

第265回 自然言語処理研究発表会

# J-RAGBench: 日本語RAGにおける Generator評価ベンチマークの構築

板井 孝樹 長谷川 駿一 山本 勇太 峰岸 剛基 大槻 真輝

株式会社 neoAI

## 目的

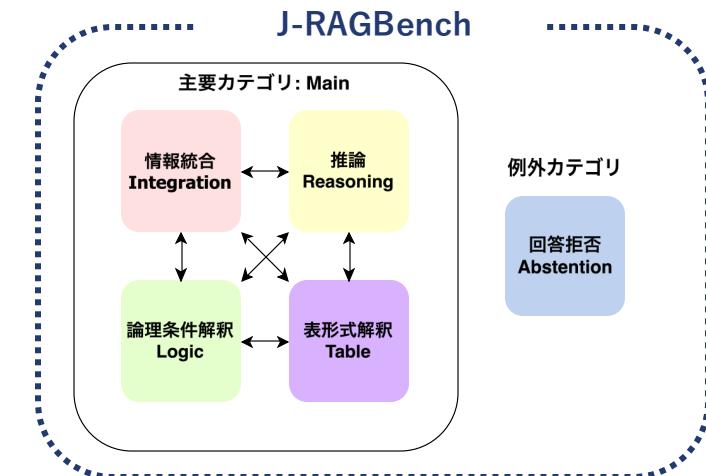
RAGのGeneratorの多様な能力を、同一条件下で体系的に評価可能な枠組みの構築

## 発表内容

RAGのGenerator評価ベンチマーク **J-RAGBench** の提案

## 貢献

- Retrieverに依存しない条件での評価により、**Generatorの純粋な外部文書参照能力の評価をするベンチマーク**を構築
- RAGの実運用に近い設計で、**Generatorの能力観点別の横断比較**を実現  
⇒ 実運用時のモデル選定指針・RAG特化モデル構築の指標を提供



# Index

- 01 背景・目的**
- 02 提案手法：J-RAGBench**
- 03 実験**
- 04 分析**
- 05 まとめ**

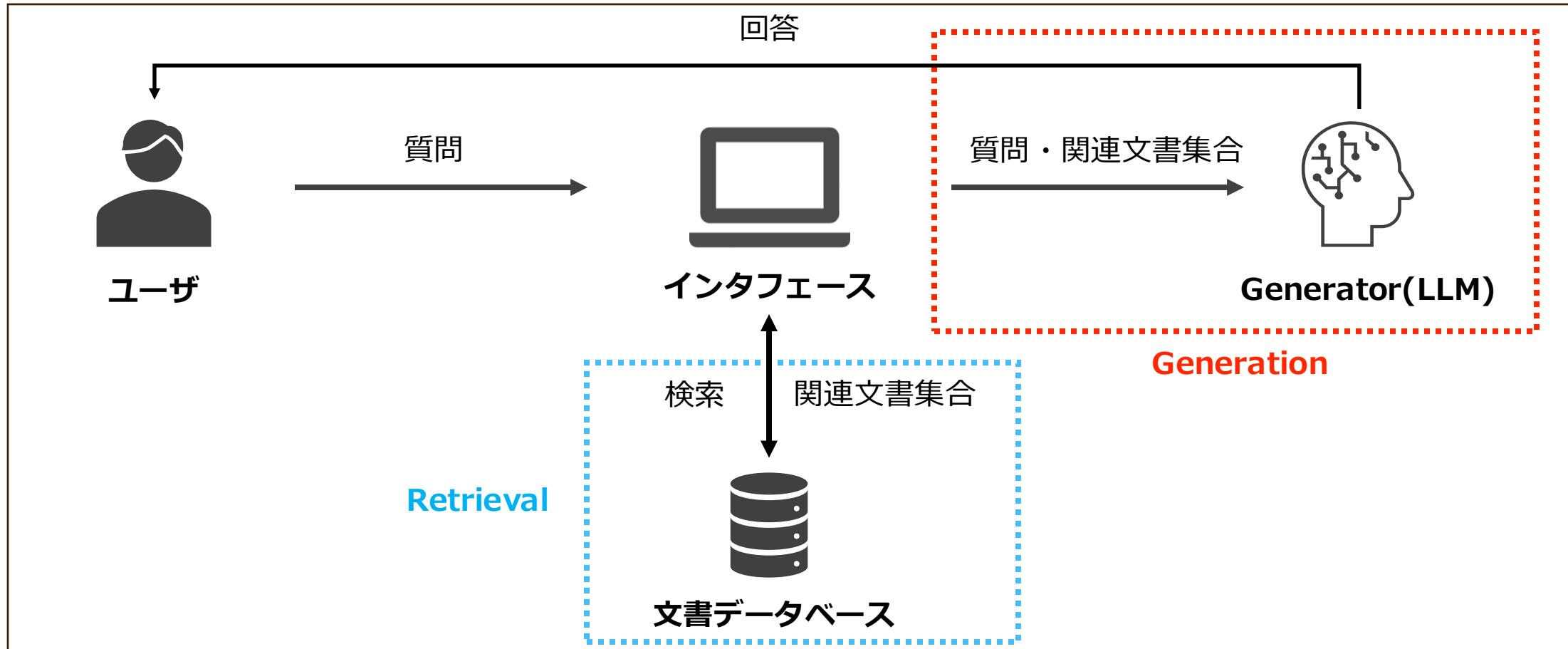
# Index

- 01 背景・目的**
- 02 提案手法：J-RAGBench**
- 03 実験**
- 04 分析**
- 05 まとめ**

# 検索拡張生成：RAG (Retrieval Augmented Generation)

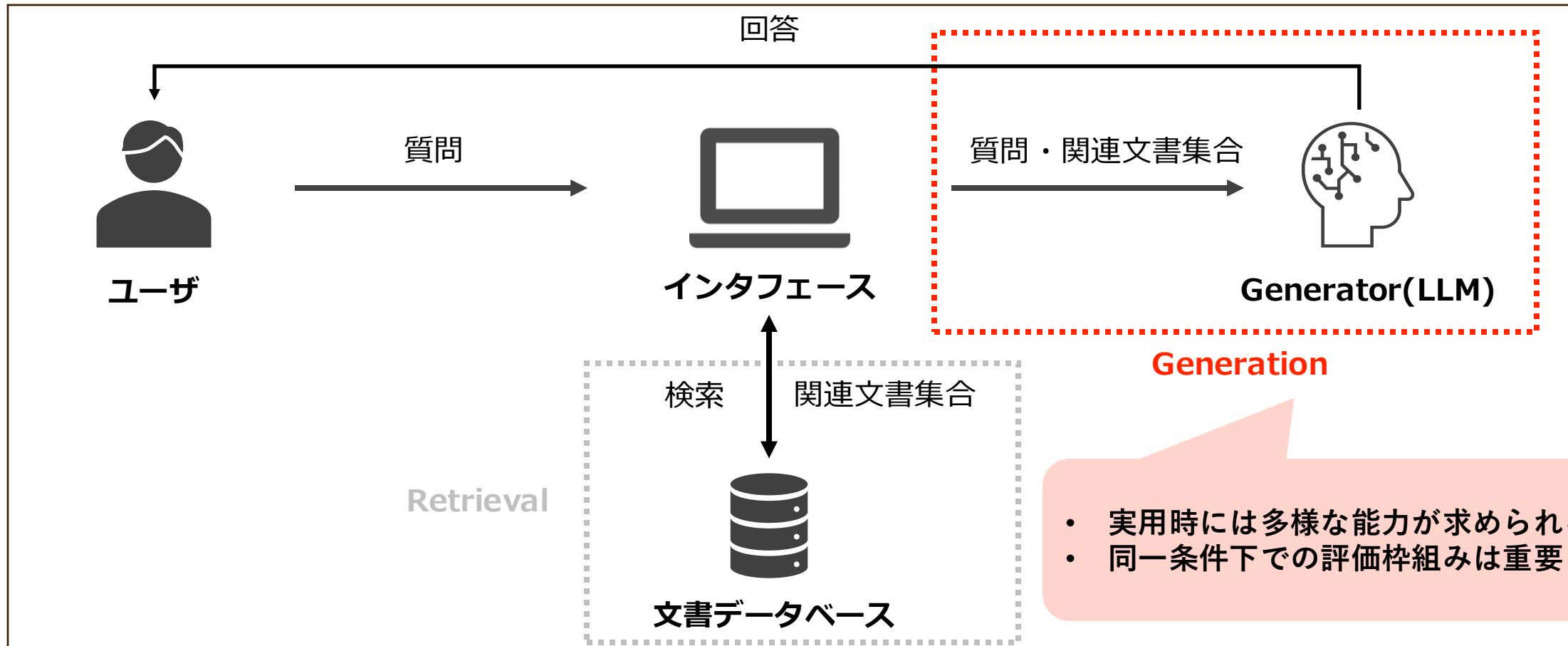
Retrieval : 外部文書集合から検索器により関連文書集合を取得

Generation : LLMが関連文書集合に基づき質問に対する回答を生成



# 検索拡張生成：RAG (Retrieval Augmented Generation)

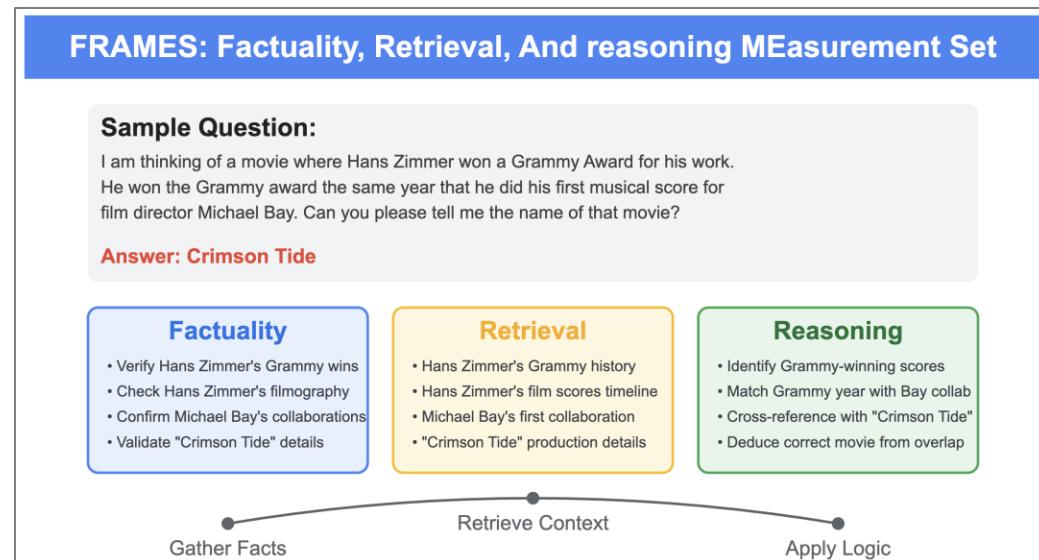
Retrieval : 外部文書集合から検索器により関連文書集合を取得  
Generation : LLMが関連文書集合に基づき質問に対する回答を生成



# 既存研究：Generator 評価ベンチマーク（英語）

## FRAMES

事実性・検索能力・推論の3観点を評価



## RGB

ノイズ頑健性・情報統合・反事実頑健性・回答拒否の4観点を評価

### ノイズ頑健性

#### Noise Robustness

**Question**  
Who was awarded the 2022 Nobel prize in literature?

**External documents contain noises**  
The Nobel Prize in Literature for 2022 is awarded to the French author Annie Ernaux, "for the courage and clinical acuity ..."

The Nobel Prize in Literature for 2021 is awarded to the novelist Abdulrazak Gurnah, born in Zanzibar and active in ...

**Retrieval Augmented Generation**  
↓  
Annie Ernaux

#### Negative Rejection

**Question**  
Who was awarded the 2022 Nobel prize in literature?

**External documents are all noises**  
The Nobel Prize in Literature for 2021 is awarded to the novelist Abdulrazak Gurnah, born in Zanzibar and active in ...

The 2020 Nobel laureate in Literature, poet Louise Glück, has written both poetry and essays about poetry. Since her...

**Retrieval Augmented Generation**  
↓  
I can not answer the question because of the insufficient information in documents

### 情報統合

#### Information Integration

**Question**  
When were the ChatGPT app for iOS and ChatGPT api launched?

**External documents contain all answers**  
On May 18th, 2023, OpenAI introduced its own ChatGPT app for iOS...

**Retrieval Augmented Generation**  
↓  
May 18 and March 1.

#### Counterfactual Robustness

**Question**  
Which city hosted the Olympic games in 2004?

**Counterfactual external documents**  
The 2004 Olympic Games returned home to New York, birthplace of the ...

After leading all voting rounds, New York easily defeated Rome in the fifth and final vote ...

**Retrieval Augmented Generation**  
↓  
There are factual errors in the provided documents. The answer should be Athens.

