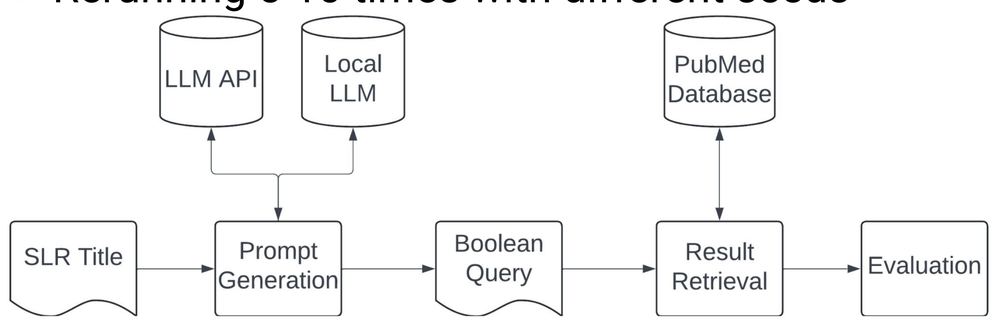
- Suitable for Data Retrieval in IR Systems, Databases and Knowledge Graphs
- Aimed at crafting a query that captures all relevant results
- Needs significant time and expert-level understanding
- ML used to expand, enrich and generate queries

Boolean Search Strategy Example

1. exp Endometrial Neoplasms/ (25288)
2. (endometri* adj5 (cancer* or tumor* or tumour* or neoplas* or malignan* or carcinoma* or adenocarcinoma* or patholog* or disease*)).mp. (53270)
3. ((uter* and lining) adj5 (cancer* or tumor* or tumour* or neoplas* or malignan* or carcinoma* or adenocarcinoma* or patholog* or disease*)).mp. (133)
4. 1 or 2 or 3 (53460)
5. exp Uterine Hemorrhage/ (23504)
6. ((post menopaus* or postmenopaus*) adj5 (bleed* or period* or discharge* or menstruat* or haem* or hemor*)).mp. (3737)
7. PMB.mp. (1144)
8. 5 or 6 or 7 (27287)
9. 4 and 8 (2101)

Experiment Design

- Generate Boolean Queries and evaluate their retrieval performance
- Usage of JSON for parsing and GPT/Mistral APIs + local setups in the pipeline
- Rerunning 5-10 times with different seeds



Results

Different prompting strategies

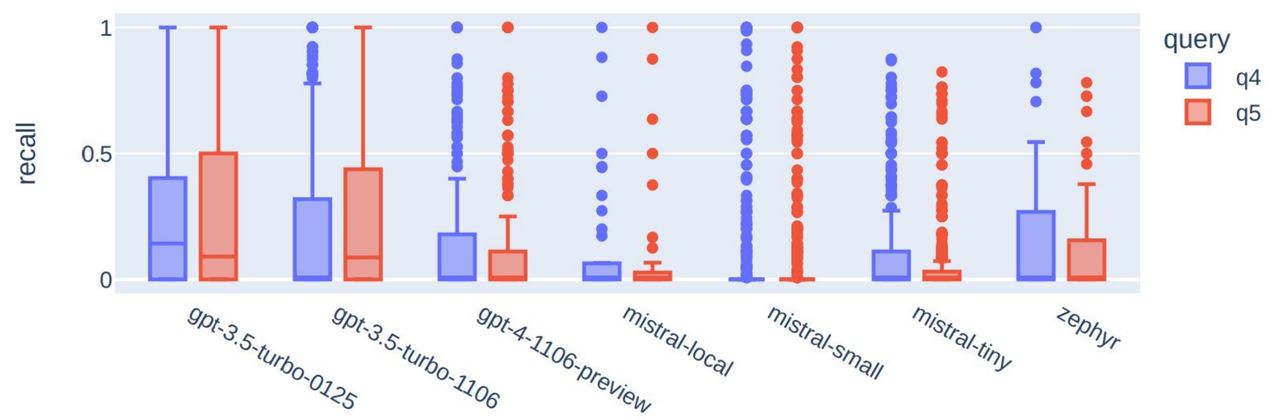
- Zero-Shot (q1-q3)
- One-Shot (q4-q5)
  - High Quality example
  - Relevant example
- Chain-of-Thought (guided)

Usage of different sections/search fields in text

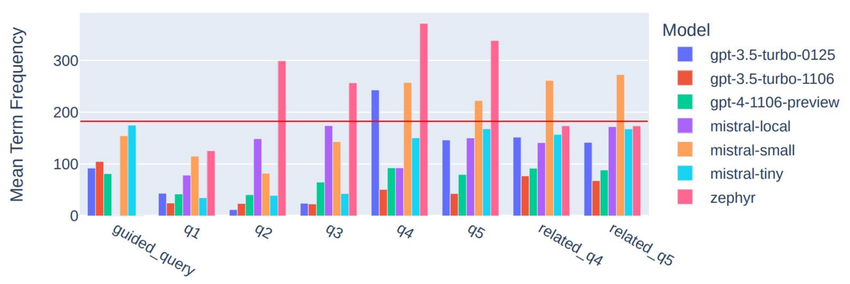
Evaluated on CLEF-TAR 2017 and Seed collection

