"So, How Much Do LLMs Hallucinate on Low-Resource Languages?" A Quantitative and Qualitative Analysis

Kushal Trivedi [1], Murtuza Shaikh[1], Sriyansh Sharma[2],
[1]Indian Institute of Information Technology, Gwalior, India
[2]B.M.S College of Engineering, Bengaluru, India
Correspondence: kushal.trivedi.2110@gmail.com

Abstract

Language models have recently gained significant attention in natural language processing, showing strong performance across a wide range of tasks such as text classification, text generation, language modeling, and question answering. Despite these advances, one of the most critical challenges faced by language models is hallucination – the generation of fluent and plausible responses that are factually incorrect, fabricated or irrelevant. This study presents preliminary work on analyzing hallucinations in Q&A tasks for low-resource languages. We evaluate model performance on the Mpox-Myanmar and SynDARin datasets using three API-accessible models: LLaMA 3.1 70B, LLaMA 3.1 8B, and Gemini 2.5 and two monolingual language models: HyGPT 10B for Armenian and SeaLLM for Burmese. Our work contributes by systematically examining hallucinations through quantitative analysis using Natural Language Inference and Semantic Similarity metrics across different model sizes and prompting strategies, as well as qualitative analysis through human verification. We further investigate whether common assumptions about model behavior hold consistently and provide explanations for the observed patterns.

1 Introduction

Large language models (LLMs) have seen a surge in both practical applications and research developments in recent years (Hadi et al., 2023). LLMs are trained on a vast diversity of data and operate on the principle of probabilistic outcomes (Brown et al., 2020). Consequently, hallucination is an inherent phenomenon in language models that cannot be fully eliminated (Xu et al., 2025).

Hallucination occurs across all modalities - text, image, video, and audio (Sahoo et al., 2024). However, hallucination is not always harmful. Hallucinations become problematic when the generated statements are factually inaccurate or conflict with universal human, societal, or cultural norms. The issue becomes critical in mission-specific, high-stakes domains such as finance, medicine, and law, where even a single biased or incorrect decision can pose significant risks and, in some cases, become a matter of life and death (Rawte et al., 2023).

Hallucination effects differ between high- and low-resource languages (Rohera et al., 2025). While high-resource languages (such as English and Chinese) dominate model training data, low-resource languages like Burmese and Armenian, which are the focus of this study, are represented at much smaller scales.

2 Background

2.1 Hallucination in Early NLP Tasks

Hallucination as a problem was first studied in neural machine translation (NMT) (Lee et al., 2018), as well as in summarization (Cao et al., 2022) and dialogue systems (Das et al., 2022; Pan et al., 2024), before gaining attention in domain-specific Q&A (Xu et al., 2024; Chen et al., 2025). For example, in the case of summarization, hallucinations are typically categorized as intrinsic, where the generated text directly contradicts the source, or extrinsic, where the output introduces facts not supported by the source content. These early findings highlighted that hallucination is not confined to a single NLP task, but rather represents a broader phenomenon inherent to probabilistic sequence generation.

2.2 Hallucination in Domain-Specific Q&A

Although hallucination is inherent in language models, prior work shows that models can sometimes correct their own incorrect claims during explanation or justification (Zhang et al., 2023). Such hallucinations occur across tasks of varying complexity, from simple counting to more complex reasoning problems (Zhang et al., 2023). While errors in simple tasks may have limited impact, hallucinations become critical in high-stakes domains such as medicine, finance, and law, as well as in low-resource language settings, where models are more likely to produce plausible but factually incorrect outputs.

Medicine: Several studies highlight the risks of LLMs in healthcare Q&A. (Siontis et al., 2023) demonstrated that ChatGPT can hallucinate in cardiology-related queries. (Pal et al., 2023) introduced *Med-HALT*, a benchmark specifically designed to evaluate hallucinations in medical reasoning and fact recall. (Zhu et al., 2025) reviewed recent progress on hallucination detection in medical LLMs and LVLMs, providing an overview of available benchmarks. Further, (Jiang et al., 2025) showed that chain-of-thought (CoT) prompting can reduce hallucinations in medical tasks.

Finance: Hallucinations in financial Q&A often involve incorrect numerical values, fabricated company data, or invalid market explanations. (Kang and Liu, 2023) evaluated LLMs on financial tasks and found hallucination rates particularly high when numerical reasoning was required.