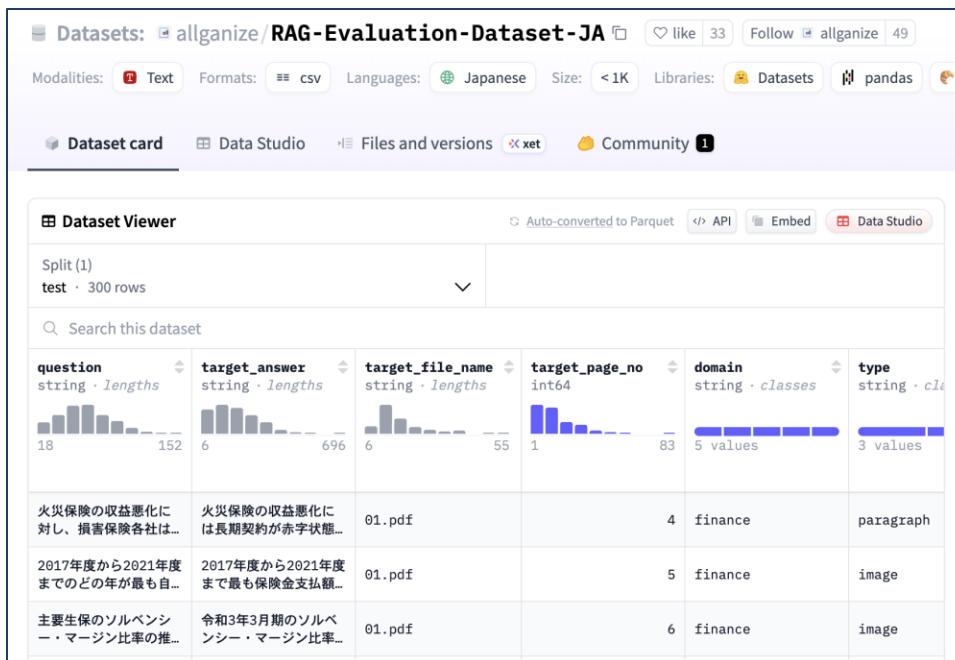
### 回答拒否

### 反事実頑健性

# 既存研究：Generator 評価ベンチマーク（日本語）

## Allganize RAG Leaderboard

図表を含む実ユースケースを想定したQA



## 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

架空シナリオに基づくQAによりLLMの事前学習  
で得た知識による回答を排除  
⇒ Generatorの純粹な外部文書参照能力を評価

リークの可能性を考慮してデータは非公開

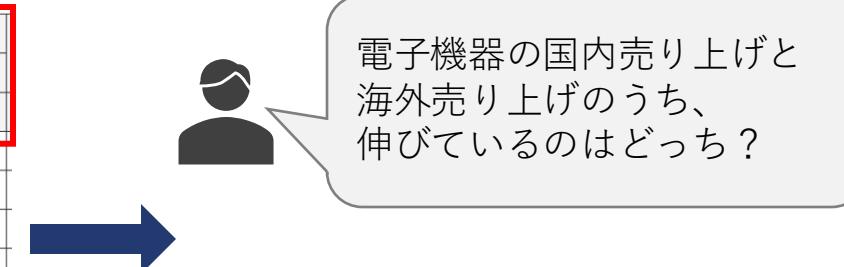
SB Intuitions

# 実運用時のRAGで頻出する誤答パターン

複雑な表を正しく解釈できない

部門	四半期	売上および利益(百万円)					
		国内		海外		合計	
		売上	利益	売上	利益	売上	利益
電子機器	Q1	3000	900	2000	600	5000	1500
	Q2	3200	950	2100	650	5300	1600
	Q3	3400	1000	2200	700	5600	1700
	Q4	3600	1050	2300	750	5900	1800
家電	Q1	1800	500	1200	400	3000	900
	Q2	1900	550	1300	450	3200	1000
	Q3	2000	600	1400	500	3400	1100
	Q4	2100	650	1500	550	3600	1200

表の読み取り × マルチホップ推論



抽出した情報 ⇒ 多段階の推論を要する

大規模な表・結合セルなど

情報抽出から発展したタスクなど

答えられないのに答えてしまう

入力コンテキスト



Generator



根拠情報が存在しない場合は  
素直に回答を拒否してほしい

# 研究目的

## 課題意識

実運用時は、Generatorには以下が求められる

- より多様な能力 (ex. 表の読み取り, 適切な回答拒否など)
- 同時に複数の能力を求められるユースケースの対応

## 既存手法の課題

- ✓ 評価する能力が限定的
- ✓ 複数能力を同時かつ網羅的に評価する枠組みの整備は不十分

## 研究目的

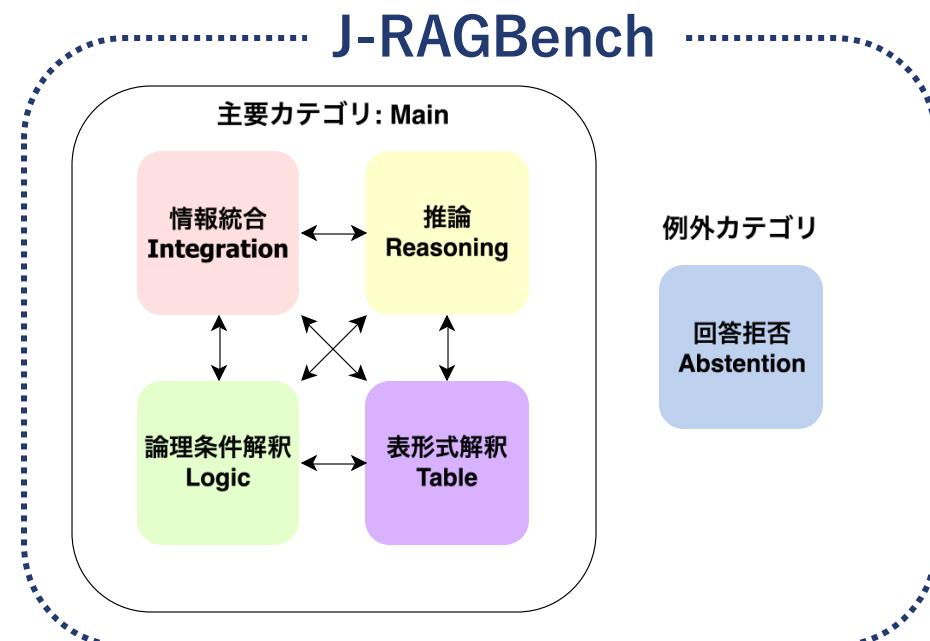
- 実運用時に得た課題やユースケースを加味した**能力観点の整理・体系化**
- **複数観点が共起**する場合を含めた評価枠組みの整備

# Index

- 01 背景・目的
- 02 提案手法：J-RAGBench
- 03 実験
- 04 分析
- 05 まとめ

## J-RAGBench (Japanese RAG Generator Benchmark)

- 5つの「評価カテゴリ」を定義：金融・製造業などの多様な業界にRAGシステムを導入する過程で直面した課題・実運用時のユースケースを反映
- 「評価カテゴリ」ごとにユースケースを想定した「評価観点」に細分化
- 架空シナリオに基づくQAで、評価観点2種までの全組み合わせを網羅する設計
- HuggingFace Hubにて公開：neoai-inc/Japanese-RAG-Generator-Benchmark



The screenshot shows the HuggingFace Dataset card for 'Japanese-RAG-Generator-Benchmark'. The card displays various details about the dataset, including its tasks (Question Answering), languages (Japanese), and size (1K<n<10K). It features a 'Dataset Viewer' section with a histogram showing the distribution of question lengths and answer lengths. Below the viewer, there are several examples of dataset entries, each consisting of a question, an answer, and a detailed description of the context or source. The interface includes tabs for 'Dataset card', 'Files and versions', 'Community', and 'Settings'.

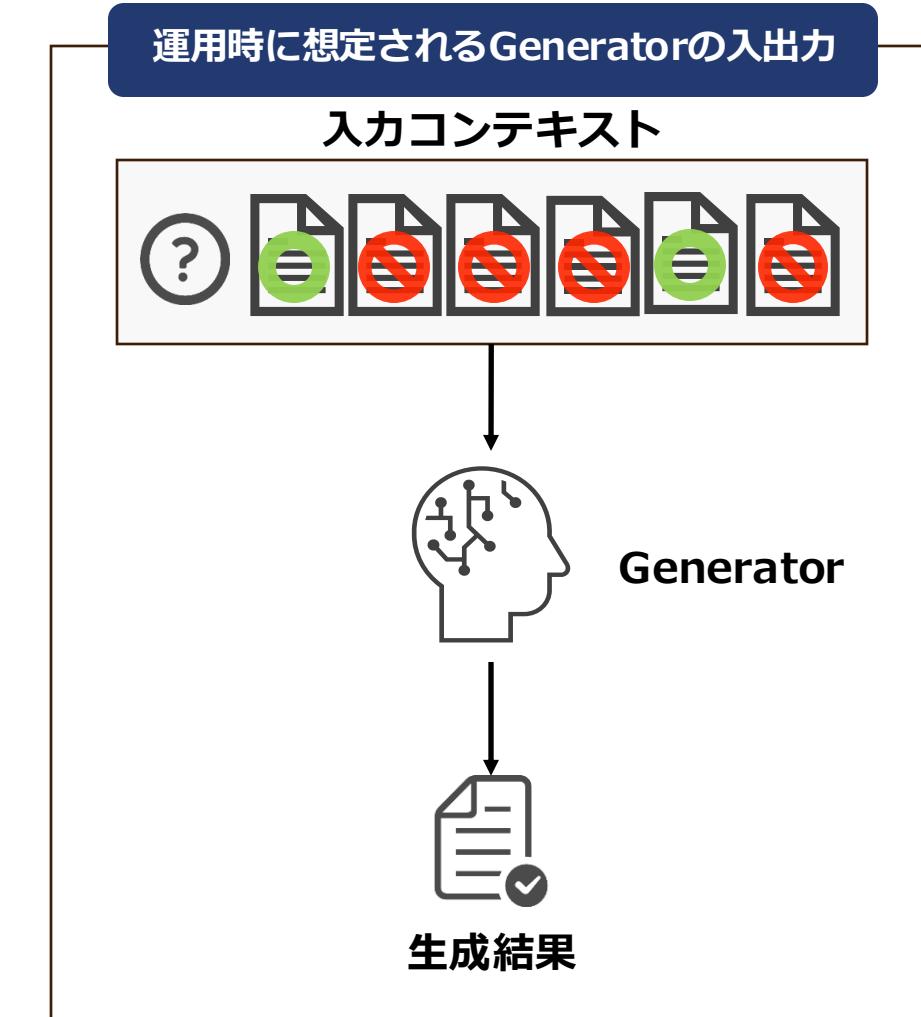
## ベンチマークの構成

評価データセットの各QAは以下で構成

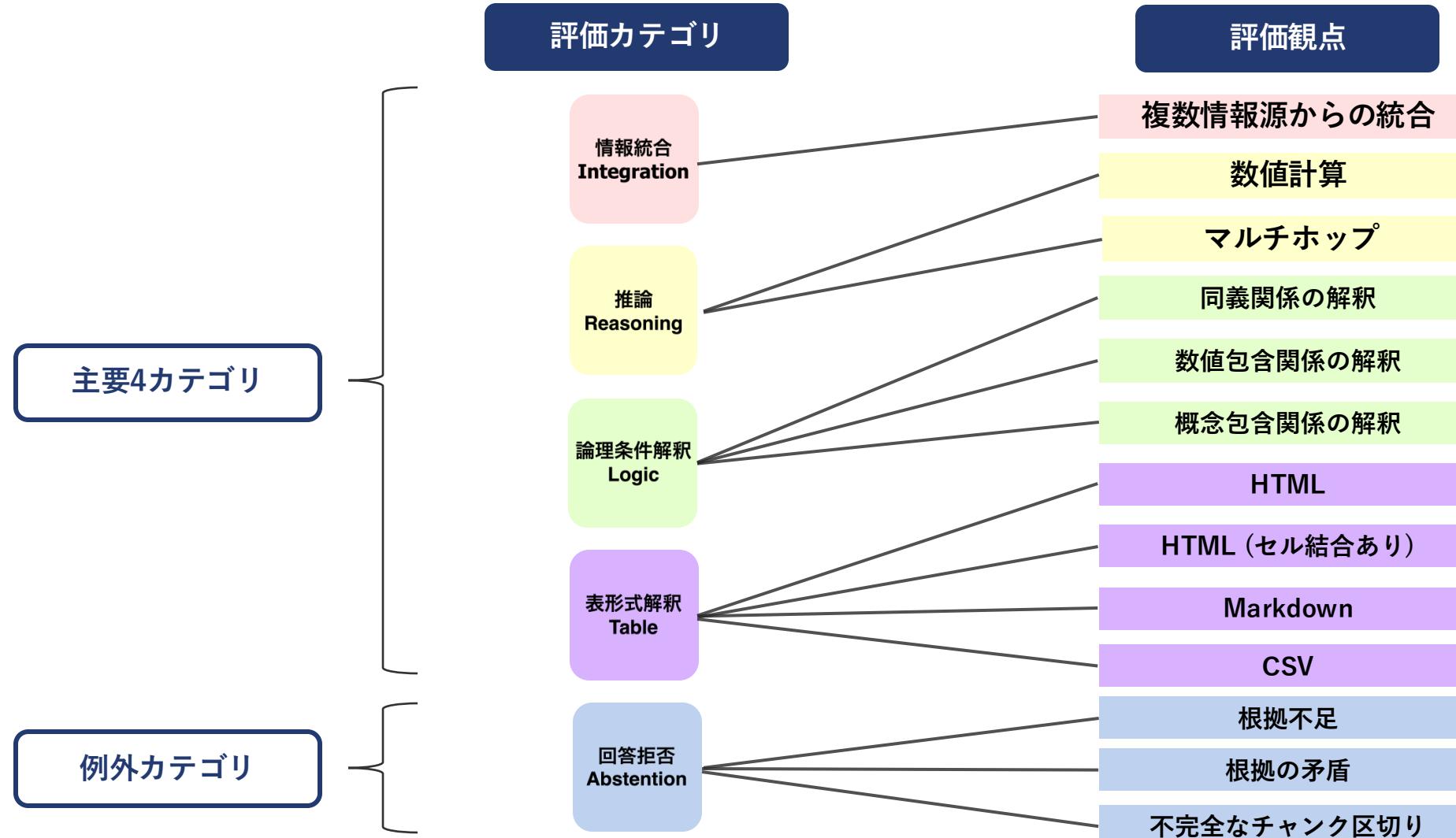
評価データセットの構成要素

構成要素	説明
評価観点集合	QAの評価観点 (1~2つ)
質問	評価観点集合に基づく質問
関連文書集合	<ul style="list-style-type: none"> <li>質問の根拠となる文書集合</li> <li>チャンク単位で分割済み</li> </ul>
非関連文書集合	<ul style="list-style-type: none"> <li>質問の根拠にならない文書集合</li> <li>質問との意味的類似度が高いor キーワードが含まれる文書</li> <li>チャンク単位で分割済み</li> <li>6~8チャンク</li> </ul>
正答	質問に対する正答

※ 1チャンク: 512トークン程度(tiktokenを用いて計測)



## J-RAGBench の評価カテゴリ・評価観点



## 主要カテゴリ①：情報統合 (Integration)

### 概要

根拠情報が複数の文書に分散 ⇒ 抽出・統合して回答する能力

- 複数の情報源からの統合 : 2 ~ 3文書からの情報源の統合を対象としたQA

#### 例：複数の情報源からの統合



Vertex Sky Digital 社とNimbus Digital 社の  
デザイナー職の新卒の月給はいくら？



【Vertex Sky Digital  
新卒採用要項】  
...  
デザイナー職34万円



【VSD社 中途採用要  
項】採用情報 募集要項  
担当いたく業務概要  
...