CLEF								
Baseline				**Recall** 0.832				
	Wang et al. [52]	**GPT-3.5-1106**	**GPT-3.5-0125**	**GPT-4**	**Mistral-tiny**	**Mistral-small**	**Mistral-local**	**Zephyr**
q1	0.129	0.084 ± 0.145	0.019 ± 0.079	0.072 ± 0.142	0.046 ± 0.119	0.062 ± 0.128	0.037	0.015
q2	0.131	0.067 ± 0.125	0.019 ± 0.081	0.093 ± 0.169	0.026 ± 0.071	0.021 ± 0.063	0.025	0.017
q3	0.118	0.115 ± 0.195	0.026 ± 0.106	0.086 ± 0.147	0.041 ± 0.108	0.063 ± 0.123	0.038	0.007
q4-HQE	0.504	0.139 ± 0.212	0.033 ± 0.131	0.086 ± 0.170	0.063 ± 0.166	0.067 ± 0.150	0.020	0.062
q5-HQE	0.334	0.150 ± 0.212	0.027 ± 0.107	0.091 ± 0.142	0.043 ± 0.129	0.060 ± 0.138	0.053	0.005
Seed				**Recall**				
Baseline				**0.711**				
Baseline-edit				0.647				
	Wang et al. [52]	**GPT-3.5-1106**	**GPT-3.5-0125**	**GPT-4**	**Mistral-tiny**	**Mistral-small**	**Mistral-local**	**Zephyr**
q1	0.053	0.148 ± 0.24†	0.203 ± 0.291†	0.132 ± 0.25†	0.132 ± 0.244	0.190 ± 0.293	0.122	0.036
q2	0.039	0.025 ± 0.108	0.147 ± 0.247†	0.141 ± 0.246†	0.047 ± 0.142	0.057 ± 0.152	0.063	0.087
q3	0.052	0.086 ± 0.206†	0.169 ± 0.255†	0.156 ± 0.244†	0.065 ± 0.168	0.214 ± 0.299	0.095	0.002
q4-HQE	0.129	0.213 ± 0.310†	0.237 ± 0.296†	0.145 ± 0.267	0.091 ± 0.184	0.092 ± 0.228	0.121	0.193
q5-HQE	0.079	0.244 ± 0.311†	0.258 ± 0.326†	0.113 ± 0.224†	0.064 ± 0.155	0.084 ± 0.210	0.092	0.153
q4-RE	0.016	0.174 ± 0.256†	0.202 ± 0.280†	0.088 ± 0.205†	0.080 ± 0.171	0.080 ± 0.212	0.055	0.066
q5-RE	—	0.178 ± 0.281	0.267 ± 0.340	0.104 ± 0.205	0.067 ± 0.160	0.075 ± 0.195	0.064	0.060
guided	0.517	0.035 ± 0.130	0.048 ± 0.109	0.125 ± 0.221	0.017 ± 0.090	0.100 ± 0.206	—	—

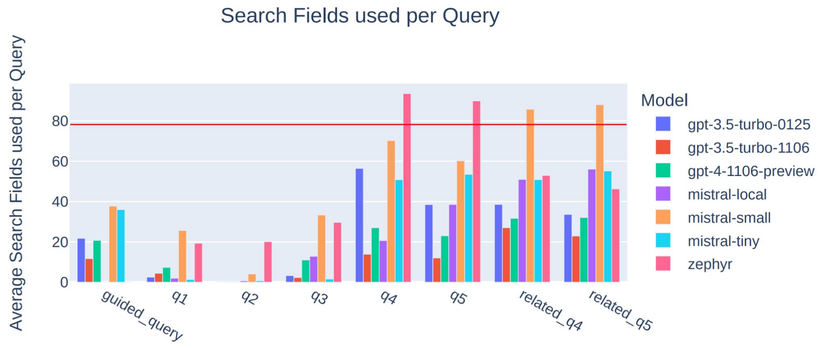
Variability of Recall of Generated Queries



Average Boolean Query Term Length



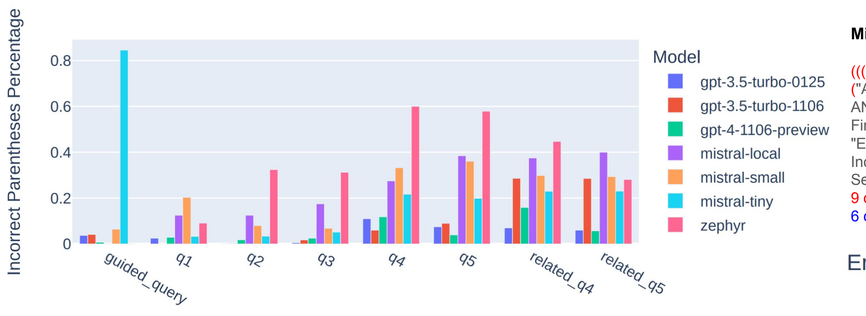
Search Fields used per Query



Analysis & Conclusion

- Quick generation without expert knowledge
- High variance between same prompts
- Manual correction / adaptation necessary
- Generation is not explainable
- Low reproducibility for the query generation

Parentheses Incorrect When Generating Queries



Mistral Tiny:

((("Thyroid Neoplasms"[MeSH] OR "Differentiated Thyroid Neoplasms"[MeSH]) AND ("Autopsy"[MeSH] OR "Postmortem"[MeSH] OR "Pathology"[MeSH] OR "Prevalence"[MeSH]) AND ((("Thyroid Gland"[MeSH] OR "Carcinoma"[MeSH] OR "Neoplasms"[MeSH] OR "Autopsy Findings"[MeSH] OR "Thyroid Cancer"[MeSH]) OR ("Environmental Factors"[MeSH] OR "Etiology"[MeSH] OR "Risk"[MeSH] OR "Cancer Epidemiology"[MeSH] OR "Cancer Incidence"[MeSH]))) AND ((("Meta-Analysis"[PubMed] OR "Study Design"[MeSH] OR "Autopsy Series"[Text] OR "Series"[Text] OR "Published"[Text] OR "Recent"[Text])
9 opening parentheses
6 closing parentheses

Errors when generating queries