実在しないエンティティや出来事に関する合成文書を用いたRAGベンチマークの構築

2025年3月14日

〇李聖哲[1, 2], 大萩雅也[1], 塚越駿[1], 福地成彦[1], 柴田知秀[1], 河原大輔[2]
[1] SB Intuitions株式会社　[2] 早稲田大学

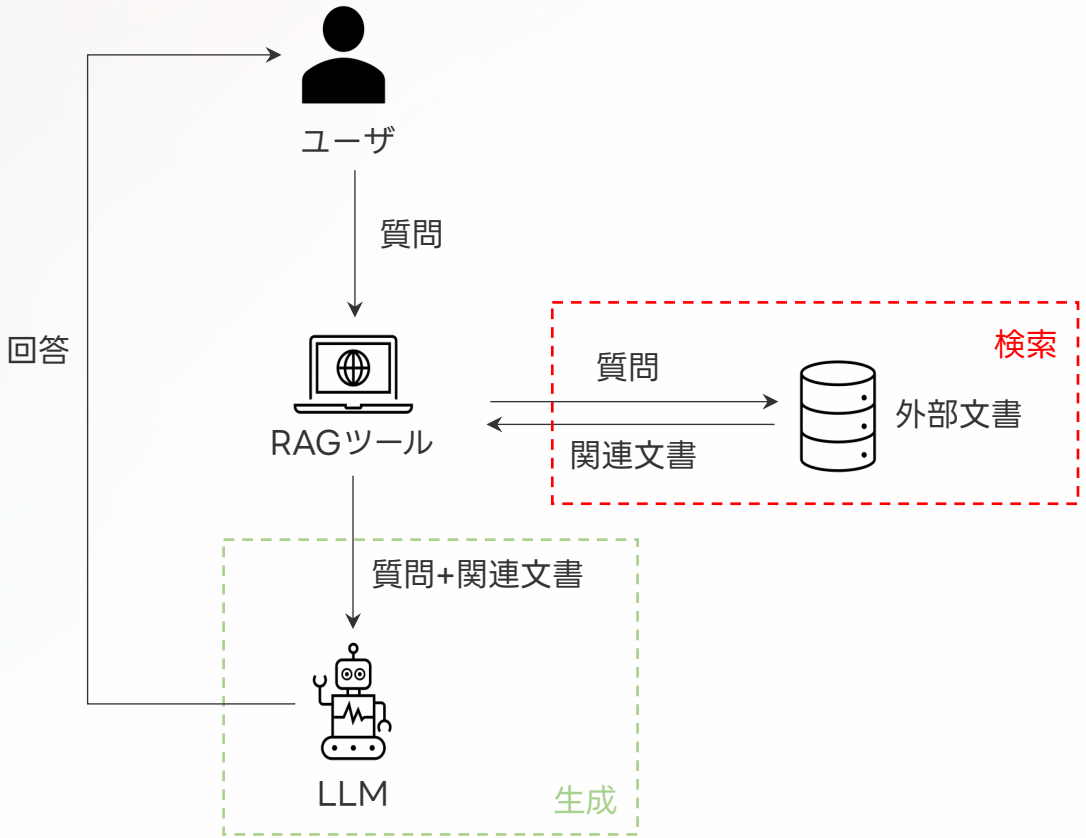
JLR 2025

SB Intuitions

RAGとは

SB Intuitions

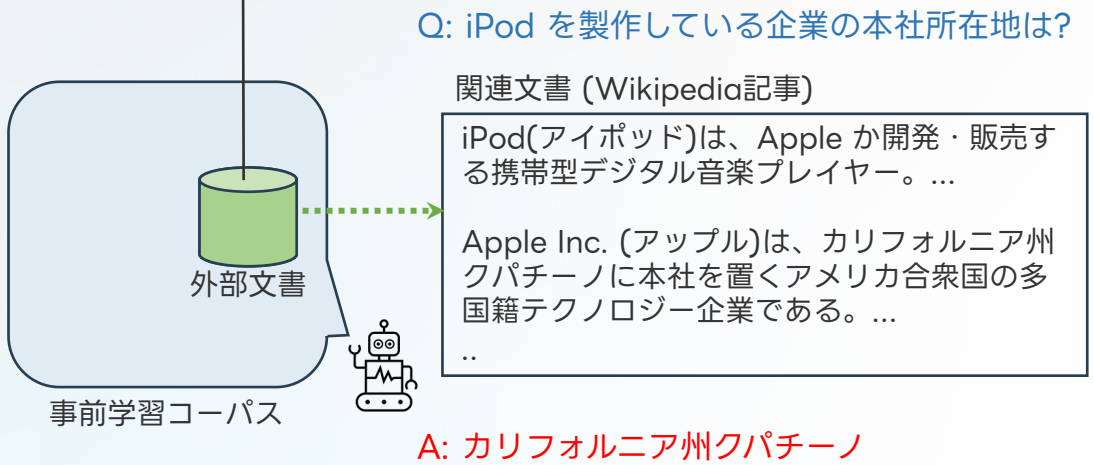
- 検索拡張生成 (RAG, Retrieval-Augmented Generation): 外部文書集合から検索により取得した関連文書に基づき，大規模言語モデル (LLM) に回答を生成させる手法