Law: In legal Q&A, hallucinations can be especially damaging due to the reliance on case law and statutes. (Mik, 2024) showed that LLMs often invent legal precedents or misinterpret statutes, leading to legal hallucinations.

2.3 Hallucination in Low-Resource Languages

Hallucination detection in low-resource languages for domain-specific Q&A remains underexplored, with most existing research focused on high-resource settings. Some recent work has begun to address this gap, including NitiBench (Akarajaradwong et al., 2025) for Thai legal Q&A systems, Better to Ask in English (Jin et al., 2024) for evaluating LLMs on healthcare queries, MedHalu (Agarwal et al., 2024), a medical hallucination benchmark covering diverse health-related topics, and (Vázquez et al., 2025) for multilingual hallucination evaluation.

Retrieval-augmented generation (RAG) models (Siriwardhana et al., 2023) improve domain-specific Q&A by grounding responses in external knowledge sources, thereby reducing hallucinations and improving factual consistency.

However, their effectiveness in low-resource domains remains limited due to scarce parallel corpora, language-specific fact-checking tools, and annotated hallucination datasets.

3 Problem Statement

Large language models (LLMs) often hallucinate in Q&A tasks. Hallucinations can occur in both cases: (i) when questions relate to information seen during training, as well as (ii) when they concern previously unseen data.

Ideally, the model should indicate that it does not know the answer; however, it frequently produces responses that are factually incorrect. This issue is critical in two scenarios:

1. **Domain-specific Q&A** in sensitive areas such as medicine, finance, and law, where incorrect information can be highly risky, and

2. **Q&A in low-resource languages**, where the training data is significantly smaller than that available for high-resource languages.

To address this problem, we evaluate the performance of large language models on two low-resource datasets: a critical domain knowledge dataset (MPox-Myanmar) (MinSiThu) and a contextual information dataset (SynDARin) (Ghazaryan et al., 2025).

Concisely, we aim to answer the following research questions:

- **RQ1:** What is the baseline hallucination rate of current LLMs when answering questions in Burmese and Armenian?

- **RQ2:** How do different prompting strategies (zero-shot, one-shot, few-shot, and chain-of-thought) affect hallucination rates in low-resource Q&A tasks across specific domain knowledge and general knowledge?

- **RQ3:** Can semantic similarity–based evaluation and Natural Language Inference (entailment, contradiction, and neutral classification) reliably detect hallucination in multilingual contexts, or is further qualitative analysis required?

- **RQ4:** Do LLMs struggle only with specific domain knowledge, or do they also hallucinate when answering general knowledge questions too?

- **RQ5:** What patterns of hallucination (both quantitative and qualitative) emerge as supporting material in the form of contextual background information is progressively removed from the LLM input?

4 Methodology

4.1 Datasets

We have utilized two datasets, namely the Mpox-Myanmar dataset and the SynDARin dataset, for the purpose of our study.

The Mpox-Myanmar dataset is a domain-specific question-answering dataset containing 99 question-answer pairs about Mpox (monkeypox) in the Burmese language. The dataset contains questions ranging from factual inquiries (symptoms, transmission) to procedural knowledge (prevention measures, treatment guidelines).

Additionally, the SynDARin dataset is a Q&A dataset with 1.2K samples for the Armenian language. The dataset provides a paragraph, a question, and four multiple-choice options as input for the LLM. Only one option is correct. The paragraph provides background information for the LLM to choose the correct option.

These datasets represent a critical use case for evaluating LLM performance on:

- **Low-resource languages**: Burmese and Armenian have limited training data compared to high-resource languages.
- **Domain-specific content**: The Mpox-Myanmar dataset contains medical and healthcare information requiring factual accuracy.
- **Contextualized information**: The SynDARin dataset contains questions that are asked on the basis of a supporting context paragraph and include specific topic-related questions.

4.2 Models, APIs, Token Size and Temperature

For Experiment 1, we evaluate Llama 3.1 8B and 70B (Grattafiori et al., 2024), Gemini 2.5 (Comanici et al., 2025), and SeaLLM (Zhang et al., 2025), covering both open-source and proprietary models. To comply with API rate limits, we introduce a 0.5-second delay for the Groq API (30 RPM) and a 4.5-second delay for Gemini and Gemma models (15 RPM). Search is disabled for all API calls.

For Experiment 2, we use Llama 3.1 70B, Gemini 2.5, and HyGPT 10B, following similar API constraints as in Experiment 1 (Section 4.3.3). Due to rate limits, evaluation is conducted on the first 50 test examples, with search disabled for all models.

