

Hallucination対策に関する 最近の動向

2024年7月30日

NII LLMC

特任研究員 薦 侑磨

大規模言語モデル（LLM）

- 翻訳・要約・対話など様々な用途に利用される



翻訳



要約



対話型サービス（ChatGPT）

- LLMの課題
 - 正しい出力を行えているか（**関連課題：ハルシネーション**）
 - プライバシーとセキュリティ
 - 膨大な計算リソース

LLMの課題：ハルシネーション

- ハルシネーションとは？
 - 生成された文が現実の事実やユーザーの入力と一致しない現象
 - 入力例 1：月に最初に着陸した人は誰？
 - 誤回答例：人類で初めて月に降り立ったのは**チャールズ・リンドバーグ**です
 - 入力例 2：唾液アミラーゼの働きは？
 - 誤回答例：
 - 唾液アミラーゼは、食べ物に含まれるでんぷんを分解し、**胃で消化されやすい**状態にする
 - 引用: <https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240306-OYT1T50080/>
- **情報の正確性の重要な分野**では重大な問題
 - 医療、法務、ニュースなど

ハルシネーションの種類

- ハルシネーションの分類方法は様々
 - メジャーな分類：内因性・外因性 [Huang+2023, Li+2022, Ji+2023]
- **事実性ハルシネーションと忠実性ハルシネーション** [Huang+ 2024]
 - 事実性ハルシネーション（**factuality hallucination**）
 - 事実と異なる情報を生成
 - 忠実性ハルシネーション（**faithfulness hallucination**）
 - ユーザーの指示に反する生成

事実性ハルシネーションの具体例

事実性ハルシネーション：事実と異なる情報を生成

- 事実的不一致（**factual inconsistency**）

- 入力: 月に最初に着陸した人は誰ですか？
- 誤回答例：チャールズ・リンドバーグ
- 正回答例：ニール・アームストロング

- 事実の捏造（**factual fabrication**）

- 入力：ユニコーンの起源について教えてください
- 誤回答例：
 - ユニコーンは、紀元前10,000年頃にアトランティスの平原を駆け抜けたとされ、王族と神聖視されていました

忠実性ハルシネーションの具体例

- 忠実性ハルシネーション: **ユーザーの指示に反する生成**
 - **指示の不一致 (instruction inconsistency)**
 - 次の英文を翻訳して “What is the capital of Japan?”
 - 誤回答例: The capital of Japan is Tokyo.
 - **コンテキストの不一致 (context inconsistency)**
 - 入力: 次の文を要約して「信濃川は日本で最も長い川で、新潟県と長野県を流れています。川の源流は長野県の山間部にあり、日本海に注ぎます。」
 - 誤回答例: 信濃川の源流は新潟県にあり、日本海に注ぎます。
 - **論理的不一致 (logical inconsistency)**
 - 入力: 次の方程式をステップバイステップで解いてください: $2x + 3 = 11$
 - 誤回答例: ステップ1: 両辺から3を引いて、 $2x = 8$ とします。
ステップ2: 両辺を2で割って、 $x = 3$ とします

ハルシネーションへの対策

- **外部の事実を利用した情報確実性の向上** [Li+ NAACL 2024]
 - 生成文を信頼できる外部の情報源と参照する
- **モデルの不確実性(uncertainty)を検証** [Manakul+ EMNLP 2023, Cohen+ EMNLP 2023]
 - 不確実性：どの程度同じ出力を一貫して行うか
 - モデルの複数の生成例から不確実性を推定しハルシネーションを評価
- **モデルの応答信頼性を利用** [Zhang+ NAACL 2024]
 - 生成時に応答信頼性を出力する

ハルシネーションへの対策： 外部の事実を利用した生成・検証

- 方法:

1. ウェブやデータベースなどの**外部情報源**から情報を収集
2. 収集した情報を**利用して生成** or 生成文と**矛盾がないか検証**

- 具体例（文生成）:

