SecHack365

外部知識なし/モデル内部秘匿/低温度固定状況下での日本語LLMを対象とした「嘘発見器」の作成

研究駆動コース
20R　中井諒馬

1. 背景/目的

LLMがつく「嘘」

LLMが嘘発見器で試験されている様子→

- ChatGPT等のLLM(大規模言語モデル)に質問をすると「嘘」を返されることがあり**hallucination(幻覚)**と呼ばれる
- 誤情報がそこから広まってしまう可能性があり、hallucinationの検出には需要がある

LLMの温度パラメタ

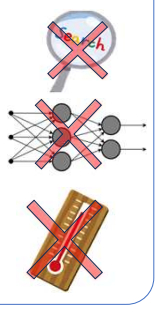
```
res = openai.Completion.create(
  engine=ultimate_engine,
  prompt=some_awosome_prompt,
  temperature=0.7, ←温度パラメタ
)
```

OpenAIのAPIで温度を設定して文章生成を行うコード

- 温度というパラメタを指定することでLLMに文章を生成する際に出力のランダムネスを制御することが可能
  - 低温にすると出力が安定し、正確性を要するタスクに有用
- 温度パラメタをユーザが指定できないLLMサービスも存在
  - 例:AI chatサービスChatGPTのWeb版、AI検索エンジンPerplexity

先行研究

- SelfCheckGPT:
  - 以下のような厳しい条件下でhallucinationの検出が可能
    - Zero-Resource : 外部知識の利用禁止
      - Google検索や独自データベースなど×(コスト大💰)
    - Black-Box: LLMの内部情報の利用禁止
      - token生成確率分布など×(未公開モデルもある🔒)



目的

- 外部知識なし/モデル内部秘匿/低温度固定状況下でのhallucination検出手法の探求

2. 先行研究手法

先行研究のアイデア

知ったかぶりをする人は聞くたびにバラバラな回答をしがち

LLMが確信をもって回答文を出力 ▶ 回答に一貫性がある

- 温度を高温へと変更後に同じ入力を何度も与えて出力をサンプリングし、**出力のばらつきやすさ**を定量化
- ユーザが検証したい出力文とサンプリングで得られた出力文群とからhallucinationリスクをスコアリング
- hallucinationが起こっているような文にのみ高いスコアを付与することが目的

先行研究でのスコアの計算方法

文と文の類似度を測る指標BERTScoreを利用し検証対象rを文単位でスコアリング

$$HallucinationScore(r_i) = 1 - \frac{1}{N}\sum_{n=1}^{N} \max_k BERTScore(r_i, s_{n,k})$$

N: サンプリングの回数 r_i: rのi文目 $s_{n,k}$: n個めのサンプルのk文目

検証対象の内容が多くのサンプルに入っているほど小さなスコアになる

3. 提案手法

問題点

低温条件下での幻覚発生時のサンプリングの様子(想定解：サルモネラ)

サンプリング方法	SelfCheckGPT	提案手法（逆翻訳）
サンプル 1	黄色ブドウ球菌。	アニマス菌 (サーモン病原性細菌) は、…
サンプル 2	黄色ブドウ球菌。	黄色ブドウ球菌 (Staphylococcus aureus) は、…
サンプル 3	黄色ブドウ球菌。	アニマス菌 (サーモン病原性細菌) は、…
サンプル 4	黄色ブドウ球菌。	エンテロコッカス・フェタミン耐性菌 (EFDB) は、…
サンプル 5	黄色ブドウ球菌。	アニマストラス菌 (別名サーモン・スクワッド) は、…

- 温度パラメタの操作が必要
  - 理由
    低温度で運用されているLLMでは出力がばらつかず同じ文ばかり出力される
    →ほぼすべてのスコアが 0 になり検出性能が大幅に悪化

