B2-17　ハルシネーションから学ぶ：内部表現への介入によるハルシネーション抑制

○門谷宙, 西田光甫, 西田京介 (NTT株式会社 人間情報研究所)

概要

課題認識：嘘つきLLM (anti-expert) を用いる手法は, ハルシネーション抑制の最高性能を達成している有望な手法だが, 同時に2つのLLMを動かすため推論コストが高い

提案手法：in-model anti-expert (IMAE) は, LLMの内部表現に介入して事実性が向上する方向に編集することで, モデル単体でのハルシネーション抑制を実現

実験結果：IMAEは従来のanti-expertよりもはるかに軽量ながら, anti-expert以外の既存手法を上回る精度を達成

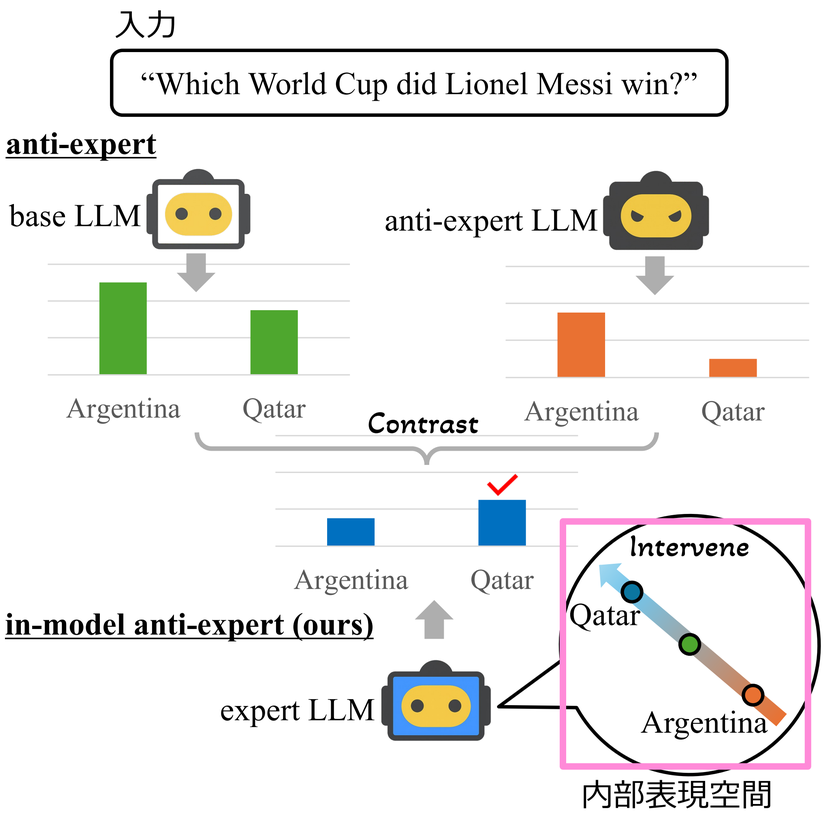
既存手法：anti-expert [Zhang+, 25]

- 事実性を直接向上させるfine-tuningはハルシネーションを増加 [Yang+, 24]
- base LLMを嘘を含むデータでfine-tuningしてanti-expert LLMを構築
- 推論時にbase LLMからanti-expert LLMの出力確率をペナルティとして差し引くことで事実性が向上した出力確率を得る

利点と課題

☺ ハルシネーション抑制の最高性能を達成, anti-expertを適用したLlama2-7B-ChatはGPT-4の精度を上回る

☹ 推論コストが高く, **2.2倍**のGPUメモリ使用量と**1.9倍**の推論時間が必要



提案手法：in-model anti-expert (IMAE)

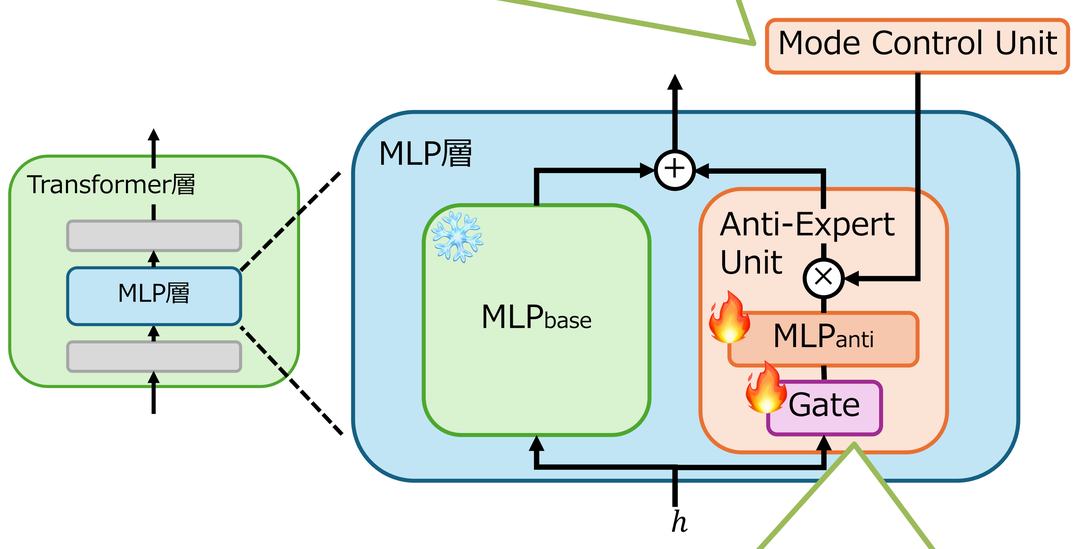
モデル構造

- パラレルアダプタ [He+, 22] に基づくモデル構造
- anti-expert unitをbase LLMの各MLP層と並列に追加, mode control unitも追加
- 3つの生成モードを導入:
  - anti-expertモード：$\sigma = 1$で事実性が低下したテキストを生成
  - baseモード：$\sigma = 0$でbase LLMと同じテキストを生成
  - expertモード：$\sigma = -1$で事実性が向上したテキストを生成
- MLP層の計算式:

$$\alpha = \mathrm{softmax}(Wh + b)_0$$

$$\mathrm{MLP}(h) = \mathrm{MLP}_{\mathrm{base}}(h) + \sigma \cdot \mathrm{MLP}_{\mathrm{anti}}(\alpha h)$$

各モードに対応するスカラ値$\sigma \in (-1, 0, 1)$を出力



損失関数

- base LLMを凍結してanti-expert unitのパラメータのみを調整
- $\mathrm{MLP}_{\mathrm{anti}}$の出力の逆ベクトルが, $\mathrm{MLP}_{\mathrm{base}}$の出力ベクトルの事実性を高める方向を向くように調整 (図のピンク部分)
- 質問と誤答のペアで構成された訓練データを分割し, 半分をanti-expertモード, もう半分をexpertモードの訓練に使用してマルチタスク学習を適用
  - anti-expertモード：cross-entropy損失
  - expertモード：Kullback-Leibler距離

$$\mathrm{L}_{\mathrm{expert}} = \sum_i \mathrm{D}_{\mathrm{KL}}\left(p_{\mathrm{target}}(x_i) || p_{\mathrm{expert}}(x_i)\right)$$

expertモードの出力確率

事実性が向上した出力確率, baseモードからanti-expertモードの出力確率をペナルティとして差し引くことで得る

評価実験

設定

- 訓練データ：HaluEval [Li+, 23] (10,000件)
- 評価データ：TruthfulQA [Lin+, 21] (817件)
- base LLM：Llama2-7B-Chat
- 評価指標：
  - 事実性：MC1
  - 推論コスト：GPUメモリ (GB), 推論時間 (ms/token)

	MC1↑	GPUメモリ↓	推論時間↓
base	36.96	13.2 (**1.0x**)	2.09 (**1.0x**)
anti-expert [Zhang+, 23]	46.32	28.6 (**2.2x**)	4.05 (**1.9x**)
ITI [Li+, 23]	37.01	16.2 (**1.2x**)	2.09 (**1.0x**)
Dola [Chuang+, 23]	32.97	15.1 (**1.2x**)	2.21 (**1.1x**)
CD [Li+, 23]	28.15	41.0 (**3.1x**)	6.42 (**3.1x**)
IMAE (ours)	40.02	18.4 (**1.4x**)	2.60 (**1.2x**)

結果

- IMAEは, MC1においてanti-expert以外の既存手法を上回る精度を達成
- IMAEは, 従来のanti-expertのGPUメモリ使用量を2.2倍から**1.4倍**, 推論時間を1.9倍から**1.2倍**に改善