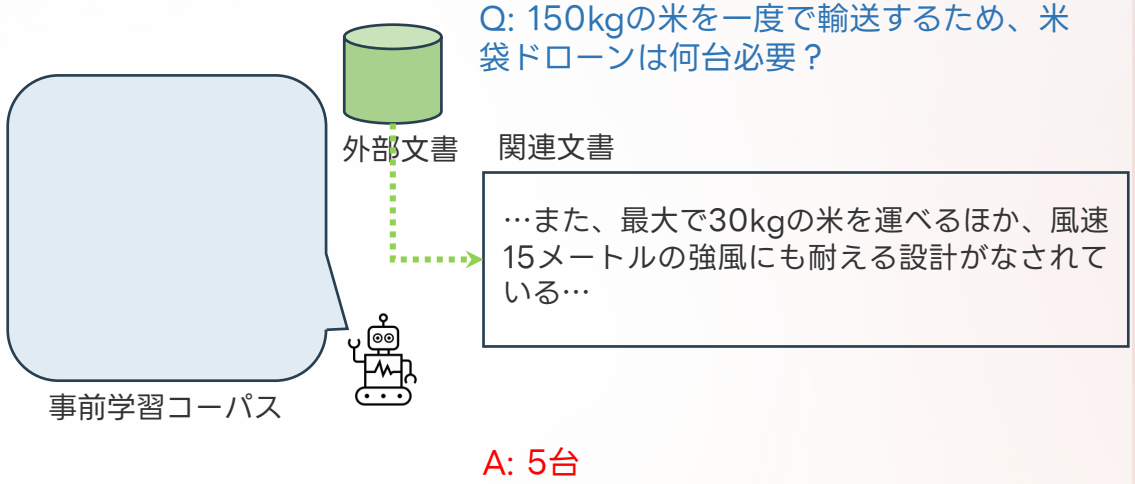
従来のベンチマークの問題点

外部文書が公開データであり，既にLLMに学習されている可能性が高い
- 新聞記事: RGB [Chen+ 2023], MultiHop-RAG [Tang+ 2024], etc.
- Wikipedia: RAGAs [Es+ 2023], ARES [Saad+ 2023], etc.

従来のRAGベンチマークでの評価



提案手法: LLMの事前学習コーパスと外部文書を分離



🤔 LLM自ら持っている知識だけで，その外部文書に関連する質問を解けてしまう
🤔 RAGにおける外部文書に基づく回答能力への評価にはならない恐れ

😀 LLMの内部知識だけでは解けず，関連文書に基づく回答生成能力をより正確に評価

提案手法

SB Intuitions

(目的) LLMの内部知識を測るのではなく，外部文書に基づく回答生成能力を測る

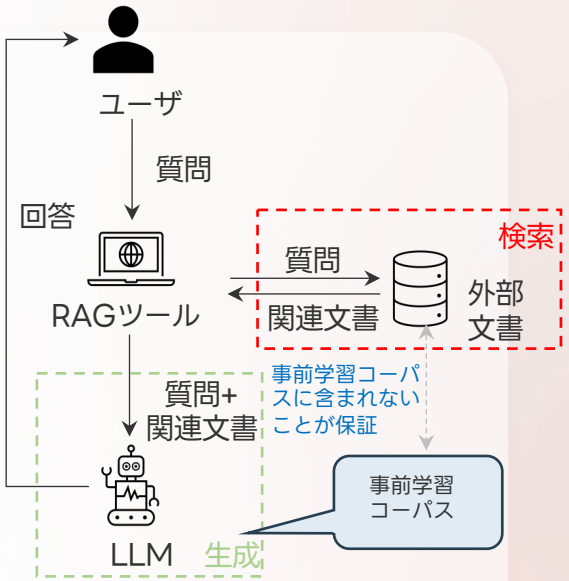
(要件) LLMの事前学習コーパスに含まれないことが保証できる外部文書が必要

- Webで公開される文書 → いつかクローリングされLLMの事前学習コーパスに入る
- 非公開データ → ベンチマークとして使用されることが難しい
- …

(手法) 実在しないエンティティや出来事に関する外部文書を作成

- このような文書は人にたくさん書いてもらうのが現実的ではない

(解決案) 外部文書は，LLMによって合成する



『週刊ザンカイ』（しゅうかんザンカイ、WEEKLY　ZANKAI）は、かつて大森出版が発行していた日本の青年週刊漫画雑誌。『未来大志』を青年誌として全面改訂する形で1993年に月2回刊誌の『ザンカイ』として誕生した。2001年に週刊化される際に誌名も『週刊ザンカイ』に変更され，読者層の拡大に成功した。…

実在しないエンティティに関する文書の一例

INDEX

SB Intuitions

ベンチマークの構築 – 概要

SB Intuitions

1. 元となる文書の収集

大阪国際空港（おおさかこくさいくうこう、IATA: ITM, ICAO: RJOO）は、大阪府豊中市、同府池田市、兵庫県伊丹市にまたがる空港… (wikipedia記事)

自然言語処理（しぜんげんごしょり、略称：NLP）は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術であり…(wikipedia記事)

…

無限に続くと思われていた円周率がついに終りを迎えた。千葉電波大学の研究グループがこれまでの円周率演算プログラムに誤りがあったことを発見…(「虚構新聞」記事)

…

…

模倣で合成

2. LLMによる合成文書の生成

- 実在しないエンティティや出来事を扱っていない
- 事実と矛盾する
- 論理展開が破綻，照応関係が欠如など

3. 基準を満たさない文書の除去

『週刊ザンカイ』（しゅうかんザンカイ、WEEKLY　ZANKAI）は、かつて大森出版が発行していた日本の青年週刊漫画雑誌。『未来大志』を青年誌として全面改訂…

オルタルは、原子番号79の第7周期に属する第6族元素である。元素記号はOr。オルタルの単体は非常に密度が高く、銀灰色を呈し、…

…

昼間のほとんどを寝て過ごす猫を「眠りこむネコ」などと称することが猫権の尊厳を損なうとして、国際猫権擁護団体が報道各社に対し、そういった表現の使用自粛…

…

…

4. 合成文書に対するQAの作成

Q: 休刊期間を含まず、『月刊ザンカイエクスプレス』およびその前身は何年続いた？
A: 18年

Q:『週刊ザンカイ』は休刊までに何回名前を変更された？
A: 1回
…

…

Q:「猫睡眠の健全さ」を啓発するキャンペーンを実施する目的は、猫に関するどんな誤解を払拭することを目指しているか？
A: 猫が怠惰の象徴であること
…

…

…

構築したベンチマークの構成

SB Intuitions

RAGの応用においてよく用いられる4種類のデータセットを選定

データセット	#Doc	#QA	概要
Pseudo Wikipedia (Wikipedia)	50	165	• 実在のエンティティに関するWikipedia記事を模倣し，架空のエンティティに関する紹介・解説 • 新たに登場した百科事典の記事を想定している
Pseudo News (News)	47	173	• 「虚構新聞」の記事を模倣して合成した，架空の出来事を報じるニュース記事 • 日々更新される最新の出来事を報じる記事を模擬することで，時系列的な新情報へ対応する能力を測ることを狙う
Pseudo Product Review (ProductReview)	40	48	• レビューサイトや個人ブログなどで見られる商品の評価記事を模倣して合成した，実在しない商品の仕様や使用感の記述 • 商品に関するチャットボットのシミュレーション
Pseudo Company Rules (CompanyRules)	24	175	• 架空の企業の社内規程や就業規則をモデルに，勤務体系・報酬体系・福利厚生などについて記述した文書 • 実在の「人事労務諸規程モデル集」を参考に，社内文書らしい文調や構成の文書 • 社内ボットが非公開の規定等の情報を取り扱うシナリオを想定

ベンチマークの構築 – 1. 元となる文書の収集

SB Intuitions

- 実在しないエンティティや出来事に関する文書の作成は，人手では現実的ではない
- そのため，実在する元となる文書 (ベース文書) をLLMに模倣させて合成文書の作成を行う

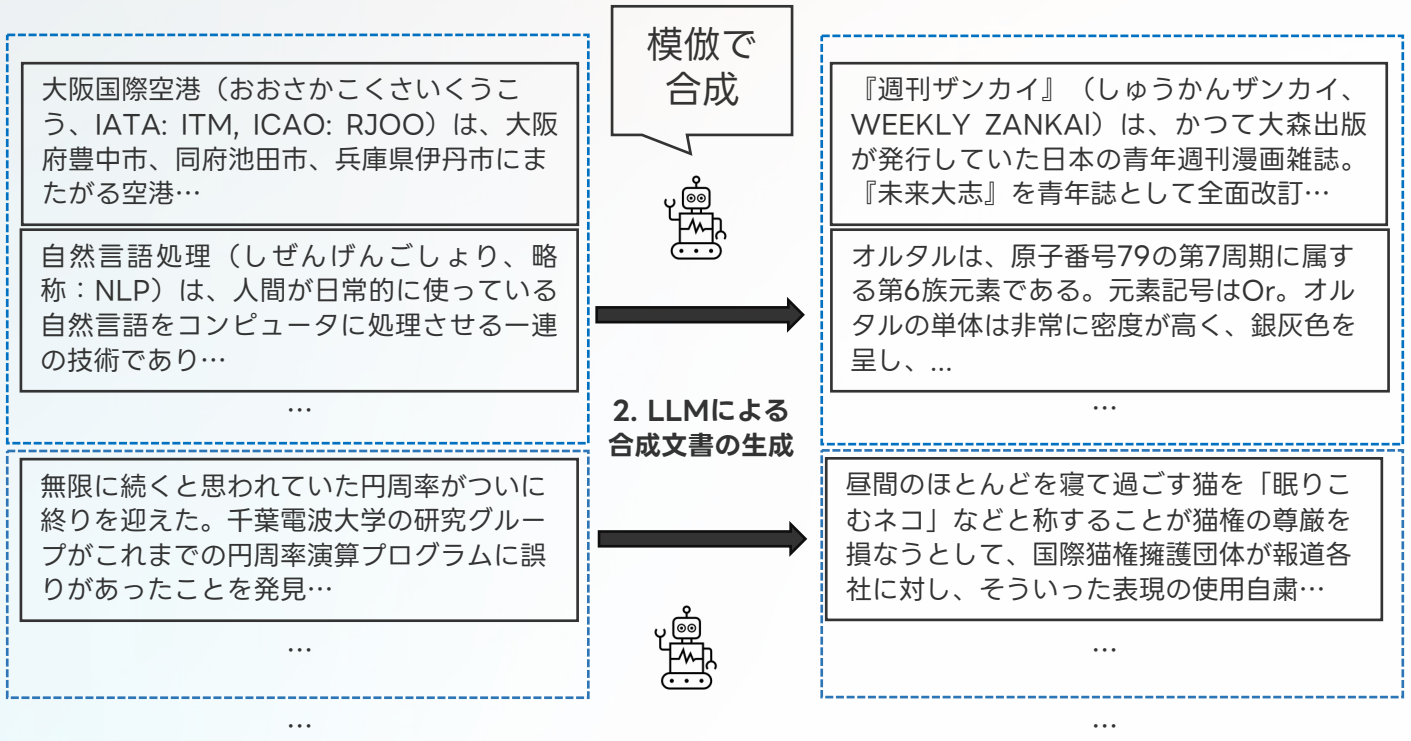
1. 元となる文書の収集
2. LLMによる合成文書の生成
3. 基準を満たさない文書の除去
4. 合成文書に対するQAの作成

