Hallucination Detection in LLMs: Fast and Memory-Efficient Fine-Tuned Models

Gabriel Y. Arteaga[1,2], Thomas B. Schön[2], and Nicolas Pielawski[2]

[1]Department of Informatics, University of Oslo [2]Department of IT, Uppsala University

UNIVERSITY OF OSLO

UPPSALA UNIVERSITET

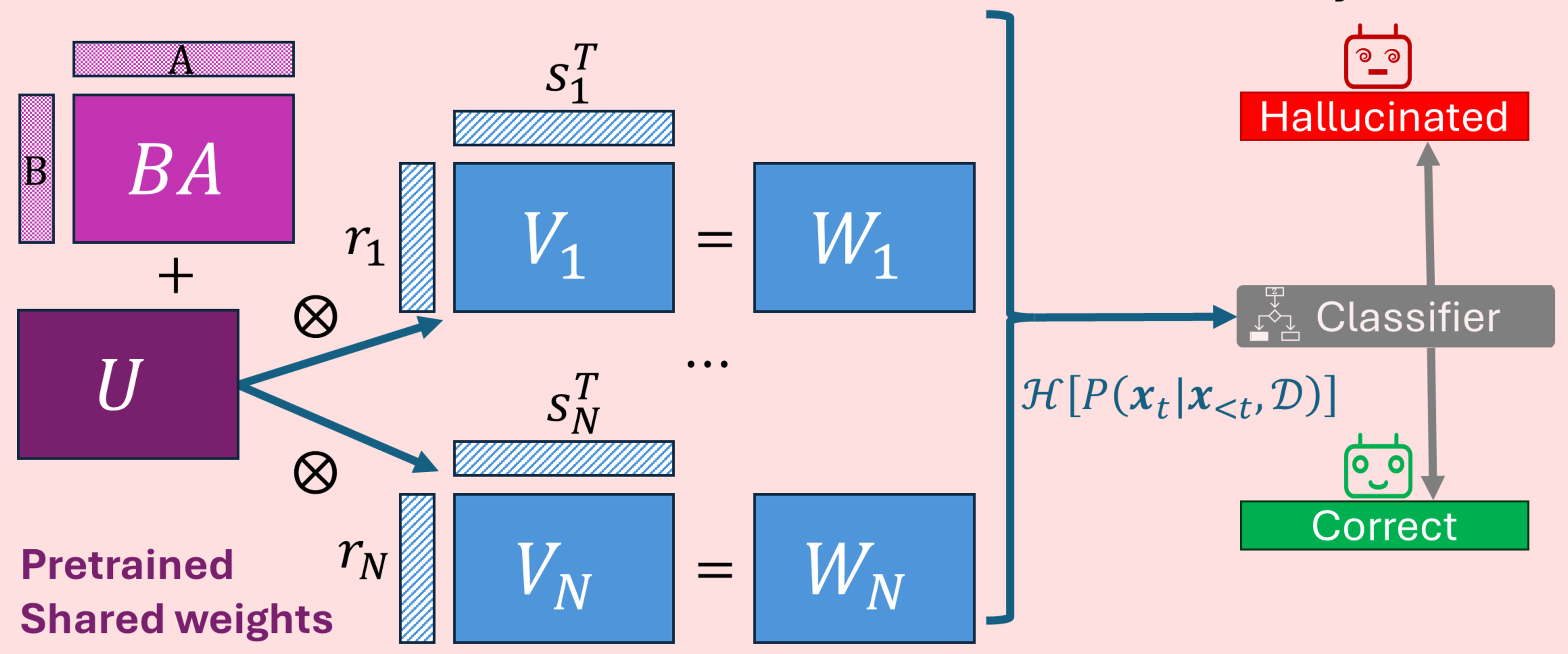
1 Motivation

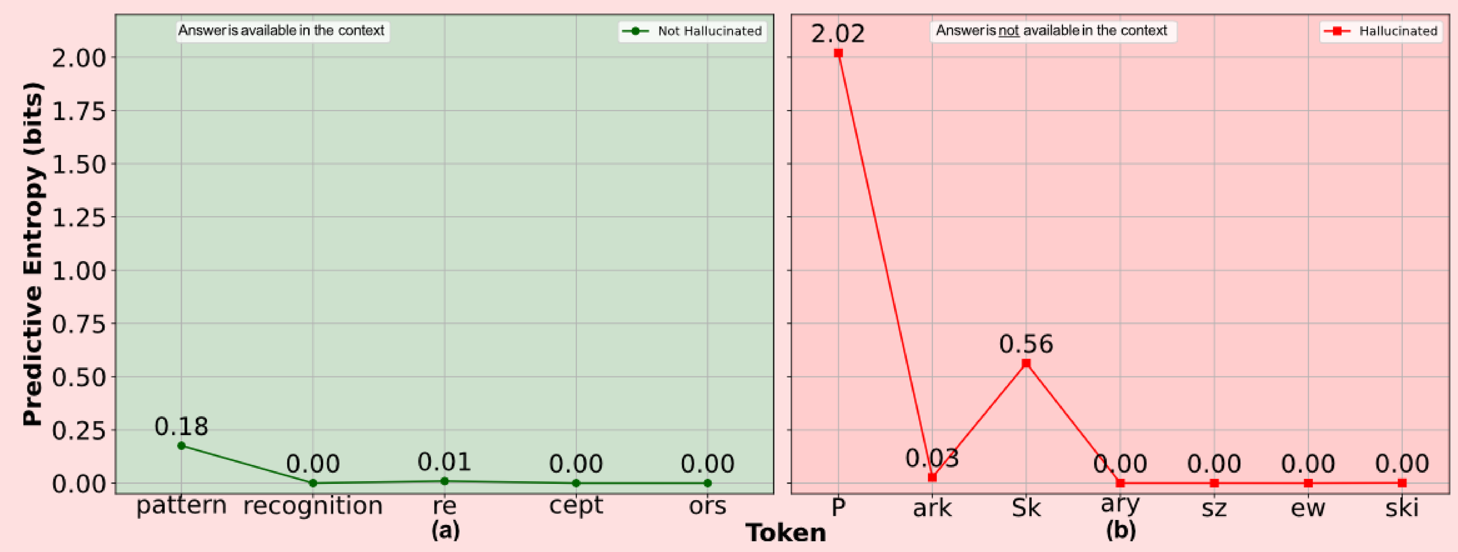
- Hallucinations in LLMs pose significant risks in safety-critical fields such as healthcare.
- Existing hallucination detection methods are often task-specific or unreliable.
- Deep ensembles are effective but computationally infeasible for large LLMs.
- Scalable, resource-efficient approaches to uncertainty estimation are needed to enable reliable hallucination detection in large-scale LLMs.

2 Problem Statement

- Faithful hallucinations occur when outputs deviate from instructions, while factual hallucinations produce content that contradicts verifiable facts; both pose distinct challenges for reliable LLM outputs.
- Existing hallucination detection methods are often tailored to a specific task, limiting their versatility.
- Current uncertainty-based approaches often rely on perturbations through sampling, which can be unreliable.
- Traditional deep ensembles scale linearly with the number of parameters, making them infeasible for LLMs with billions of parameters.

3 Method





Memory-Efficient Ensemble

BatchEnsemble

$W_i = U \odot V_i$, where $V_i = r_i s_i^T$ and $r_i \in \mathbb{R}^{m\times 1}, s_i \in \mathbb{R}^{n\times 1}$

Savings in Memory Complexity

$O(Mmn) \rightarrow O(mn + M(m+n))$ per layer where M is the ensemble size

Low-Rank Adaptation (LoRA)

$U = U_0 + BA$, where $B \in \mathbb{R}^{m\times r}, A \in \mathbb{R}^{n\times r}$ and U_0 is a pre-trained model