ユーザが温度を高温に変更できないようなLLMサービスでは適用不可

提案手法

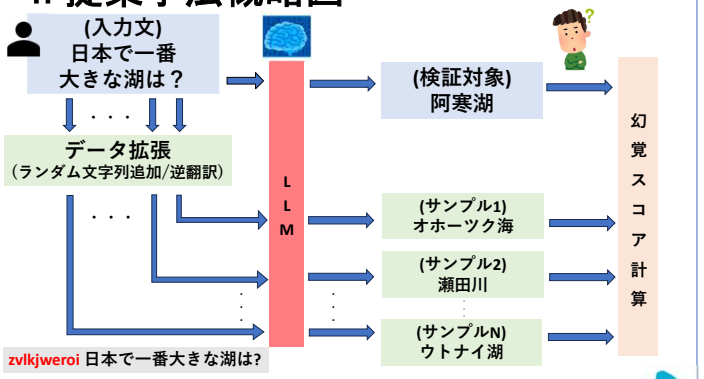
聞き方が変わっても答えは同じ

良く学習できている人は少しぐらい問題文が変わっても同様に回答可能

- サンプリング時に利用するクエリを毎回変更する
  - ランダム文字列の追加(alphabet or 平仮名)
  - 逆翻訳を利用した言い換え with gpt-3.5-turbo (日→英→他言語→英→日)
- クエリのデータ拡張に対する出力の変化を定量化

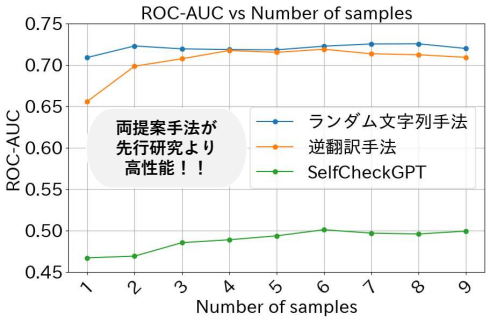
各種パラメータへのアクセスが制限されている場合でも
入力文自体を変更したサンプリングにより
文章入出力インタフェースのみの提供でも
hallucinationの検出ができる可能性😎

4. 提案手法概略図

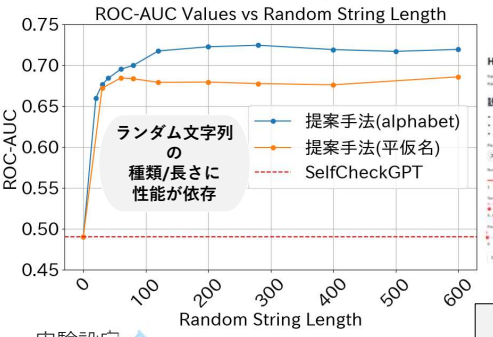


このように検証対象に対してサンプルが一貫していない場合は
hallucinationのリスクが高いと判定

5. 実験結果



逆翻訳に利用した言語
英語
中国語
フランス語
ドイツ語
イタリア語
韓国語
ポルトガル語
ロシア語
スペイン語

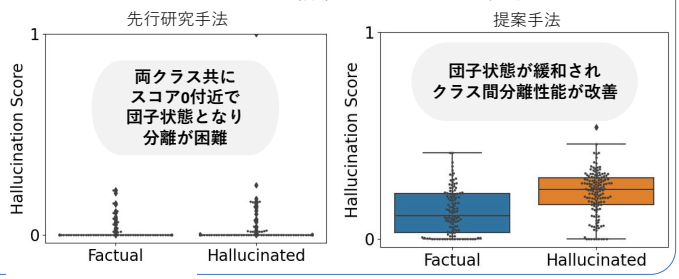


デモサイト

- 実験設定
  - 低温度(0.01)固定状況の国産LLMを想定
  - 入力クエリにはクイズを利用
  - 二値分類の評価指標ROC-AUCで評価
- ROC-AUC⤴で先行研究手法が改良

利用データセット詳細
入力200クエリ
出力263文
嘘：56%，事実：44%
品質
(二人の作業者のラベル一致度合い)
Cohen's kappa = 0.749
Substantial Agreement
(かなりの一致)

スコアリング結果のクラスごとの分布



6. まとめ

- **サンプリング時クエリの変更**により
  外部知識なし/モデル内部秘匿/低温度固定状況下における
  hallucination検出性能を改善することに成功
- **文章入出力インタフェースのみを用いた**
  検出手法の開発への貢献💪
- 2024年5月に人工知能学会全国大会(第38回)で発表予定

https://github.com/ryoryon66/variable_prompt_selfcheckgpt