1. 質問：ヒマラヤ山脈の最高峰は何ですか？
2. 検証：「エベレスト」（**外部情報源を参照**）
3. 出力：ヒマラヤ山脈の最高峰は**エベレスト**です。

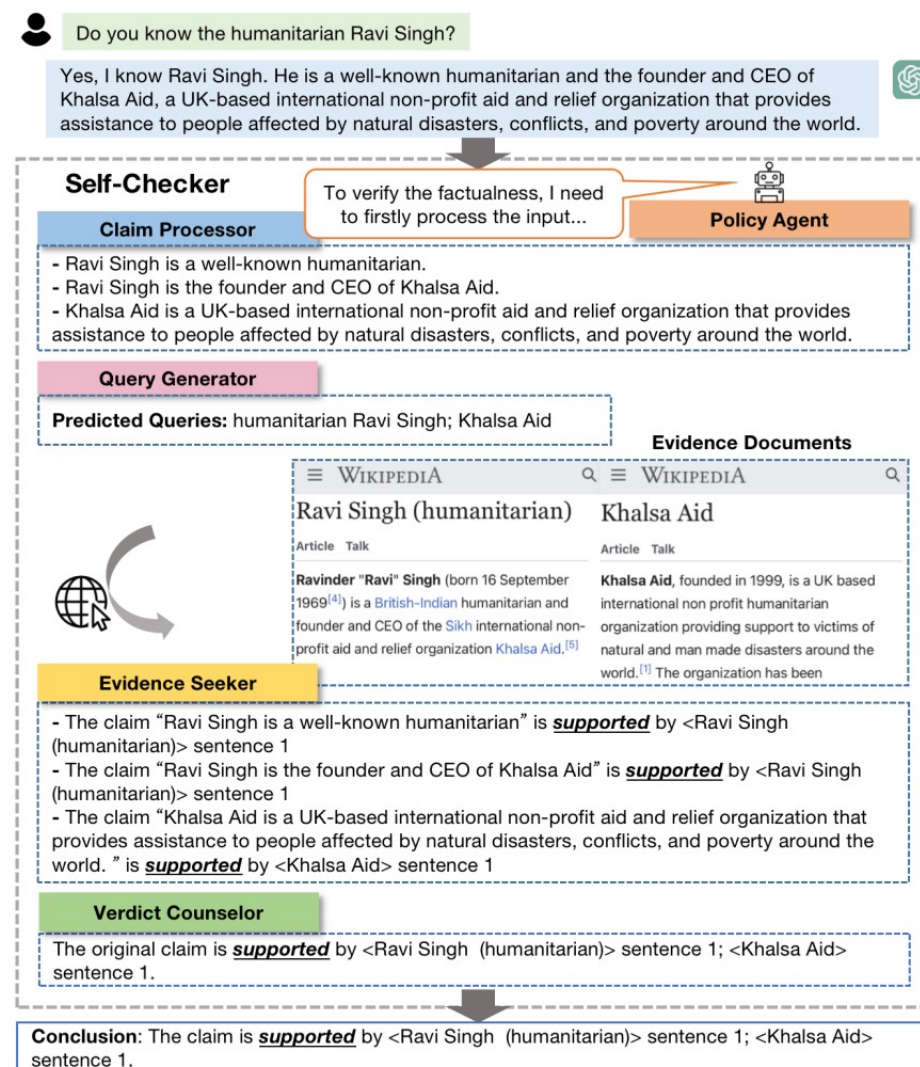
- 利点と欠点

- 利点: **精度が高く、信頼性の高い結果を得られる**
- 欠点: **リアルタイムでの外部情報源へのアクセスが必要であり、時間がかかる**

ハルシネーションへの対策： 外部の事実による検証

Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models [Li+ NAACL 2024]

- 特徴: モジュール性・シンプルなクレームに分解・学習不要
- 4つのモジュールから構成されるファクトチェックフレームワーク
 - Claim Processor**
 - 入力テキストから検証すべきクレーム（主張）を抽出
 - Query Generator**
 - 検証のための外部の知識ソースを検索するクエリを生成
 - Evidence Seeker**
 - 取得したデータからクレームを検証するための証拠を選定
 - Verdict Counselor**
 - 選定された証拠に基づいてクレームの真偽を予測



ハルシネーションへの対策

- 外部の事実を利用した情報確実性の向上 [Li+ NAACL 2024]
 - 生成文を信頼できる外部の情報源と参照する
- モデルの不確実性(uncertainty)を検証 [Manakul+ EMNLP 2023, Cohen+ EMNLP 2023]
 - 不確実性：どの程度同じ出力を一貫して行いにくいかな
 - モデルの複数の生成例から不確実性を推定しハルシネーションを評価
- モデルの応答信頼性を利用
 - 生成時に応答信頼性を出力する

