

[説明資料(提出ファイル)] 発明・工夫作品コンテスト 製作の動機または目的, 利用方法, 作品 자체やその製作過程で工夫したことを, 文章, 写真, 図などで説明。この用紙1枚に記入し, PDFに変換した後, web提出フォームにて提出する。

個人・グループ名	小林颯太	大学名	大阪電気通信大学
作品名	AIの多数決チェックによる生成 AI誤回答検出システム	人数	1人

開発背景と目的

生成AIを使用する際に、もっともらしい誤情報(ハルシネーション)が含まれてしまうという問題がある。そこで、メインのAIとのやり取りを異なる種類のAIが観察し、事実関係の誤りや計算誤りなどの検証可能な誤りをユーザーに伝えるシステムを開発し、AIに対する信頼性を高めることを目標とした。

システム概要

本システムでは会話用AIとチェック用AIの2種類を使用している。チェック用AIは会話用AIとの会話を監視し、AIの回答に誤りが見つかった際にユーザに警告を表示する。図1に誤回答検出の例を示す。 「strawberryに含まれるrの数を数えて」という質問に対する会話用のAIからの「2つ」という誤回答をチェックAIが検出し、警告メッセージを表示している。図2にシステム構成図を示す。

1. ユーザーは画面から質問(プロンプト)を入力し会話用AIのAPIを利用して会話する。
2. システムは質問と回答をチェックAIに送信し会話を確認する。問題があれば表示する。
3. 多数決方式により、2つ以上のAIが誤りを指摘した場合に警告を表示する。



図1 誤回答検出の例

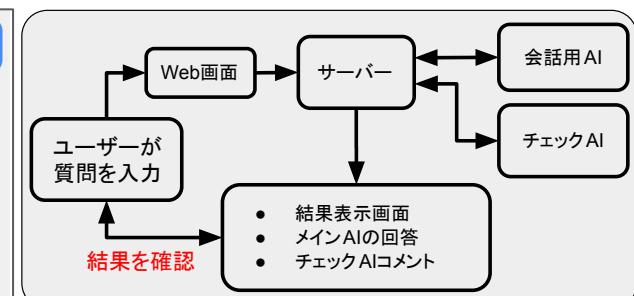


図2 システム構成図

工夫点

- ローカルのサーバーにAIをインストールして使用した
- 次の4種類から1個を会話用に選択し、他の3種類で監視を行う
(1)qwen3:8b、(2)gemma3:12b、(3)gpt-oss:20b、(4)deepseek-r1:32b
- チェックは多数決方式を採用し、AIごとの知識の違いによる多面的なチェックを可能にした

評価実験

実験では(1)のAIを会話用に使用した。また、(3)のAIを単独で使った場合と比較した。図1のような12個の質問を使用した。

結果を表1に示す。誤回答検出率はチェック用AIが検出できた誤回答の割合で、正答回答率は結果として正しい情報を回答できた割合である。本システムを使うことで、AIを単独で使うときと比べて、正答回答率を17ポイント向上させることに成功した。

最近の東京オリンピックの開催日は?
20%引きのさらに15%引きの値段は?

図1 実験に使用した質問例

表1 本システムと単独AIの比較

	本システム	単独AI
誤回答検出率	92%	-----
正答回答率	84%	67%

まとめと今後の課題

複数のAIによる多数決チェック方式により、単独AIと比較した実験でAIのハルシネーション問題に対する有効性を確認できた。今後は教育現場での活用を想定し、生徒に正確な情報を提供しながら学習支援を行う教育用AIチャットボットへの応用を研究していきたい。