

Hypothetical Documents or Knowledge Leakage? Rethinking LLM-based Query Expansion

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon,
Kunwoo Park

In Findings of the Association for Computational Linguistics:
ACL 2025, pages 19170–19187, Vienna, Austria. Association for
Computational Linguistics.

S2310157 谷口

検索器の需要

ゼロショット検索

- クエリから関連する文書を特定
 - 特定タスク専用の訓練不要
 - 様々なタスクで活用できる汎用性

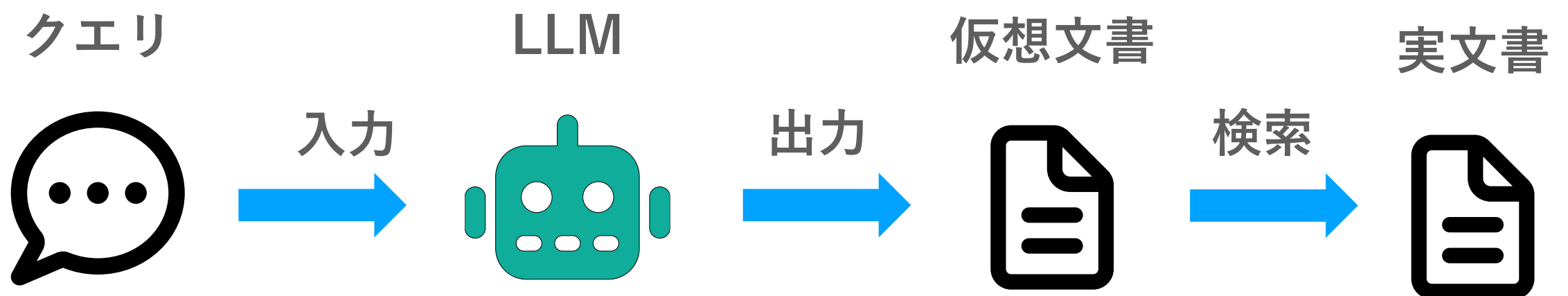
ゼロショット検索の有用性

- 知識を必要とする応用で活用
 - 質問応答
 - 事実検証

クエリ拡張

LLMを用いたアプローチ

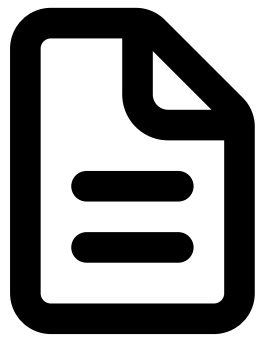
- クエリから「仮想文書」を生成させる
- 「仮想文書」をもとに検索を行う
- 「幻覚を含んでいても検索を改善する」という仮定
- 様々なベンチマークで性能向上を果たした



データリーク

仮想文書が正しい知識を出力するケース

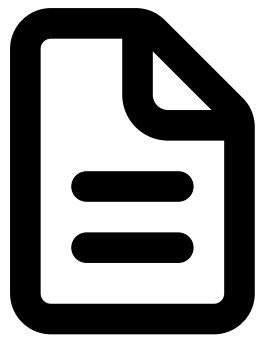
- ・クエリ: トランプは極右デモで「両陣営に立派な人がいた」と発言した



仮想文書

この主張は、**2017年にバージニア州シャーロットツビルで開かれた「ユナイト・ザ・ライト」集会**の後に、ドナルド・トランプ前大統領が行った発言を…

検索するまでもなく答えを知っている
このようなケースで正解するのは当然



主張を証明する
正解の文書

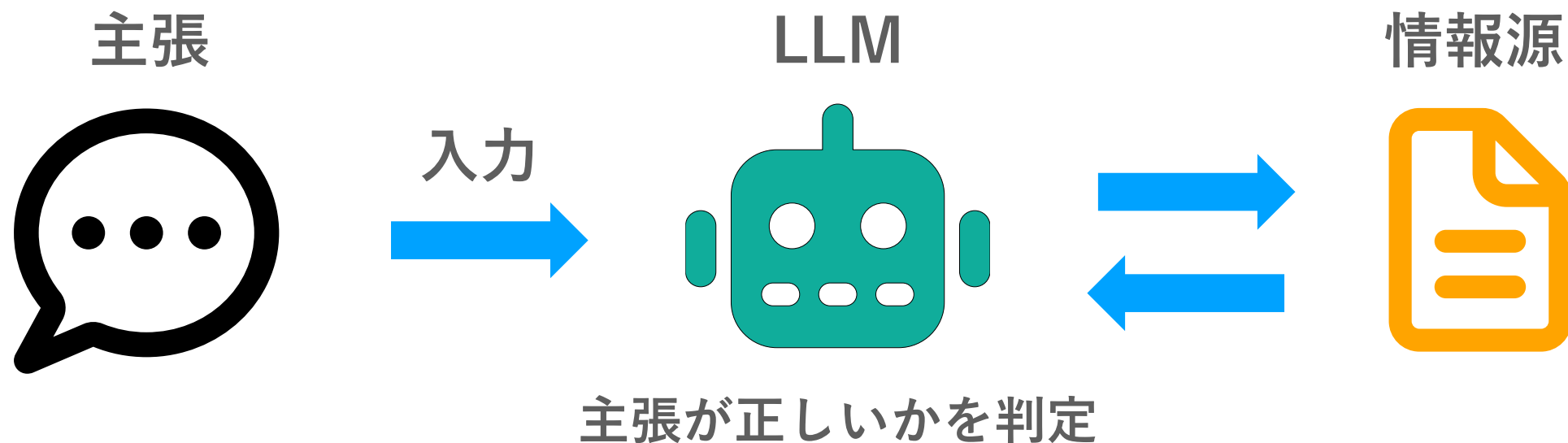
トランプ氏は、**2017年にバージニア州シャーロットツビルで開かれた「ユナイト・ザ・ライト」集会**の後、この発言をしました。その際、彼は……

クエリ拡張の効果はデータリークによるものという可能性

実験

事実検証タスク

- ある主張の真偽を判定
- 検索によって主張を証明/反証する根拠を収集
- 現実では新しい知識のファクトチェックが重要
- 分類タスクは性能評価がしやすい



実験

データセット

データセット	主張数	1主張の検証に必要な正解根拠	情報源(文書数)
FEVER	6,666	1.66	Wikipedia (5,416,536)
SciFact	693	1.8	論文の概要 (5,183)
AVeriTeC	3,563	2.77	Web文書 (2,623,538)

実験

クエリ拡張手法

- Query2doc

- クエリをn回繰り返したものとクエリから生成した仮想文書
- BM25で検索

- HyDE

- クエリから生成した仮想文書n個 / クエリ
- それぞれ埋め込みし平均
- Contriever

実験

マッピング判定

- 仮想文書を文単位で分割
 - クエリの言い換えでしかない文は除外
 - spaCy及びROUGEを使用
- 各文が正解の文書の内容を含意しているかをチェック
 - 一致 : matched
 - 不一致 : unmatched
 - gpt-4o-mini

実験結果

クエリ拡張 vs 通常の手法

Query2doc

Method	FEVER			SciFact			AVeriTeC		
	Recall @5	NDCG @5	F1	Recall @5	NDCG @5	F1	MET EOR	BERT Score	F1
BM25	31	25	54	51.2	45.5	48.6	17.8	11.6	32
GPT-4o -mini	36.4 ±0.1	29.3 ±0.1	55.6 ±0.1	55.1 ±0.2	47.9 ±0.1	<u>52.5</u> ±0.5	19.1 ±0.0	12.4 ±0.0	32.6 ±0.1
Llama-3.1 -70b-it	<u>38.3</u> ±0.1	<u>31</u> ±0.1	<u>56.1</u> ±0.1	<u>56.4</u> ±0.3	<u>49.2</u> ±0.1	52.4 ±0.7	<u>19.3</u> ±0.1	<u>12.7</u> ±0.0	<u>33.4</u> ±0.2

Query2docはどのLLMモデルでも高い評価

実験結果

クエリ拡張 vs 通常の手法

HyDE

Method	FEVER			SciFact			AVeriTeC		
	Recall @5	NDCG @5	F1	Recall @5	NDCG @5	F1	MET EOR	BERT Score	F1
Contriever	26.8	20.2	53.1	55.1	47.3	51.2	17.6	12.6	33.9
GPT-4o -mini	37.3 ±0.1	28.8 ±0.0	55.6 ±0.1	61.2 ±0.2	53.1 ±0.1	<u>54</u> ±0.5	18.7 ±0.0	13.2 ±0.0	35.7 ±0.6
Llama-3.1 -70b-it	<u>40.4</u> ±0.2	<u>31.7</u> ±0.2	<u>55.9</u> ±0.1	<u>61.9</u> ±0.3	<u>54.1</u> ±0.2	53.6 ±0.5	<u>19</u> ±0.2	<u>13.7</u> ±0.1	<u>35.4</u> ±0.7

HyDEも同様にどのLLMモデルでも高い評価
クエリ拡張は性能を高めている

実験結果

仮想文書と正解の文書のマッチ率(%)

Method	FEVER		SciFact		AVeriTeC	
	Query 2doc	HyDE	Query 2doc	HyDE	Query 2doc	HyDE
GPT-4o-mini	75.8 ±0.1	83.5 ±0.1	52.8 ±0.4	59.1 ±0.7	40.4 ±0.2	68 ±0.3
Llama-3.1-70b-it	78.3 ±0.7	71.7 ±0.7	57.5 ±0.3	55 ±0.8	48.1 ±0.8	47 ±4.8

ほとんどの場合40%以上が正解の根拠と一致する文書を出力
(最低値はgemini-1.5-flash/SciFact/HyDEで27.6%)

実験結果

matched/unmatchedの仮想文書生成時の性能

Gpt-4o-mini Query2doc

Method	Data	FEVER			SciFact			AVeriTeC		
		Recal l @5	NDC G@5	F1	Recal l@5	NDC G@5	F1	MET EOR	BERT Score	F1
BM25	ALL	31	25	54	51.2	45.5	48.6	17.8	11.6	32
Query2 doc	ALL	36.4 ±0.1	29.3 ±0.1	55.6 ±0.1	55.1 ±0.2	47.9 ±0.1	52.5 ±0.5	19.1 ±0.0	12.4 ±0.0	32.6 ±0.1
	M	<u>40.5</u> ±0.1	<u>32.8</u> ±0.1	<u>58.4</u> ±0.0	<u>63.3</u> ±0.4	<u>57.1</u> ±0.3	<u>53.7</u> ±0.3	<u>21.6</u> ±0.1	<u>17.6</u> ±0.1	<u>38.3</u> ±0.1
	¬M	<u>23.8</u> ±0.3	<u>18.5</u> ±0.2	<u>44.9</u> ±0.1	<u>45.9</u> ±0.4	<u>37.6</u> ±0.3	<u>49</u> ±0.4	<u>17.4</u> ±0.0	<u>9</u> ±0.0	<u>27.6</u> ±0.1

実験結果

matched/unmatchedの仮想文書生成時の性能

Gpt-4o-mini HyDE

Method	Data	FEVER			SciFact			AVeriTeC		
		Recal l @5	NDC G@5	F1	Recal l@5	NDC G@5	F1	MET EOR	BERT Score	F1
Contriever	ALL	26.8	20.2	53.1	55.1	47.3	51.2	17.6	12.6	33.9
HyDE	ALL	37.3 ±0.1	28.8 ±0.0	55.6 ±0.1	61.2 ±0.2	53.1 ±0.1	54 ±0.5	18.7 ±0.0	13.2 ±0.0	35.7 ±0.6
	M	<u>40</u> ±0.1	<u>30.9</u> ±0.1	<u>57.8</u> ±0.0	<u>68.4</u> ±0.3	<u>61.4</u> ±0.3	<u>57.1</u> ±0.2	<u>19.8</u> ±0.0	<u>15.5</u> ±0.0	<u>37</u> ±0.2
	¬M	<u>23.4</u> ±0.4	<u>17.9</u> ±0.3	<u>41.9</u> ±0.2	<u>50.8</u> ±0.5	<u>41.2</u> ±0.4	<u>48.9</u> ±0.4	<u>16.4</u> ±0.0	<u>8.3</u> ±0.1	<u>30.3</u> ±0.4

実験結果

まとめ

- クエリ拡張は性能を高めている
- 正解の根拠と一致するような仮想文書が頻出
- 性能の改善はそのようなmatched時のみ
 - unmatched時はむしろ悪化

考察

- LLMは学習済みの知識を再現しただけの可能性
- ⇒最新情報やニッチな分野ではクエリ拡張が有効でないかも

(明確にデータリークのせいと断言はできない)

まとめ

クエリ拡張とデータリーク

- クエリ拡張は知識を再現しているだけという仮説
- 3つの事実検証ベンチマークで実験
 - 正解となる根拠と一致する仮想文書を生成したときのみ性能向上
- ベンチマークにデータリークが存在する可能性
- クエリ拡張という手法の過大評価
- 最新情報やニッチな分野では通用しない可能性

所感

クエリ拡張への疑念

- 「クエリ拡張の精度向上はデータリークによるものでデータリークが起きてないときは精度を下げる」
⇒面白い、クエリ拡張の幻覚は普通に有害
- 正解根拠と一致する内容が生成できたときは精度向上できなかったときは精度低下
⇒当たり前の気がする
- この論文はNLタスクでの話で、SEタスクでも同様の検証が為されてるのは気になる