Learning from Hallucinations: Mitigating Hallucinations in LLMs via Internal Representation Intervention

○Sora Kadotani, Kosuke Nishida, Kyosuke Nishida (NTT, Inc.)

Summary

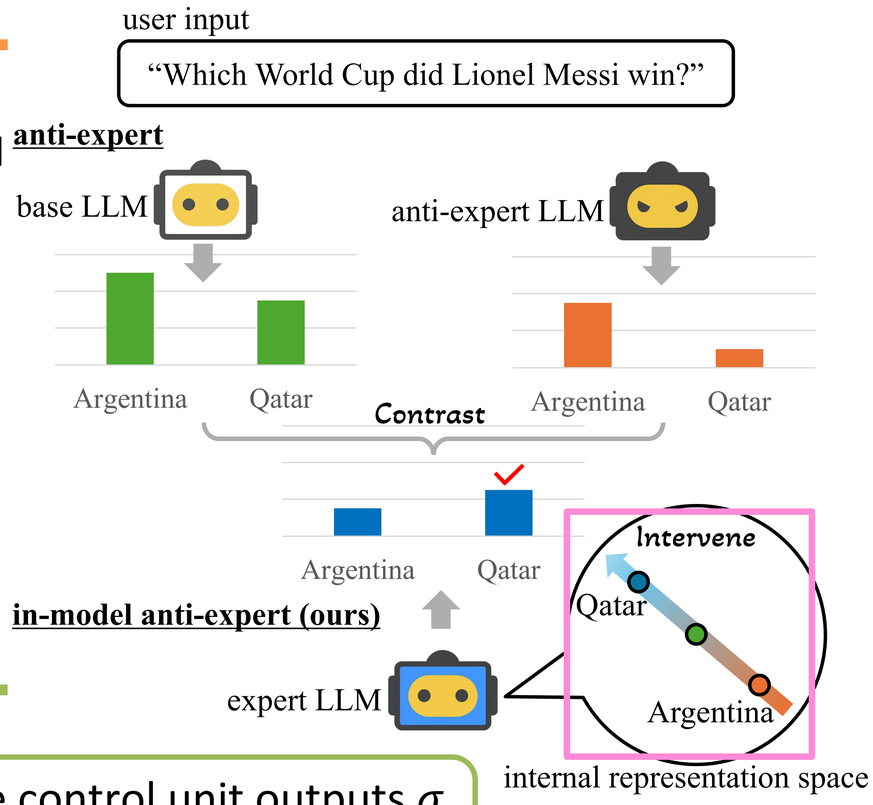
Problem: Hallucination mitigation methods using non-factual LLMs (anti-expert) are effective but require high computational costs because the two LLMs are run simultaneously

Proposal: Our in-model anti-expert (IMAE) mitigates hallucinations with a single LLM by intervening to change the internal representations in the direction of improving factuality

Results: IMAE was less costly than the conventional anti-expert method and outperformed baselines

Existing Method: Anti-expert [Zhang+, 25]

- Fine-tuning using factual answers make LLMs to hallucinate [Yang+, 24]
- They created an anti-expert LLM using hallucinated answers
- They obtained the output distribution of a factual LLM (expert) by contrasting the output distribution of the base and anti-expert LLM



Pros and Cons

- ☺ Anti-expert has achieved state-of-the-art performance
- ☹ Anti-expert requires **2.2x** more GPU memory usage and incurs **1.9x** higher latency

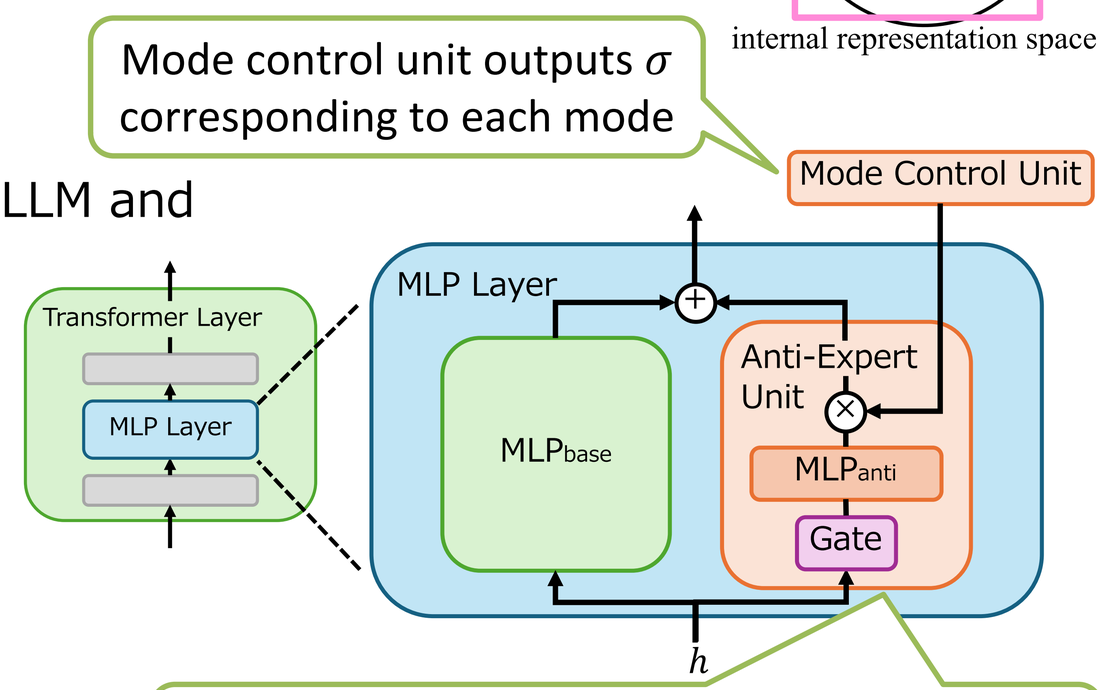
Proposed Method: In-model Anti-expert (IMAE)

Model Architecture

- Our architecture is based on parallel adapter [He+, 22]
- We add an anti-expert unit to each MLP layer of the base LLM and a mode control unit
- IMAE operates in three modes:
  - ☐ Anti-expert mode: $\sigma = 1$ for generating non-factual text
  - ☐ Base mode: $\sigma = 0$ for replicating the base LLM output
  - ☐ Expert mode: $\sigma = -1$ for generating factual text
- Computation of the output of the MLP layer:

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

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



Loss Function

- We fine-tune the anti-expert unit so that the opposite vector of $\mathrm{MLP}_{\mathrm{anti}}$ output points in the direction that the factuality of the output vector of $\mathrm{MLP}_{\mathrm{base}}$ (as illustrated in the pink area)
- We use a dataset in which each sample consists of a question and its hallucinated answer and apply multi-task learning
  - ☐ Anti-expert mode: cross-entropy
  - ☐ Expert mode: Kullback-Leibler divergence

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

Output probability of the expert mode

Factual probability calculated by contrasting the output distributions of the base and anti-expert mode (like [Zhang+, 25])

Evaluation

Settings

- Train data: HaluEval [Li+, 23] (10,000 samples)
- Test data: TruthfulQA [Lin+, 21] (817 samples)
- Base LLM: Llama2-7B-Chat
- Evaluation metrics:
  - ☐ Factuality: MC1
  - ☐ Cost: GPU memory usage (GB), latency (ms/token)

	MC1↑	memory↓	latency↓
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**)

Results

- IMAE outperformed the existing methods in MC1, except for the conventional ant-expert method
- IMAE improved GPU memory usage form 2.2x to **1.4x** and latency from 1.9x to **1.2x**