【初任給】 初任給（しょ  
にんきゅう）は、学校を  
卒業して正規雇用される  
ようになった人が ...



【Nimbus Digital 新卒  
採用要項】  
アニメータ/デザイナー:  
月額基本給：26 万円



Vertex Sky Digital 社は34万円,  
Nimbus Digital 社は26万円です。

### J-RAGBench

#### 主要カテゴリ: Main

情報統合  
**Integration**

推論  
**Reasoning**

論理条件解釈  
**Logic**

表形式解釈  
**Table**

#### 例外カテゴリ

回答拒否  
**Abstention**

## 主要カテゴリ②：推論 (Reasoning)

### 概要

**抽出された情報を踏まえて、多段階推論や数値計算などを実行する能力**

- **数値計算:** 四則演算・利益率等の指標計算
- **マルチホップ推論:** 直接的な記載がない結論を導く (関連研究: HotpotQA, JEMHopQA)

### 例：マルチホップ推論



映画「永遠の風見鶏」の主演を務めた俳優の妻は？



【永遠の風見鶏】永遠の風見鶏は、（えいえんのかざみどり）は、2018年に公開された日本の映画… **主演: 鈴木陽一。**



【鈴木陽一】鈴木陽一（すずき よういち、昭和60年6月15日-）は、日本の俳優。2012年に女優の**新木春菜**氏と結婚



新木春菜です。



### J-RAGBench

#### 主要カテゴリ: Main

情報統合  
Integration

推論  
Reasoning

論理条件解釈  
Logic

表形式解釈  
Table

#### 例外カテゴリ

回答拒否  
Abstention

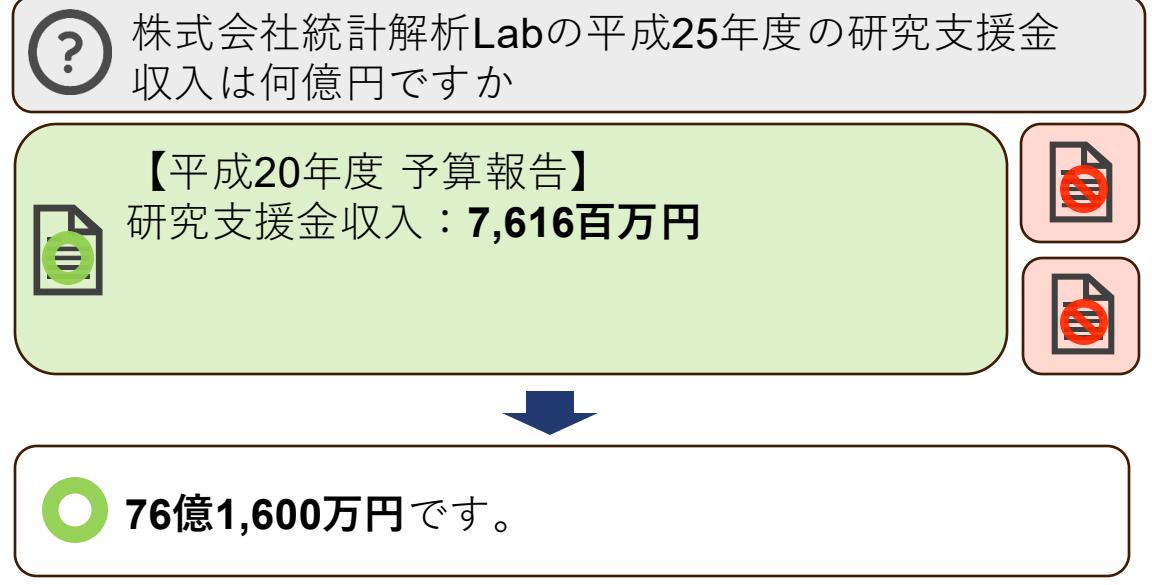
## 主要カテゴリ③：論理関係解釈 (Logic)

**概要**

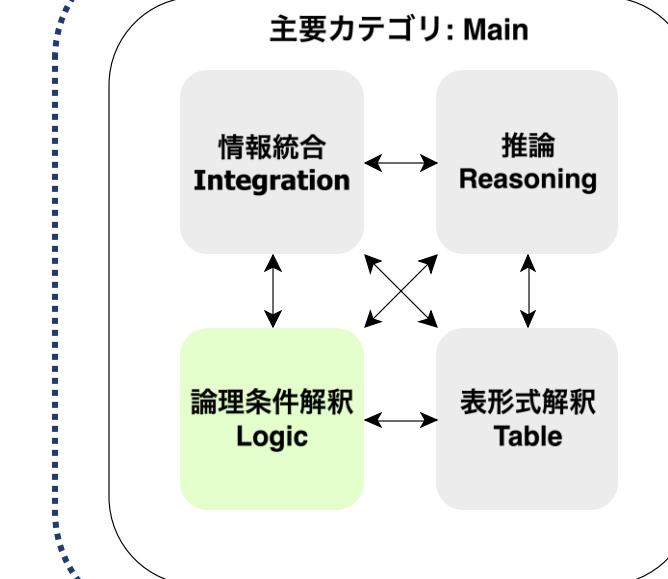
質問・関連文書間での語彙や表現の差異を解釈し情報を抽出・回答する能力

- ▶  **同義関係の解釈** : 質問と関連文書の間で異なる表現を同義であると解釈
- 数値包含関係の解釈** : 質問と関連文書の間での数値的な包含関係を解釈
- 概念包含関係の解釈** : 質問と関連文書の間での概念的な包含関係を解釈

### 例：同義関係の解釈



### J-RAGBench



例外カテゴリ  
回答拒否  
Abstention

## 主要カテゴリ③：論理関係解釈 (Logic)

**概要**

質問・関連文書間での語彙や表現の差異を解釈し情報を抽出・回答する能力

□ 同義関係の解釈 : 質問と関連文書の間で異なる表現を同義であると解釈

▶ □ 数値包含関係の解釈 : 質問と関連文書の間での数値的な包含関係を解釈

□ 概念包含関係の解釈 : 質問と関連文書の間での概念的な包含関係を解釈

例：数値含関係の解釈



27歳で2件の社内プロジェクト提案経験を有する  
佐藤主任研究員は参加資格を満たすか？



【参加資格】

年齢: **満24歳以上**

社内プロジェクト提案経験: **3件以上**

...



参加資格を満たさない。参加には社内プロジェクト提案経験が3件以上必要であるが、2件で満たさない。

J-RAGBench

主要カテゴリ: Main

情報統合  
Integration

推論  
Reasoning

論理条件解釈  
Logic

表形式解釈  
Table

例外カテゴリ

回答拒否  
Abstention

## 主要カテゴリ③：論理関係解釈 (Logic)

### 概要

**質問・関連文書間での語彙や表現の差異を解釈し情報を抽出・回答する能力**

- 同義関係の解釈 : 質問と関連文書の間で異なる表現を同義であると解釈
- 数値包含関係の解釈 : 質問と関連文書の間での数値的な包含関係を解釈

### ▶ □ 概念包含関係の解釈 : 質問と関連文書の間での概念的な包含関係を解釈

#### 例：概念包含関係の解釈

?(A325室でオンライン会議を行うことは可能ですか？)

【A325室禁止事項】…電子機器の持ち込み  
及び使用は厳禁である。



○ A325室では電子機器の使用が禁止されており、  
オンライン会議を行うことはできません。

### J-RAGBench

#### 主要カテゴリ: Main

情報統合  
**Integration**

推論  
**Reasoning**

論理条件解釈  
**Logic**

表形式解釈  
**Table**

#### 例外カテゴリ

回答拒否  
**Abstention**

## 主要カテゴリ④：表形式解釈 (Table)

概要

関連文書の根拠の記述が表形式 ⇒ 根拠の情報を解釈・抽出する能力

HTML形式

Markdown形式

HTML形式(セル結合あり)  CSV形式

例 : Markdown形式



2019年から2021年にかけて、グローバリンク  
社の海外支店からの帰任数は増加したか？



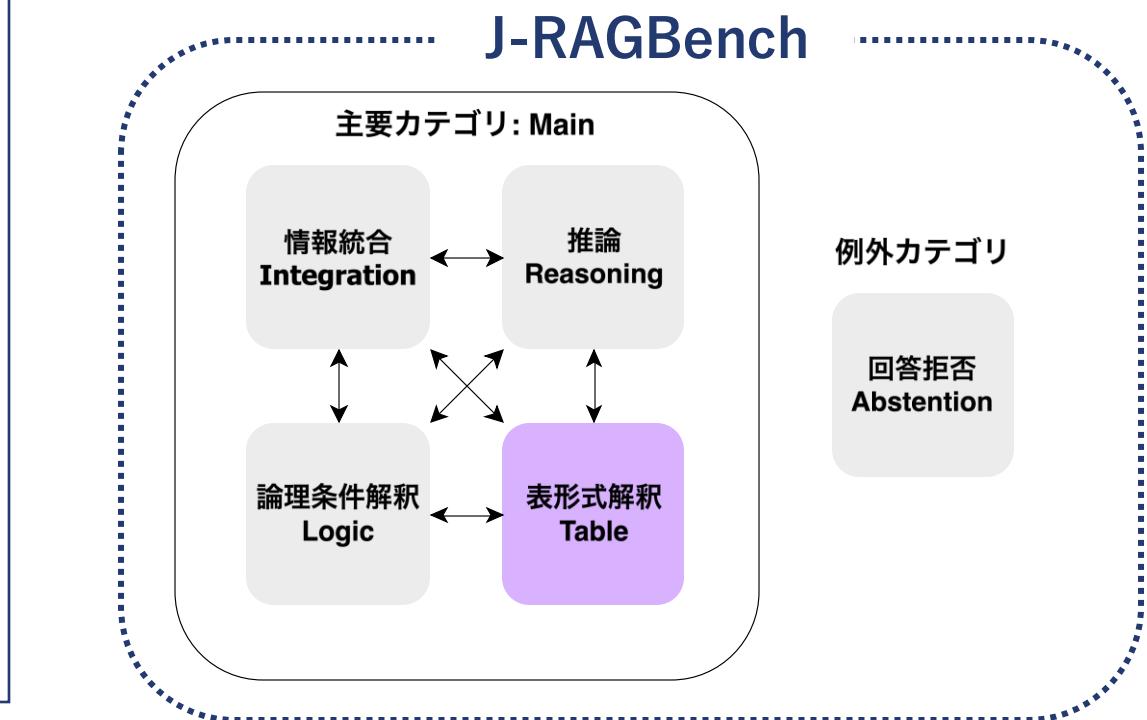
【グローバリンク社 帰任者数の推移】

本レポートでは、…

期間	2017	2018	2019	2020	2021
総数(人)	1,788	1,891	2,003	368	50



いいえ、2,003人から50人へ減少しています。



## 例外カテゴリ：回答拒否 (Abstention)

### 概要

回答不可である特定の状況下において、適切に回答を拒否する能力

▶ □ **根拠不足**: 関連文書集合が得られなかった場合

□ **根拠の矛盾**: 根拠になり得る情報が複数存在し、それらが矛盾している

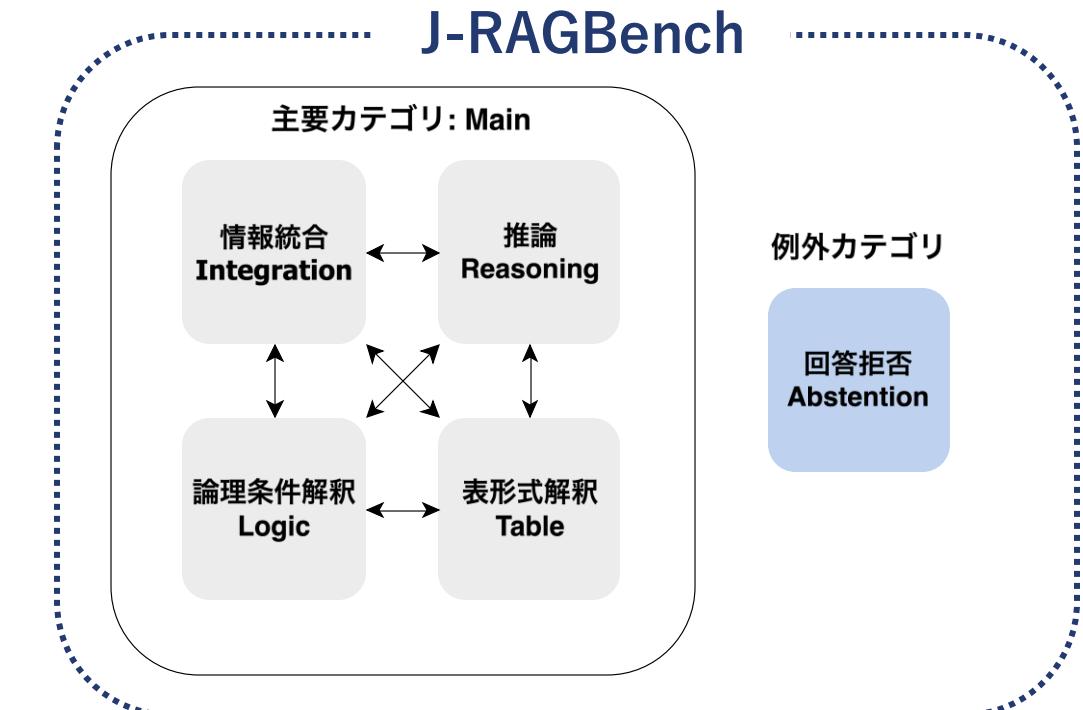
□ **不完全なチャック区切り**: 根拠がチャック分割の境界に存在 ⇒ 不完全な根拠情報

**例：根拠不足**

?(?) 2019年から2021年にかけて、グローバリンク  
社の海外支店からの帰任数は増加したか？

四つの文書アイコン（各に赤い斜線）

与えられた文書内にグローバリンク社に関する  
情報が無いため、回答できません。



## 例外カテゴリ：回答拒否 (Abstention)

### 概要

回答不可である特定の状況下において、適切に回答を拒否する能力

□ 根拠不足：関連文書集合が得られなかった場合

▶ □ 根拠の矛盾：根拠になり得る情報が複数存在し、それらが矛盾している

□ 不完全なチャunk区切り：根拠がチャunk分割の境界に存在 ⇒ 不完全な根拠情報

### 例：根拠の矛盾



新人研修の参加に必要な統一英語力試験のスコアは？



【新人研修 2021】参加に資する英語力の指標として、統一英語力試験650点以上の取得が必須…



【研修 概要】参加にあたっては、統一英語力試験550点以上が条件となっている…



根拠に当たる情報が複数存在し、これらが矛盾した内容でのため正確にお伝えできません。参考として…

### J-RAGBench

#### 主要カテゴリ: Main

情報統合  
Integration

推論  
Reasoning

論理条件解釈  
Logic

表形式解釈  
Table

#### 例外カテゴリ

回答拒否  
Abstention

## 例外カテゴリ：回答拒否 (Abstention)

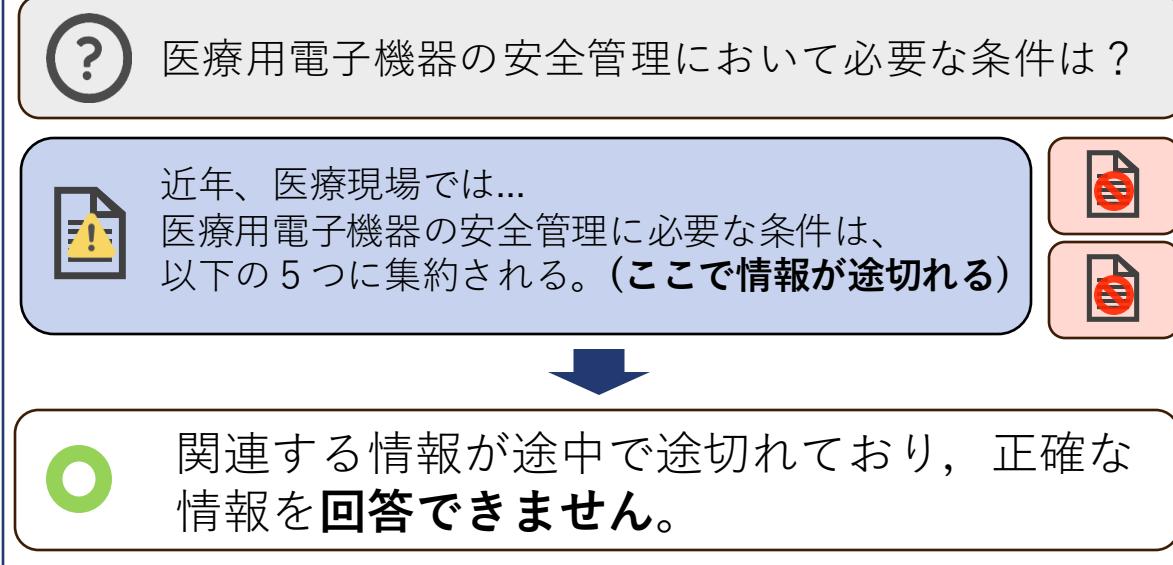
### 概要

回答不可である特定の状況下において、適切に回答を拒否する能力

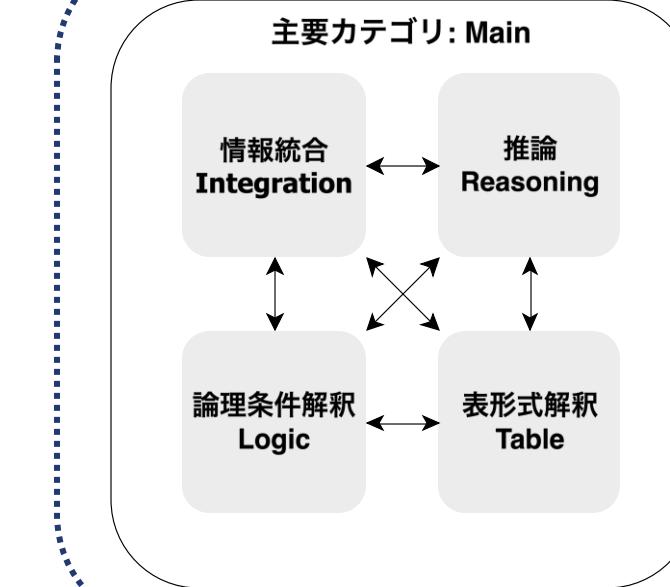
- 根拠不足：関連文書集合が得られなかった場合
- 根拠の矛盾：根拠になり得る情報が複数存在し、それらが矛盾している

### ▶ □ 不完全なチャック区切り：根拠がチャック分割の境界に存在 ⇒ 不完全な根拠情報

#### 例：不完全なチャック区切り



### J-RAGBench



## 作問手順

### 1. 架空QAシナリオの決定

- 架空の企業・製品・人物に基づくシナリオを作成

### 2. 質問・関連文書・正答の作成

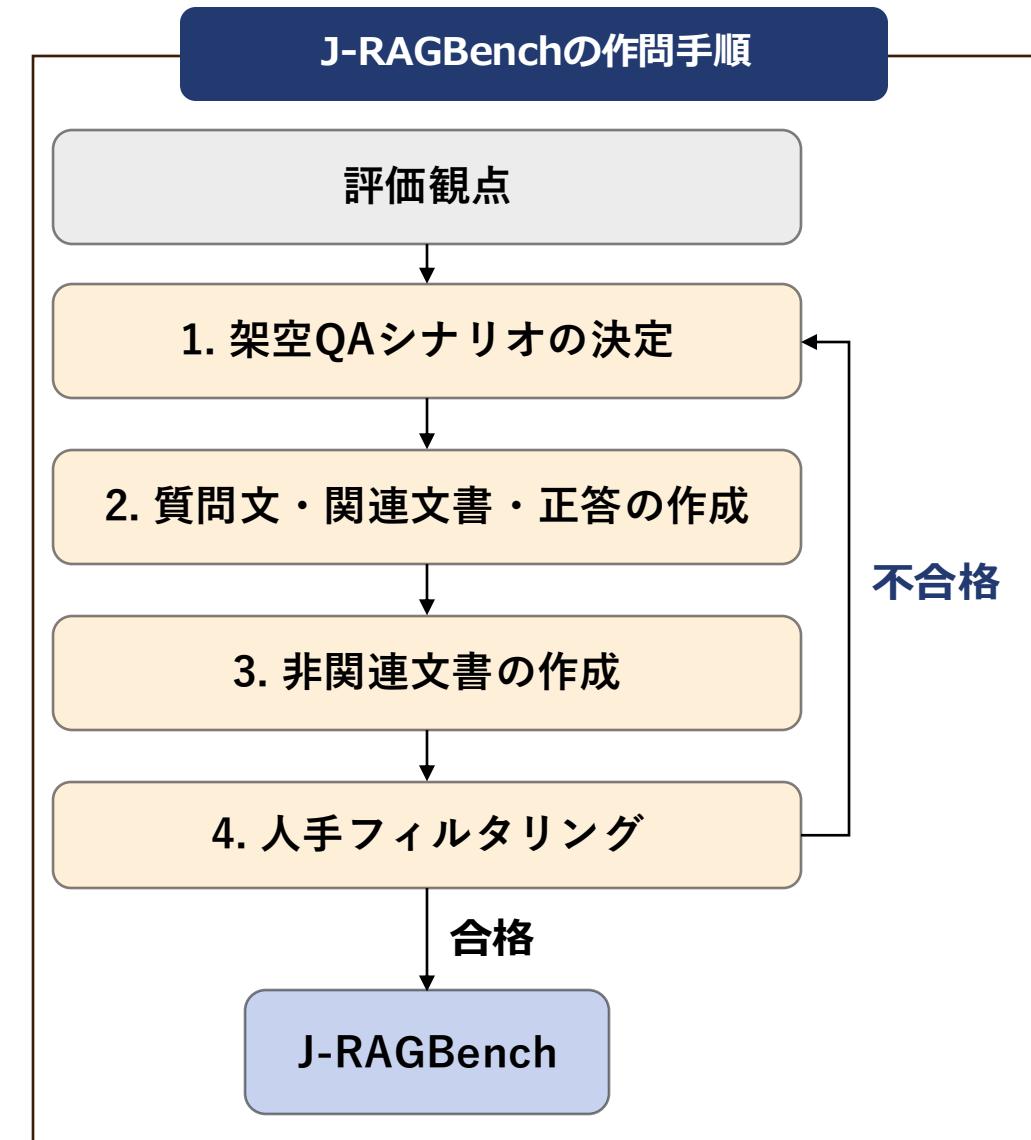
- 評価観点を反映した質問・関連文書・正答を作成
- LLMを用いた合成および人手による整形

### 3. 非関連文書の作成

- 質問の直接的な根拠にはならないが、キーワードや意味的類似度の高い非関連文書を作成
- LLMを用いた合成および人手による整形

### 4. 人手フィルタリング

- 評価観点を適切に反映した質問であるか
- 関連文書のみを根拠として回答可能か
- LLMの事前知識のみで回答不可能か
- 架空情報が実在情報と矛盾していないか



## ベンチマークの統計

合計114問の評価データセットを構築

- Mainカテゴリ：54問
- Abstention(根拠不足)：  
MainのQAデータから関連文書集合を  
空集合に変換

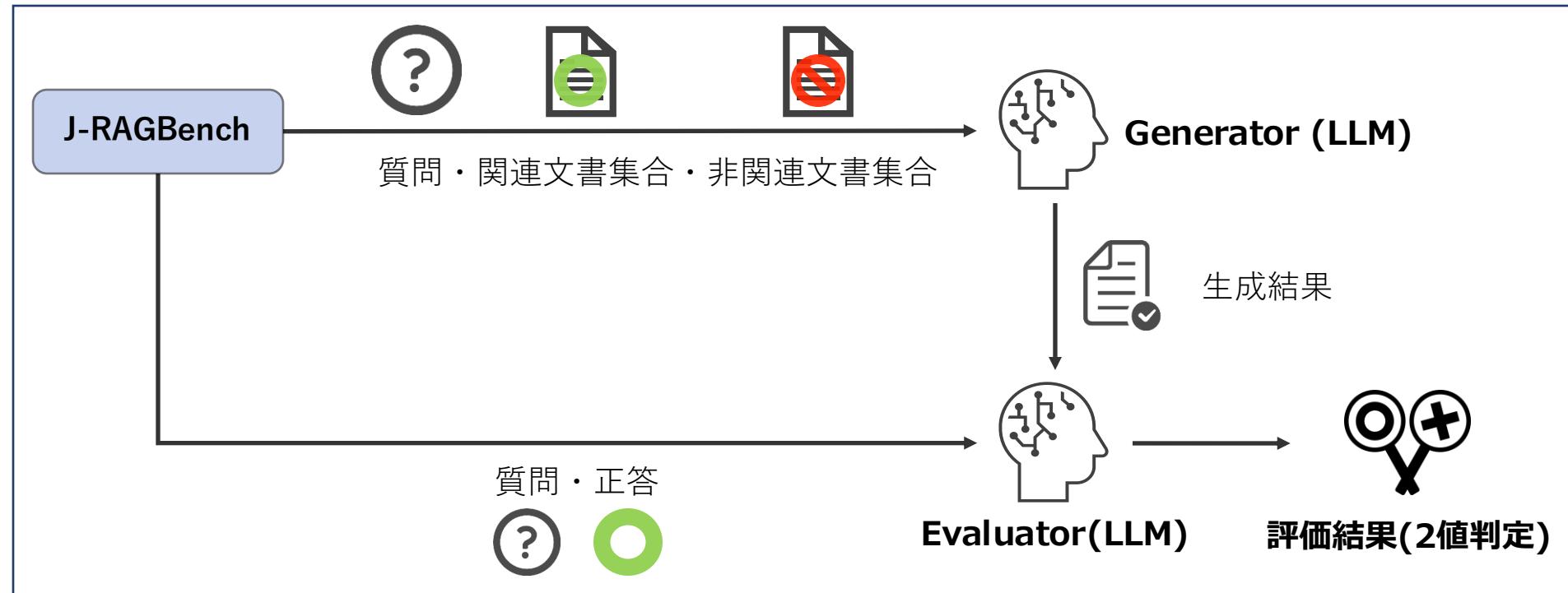
J-RAGBenchの評価観点別の問題数

評価カテゴリ	評価観点	問題数
Integration	複数情報源からの統合	12問
Reasoning	マルチホップ推論 数値計算	12問 11問
Logic	同義関係の解釈 数値包含の関係 概念包含関係の解釈	11問 10問 9問
Table	HTML形式 HTML形式 (セル結合) Markdown形式 CSV形式	8問 9問 7問 7問
Abstention	根拠不足 根拠の矛盾 不完全なチャunk区切り	54問 3問 3問

## 評価方法

LLM-as-a-Judge : LLMを用いた自動評価手法

⇒ Generatorの生成結果と正答の一致性に基づき正誤判定



→ 評価セット全体・評価観点別の**Accuracy (正解率)**を算出

# Index

- 01 背景・目的
- 02 提案手法：J-RAGBench
- 03 実験
- 04 分析
- 05 まとめ

## 実験概要

J-RAGBenchを用いて日本語を生成可能なAPI提供・オープンウェイトのLLMを評価

評価対象のLLM (左 : API提供モデル・右 : オープンウェイトモデル)