データセット	ベース文書ソース	ベース文書例
Wikipedia	日本語Wikipediaの第一段落をエンティティの定義文として抽出	大阪国際空港（おおさかこくさいくうこう、IATA: ITM, ICAO: RJOO）は、大阪府豊中市、同府池田市、兵庫県伊丹市にまたがる空港…
News	ウェブ媒体「虚構新聞」の記事	無限に続くと思われていた円周率がついに終りを迎えた。千葉電波大学の研究グループがこれまでの円周率演算プログラムに誤りがあったことを発見…
ProductReview	価格比較サイト「価格.com」の記事	家で外でも作業の主役！ デル「XPS 13」は小型・高性能な「Copilot+ PC」 デルの高性能ノートパソコンが「XPS」シリーズ。そのなかでも持ち歩きに適した13.4型で、最新のCPUインテル「Core Ultra 7 258V」を搭載する第2世代の「Copilot+ PC」が、…
CompanyRules	「人事労務諸規程モデル集」	第１条　この就業規則（以下「規則」という。）は、株式会社〇〇〇〇（以下「会社」という）の社員の服務規律、労働条件その他の就業に関する事項を定めたものである。 ２．この規則に定めのない事項については、労働基準法その他関係法令の定めるところによる。…

ベンチマークの構築 – 2. LLMによる合成文書の生成

SB Intuitions

- 収集したベース文書をLLMに模倣させ，実在しないエンティティや出来事に関する文書を合成



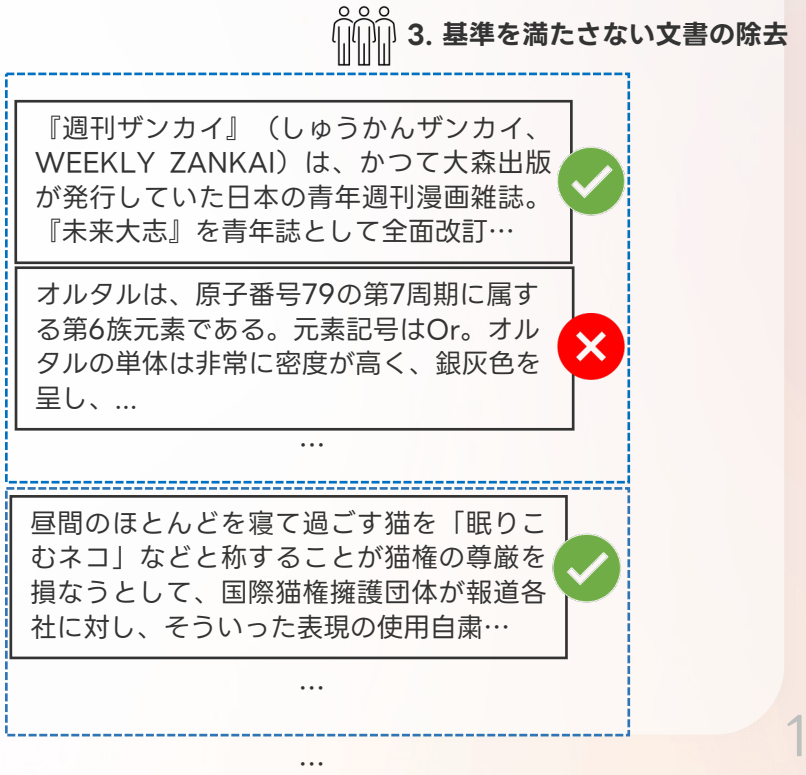
1. 元となる文書の収集
2. LLMによる合成文書の生成
3. 基準を満たさない文書の除去
4. 合成文書に対するQAの作成

ベンチマークの構築 – 3. 基準を満たさない合成文書の除去

SB Intuitions

- 次のいずれかを満たさない文書を人手で除去
  - 実在しないエンティティや出来事を扱っている
  - 事実と矛盾しない
  - 論理展開の破綻・主題と無関係な情報の過度な混入・照応関係の明らかな欠如などを含まない

1. 元となる文書の収集
2. LLMによる合成文書の生成
3. 基準を満たさない文書の除去
4. 合成文書に対するQAの作成



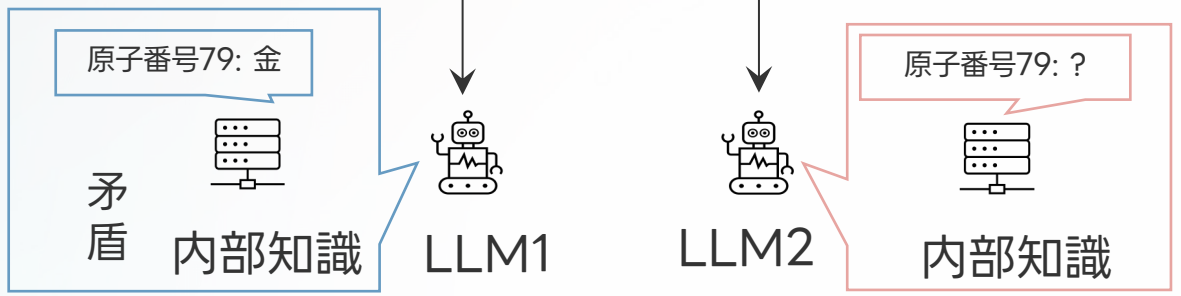
なぜ「事実と矛盾する」文書を除去するのか？

SB Intuitions

実在しないエンティティに関する関連文書

オルタルは、原子番号79の第7周期に属する第6族元素である。元素記号はOr。オルタルの単体は非常に密度が高く、銀灰色を呈し、...

Q: 原子番号79の元素は何ですか？



A: 金

A: オルタル

原子番号79の元素がAuという内部知識の有無により，評価時にLLM1とLLM2の動作の差異が出ている

ベンチマークの構築 – 4. 合成文書に対するQAの作成

SB Intuitions

文書から情報を抽出するのみで回答可能な簡単なQA

- **EXTRACTING**: 質問に対する正解が文書内にほとんどそのまま記載されており，モデルは抽出により答えを導く

抽出した情報をもとに更に推論を行うより高度なQA

- **REASONING**: 文書中に明示されていない情報を，常識的あるいは比較的簡単な論理推論を用いて答えを導く
- **COMPUTING**: 文書中の数値や時間などを加工・計算して答えを導く
- **INFORMATION INTEGRATING**: 文書や文書群内の複数の箇所からの情報を整理することで答えを導く

1. 元となる文書の収集
2. LLMによる合成文書の生成
3. 基準を満たさない文書の除去
4. 合成文書に対するQAの作成

作成した各タイプの質問の例

SB Intuitions

- EXTRACTING: 質問に対する正解が文書内にほとんどそのまま記載されており，モデルは抽出により答えを導ける

[関連文書抜粋] …クアドラ技研の広報担当者は、これまで農家が収穫後に重労働を強いられていた米袋の運搬作業が、このドローンによって一気に効率化すると話す。「空を飛ぶことで、交通の混雑や道路の起伏による影響を受けず、短時間で安全に目的地まで届けることができるのが最大の特長です」と自信を見せた。…

[質問] 米袋ドローンには、農家を収穫後に重労働から解放できる以外に、何のメリットがある？

[正解] 空を飛ぶことで、交通の混雑や道路の起伏による影響を受けず、短時間で安全に目的地まで届けられること

作成した各タイプの質問の例 (cont.)

SB Intuitions

- REASONING: 文書中の情報と文書中に明示されていない情報（常識的あるいは比較的簡単な論理推論）を合わせて答えを導く

+夜に太陽光がない

[関連文書抜粋] また、太陽光電池を装備しており、日中の稼働時間を無制限に延ばせることから、エコフレンドリーな製品としても注目されている…

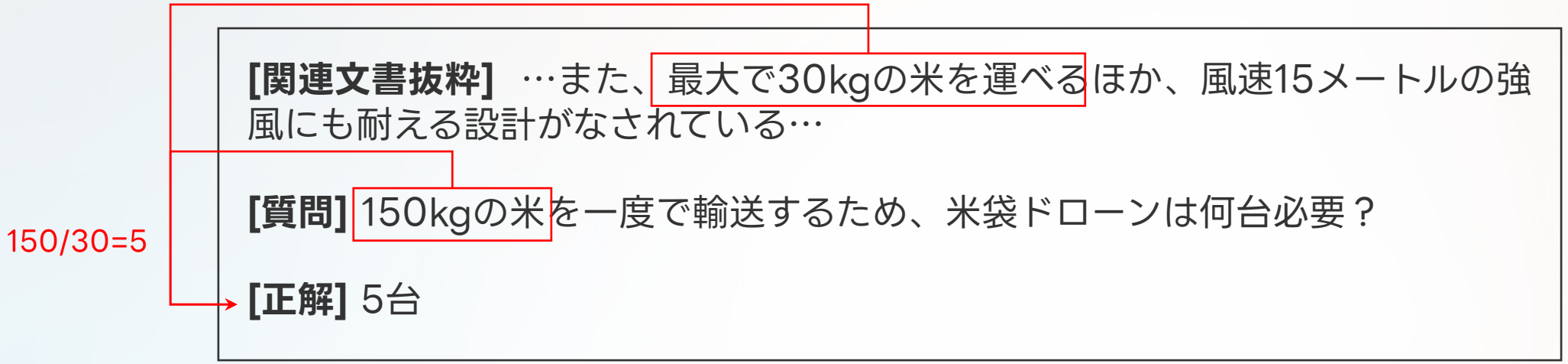
[質問] 米袋ドローンが、夜に無限に動くことができない理由は？

[正解] 太陽光電池で稼働するため、夜には太陽光がなく電池が切れるから

作成した各タイプの質問の例 (cont.)

SB Intuitions

- COMPUTING: 文書中の数値や時間などを加工・計算して答えを導く



作成した各タイプの質問の例 (cont.)

SB Intuitions

- INFORMATION INTEGRATING: 文書や文書群内の複数の箇所からの情報を整理することで答えを導く

[関連文書抜粋] 1993年に月2回刊誌の『ザンカイ』として誕生した。2001年に週刊化される際に誌名も『週刊ザンカイ』に変更され、… 2009年に編集方針の転換を余儀なくされ、『週刊ザンカイ』は一時的に休刊となった…

[質問] 『週刊ザンカイ』は休刊までに何回名前を変更された？

[正解] 1回

データセットの統計と分析

SB Intuitions

各データセットの統計値

データセット	文書長の最小値	文書長の中央値	文書長の最大値
Wikipedia	485	727	1,055
News	445	785	1,150
ProductReview	564	1,379	2,838
CompanyRules	979	1,607	11,478

文書が相対的に短い

文書が相対的に長い

各データセットの質問タイプ分布 (質問が複数タイプに該当する場合があるため，各データセットの割合の和が100%であるとは限らない)

	Wikipedia	News	ProductReview	CompanyRules
#QA	165	173	48	175
EXTRACTING	**95 (57.6%)**	**90 (52.0%)**	13 (27.1%)	**119 (68.0%)**
REASONING	36 (21.8%)	36 (20.8%)	15 (31.2%)	45 (25.7%)
COMPUTING	21 (12.7%)	18 (10.4%)	0 (0.0%)	2 (1.1%)
INFORMATION INTEGRATING	21 (12.7%)	31 (17.9%)	**24 (50.0%)**	11 (6.3%)

INDEX

SB Intuitions

実験の目的

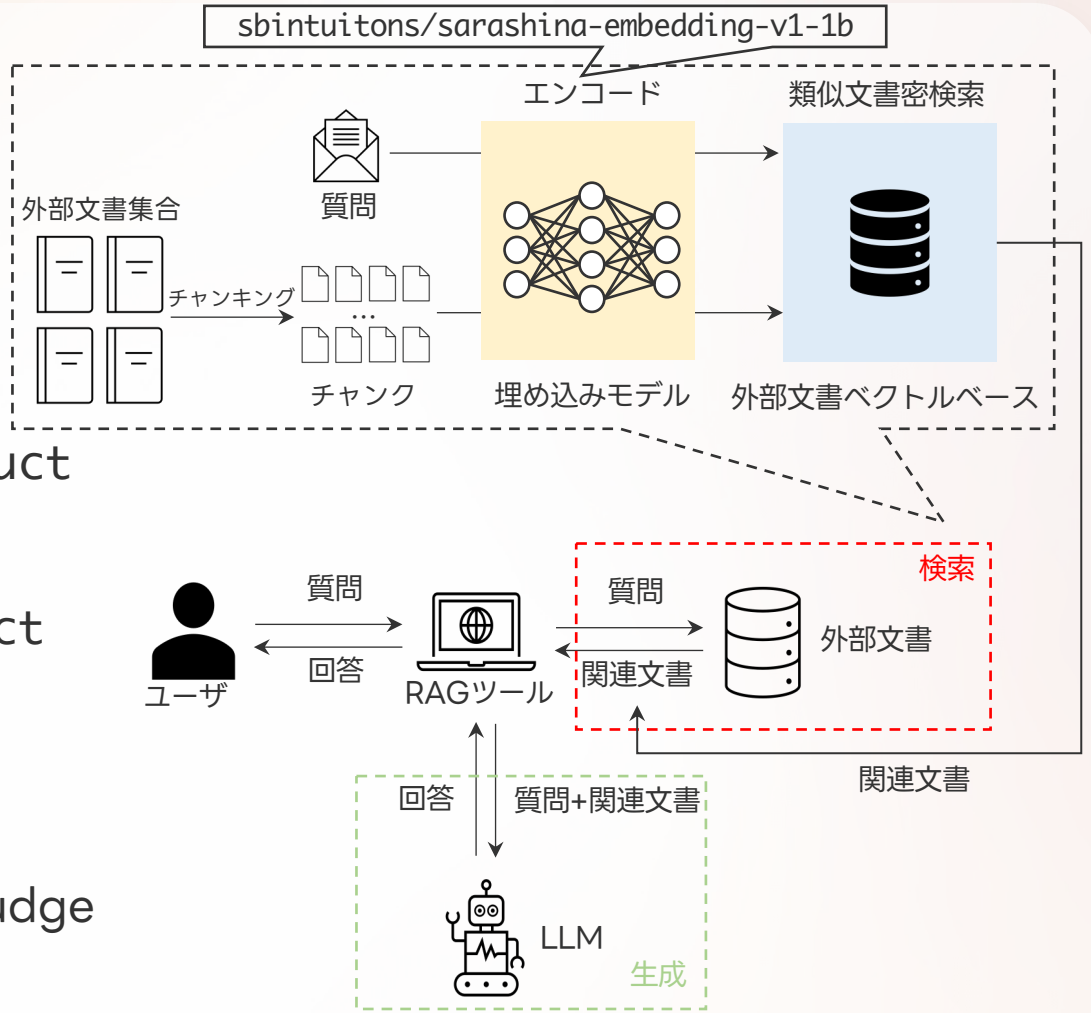
SB Intuitions

- 本研究で構築したベンチマークを用いて，以下の検証を行う
  - 様々なLLMが，本ベンチマークをどのくらい解けているか
  - 本ベンチマークが，LLM内部知識の評価にならず，外部情報活用の能力への評価になっているか
    - 具体的には，検索文書を与えない場合に解けないベンチマークになっているか

実験設定

SB Intuitions

- 評価対象LLM
  - OpenAI/GPT-4o-2024-08-06 (GPT-4o)
  - tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3 (Swallow-8B)
  - llm-jp/llm-jp-3-{1.8b, 3.7b, 13b}-instruct (llm-jp-{1.8B, 3.7B, 13B})
  - Qwen/Qwen2.5-{1.5B, 3B, 7B, 14B}-Instruct (Qwen-{1.5B, 3B, 7B, 14B})
- 評価器
  - OpenAI/GPT-4o-2024-08-06によるLLM-as-a-judge
  - 正解と一致している・一致していないと二値判定



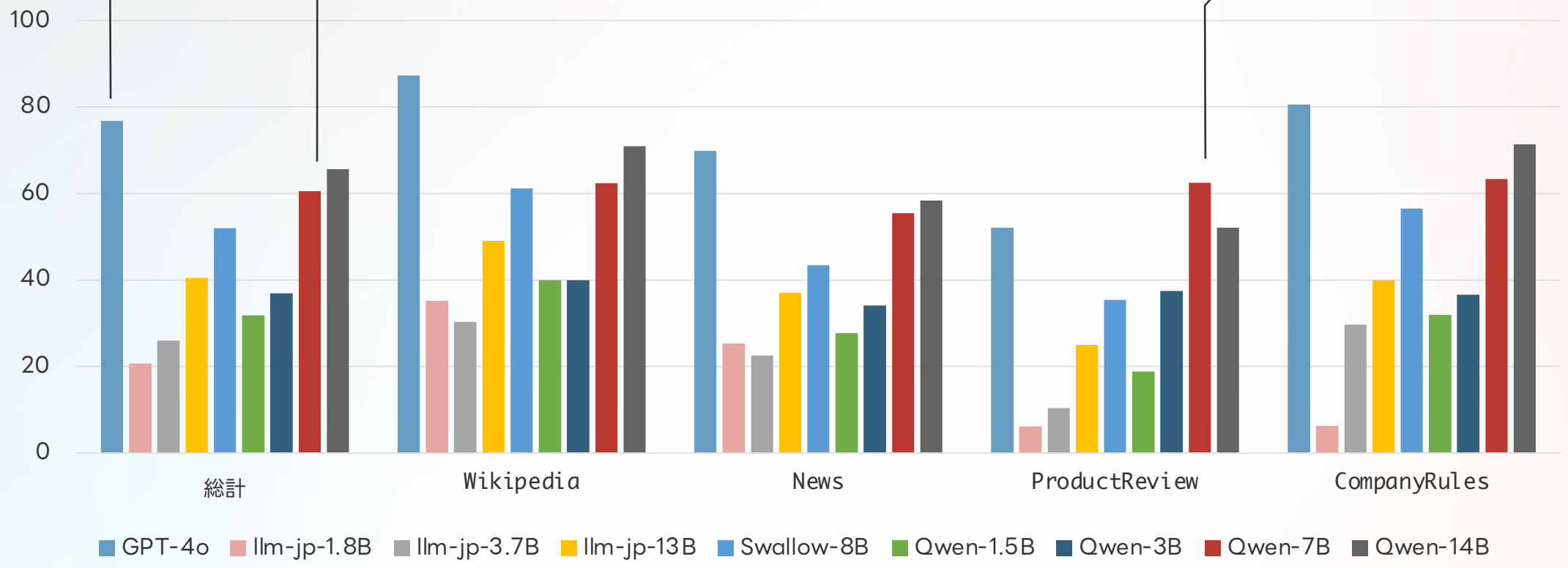
全体の評価結果

SB Intuitions

全体的には，GPT-4oが一番高い性能を示した

Qwen-14BとQwen-7Bが軽量でありながら優れた性能を示している

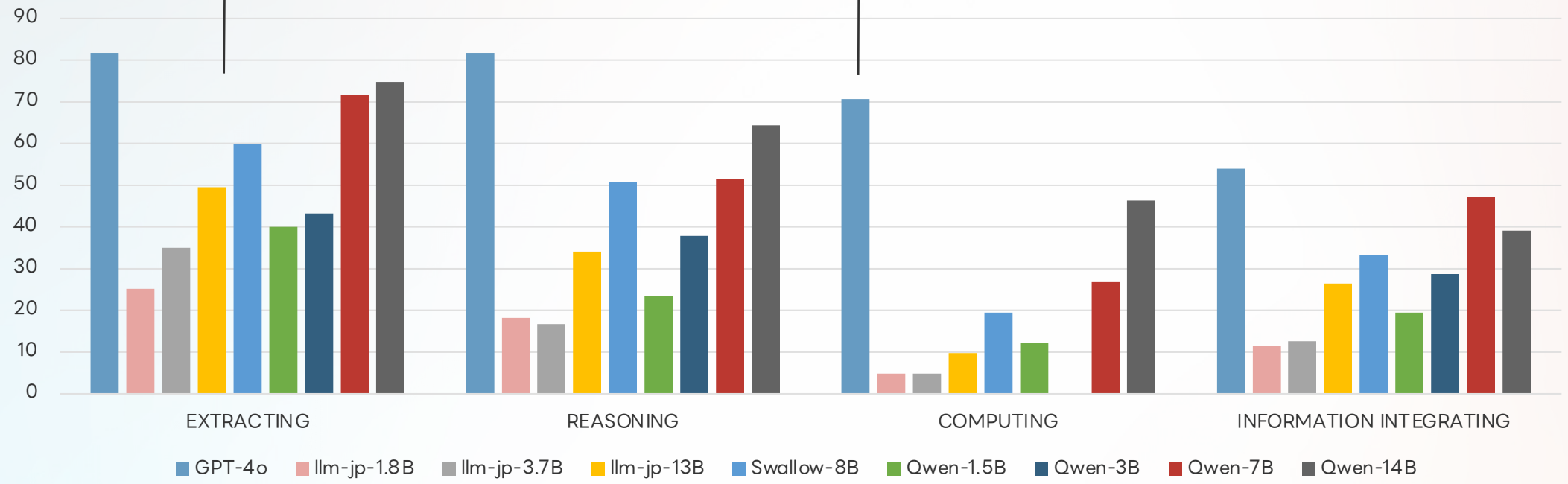
Qwen-14BとQwen-7BがProductReviewではGPT-4oと同等以上の性能



質問タイプ別の評価結果

どのLLMもEXTRACTINGタイプにおいて，4つの質問タイプの中で最も高いスコアを示した

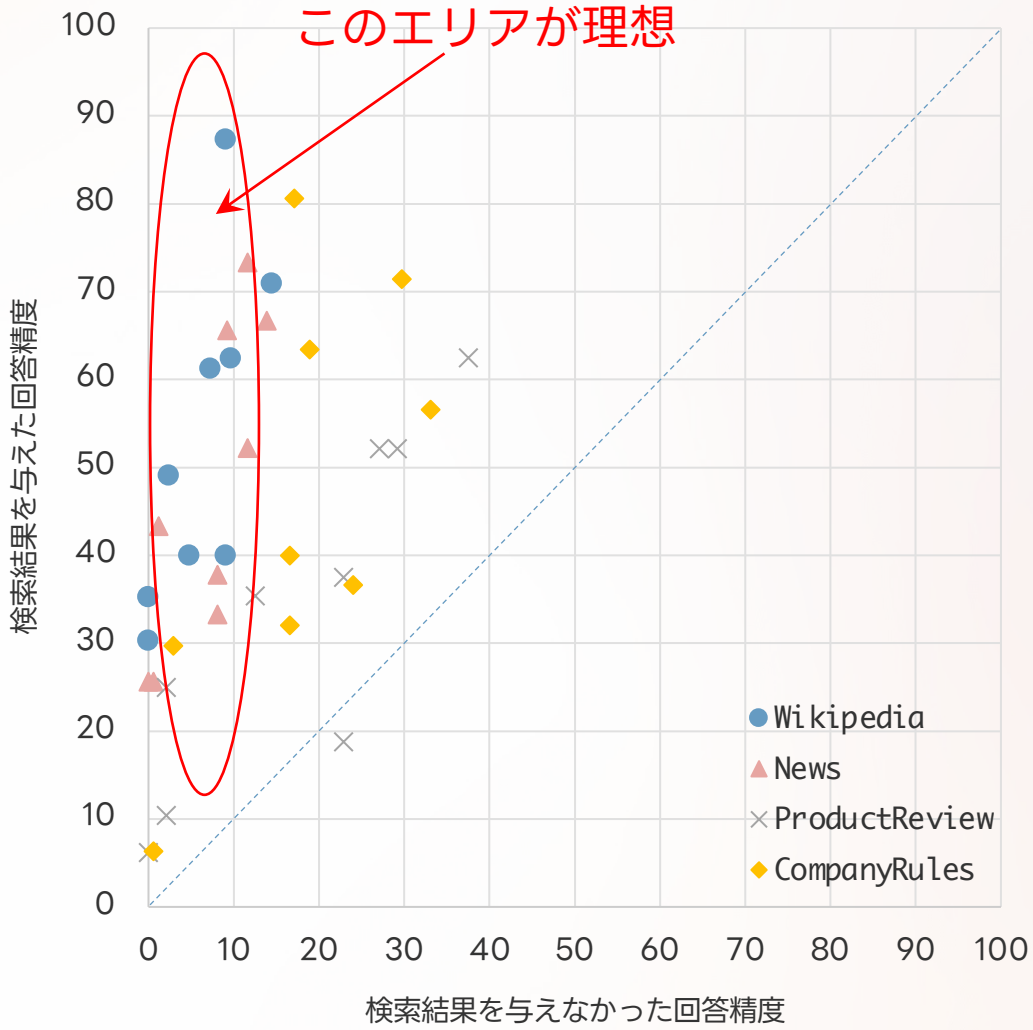
COMPUTINGにおいてはGPT-4oが他のLLMを特に大きく上回った



検索結果有無による比較分析

SB Intuitions

- 同じ質問を以下2つの方法で検証
  - RAGなし: 検索結果を与えず回答を求める
  - RAGあり: 検索結果を与え回答を求める
- 本ベンチマークは外部情報活用能力を測る目的で構築
  - 理想:外部情報必須なため，RAGなしで正解率は0
- 実際は二者択一質問があり偶然正解の可能性あり (チャンスレート) 分析の結果，偶然正解以外はほぼ無し
- 結論: 全体として「検索結果がないと回答できない」は概ね達成



ケース分析

SB Intuitions

2005年年始

2012年年末

[関連文書抜粋] 『レヴェナント・クロニクル』（Revenant Chronicle）は、井端圭史による日本の漫画作品である。...『週刊クリエイトタイムズ』（晴耕社）の創刊号である2005年1号から2012年50号まで連載され、 ...
[質問] 『レヴェナント・クロニクル』は『週刊クリエイトタイムズ』で約何年間連載された？
[正解] 約8年間
[GPT-4oによる回答] 約8年間
[Qwen-14Bによる回答] 7年間

2012-2005=7

外部情報を活用する能力において，GPT-4oが計算，推論，情報の整理に優れた性能を持つと評価ができる

「川が国の東にある」のであれば「国が川の西にある」

[関連文書抜粋] アルフォニアは、バレテス海南岸に位置しており、北はケリウム山脈を挟んでベルタニアと接し、南はエヴァラン共和国と長い国境線を共有している。東はアンデス川が自然の境界となっており、それを越えるとテレンティア王国が広がっている。...
[質問] アルフォニアはアンデス川の西岸と東岸のどちらにあるか？
[正解] 西岸
[GPT-4oによる回答] 東岸
[Qwen-14Bによる回答] アンデス川の東岸にある。

GPT-4oもQwen-14Bも情報の抽出ができているとみられるが，上記の推論ができていない

INDEX

SB Intuitions

結論と今後の展望

SB Intuitions

- 本研究ではRAGベンチマークを構築
  - 従来のRAGベンチマークはLLMが学習済みの可能性が問題になる
  - RAG関連文書に基づく回答生成をより正確に評価するため，実在しないエンティティや出来事をLLMに合成
  - 関連文書に対するQAは人手で作成
  - 4種類のデータセットを構築
- 構築ベンチマークでLLMのRAG性能を評価
  - 外部情報の抽出能力
  - 抽出情報を用いた推論能力
- Web公開はリークの可能性があり，公開方法は検討

直感を、知性へ