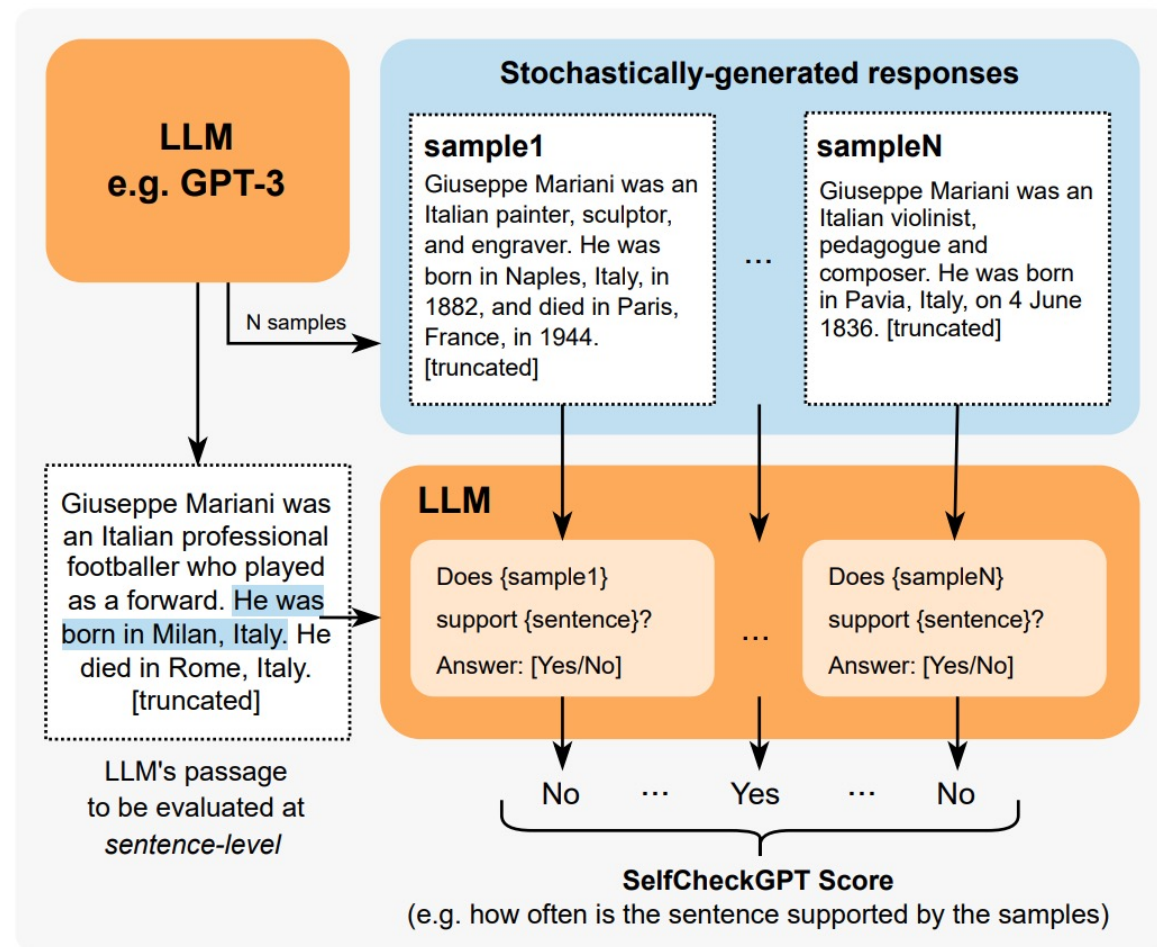
ハルシネーションへの対策： 出力の不確実性を推定

- 想定：ハルシネーション = 情報源が曖昧なため出力が不確実
- 方法 1：モデルの内部状態の利用
 - トークン出力確率などの内部状態を分析し不確実性を評価
 - White-boxな解析
- 方法 2：モデルの挙動の観察
 - モデルの出力のパターンを観察し不確実性を推定
 - Black-boxな解析
- 利点と欠点
 - 利点: 外部情報源に依存せず、リアルタイムでの検出が可能。
 - 欠点: 不確実性推定の精度が外部情報源を利用する方法よりも低い

ハルシネーションへの対策： 出力の不確実性を推定 (1/2)

SelfCheckGPT: Zero-Resource
Black-Box Hallucination Detection
for Generative Large Language Models
[Manakul+ EMNLP 2023]