モデル名	バージョン	開発元	推論モデル	モデル名	開発元	推論モデル
GPT5	2025-08-07	OpenAI	✓	Llama 3.1 8B Instruct	Meta	
GPT5 mini	2025-08-07	OpenAI	✓	Llama 3.3 70B Instruct	Meta	
GPT5 nano	2025-08-07	OpenAI	✓	Gemma 3 27B Instruct	Google	
o3	2025-04-16	OpenAI	✓	Qwen3 235B A22B Instruct	Alibaba	
o4 mini	2025-04-16	OpenAI	✓	Qwen3 235B A22B Thinking	Alibaba	✓
GPT 4.1	2025-04-14	OpenAI				
GPT 4.1 mini	2025-04-14	OpenAI				
Gemini 2.5 Flash	2025-05-17	Google				
Gemini 2.5 Pro	2025-05-17	Google	✓			
Claude Sonnet 4	2025-05-17	Anthropic				

## 実験設定: 生成パラメータ

### Generator: 実験対象モデル

- 推論モデルの思考トークンの生成長はいずれも最長に設定
  - 例: reasoning effort: high
- サンプリングパラメータが設定可能なモデルはいずれも以下を設定
  - temperature: 0.0
  - top\_p: 1.0
- seed値はいずれも42

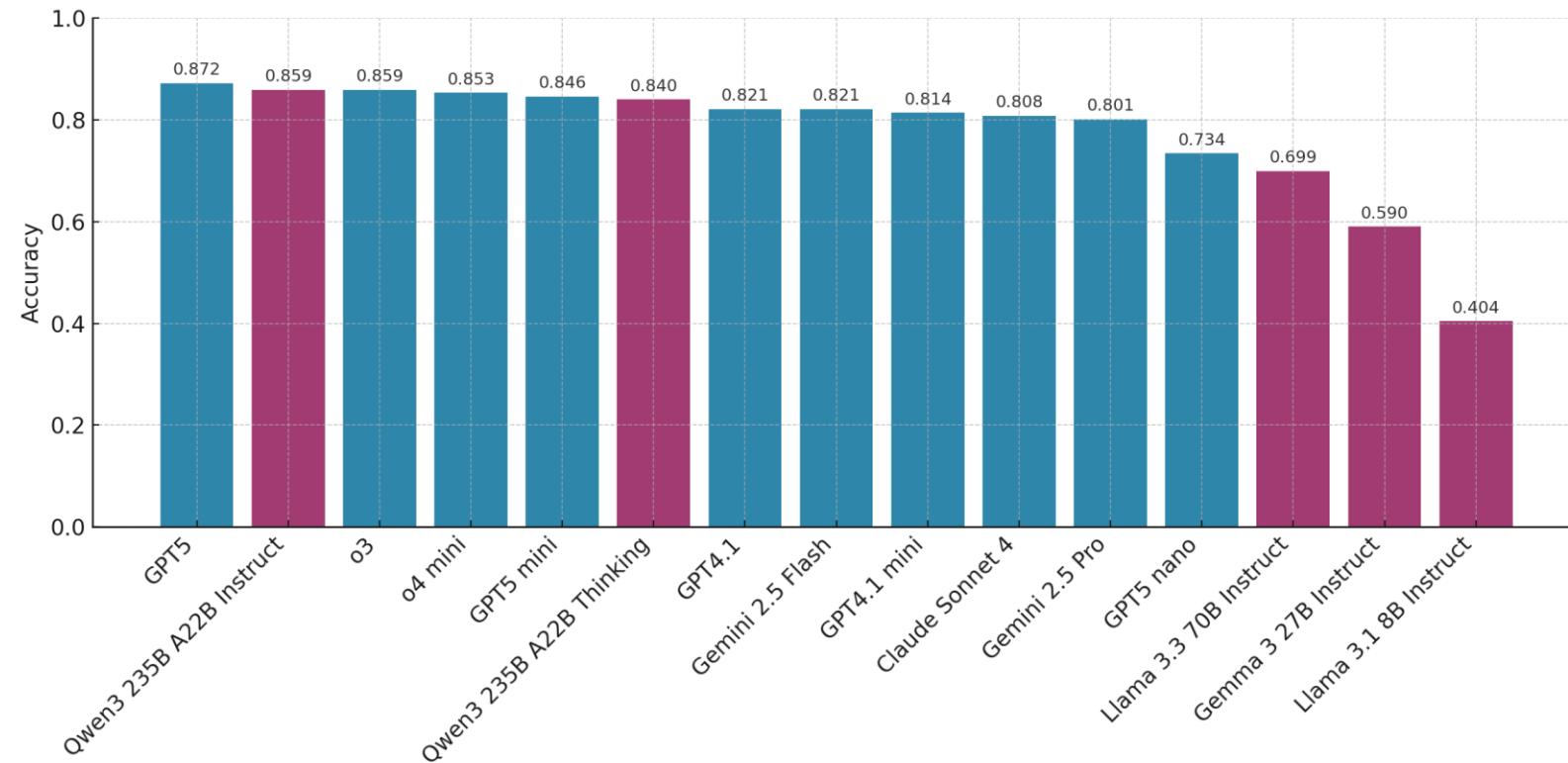
### Evaluator: 評価器モデル

- GPT 4.1
- 生成パラメータは同上

## 評価結果: 総合評価

総合正解率 : 5つの評価カテゴリのAccuracy

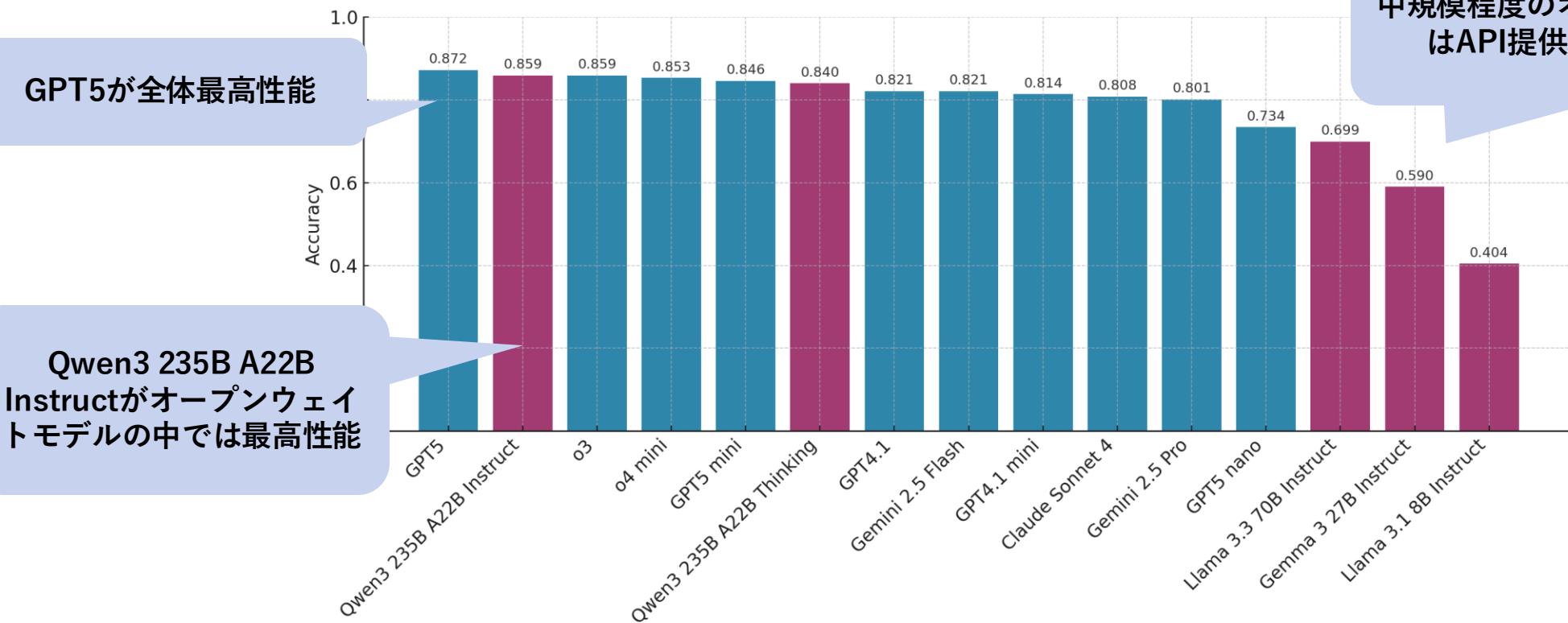
- ✓ Accuracy > 0.9 のモデルは無い
- ✓ 全体的に推論モデルが良好な傾向



## 評価結果: 総合評価

総合正解率 : 5つの評価カテゴリのAccuracy

- ✓ Accuracy > 0.9のモデルは無い
- ✓ 全体的に推論モデルが良好な傾向



中規模程度のオープンウェイトモデル  
はAPI提供モデルに比べて劣る

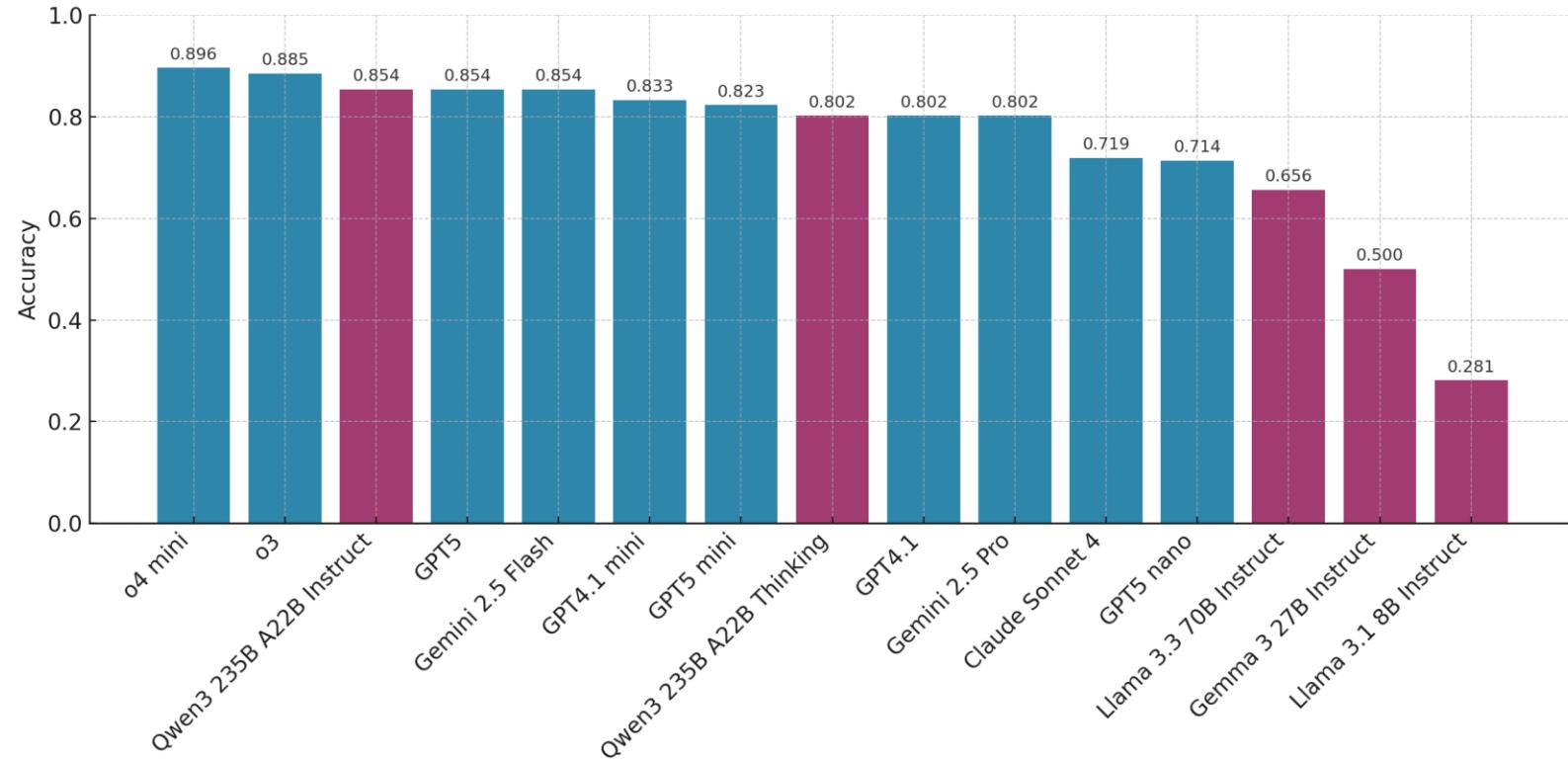
GPT5が全体最高性能

Qwen3 235B A22B Instructがオープンウェイ  
トモデルの中では最高性能

## 評価結果: Main

主要4カテゴリ(回答拒否を除く)の評価結果

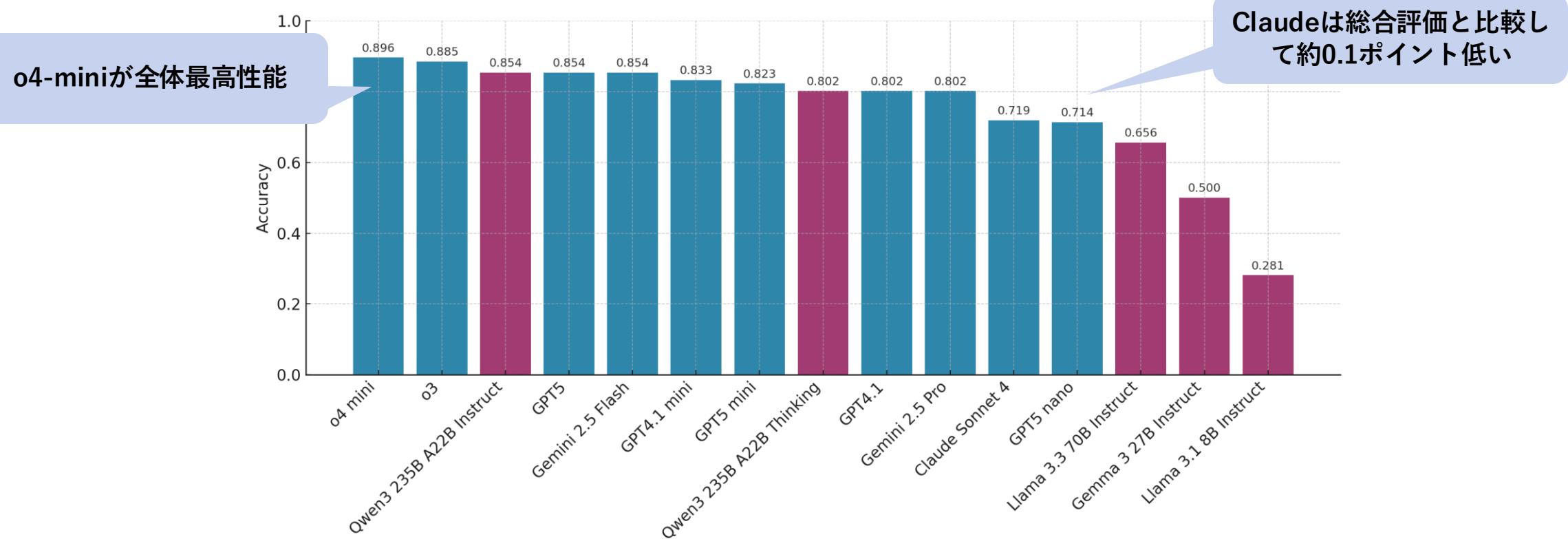
✓ 総合正解率とは順位が変動するも、全体的な傾向は総合評価と共通の傾向



## 評価結果: Main

主要4カテゴリ(回答拒否を除く)の評価結果

✓ 総合正解率とは順位が変動するも、全体的な傾向は総合評価と共に傾向



## 評価結果: 評価カテゴリ別

カテゴリ別に定量化 ⇒ モデル間で各能力の比較が可能に

J-RAGBenchの評価カテゴリ別の評価結果					
モデル名	Integration	Reasoning	Logic	Table	Abstention
GPT5	0.833	0.870	0.867	0.839	0.900
GPT5 mini	<b>0.917</b>	0.826	0.867	0.774	0.833
GPT5 nano	0.750	0.565	0.733	0.677	0.767
o3	0.833	<b>0.957</b>	<b>0.900</b>	0.839	0.817
o4 mini	<b>0.917</b>	0.913	<b>0.900</b>	<b>0.871</b>	0.873
GPT 4.1	0.833	0.739	0.800	0.839	0.850
GPT 4.1 mini	<b>0.917</b>	0.870	0.800	0.806	0.783
Gemini 2.5 Flash	<b>0.917</b>	0.783	0.867	<b>0.871</b>	0.783
Gemini 2.5 Pro	0.667	0.870	0.833	0.774	0.800
Claude Sonnet 4	0.750	0.783	0.700	0.677	<b>0.950</b>
Llama 3.1 8B Instruct	0.167	0.130	0.367	0.355	0.600
Llama 3.3 70B Instruct	0.750	0.478	0.733	0.677	0.767
Gemma 3 27B Instruct	0.667	0.348	0.567	0.484	0.733
Qwen3 235B A22B Instruct	<b>0.917</b>	0.870	0.833	0.839	0.867
Qwen3 235B A22B Thinking	<b>0.917</b>	0.826	0.767	0.774	0.900

o3, o4-mini:  
Mainで良好

GPT5: 全体的に好スコア  
バランスが良いモデル

Qwen 235B:  
API提供に肉薄

GPT 4.1 mini:  
小型モデルでも比較的良好

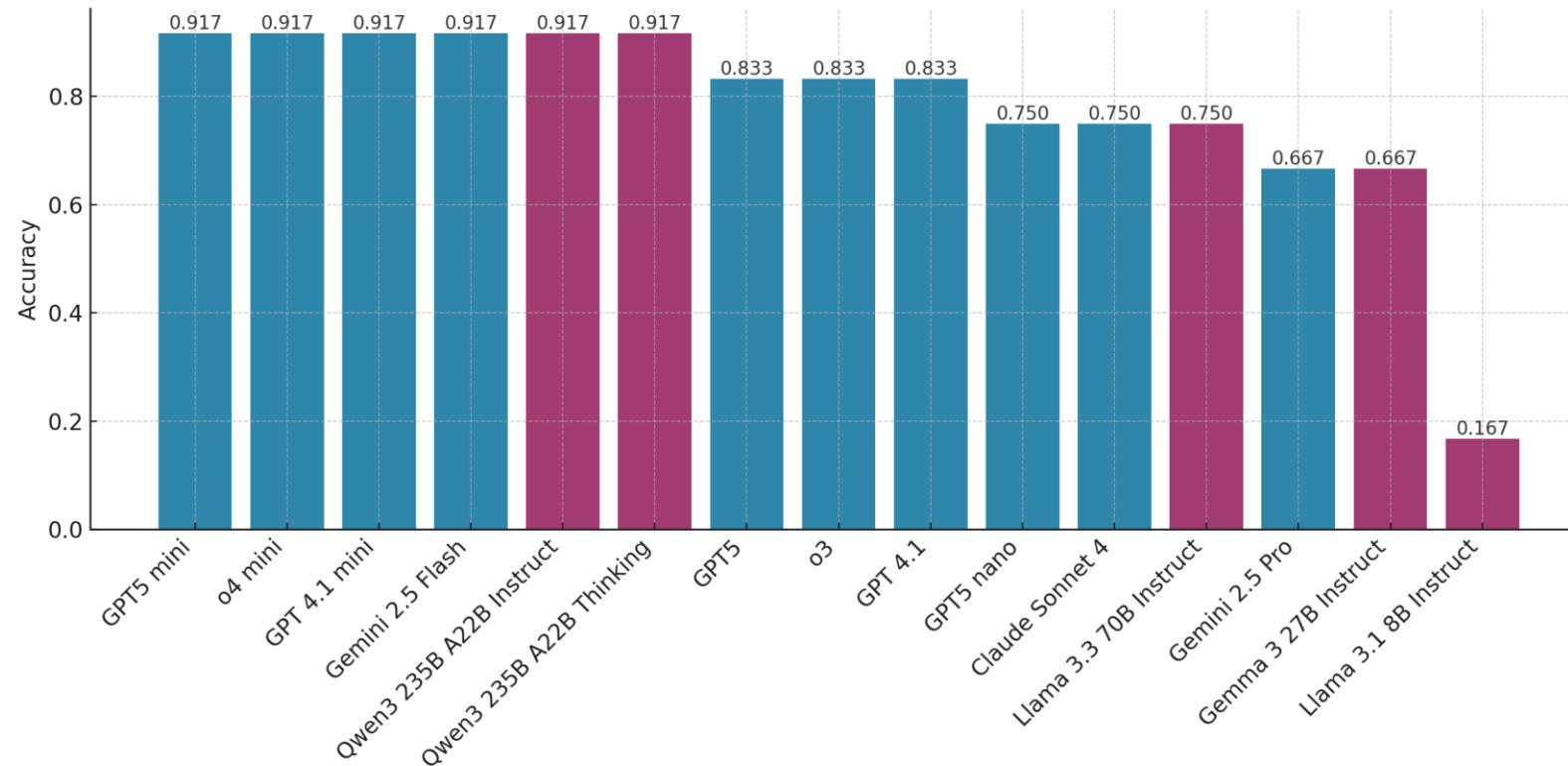
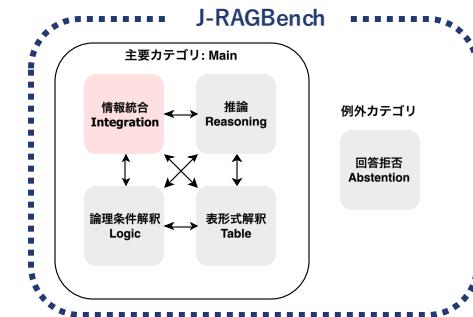
Claude: 回答拒否に強み

# Index

- 01 背景・目的**
- 02 提案手法：J-RAGBench**
- 03 実験**
- 04 分析**
- 05 まとめ**

## 主要カテゴリ①：情報統合 (Integration)

- ✓ 多くのモデルが他カテゴリと比較して全体的に高いスコアを獲得
- ✓ 小型モデル (GPT4.1 mini, Gemini 2.5 Flash等) が比較的良好なスコア



## 主要カテゴリ①：情報統合 (Integration)

誤答ケース

- ・ 関連文書に直接的な語彙的手がかりが存在しない場合
- ・ 複数の情報が並列して記載 ⇒ 情報の粒度が一部異なる場合

例: 粒度が一部異なる

注釈を加味する必要 &  
異なる単位の記載形式



A社、B社、C社の3つの会議室の料金をまとめて。土曜日に4時間利用したい。



【A社 会議室貸出サービス】

基本料金 : 300円 / 1h

※ 土日祝日の場合は +50円 / 1h となります



【会議室 B社】

基本料金 : 400円 (1時間あたり)



【C社 レンタル会議室】

基本料金 : 320円 (1時間あたり)



A社は1,400円(350円 × 4), B社は1,600円(400円× 4), C社は 1,280円(320円 × 4)です



A社は1,200円(300円 × 4), B社は1,600円(400円× 4), C社は 1,280円(320円 × 4)です



※ 説明のためにQA内容を一部簡易化

## 主要カテゴリ①：情報統合 (Integration)

誤答ケース

- ・ 関連文書に直接的な語彙的手がかりが存在しない場合
- ・ 複数の情報が並列して記載 ⇒ 情報の粒度が一部異なる場合

例: 粒度が一部異なる

注釈を加味する必要 &  
異なる単位の記載形式



A社、B社、C社の3つの会議室の料金をまとめて。土曜日に4時間利用したい。



【A社 会議室貸出サービス】

基本料金 : 300円 / 1h

※ 土日祝日の場合は +50円 / 1h となります



【会議室 B社】

基本料金 : 400円 (1時間あたり)



【C社 レンタル会議室】

基本料金 : 320円 (1時間あたり)



A社は1,400円(350円 × 4), B社は1,600円(400円× 4), C社は 1,280円(320円 × 4)です



A社は1,200円(300円 × 4), B社は1,600円(400円× 4), C社は 1,280円(320円 × 4)です

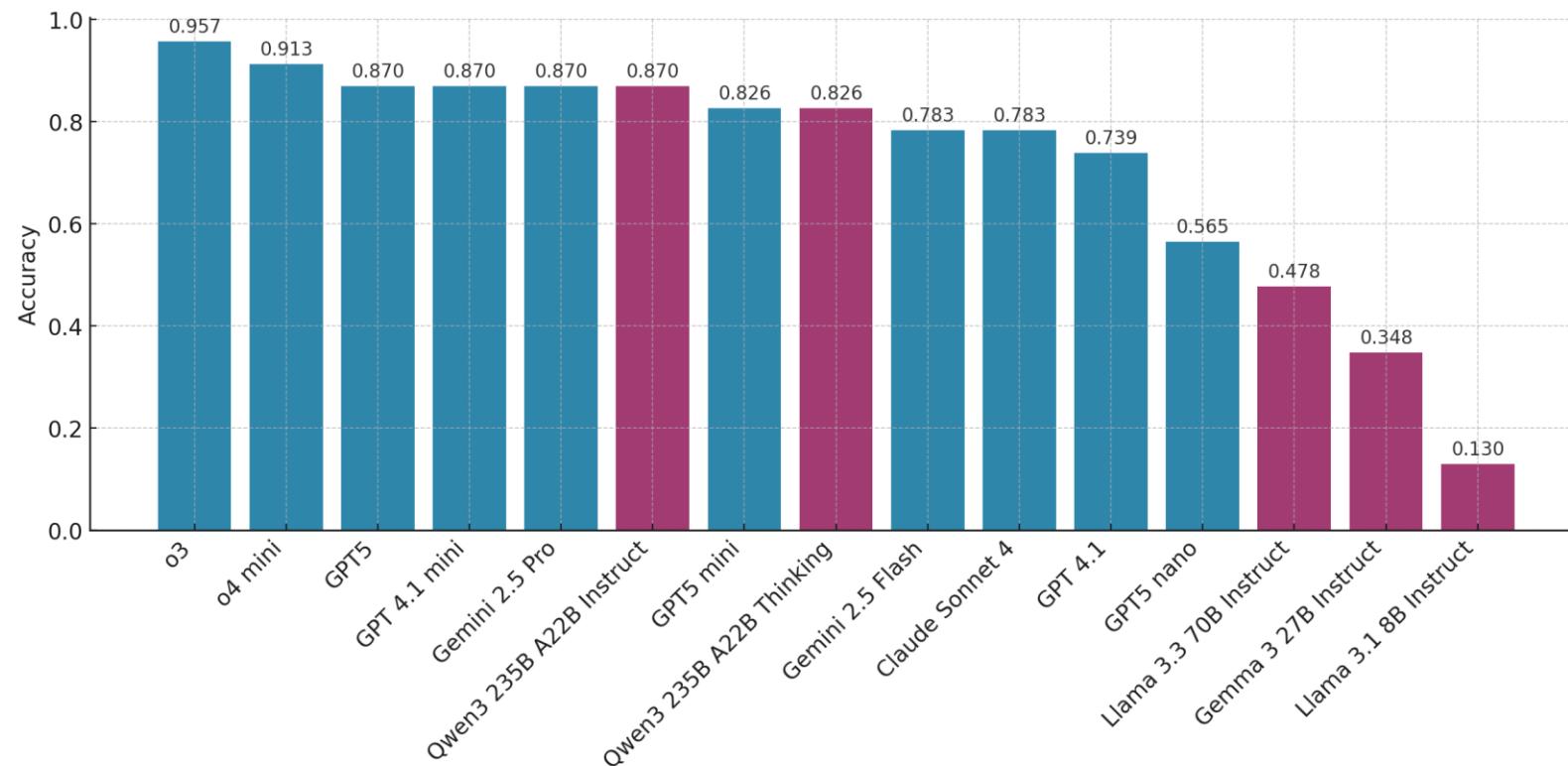
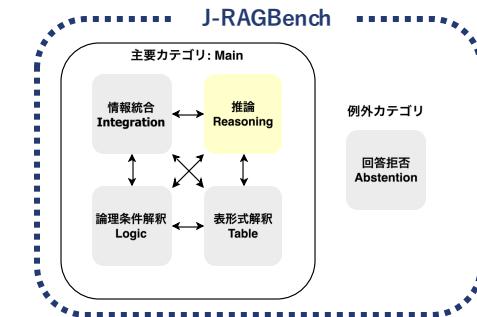


※ 説明のためにQA内容を一部簡易化

⇒ 実運用時では、文書の構造化や正規化といった前処理の重要性を改めて確認

## 主要カテゴリ②：推論 (Reasoning)

- ✓ o3を筆頭に推論モデルが全体的に高い性能
- ✓ モデルによる性能のばらつきが大きい評価カテゴリ



## 主要カテゴリ②：推論 (Reasoning)

- 誤答ケース
- 計算ミス：多段階での計算過程における途中または最終生成時
  - マルチホップ：中間エンティティに関する直接的な語彙的手がかりが紐づけられない ⇒ 情報不足と判断して回答拒否をする傾向

## 主要カテゴリ②：推論 (Reasoning)

- 誤答ケース
- ・ 計算ミス：多段階での計算過程における途中または最終生成時
  - ・ マルチホップ：中間エンティティに関する直接的な語彙的手がかりが紐づけられない ⇒ 情報不足と判断して回答拒否をする傾向

評価観点：数値計算

立式までは正解  
加算時に計算ミス

商品Aを作る全材料費の合計はいくらか？

【商品A レシピ】  
パラフィンワックス: 200g, キャンドル芯: 1本, アロマオイル: 5ml, ガラスジャー: 1個

【仕入れリスト】  
パラフィンワックス: 80円/100g キャンドル芯: 150円/10本セット, アロマオイル: 1,200円/50ml, ガラスジャー: 180円/個

↓

材料費は375円 (160円 + 15円 + 120円 + 180円)

材料費は385円 (160円 + 15円 + 120円 + 180円)

- 以下のモデルでは確認されなかった
- GPT5
  - o3
  - o4-mini
  - Gemini 2.5 Pro

## 主要カテゴリ②：推論 (Reasoning)

- 誤答ケース
- ・ 計算ミス：多段階での計算過程における途中または最終生成時
  - ・ マルチホップ：中間エンティティに関する直接的な語彙的手がかりが紐づけられない→ 情報不足と判断して回答拒否をする傾向

評価観点：マルチホップ

隠れた中間エンティティ：  
ブルーイノベーション研究助成

中間エンティティの関係を  
推論できずに回答拒否



グリーンウェーブ社が海洋プラスチック問題で採択された研究テーマは何か？



【グリーンウェーブ社 プレスリリース】  
海洋プラスチック問題に関する研究がブルーイノベーション研究助成に採択



【ブルーイノベーション研究助成 採択一覧】  
グリーンウェーブ: マイクロプラスチックを分解するバイオポリマーの開発

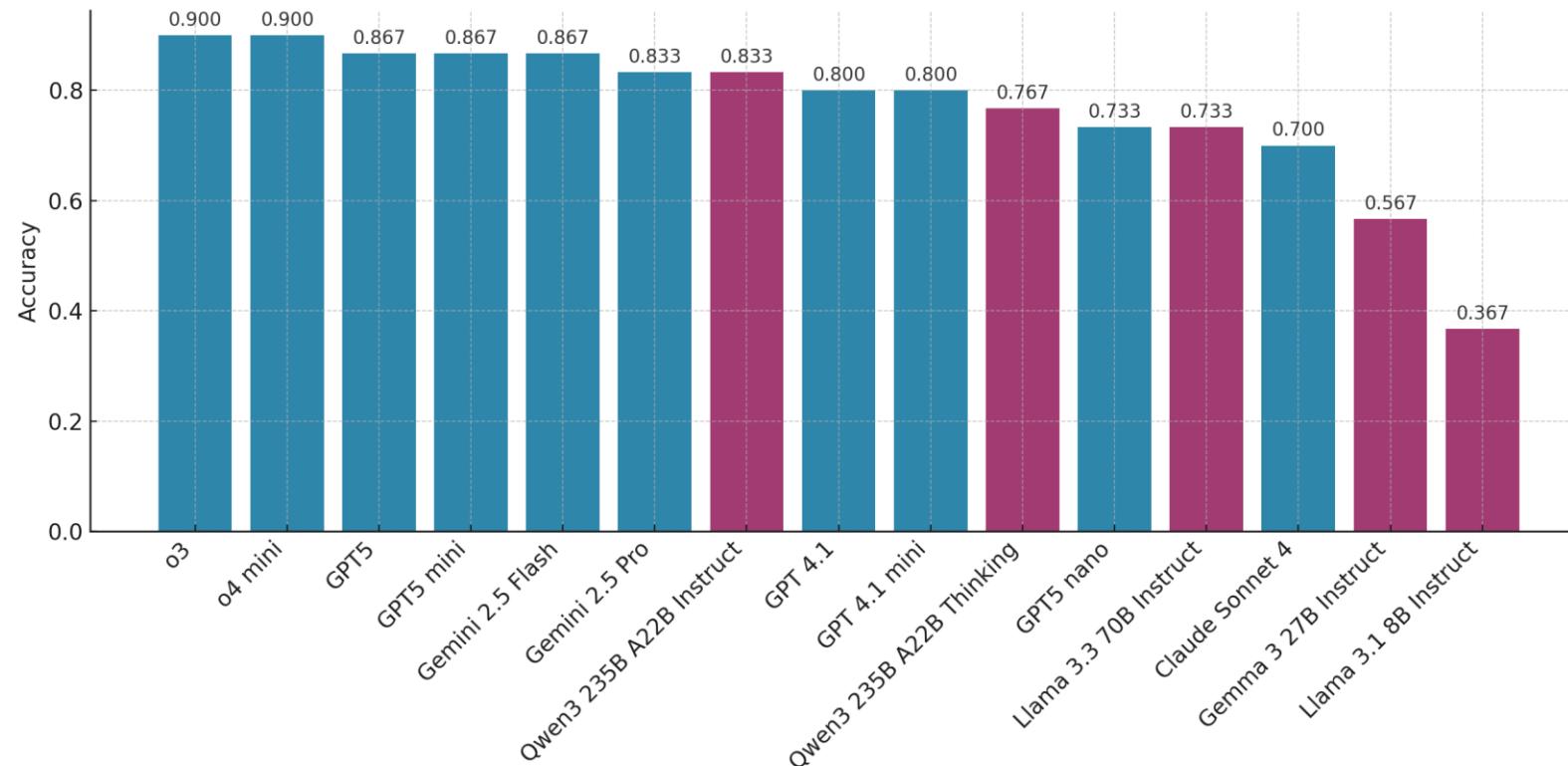
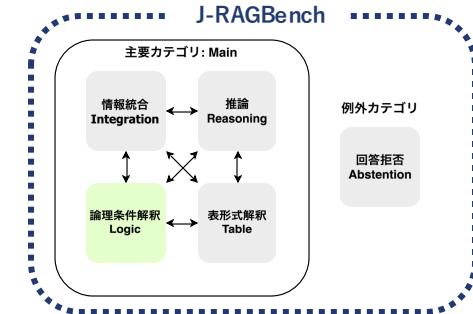


明確な情報は見つかりませんでした。関連する情報として、ブルーイノベーション研究助成に採択された

非推論特化モデルの場合に散見

## 主要カテゴリ③：論理関係の解釈 (Logic)

- ✓ 全体的には推論モデルが良い性能を示した
- ✓ 「7,616 百万円 ⇔ 76億1600万円」等の単位変換：  
GPT 4.1等の上位モデルにおいても7.6億円などの誤答を確認



## 主要カテゴリ③：論理関係の解釈 (Logic)

- 誤答ケース
- 日本語特有の単位変換：  
「7,616 百万円 ⇄ 76億1600万円」であるが、7.6億円と回答

評価観点：同義関係の解釈

① 株式会社統計解析Labの研究支援金収入は何億円ですか？

<予算計画>  
(単位：百万円)

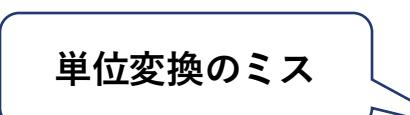
区分	内容	中期計画予算（平成25～29年度）	平成25年度予算
収入	研究支援金収入	41,339	7,616
収入	受託製表収入	97	22

⋮

7.6億円です。

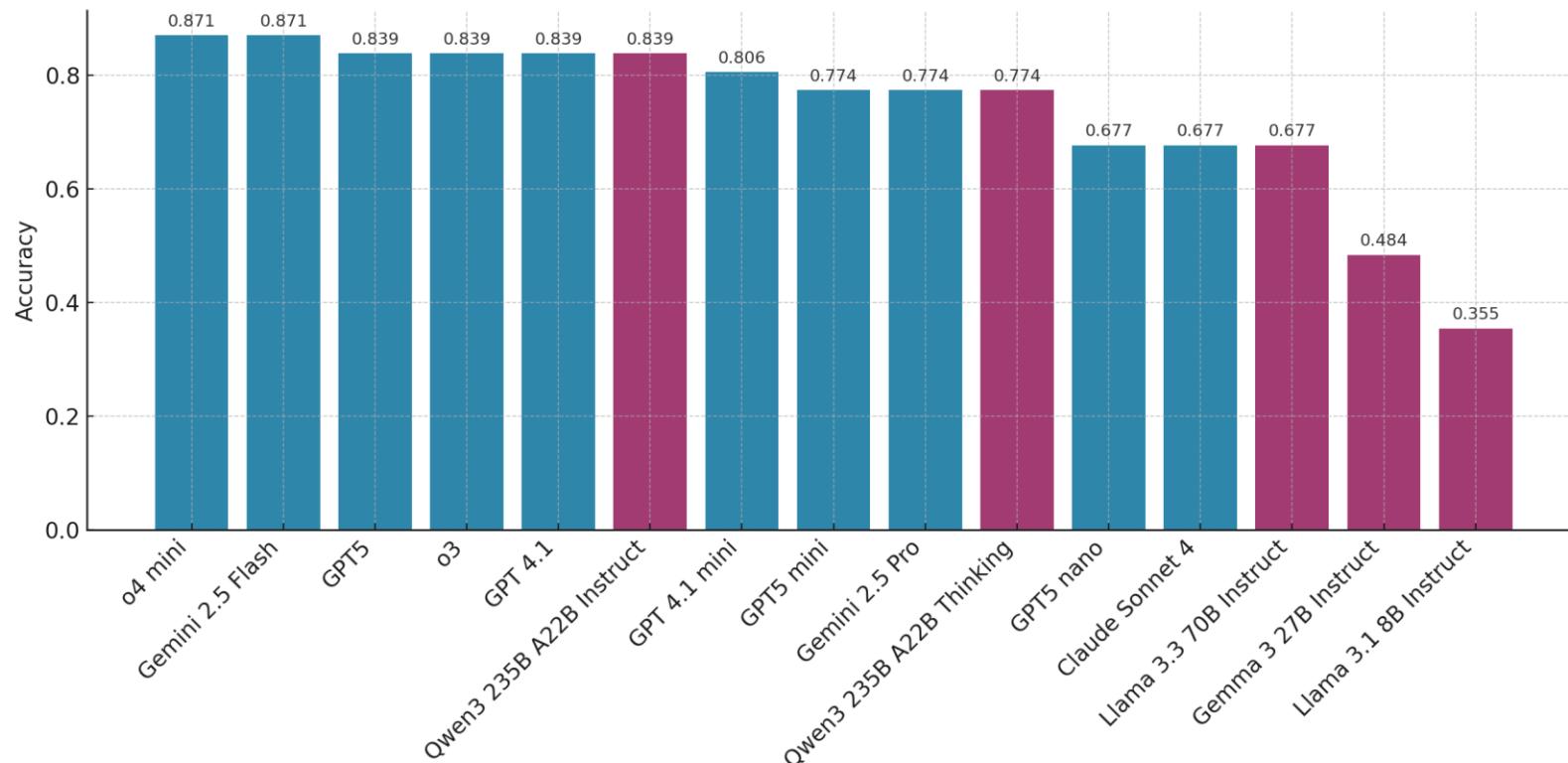
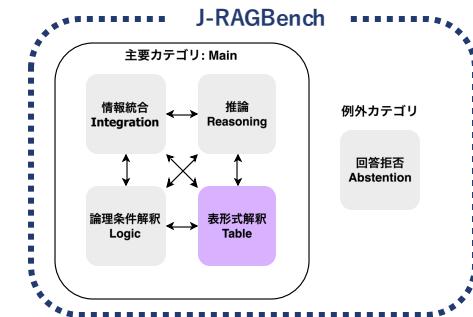
76億1600万円です

単位変換のミス



## 主要カテゴリ④：表形式の解釈 (Table)

GPT 4.1 mini等の小型モデルが比較的良好なスコア



## 主要カテゴリ④：表形式の解釈 (Table)

- 誤答ケース
- セル結合時の情報抽出
  - 大規模な表が複数チャunkに分割

## 主要カテゴリ④：表形式の解釈 (Table)

誤答ケース

- セル結合時の情報抽出
- 大規模な表が複数チャンクに分割

HTML セル結合

(?) エコアンドフロンティアの家電部門、IoT部門のQ3の海外利益は？

表のヘッダー部分がセル結合

部門	四半期	売上および利益 (百万円)					
		国内		海外		合計	
		売上	利益	売上	利益	売上	利益
電子機器	Q1	3000	900	2000	600	5000	1500
	Q2	3200	950	2100	650	5300	1600
	Q3	3400	1000	2200	700	5600	1700
	Q4	3600	1050	2300	750	5900	1800
家電	Q1	1800	500	1200	400	3000	900
	Q2	1900	550	1300	450	3200	1000
	Q3	2000	600	1400	500	3400	1100
	Q4	2100	650	1500	550	3600	1200

IoT部門も下に続く



情報を特定できずに回答拒否



海外利益に関する明確な情報は見つかりませんでした。

## 主要カテゴリ④：表形式の解釈 (Table)

- 誤答ケース
- セル結合時の情報抽出
  - 大規模な表が複数チャunkに分割

評価観点 : Markdown

チャunk分割後にヘッダーを挿入して保存

① キャッシュレス決済に対応している店舗名を列挙して

常慶商店街 店舗 ID	店舗名	現金	電子決済	クレジットカード決済	QR決済
T001	灯凧コーヒー舎	○	○	○	○
...					
T025	笹鳴うどん屋	○	○	×	×

常慶商店街 店舗 ID	店舗名	現金	電子決済	クレジットカード決済	QR決済
T026	曙ミルクスタンド	○	○	○	○
...					
T050	海鳴レモネード店	○	○	○	×

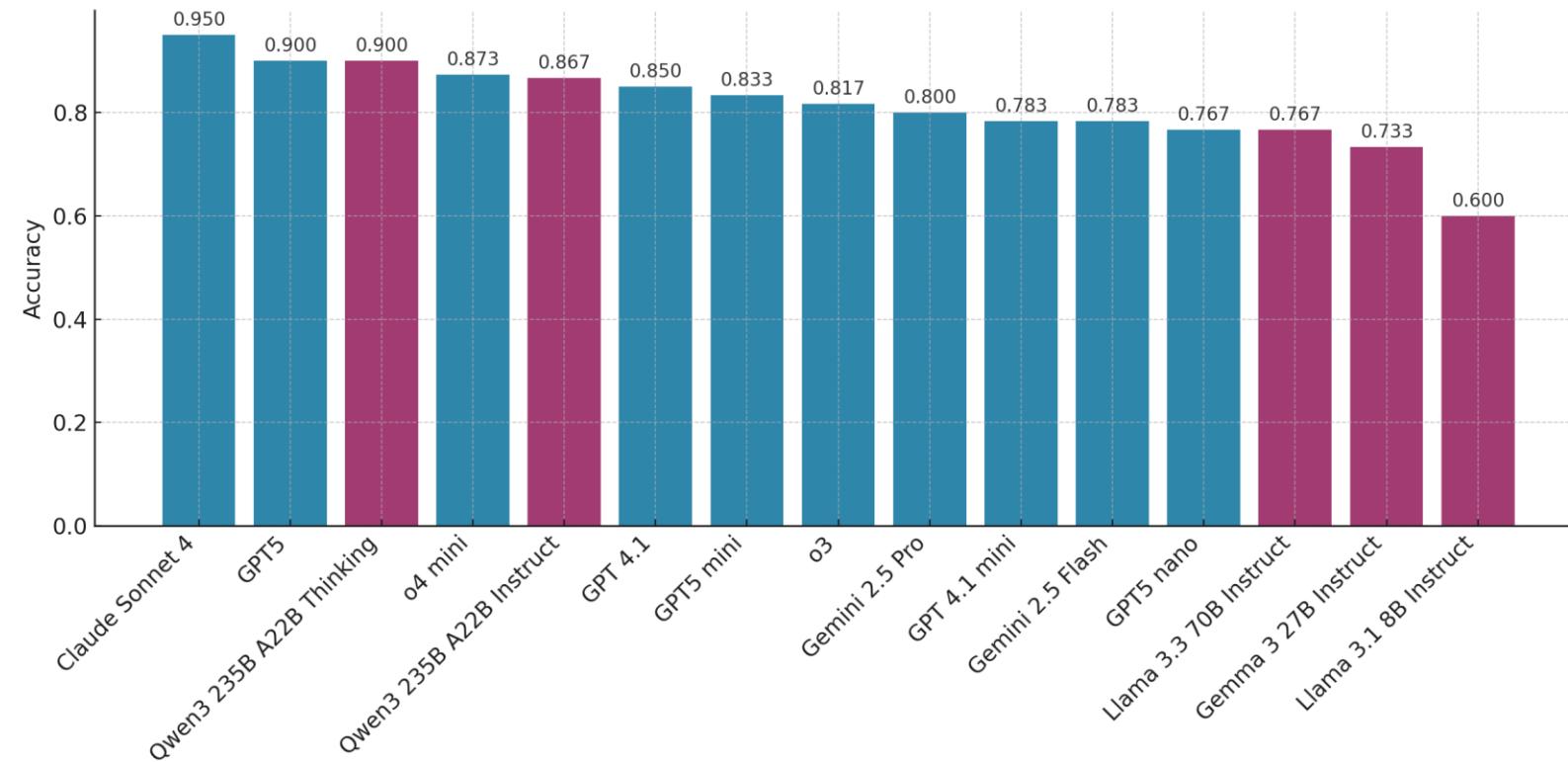
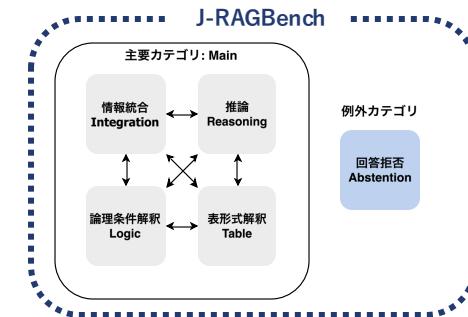
常慶商店街 店舗ID	店舗名	現金	電子決済	クレジットカード決済	QR決済
T001	灯凧コーヒー舎	○	○	○	○
T002	露霞ベーカリー	○	×	×	○
T003	翠雨書房	○	○	×	×
T004	麗月雑貨店	○	○	○	×
T005	桔梗ラーメン工房	○	×	○	×
T006	風燈スイーツ堂	○	○	○	○
T007	朱雀ハーブ茶屋	○	×	○	×
T008	青藍カレー食堂	○	○	×	○
T009	雷渓アウトフィッターズ	○	○	○	×
T010	木端れ日文具館	○	×	○	○
T011	霜夜キャンドル工房	○	○	×	×
T012	紫苑トーストスタンド	○	○	○	○
T013	葉陰フランチャリエ	○	×	×	○
T014	竹葉レコード店	○	○	○	×
T015	銀砂バスク食堂	○	×	○	○
T016	風待ちブックス	○	○	×	×
T017	氨基姫スマイル商会	○	○	○	○
T018	露草フェラート舗	○	×	×	○
T019	白濱ギフトサロン	○	○	○	×
T020	星屑ティーラボ	○	○	×	○
T021	木雲アウトドア商店	○	×	○	×
T022	南天ファンボイケ	○	○	○	○
T023	砂紋ファブリック館	○	×	×	×
T024	月灯り珈琲本店	○	○	×	○
T025	笹鳴うどん屋	○	×	○	×
T026	曙ミルクスタンド	○	○	○	○
T027	海霧マリン雑貨	○	○	×	×
T028	風薰ベーカリーラボ	○	○	○	○
T029	桃色クラフト酒場	○	×	○	○
T030	天羽キャラフギア	○	○	×	×
T031	柳じぞく甘味店	○	○	○	○
T032	月白インテリア堂	○	×	×	○
T033	山崎ロースターズ	○	○	○	×
T034	波灯食堂	○	×	○	×
T035	相模モバイル茶屋	○	○	×	○
T036	駒鳥バーサンド舗	○	○	○	○
T037	砂炒テキスタイル	○	×	×	×
T038	花霞ソーダバー	○	○	×	○
T039	雪解けバティスリー	○	○	○	×
T040	萬空ブランジュ	○	×	○	○
T041	霧雨リビング用品店	○	○	×	×
T042	音羽ペンと紙	○	○	○	○
T043	露原スープ食堂	○	×	×	○
T044	白楓ワークス	○	○	○	×
T045	風致せんべい専門	○	×	○	×
T046	月雲ティールーム	○	○	○	○
T047	朔風プランツショップ	○	○	×	×
T048	瑞鶴パワーエンジニア	○	○	○	○
T049	木庵レザーグラフト	○	×	×	○
T050	海鳴レモネード店	○	○	○	×
T051	ひより和菓子庵	○	○	○	×
T052	雪ノ湯サロン	○	○	×	○
T053	石畠キッチン	○	○	○	○
T054	鶴色パン研究所	○	×	×	×
T055	霜花ピート万年筆	○	○	○	×
T056	真砂スムージー舗	○	○	×	○
T057	灯台クラフトビール	○	×	○	○
T058	達ブティック	○	○	×	×
T059	銀の林古道酒店	○	○	○	○
T060	茶燭カラーフジ焙煎	○	×	×	○
T061	露立フーラーメン横丁	○	○	○	×
T062	雨音フルーツ工房	○	○	×	○
T063	薄荷トートと雑貨	○	×	○	×
T064	月代サンエーテック	○	○	○	○
T065	霧笛オーブンスタジオ	○	○	×	×
T066	銅葉バスクテラ	○	○	○	○
T067	春抱デリカテッセン	○	×	×	○
T068	雨燕グッドアンド美術	○	○	○	×
T069	黒鳩ステーションオフィス	○	×	○	○
T070	白雲シリアルズパー	○	×	×	×

## 例外カテゴリ：回答拒否 (Abstention)

関連する尤もらしい知識の補完による誤答を確認：知識の捏造

Claude Sonnet 4が最も高いスコア：知識の捏造は0件

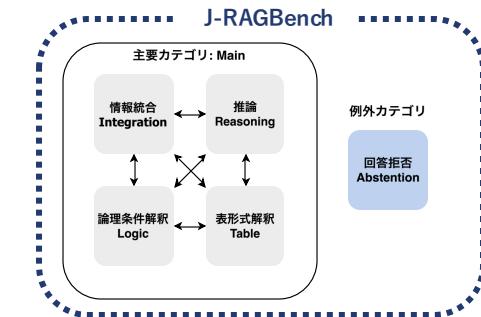
一部の推論モデル（o3やGemini2.5 Pro等）がやや低いスコアの傾向



## 例外カテゴリ：回答拒否 (Abstention)

### Claude Sonnet 4

- 回答拒否能力が他モデルと比較して高い
- 回答可能なMainカテゴリのQAに対しても回答拒否をして誤答する傾向  
⇒ 回答拒否の閾値設定が保守的である可能性を示唆



## 例外カテゴリ：回答拒否 (Abstention)

- 誤答ケース
- 根拠の捏造：根拠は含まれないが、関連するキーワードに反応して誤った事実を回答
  - LLMの事前知識に基づくハリシネーション：ユーザの曖昧な質問等に対して関連文書が存在しない場合に事前知識に基づく回答を提示

## 例外カテゴリ：回答拒否 (Abstention)

### 誤答ケース

- 根拠の捏造：根拠は含まれないが、関連するキーワードに反応して誤った事実を回答
- LLMの事前知識に基づくハリシネーション：ユーザの曖昧な質問等に対して関連文書が存在しない場合に事前知識に基づく回答を提示

評価観点：根拠不足



北祭京シェアサイクルを4時間当日利用した場合の料金はいくらですか？

【掛橋市 シェアサイクル】



区分	当日利用（24時間）	定期利用（1か月）
普通自転車	210円	2,100円
電動アシスト自転車	340円	4,380円



入力コンテキストに北祭京市や4時間利用の場合についての情報は存在しない



北祭京市シェアサイクルの当日利用料金は1日あたり210円であり、  
4時間の場合も210円で利用可能です



## 例外カテゴリ：回答拒否 (Abstention)

### 誤答ケース

- 根拠の捏造：根拠は含まれないが、関連するキーワードに反応して誤った事実を回答
- LLMの事前知識に基づくハリシネーション：ユーザの曖昧な質問等に対して関連文書が存在しない場合に事前知識に基づく回答を提示

評価観点：根拠不足

「海洋分解性PLAブレンド樹脂」という単語は入力コンテキストに存在しない



グリーンウェーブ社が2024年に海洋プラスチック問題で正式に採択された主要研究テーマは何？



波力利用スタートアップ「グリーンウェーブ社」および豪州エネ庁は共同で、生分解性漁具パイロットプロジェクトを2024年9月からタスマン海で開始すると発表した。



UNEPの2024年報告書によれば、一次マイクロプラスチック（樹脂ペレット・合成纖維など）が全体流出量の31%を占め、二次マイクロプラスチック（波浪摩耗由来）が69%と推計され



本件について回答できる情報はありません。ただし、関連する取り組みとしては、**海洋分解性PLAブレンド樹脂の開発にも取り組んでいる**



## 例外カテゴリ：回答拒否 (Abstention)

### 誤答ケース

- 根拠の捏造：根拠は含まれないが、関連するキーワードに反応して誤った事実を回答
- LLMの事前知識に基づくハリシネーション：ユーザの曖昧な質問等に対して関連文書が存在しない場合に事前知識に基づく回答を提示

評価観点：根拠の矛盾

入力コンテキストにこの情報は含まれておらずLLMの事前知識に基づいて回答をしている



医療管理用電子システムの安全管理において必要な条件を教えてください。



医療情報は、インフォームド・コンセントの観点からも、医療機関等と患者等との信頼関係に基づいて取り扱われるものであるため



1. 安全管理措置：医療・介護関係事業者は、その取り扱う個人データの漏えい、滅失又は毀損の防止その他の個人データの安全管理のため、組織的、人的、物理的、及び技術的安全管理措置等を講じなければならない。



安全管理に必要な条件は以下の5点に集約されます

1. 医療機器の…



# Index

- 01 背景・目的
- 02 提案手法：J-RAGBench
- 03 実験
- 04 分析
- 05 まとめ

## まとめ

- 日本語RAG の Generator 評価のための **J-RAGBench** を構築・公開
  - ✓ 実運用時の課題から評価範囲を5つのカテゴリに体系化
  - ✓ 評価観点の組み合わせパターンを網羅
- 各種LLM (API提供, オープンウェイト) の評価実験
  - ✓ 観点別のモデル間性能を定量的に比較可能
  - ✓ 用途に応じたモデル選択の重要性を示唆

→ 実運用時のLLM選定・RAG特化LLM構築の指標

## 今後の展望

- 問題数の拡充等の定期アップデートによる  
コミュニティ貢献の
- 多言語化による適用範囲の拡大
- Agentic RAG等への評価ベンチマークの拡張