Hallucination Detection

Predictive Entropy

$$\mathcal{H}[P(x_t|x_{<t};\mathcal{D})] = -\sum_{x_t} P(x_t|x_{<t};\mathcal{D}) log P(x_t|x_{<t};\mathcal{D})$$

Ensemble Approximation

$$P(x_t|x_{<t};\mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M} P(x_t|x_{<t};\mathcal{D})$$

Binary Classification

$$f(\mathcal{H}[P(x_t|x_{<t};\mathcal{D})]) = \hat{y}, \text{ where } \hat{y} \in \{0,1\}$$

4 Results

Classification Accuracy on Hallucination Detection

Method	Faithfulness ↑	Factual ↑	OOD ↑
(Ours) BatchEnsemble	**97.8**	68.0	62.4
(Ours) BatchEnsemble + NI	96.5	66.9	61.9
LoRA Ensemble	92.5	**73.9**	**63.3**
Sample-Based	92.1	69.6	62.2

Predictive Performance

Dataset	SQuAD		MMLU
Metric	Exact Match ↑	F1 Score ↑	Accuracy ↑
Single Model	85.1	92.1	56.3
(Ours) BatchEnsemble	**85.9**	**93.4**	**56.7**
(Ours) BatchEnsemble+NI	85.4	92.6	53.2
LoRA Ensemble	68.4	84.4	44.6

5 Conclusions

- **Hallucination Detection:** Developed an uncertainty-based method capable of detecting both factual and faithful hallucinations while maintaining effective performance.
- **Memory-Efficient Ensemble:** Demonstrated the feasibility of using BatchEnsemble for large-scale LLMs with over 7B parameters, optimizing memory usage.
- **Cost-Effective Training:** Achieved significant reductions in training overhead by integrating LoRA with BatchEnsemble, enabling the training of a 4-member 7B parameter ensemble on a single A40 GPU.
- **Future Directions:** Investigate the relationship between aleatoric uncertainty and faithful hallucinations, and epistemic uncertainty and factual hallucinations, to improve detection strategies.

Paper Github LinkedIn

European Research Council
BEIJERSTIFTELSEN
Knut and Alice Wallenberg Foundation
NAISS
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM