Instruction-tuned Quantized Small Language Models (SLMs): A Study on Hallucination Detection

Elijah Soba[1,†], Harika Abburi[2,†], Nirmala Pudota[2,†], Jain Aayush[2], Balaji Veeramani[1], Edward Bowen[1] and Sanmitra Bhattacharya[1]

[1]*Deloitte & Touche LLP, USA*
[2]*Deloitte & Touche Assurance and Enterprise Risk Services India Private Limited, India*

Abstract

Large Language Models (LLMs) have greatly advanced the field of Natural Language Generation (NLG). Despite their remarkable capabilities, their tendency to generate hallucinations—a phenomenon where models generate inaccurate or misleading information continues to be a significant challenge to their broader adoption across various domains. In this paper, we investigate the impact of instruction-tuned quantized Small Language Models (SLMs) (defined as models with fewer than 15 billion active parameters), specifically trained on a subset of Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (SHROOM) dataset for hallucination detection. We focus on SLMs to achieve a balance between computational efficiency and performance in detecting hallucinations. The instruction-tuned quantized models are compared against the Generative Pre-trained Transformer (GPT-4) and traditional "textual entailment" (entailment) based methods across various datasets. Our findings demonstrate that the optimized SLMs achieve performance comparable to LLMs like GPT-4 and outperform traditional textual entailment-based methods in hallucination detection. This research highlights the potential of smaller, instruction-tuned language models as practical and efficient solutions for improving the reliability of language models, especially in resource-constrained environments.

Keywords

Hallucination Detection, Small Language Models, Large Language Models, Instruction Tuning

1. Introduction

The domain of Natural Language Generation (NLG) is witnessing a remarkable transformation with the emergence of Large Language Models (LLMs) [1, 2]. LLMs have been shown to outperform traditional Natural Language Processing (NLP) approaches across a wide range of applications [3, 4]. Despite the rapid advancements in LLMs, a concerning trend has emerged where these models generate hallucinations [5, 6], resulting in content that appears plausible but is factually unsupported. This issue is particularly critical in sensitive domains such as healthcare, finance, and legal services, where the accuracy of generated content is paramount. Hence, the automatic detection of hallucinated content has become an active area of research, aiming to enhance the reliability and trustworthiness of LLM-generated content [7, 8].

Diverse modeling strategies, ranging from Black-Box, White-Box to evidence-based approaches [8, 7], have been investigated to develop solutions for detecting hallucinated content. Black-Box methods analyze the consistency of LLM's outputs through follow-up questions with other LLMs [9] or prompting the LLM for self-evaluation [10]. [11] proposed semantic-aware cross-check consistency (SAC^3), a sampling-based approach that builds upon self-consistency checks by incorporating semantically equivalent question perturbations and cross-model response consistency verification techniques. Similarly, [12] introduced SelfCheckGPT, which detects inconsistencies by evaluating the stability of

Disinformation, Misinformation and Learning in the Age of Generative AI: Joint Proceedings of the 1st International Workshop on Disinformation and Misinformation in the Age of Generative AI (DISMISS-FAKE'25) and the 4th International Workshop on Investigating Learning during Web Search (IWILDS'25) co-located with 18th International ACM WSDM Conference on Web Search and Data Mining (WSDM 2025)

[†]These authors contributed equally.

elsoba@deloitte.com (E. Soba); abharika@deloitte.com (H. Abburi); npudota@deloitte.com (N. Pudota); aayushjain58@deloitte.com (J. Aayush); baveeramani@deloitte.com (B. Veeramani); edbowen@deloitte.com (E. Bowen); sanmbhattacharya@deloitte.com (S. Bhattacharya)

generated responses. These methods assume that inconsistencies arise when LLM is uncertain about a concept. However, both approaches require multiple response generations from LLMs, making them computationally expensive for practical applications.

The White-Box approaches explore the internal workings of LLMs to analyze factual recall. [13] analyzed how LLMs encode factual statements with a specific structure. They proposed the multi-layer perceptron layers store facts, and transferred through attention layers that focus on subject tokens. Similarly, [14] leveraged the activations of hidden layers as inputs to a classifier designed to assess the truthfulness of statements. [15] proposed constraint SATisfaction (SAT) Probe, a method probing attention patterns, to predict factual errors and allow early error identification. While these approaches are promising for hallucination detection, their implementation remains challenging as access to the inner workings of LLMs is not always feasible.

Recently evidence-based fact-checking gained significant attention as an essential tool for combating misinformation. Factual precision in Atomicity Score (FACTSCORE) by [16] evaluated the correctness of individual facts within the generated text by referencing a knowledge source. [17] introduced a real-world claim and evidence dataset specifically designed to enhance textual entailment models by reducing the complexity of claims through a decomposition process. By breaking down claims into simpler components, this approach aims to facilitate more effective entailment evaluation and thereby improve overall model performance. [18] presented an automated pipeline for fact-checking real-world claims by retrieving raw evidence from the web. This method retrieves a fixed number of documents for each claim. But this predetermined approach may not always provide sufficient evidence, potentially resulting in incomplete or biased fact-checking. To address this limitation, [19] proposed a framework that leverages statistical decision theory and Bayesian sequential analysis, which eliminates the need for a predetermined number of observations. The analysis proceeds sequentially, enabling a quick decision-making process through a stop-or-continue strategy. While these evidence-based approaches benefit from real-world knowledge, they may introduce additional sources of error and are often limited to addressing only the fact-checking form of hallucinations.

This paper examines a specific scenario of hallucination detection, where the objective is to predict which hypothesis is a hallucination given a triplet consisting of a source input and two hypotheses. The contribution of this study is twofold.

- We explore the impact of instruction-tuned, quantized SLMs and compare their performance against both textual entailment models and GPT-4.
- Our results demonstrate that instruction-tuned, quantized SLMs achieve performance comparable to GPT-4 while offering significant advantages in terms of computational efficiency.

2. Datasets

This section describes the datasets used for instruction-tuning and evaluating our hallucination detection model. The number of training and testing samples are shown in Table 1.

Table 1
Training and testing data splits across all the datasets

Dataset	Training	Testing
SHROOM	538	115
HaluEval	–	1000

2.1. SHROOM

The SHROOM dataset is released as part of the SemEval-2024 shared task for hallucination detection. It contains data from three distinct NLG tasks: Machine Translation (MT) and Paraphrase Generation (PG).

Table 2
Examples of source and hypotheses triplets from the SHROOM dataset.

Input	Label
source: *I didn't give you enough credit.* ***hypothesis 1:*** *I didn't give you enough credit.* ***hypothesis 2:*** *I gave you enough credit.*	hypothesis 2
source: *Tokyo ekozala engumba moko pamba ya Asie oyo eyambi masano ya Oympique ya eleko ya mibale, eyambaki ya liboso na 1964.* ***hypothesis 1:*** *Tokyo will be the only Asian city to have hosted two summer Olympics, having hosted the games in 1964.* ***hypothesis 2:*** *Tokyo will be the only Asian city to host the second Olympic Games, the first being in 1964.*	hypothesis 2
source: *Medas de sas traditziones a inghÃ¬riu de sa festa sunt istadas adotadas fintzas dae sos chi non creent in sos paisos cristianos e dae sos non cristianos in totu su mundu.* ***hypothesis 1:*** *Many of the traditions surrounding the festival have been adopted by non-Christian people in their Christian countries and by non-Christian people around the world.* ***hypothesis 2:*** *Many of the traditions surrounding the holiday have been adopted also by non-believers in Christian countries and non-Christians around the world.*	hypothesis 1
source: *James, we shouldn't be here.* ***hypothesis 1:*** *James, we're supposed to be out of here.* ***hypothesis 2:*** *We shouldn't be in this situation.*	hypothesis 1

More details about the dataset can be found in the SemEval-2024 shared task 6 overview paper [20]. For this work, we consider data from MT and PG tasks with source, target, hypothesis, and label details . To enable the model to simultaneously learn the characteristics of hallucinations while also identifying the patterns that differentiate them from non-hallucinations, we transform the data into triplets. Each triplet consists of an original input sentence (source) paired with two hypotheses (hypothesis 1, hypothesis 2): one representing the correct output (target) and the other a hallucinated output (hypothesis labeled as a hallucination in the original data). The order of the hypotheses is randomized to prevent bias. This transformation resulted in a training set of 538 samples and a testing set of 115 samples. Table 2 shows few samples from training set. This is the only data we used to instruction-tune SLMs in our approach.

2.2. HaluEval

HaluEval [21] is a large-scale hallucination evaluation benchmark that offers a collection of generated and human-annotated hallucinated samples to evaluate the performance of LLMs in detecting hallucinations. It includes data from three NLP tasks: question answering, knowledge-grounded dialogue, and text summarization.

To test our approach, we exclusively focused on data from the text summarization task as it is inline with the PG data used in the SHROOM training set. This dataset is comprised of columns such as document, right summary, and hallucinated summary. As the dataset contains more than 10k samples, we randomly sampled 1,000 examples for our experiments. To create triplets, we used the document as the source, and included the right summary and hallucinated summary as the hypotheses.

3. Approach

The choice of SLMs in this study is motivated by the necessity for resource efficiency. Smaller models provide significant benefits in terms of reduced computational cost, lower memory requirements, and faster inference speed. These advantages make them more feasible for practical applications, particularly in resource-constrained environments, while maintaining competitive performance.

We explored several SLMs and finally selected Mixtral 8x7B [22] and SOLAR 10.7B [23] as the base models in our approach as illustrated in Figure 1. These models were chosen due to their strong

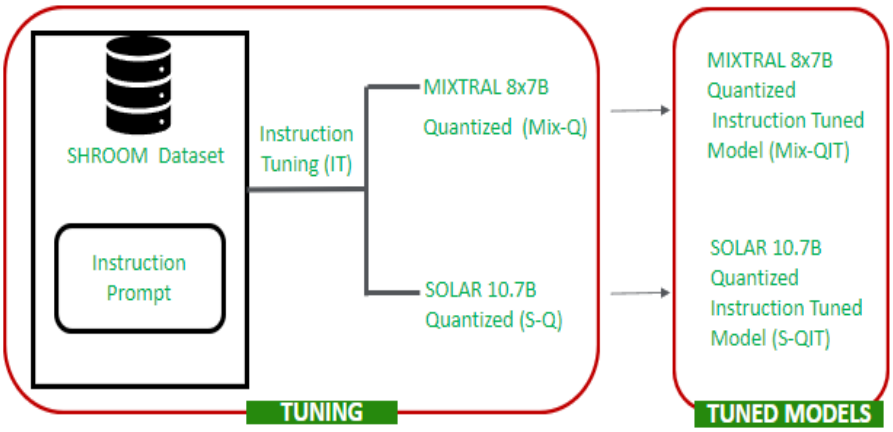

Figure 1: Instruction tuning LLMs

Table 3
Instruction tuning prompt

Given the following information: **source:** original input sentence **hypothesis:** manipulation of the source Determine which hypothesis is a hallucination. A hallucination is any hypothesis that contains information not supported by the source. source: {} **hypothesis 1:** {} **hypothesis 2:** {} Answer 1 if hypothesis 1 is a hallucination, or 2 if hypothesis 2 is a hallucination: label: {}

performance on the SHROOM test set. Mixtral 8x7B uses a Mixture of Experts (MoE) architecture. This design allows the model to dynamically select different subsets of parameters for different inputs, enhancing its ability to handle diverse linguistic tasks efficiently. Additionally, the model has been trained on a multilingual dataset, enhancing its ability to capture language nuances and understand semantic relationships across languages. SOLAR 10.7B on the other hand, utilizes Depth Up-Scaling (DUS), which combines multiple base models into a unified framework. This approach enhances the model's capacity for complex language analysis, making it particularly effective for detecting hallucinations and other intricate language phenomena.

We performed instruction-tuning on the quantized versions of both Mixtral and SOLAR to further optimize their computational efficiency. Both models were quantized to 4-bits significantly lowering the computational requirements and subsequently instruction-tuned using Quantized Weight-Decomposed Low-Rank Adaptation (QDoRA) technique [24]. We selected QDoRA due to the greater efficiency it offers in terms of speed, robustness to rank selection, and faster learning. It accelerates the fine-tuning process, allowing for quicker adaptation to specific tasks, and is less sensitive to the choice of rank during the decomposition process, ensuring stable performance across different configurations. Each LLM was instruction-tuned with the prompt shown in Table 3.

4. Results

This section details the experimental evaluation of our approach. To assess the effectiveness of our method, we employed established classification metrics like accuracy (Acc), macro F1 score (F_{mac}), precision ($Prec$), and recall (Rec). Additionally, we compared our model's performance against GPT-4

Table 4
Impact of quantization and instruction-tuning across various LLMs on SHROOM testset

Model	**Base**			**Quantized (Q)**			**Quantized Instruction-Tuned (QIT)**		
	$Prec$	Rec	F_{mac}	$Prec$	Rec	F_{mac}	$Prec$	Rec	F_{mac}
Mixtral 8x7B	0.66	0.64	0.64	0.52	0.52	0.49	0.88	0.88	0.88
SOLAR 10.7B	0.39	0.47	0.35	0.36	0.45	0.35	0.87	0.87	0.87

and two baseline entailment models on all test sets: i) SelfcheckGPT-NLI [12] which is a sample-based detection method that relies on the consistency of generated responses ii) Hughes Hallucination Evaluation Model (HHEM) [25] which examines the structure, logic, and factual grounding within the text that identify instances where the LLM might have generated incorrect or unsupported claims. We specifically chose entailment models because their training objective aligns closely with the type of hallucination we targeted in this work. To adapt these models to our triplet setting, we calculated the entailment score between the source sentence and each hypothesis. The hypothesis with the lowest entailment score was then classified as the hallucination.

Table 5
Performance comparison of baselines and instruction-tuned SLMs on various datasets

Model	**SHROOM**			**HaluEval**		
	$Prec$	Rec	F_{mac}	$Prec$	Rec	F_{mac}
SelfcheckGPT-NLI	0.65	0.65	0.65	0.64	0.64	0.64
HHEM	0.70	0.70	0.70	0.62	0.62	0.62
GPT-4	0.80	0.80	0.80	**0.79**	**0.74**	**0.75**
Mix-QIT	**0.88**	**0.88**	**0.88**	0.70	0.67	0.66
S-QIT	0.87	0.87	0.87	0.65	0.65	0.65

To justify the emphasis on smaller language models, it is essential to evaluate their resource efficiency in comparision to larger models like GPT-4. With an estimation of 1.8 trillion parameters, GPT-4 requires substantial computational resources for training and inference [1]. In contrast, the smaller language models examined in this study, Mixtral 8x7B and SOLAR 10.7B, contain fewer parameters (less than 15 billion active parameters). This significant reduction in model size results in lower computational requirements, making these smaller models more practical for deployment in resource-constrained settings.

We compared the performance of Mixtral 8x7B and SOLAR 10.7B across three configurations: Base (B), Quantized (Q), and Quantized Instruction-Tuned (QIT) as shown in the Table 4. From the results, it is observed that the F_{mac} scores of the quantized models are lower compared to their base models. However, after performing instruction-tuning on the quantized models, we observed a significant improvement in F_{mac} scores of 0.88, 0.87 for Mixtral 8x7B + QIT (Mix-QIT), SOLAR 10.7B + QIT (S-QIT) respectively. These scores represent an increase of 20% to 50% compared to the base model's F_{mac} scores, highlighting the effectiveness of instruction-tuning in enhancing the ability of quantized LLMs to detect hallucinations.

To benchmark our approach against other established methods, we compared its performance with two entailment baselines as shown in Table 5. The results demonstrate that our instruction-tuned SLMs consistently outperformed both the SelfCheckGPT-NLI and HHEM baselines across the datasets. This highlights the effectiveness of instruction-tuning for hallucination detection across different domains. Further to evaluate our approach and highlight the efficiency with SLMs, we compared the results with the standard, non-fine-tuned GPT-4 model rather than fine-tuned version of GPT-4. Fine-tuning larger models like GPT-4 is a highly resource-intensive process, often require several days of computation on

high-end hardware due to their larger parameter size [1]. On the other hand, fine-tuning smaller models like Mixtral 8x7B and SOLAR 10.7B is more efficient, both in terms of time and resource consumption. Having fewer parameters (less than 15 billion active parameters), it is quicker to train them with lower memory footprint and reduced energy usage.

We also note the results are not consistent across the datasets when we compare instruction-tuned SLMs with GPT-4. On the SHROOM dataset, both Mix-QIT and S-QIT achieved impressive F_{mac} scores of 0.88 and 0.87, exceeding GPT-4 by 8%. These results show that, inorder to detect the hallucinations, instruction-tuning the smaller models can achieve performance comparable to a larger model like GPT-4. However, the performance was not consistent on HaluEval dataset where both Mix-QIT and S-QIT F_{mac} scores (0.66 and 0.65) fell short of GPT-4 by around 10%. While GPT-4 offers superior performance due to its size, the trade-off in computational efficiency makes smaller language models a viable alternative for many use cases.

5. Conclusion

In this paper, we explored the effectiveness of instruction-tuning on the quantized versions of SLMs for hallucination detection. We compared these instruction-tuned models against established methods, including GPT-4 and entailment models, and found consistent improvement across various datasets. While our instruction-tuned models achieved performance comparable to GPT-4 on SHROOM datasets, a discrepancy emerged on the HaluEval dataset. This highlights the need for further research to enhance the robustness and generalizability of instruction tuning for hallucination detection. Smaller language models, defined as those with fewer than 15 billion active parameters, offer significant advantages in terms of computational cost, memory usage, and inference speed, making them more accessible for practical applications, especially in resource-constrained environments.

As future work, we plan to investigate methods not only to detect hallucinations but also to understand the underlying reasoning behind them, potentially leading to effective correction strategies.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[2] J. Manyika, S. Hsiao, An overview of bard: an early experiment with generative ai, AI. Google Static Documents 2 (2023).

[3] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, PLoS digital health 2 (2023) e0000198.

[4] S. M. Mousavi, S. Caldarella, G. Riccardi, Response generation in longitudinal dialogues: Which knowledge representation helps?, 2023. arXiv:2305.15908.

[5] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, arXiv preprint arXiv:2302.04023 (2023).

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[7] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, S. Shi, Siren's song in the ai ocean: A survey on hallucination in large language models, 2023. arXiv:2309.01219.

[8] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, M. Z. Shou, Hallucination of multimodal large language models: A survey, arXiv preprint arXiv:2404.18930 (2024).

[9] R. Cohen, M. Hamri, M. Geva, A. Globerson, Lm vs lm: Detecting factual errors via cross examination, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12621–12640.
[10] M. Zhang, O. Press, W. Merrill, A. Liu, N. A. Smith, How language model hallucinations can snowball, arXiv e-prints (2023) arXiv–2305.
[11] J. Zhang, Z. Li, K. Das, B. Malin, S. Kumar, Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 15445–15458.
[12] P. Manakul, A. Liusie, M. J. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, arXiv preprint arXiv:2303.08896 (2023).
[13] M. Geva, J. Bastings, K. Filippova, A. Globerson, Dissecting recall of factual associations in auto-regressive language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12216–12235.
[14] A. Azaria, T. Mitchell, The internal state of an llm knows when it's lying, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 967–976.
[15] M. Yuksekgonul, V. Chandrasekaran, E. Jones, S. Gunasekar, R. Naik, H. Palangi, E. Kamar, B. Nushi, Attention satisfies: A constraint-satisfaction lens on factual errors of language models, in: The Twelfth International Conference on Learning Representations, 2023.
[16] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, arXiv preprint arXiv:2305.14251 (2023).
[17] R. Kamoi, T. Goyal, J. D. Rodriguez, G. Durrett, Wice: Real-world entailment for claims in wikipedia, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7561–7583.
[18] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, arXiv preprint arXiv:2305.11859 (2023).
[19] X. Wang, Y. Yan, L. Huang, X. Zheng, X.-J. Huang, Hallucination detection for generative large language models by bayesian sequential estimation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15361–15371.
[20] T. Mickus, E. Zosa, R. Vázquez, T. Vahtola, J. Tiedemann, V. Segonne, A. Raganato, M. Apidianaki, Semeval-2024 shared task 6: Shroom, a shared-task on hallucinations and related observable overgeneration mistakes, 2024. arXiv:2403.07726.
[21] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. URL: https://arxiv.org/abs/2305.11747.
[22] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.
[23] D. Kim, C. Park, S. Kim, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim, C. Ahn, S. Yang, S. Lee, H. Park, G. Gim, M. Cha, H. Lee, S. Kim, Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling, 2024. arXiv:2312.15166.
[24] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, Dora: Weight-decomposed low-rank adaptation, arXiv preprint arXiv:2402.09353 (2024).
[25] S. Hughes, Cut the bull.... detecting hallucinations in large language models, ????