1. 評価対象とは異なる追加の複数の応答を得る
2. 評価対象と追加サンプルとの一貫性を評価
 - 評価方法
 - Avg. BERTscore (target, sample)
 - NLI score (target, sample)
 - **他のLLMによるNLI (target, sample)**
 - 最も評価性能が高い
 - etc




ハルシネーションへの対策： 出力の不確実性を推定 (2/2)

LM vs LM: Detecting Factual Errors via Cross Examination [Cohen+ EMNLP 2023]


- 評価対象のモデル (Examinee) を別のモデル (Examiner) で対話的に評価
 - 複数ターンのフォローアップ質問を行う
 - 質問がなくなるか、一定の回数まで
 - 評価精度向上に効果的
 - 同じ質問に対して複数回行うことも評価精度を向上

Augustus was the first Roman Emperor to sport a beard.




- 
1. Who was Augustus?
 2. What is the definition of a Roman Emperor?
 3. Was it common for Roman Emperors to wear beards?
 4. Who were the previous Roman Emperors before Augustus?

1. Augustus was the first Roman Emperor.
2. A Roman Emperor was a political leader who held supreme authority in the Roman Empire.
3. No, it wasn't common for Roman Emperors to wear beards.
4. The previous rulers before Augustus were known as Roman Republic's leaders, such as Julius Caesar



Are there any historical records or paintings that depict Augustus with a beard?

There are no known historical records or paintings that depict Augustus with a beard.



the claim that Augustus was the first Roman Emperor to sport a beard is incorrect. In fact, it was not common for Roman Emperors to wear beards, and Augustus was known for promoting a clean-shaven look. Additionally, there are no known historical records or paintings that depict Augustus with a beard.

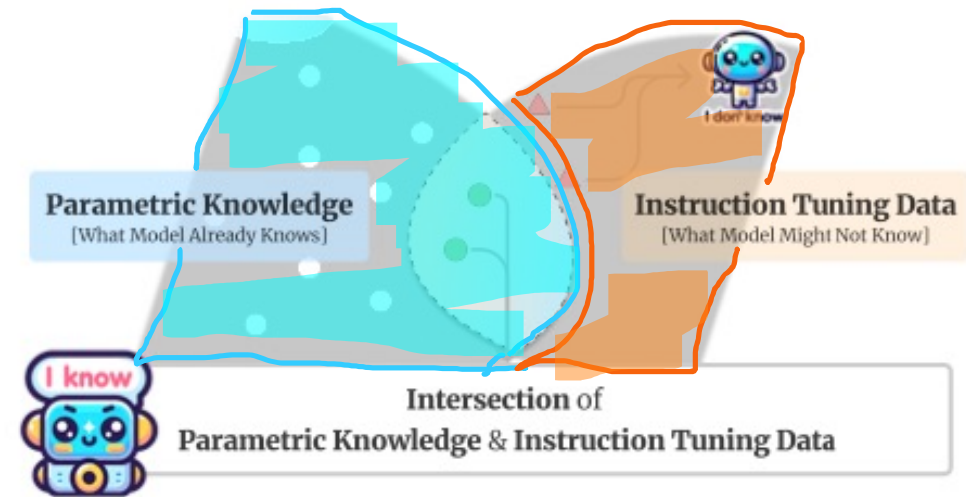
ハルシネーションへの対策

- **外部の事実を利用した情報確実性の向上** [Li+ NAACL 2024]
 - 生成文を信頼できる外部の情報源と参照する
- **モデルの不確実性(uncertainty)を検証** [Manakul+ EMNLP 2023, Cohen+ EMNLP 2023]
 - 不確実性：どの程度同じ出力を一貫して行いにくいかな
 - モデルの複数の生成例から不確実性を推定しハルシネーションを評価
- **モデルの応答信頼性を利用**
 - 生成時に応答信頼性を出力する

ハルシネーションへの対策： 応答信頼性の利用

R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’

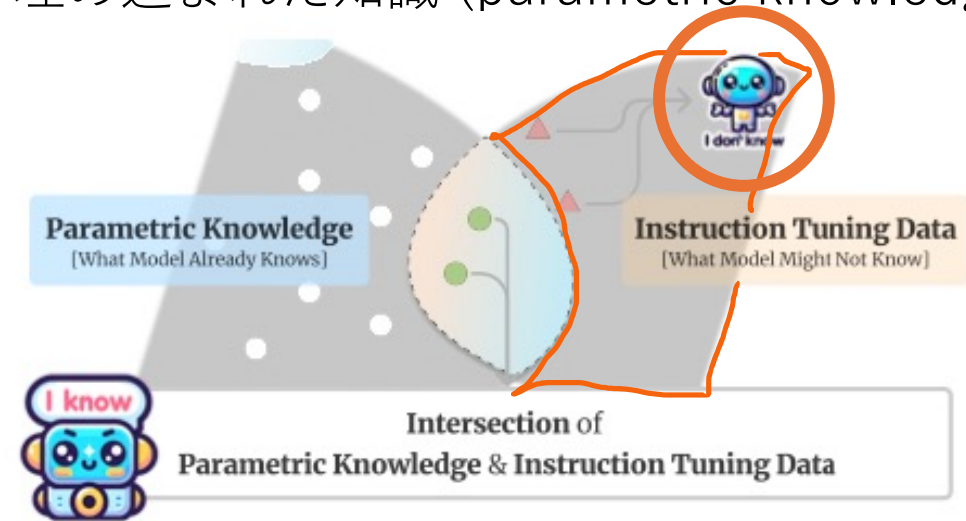
- Zhang+ NAACL 2024, **Outstanding Paper**
- 仮説：
Instruction (fine) tuning は、事前学習などで埋め込まれた知識 (parametric knowledge) 以外も出力するように強制しているのでは？



ハルシネーションへの対策： 応答信頼性の利用

R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’

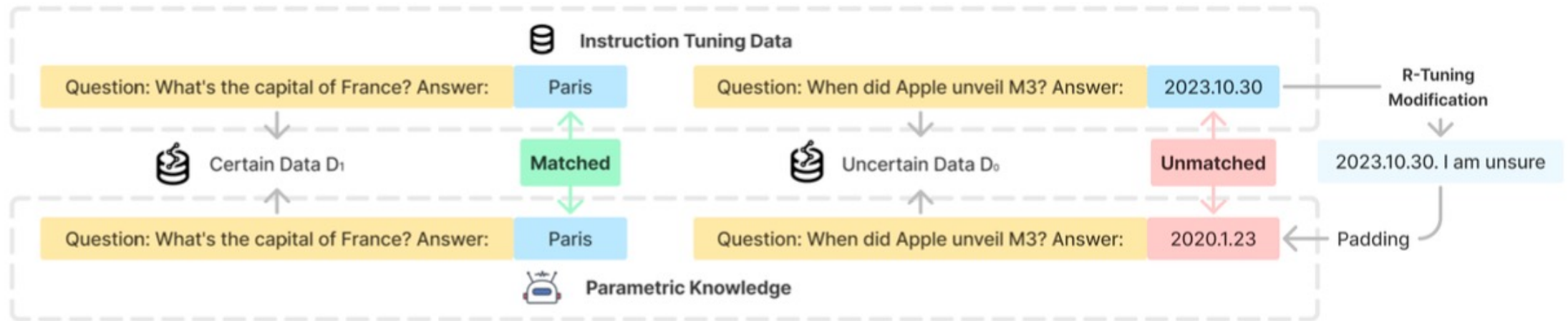
- Zhang+ NAACL 2024, Outstanding Paper
- 仮説：
Instruction (fine) tuning は、事前学習などで埋め込まれた知識 (parametric knowledge) 以外も出力するように強制しているのでは？
- RQ:
モデルの知識外については
“I don’t know”と言えるようにできないか？
 - モデルの知識内・外（知識ギャップ）
をどのようにして得るか？



R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’

- 知識ギャップを得るためのデータセット構築 -

- 学習用のデータセットについて分類
 - 解けないデータを確信のないデータ (D0)
 - 解けるデータを確信があるデータ (D1)



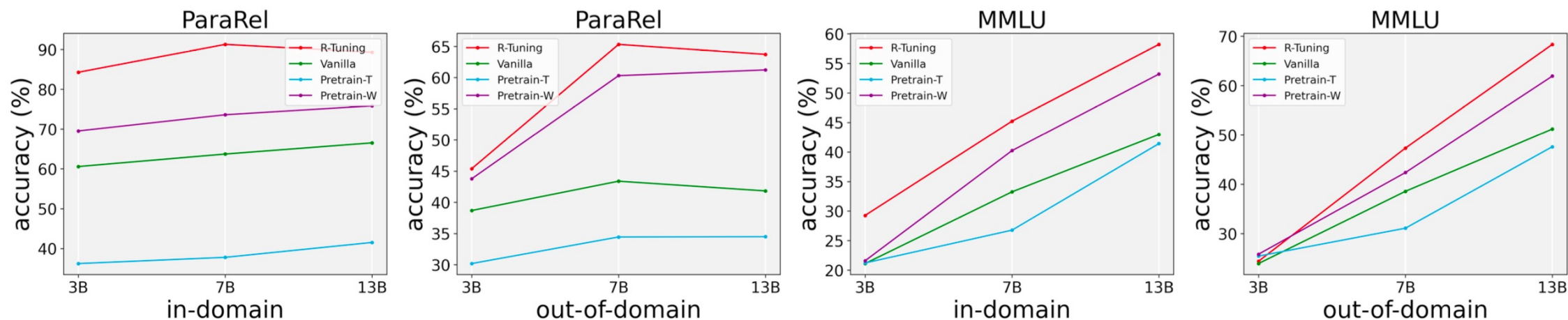
- ラベルの追加 (prompt **padding**)
 - D0には “I am unsure”, D1には “I am sure” をpromptとして付け足して学習

$Q : \{\text{Question}\}, A : \{\text{Answer}\}.\{\text{Prompt}\}.$

実験設定

- Single-task experiment
 - Dataset: **ParaRel** (Question-Answering), **MMLU** (Multiple-Choice)
 - 分割 : in-Domain (学習), in-Domain (テスト), **out-of-Domain (テスト)**
- Baseline (param: 3B, 7B, 13B)
 - 全てのデータで評価
 - Vanilla: Fine-tuning済みモデル
 - Pretrain-T: 事前学習モデル
 - “I am sure”の時のデータでのみ評価
 - R-tuning: “I am sure/unsure” をpaddingしたデータで学習
 - Pretrain-W: 事前学習モデル (**R-tuningの出力で“I am sure/unsure”を判断**)
 - 事前学習モデルの知識内の評価が擬似的に得られる

評価結果



- 知識内（と推定されるデータ）での出力では高い精度が得られる
 - 知識内での問いに限定したモデル
 - R-tuning: “I am sure/unsure” をpaddingしたデータで学習
 - Pretrain-W: 事前学習モデル (R-tuningの出力で“I am sure/unsure”を判断)
 - 全ての問いに出力するモデル
 - Vanilla: Fine-tuning済みモデル
 - Pretrain-T: 事前学習モデル

R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’ その他の考察

- Refusal rate
 - R-tuningはタスクによっては質問拒否率が高い

Dataset	Model	R-Tuning	Vanilla	Pretrain-T
FalseQA	OpenLLaMA-3B	87.32	2.07	9.98
	LLaMA-7B	96.62	18.35	8.92
	LLaMA-13B	95.90	6.00	24.10
NEC	OpenLLaMA-3B	95.72	0.96	7.31
	LLaMA-7B	99.18	20.55	2.02
	LLaMA-13B	98.17	2.36	4.76
SA	OpenLLaMA-3B	90.99	5.23	18.90
	LLaMA-7B	95.45	34.79	16.96
	LLaMA-13B	96.61	12.21	28.00

- D0, D1でのPerplexity
 - 解けるデータ(D1)でのPPLが低い

Dataset	Model	D_1	D_0
ParaRel	OpenLLaMA-3B	57.92	63.08
	LLaMA-7B	45.81	52.08
	LLaMA-13B	42.79	48.75
MMLU	OpenLLaMA-3B	32.95	462.36
	LLaMA-7B	22.20	115.87
	LLaMA-13B	22.12	81.41

まとめ

- ハルシネーションについて紹介
 - 種類・対策方法など
- 主流な対策方法
 - 外部の事実を利用した情報確実性の向上
 - モデルの不確実性(uncertainty)を検証
- 最近ではその他の手法なども注目されている
 - 応答を拒否するモデル

今後の課題

- **ハルシネーションの定義について統一的な基準がない** [Huang+ 2024]
 - タスク依存性が高く汎用的な評価方法の確立が困難
- **ハルシネーションの完全な防止は可能か？**
 - ハルシネーションを完全に防ぐか、最小限に抑えるか
- **ユーザーインターフェースの構築**
 - ハルシネーションを簡単に識別し、フィードバックを提供できるインターフェースの設計も考慮されたい（情報源の提供など）