The temperature is fixed at 0.10 to allow minimal creativity and focus on plain facts. The maximum token size is set to 1024 tokens.

4.3 Experimental Setup 1 – Mpox-Myanmar Dataset

4.3.1 Experiment Overview

The overview of this experiment is to evaluate the language models capabilities on two tasks: (i) pre-training knowledge and (ii) answering capability in a low-resource language. We query the LLM on the 99 questions present in the Mpox-Myanmar dataset and compare the responses against the ground truth. The evaluation is conducted using both quantitative as well as qualitative methods. Quantitative metrics include NLI, which involves determining the inference relationship between the ground truth and the LLMs response and classifying it into entailment, neutral, or contradiction. Qualitative analysis includes human verification and the citation of examples that are adversarial or contradictory in automatic evaluation.

4.3.2 Prompting Strategies

We employed the following prompting strategies in our experiments:

- Zero-shot prompting:
- One-shot prompting:
- Few-shot prompting (three-shot)
- Chain-of-thought prompting combined with few-shot prompting

4.3.3 Experiment Details

- **Semantic Similarity**: We calculate the semantic similarity between the generated and expected responses using the paraphrase-multilingual-MiniLM-L12-v2 transformer (Reimers and Gurevych, 2019). This model was selected for its high computational efficiency to map semantically similar multilingual phrases into a shared 384-dimensional vector space. Based on this score, responses are classified as *Correct* (> 0.75), *Partially Correct* (≥ 0.50), or *Hallucinated* (< 0.50). To the best of our knowledge, there is no prior work that defines standard thresholds for this task; therefore, we set these thresholds empirically based on qualitative inspection of Burmese paraphrases in the dataset. A score above 0.75 indicates that the response is closely aligned in meaning with the ground truth. The 0.50 threshold for partial correctness allows for answers that capture the main idea but differ in wording or style. Scores below 0.50 generally reflect a large deviation from the reference answer, indicating weak factual support.

- **Natural Language Inference**: We employ Natural Language Inference (NLI) (Bowman et al., 2015) as our primary measure of logical consistency. A response is categorized as *Entailment* if it logically follows from the correct answer, *Contradiction* if it conflicts with it, and *Neutral* if the logical relationship cannot be determined. For this classification, we use the mDeBERTa-v3 model (He et al., 2021), chosen for its support for the Burmese language. Based on these classifications, we define the Hallucination Rate (H) as:

$$\text{H} = \frac{N_c}{N_e + N_n + N_c} \tag{1}$$

  where N_c, N_n, and N_e represent counts of Contradictions, Neutral, and Entailment classifications, respectively.

- **Sentence Language Classification**: Language distribution is determined by calculating the ratio of Burmese characters to total characters in the response. In the absence of established thresholds in prior work, a response is classified as *Burmese* if the ratio exceeds 0.80, and as *Mixed* if it falls between 0.50 and 0.80. Responses below 0.50 are categorized as *Language Failures* and excluded from similarity analysis. These values are not pre-established and were selected based on empirical observation.

- **Prompting Templates**: Fixed prompting templates each strategy were chosen for uniformity. The detailed structural layouts for these templates are provided in Figures 4 and 5.

- **Note 1:** Comparing smaller models such as SeaLLM 1.5B and HyGPT 10B with larger models, although less directly justified, is practical for both Experiment 4.3.3 and Experiment 4.4.5. This is because regional models are typically trained on a single language, and the limited availability of large low-resource corpora restricts the development of large monolingual language models. As a result, we evaluate the available monolingual models, which, although they underperform compared to LLMs trained on larger corpora, still achieve comparable performance.

  Note 2: Semantic similarity is used only as a surface-level signal to capture general overlap between the generated response and the ground truth. However, similarity scores do not guarantee factual correctness, as responses can be semantically similar yet contradictory. This limitation is more pronounced in low-resource languages with high paraphrasing variability. Therefore, we use Natural Language Inference (NLI) as a more reliable metric, as it explicitly models factual consistency through entailment, neutral, and contradiction labels.

4.4 Experimental Setup 2 – SynDARin Dataset

4.4.1 Experiment Overview

This experiment examines hallucination trends in language models as supporting contextual information is progressively removed. The task is similar to Experiment 1, but uses more general questions. We conduct both quantitative and qualitative analyses across three experimental settings. The same four prompting strategies used in Section 4.3.2 are also applied in the second and third parts of this experiment. However, we observe that changing the prompting strategies results in negligible variation in the outcomes, unlike in the first experiment. This is due to the lower subjectivity of the ground-truth answers in this dataset.

4.4.2 Part 1: Paragraph and Options Provided

In the first part, the LLM is provided with both the supporting paragraph and four answer options. Performance is evaluated quantitatively by computing the percentage of correct answers.

4.4.3 Part 2: Paragraph Only

In the second part, the answer options are removed and the LLM is prompted to generate the correct answer objectively based only on the paragraph. Evaluation includes semantic similarity and NLI-based metrics. In addition, qualitative analysis is performed via human annotation using the labels *TRUE*, *FALSE*, and *DOES NOT KNOW*.

4.4.4 Part 3: No Paragraph, No Options

In the final part, both the paragraph and answer options are removed. The LLM responds without any contextual support, and evaluation follows the same semantic similarity, NLI, and human annotation procedures as in Part 2.

4.4.5 Experiment Details

- **Semantic Similarity, NLI and Other Details**: Since Armenian is not supported with the MiniLM tokenizer, we have used the BERT multilingual base model (Devlin et al., 2018) for this experiment. The temperature is fixed at 0.10 for minimal creativity and focus on plain facts. The token size is varied across different parts of the experiment to keep the LLM responses concise. For Part 1 of the experiment, the maximum token size is set to 50 tokens, whereas for Parts 2 and 3, the maximum token size is set to 256 tokens.

5 Results and Discussion

The results for Experiments 1 and 2 have been explained and discussed below:

5.1 Experiment 1:

1. Table 1 shows the hallucination and average semantic similarity across all combinations of language models and prompting strategies. It is observed that SeaLLM 1.5B performs the best, with the least hallucination percentage reported for zero-shot and chain-of-thought few-shot (three-shot) prompting. It is interesting to note that a monolingual model with 1.5B parameters can perform comparably or seldom outperform models trained with a higher number of parameters, such as

Model	Prompting	Hallucination (%)	Avg. Semantic Similarity
Llama 3.1 8B	Zero-shot	19.20	0.543
Llama 3.1 8B	One-shot	13.10	0.562
Llama 3.1 8B	Few-shot	7.60	0.631
Llama 3.1 8B	CoT-few-shot	6.20	0.535
Llama 3.3 70B	Zero-shot	8.10	0.560
Llama 3.3 70B	One-shot	5.60	0.626
Llama 3.3 70B	Few-shot	17.10	0.620
Llama 3.3 70B	CoT-few-shot	9.10	0.530
Gemini-2.5	Zero-shot	11.40	0.612
Gemini-2.5	One-shot	9.20	0.597
Gemini-2.5	Few-shot	8.30	0.614
Gemini-2.5	CoT-few-shot	5.00	0.521
SeaLLM 1.5B	Zero-shot	5.05	0.447
SeaLLM 1.5B	One-shot	6.06	0.508
SeaLLM 1.5B	Few-shot	9.09	0.456
SeaLLM 1.5B	CoT-few-shot	5.05	0.421

Table 1: Performance comparison of different models and prompting strategies on Burmese MPox Q&A in **Experiment 1** .

Model	Accuracy (P1%)	Hall. (P2 %)/ S.S.	Hall. (P3 %)/ S.S.
Llama 3.1 70B	64.00	26.00/ 0.427	32.00/ 0.417
Gemini-2.5	76.00	30.00/ 0.458	42.00/ 0.544
HyGPT 10B	64.00	24.00/ 0.606	56.00/ 0.488

Table 2: Performance comparison of different models on Armenian SynDARin Q&A. Abbreviations *P1*, *P2*, *P3*, *Hall.*, *S.S.* stand for Part 1, Part 2, Part 3 of **Experiment 2**, Hallucination, and Semantic Similarity, respectively.

Llama 3.1 70B. The model sizes can be observed in Fig. 3.

2. Across all models, it can be observed from Table 1 that the percentage of hallucination drops from zero-shot to one-shot to few-shot to chain-of-thought few-shot prompting. This trend is observed smoothly for Llama 3.1 8B and Gemini 2.5, as seen in Fig. 1. However, an abnormality is observed for Llama 3.1 70B and SeaLLM 1.5B, where the hallucination percentage increases during one-shot or few-shot prompting and drops again for chain-of-thought few-shot prompting. The first behaviour can be explained by the fact that providing more examples allows the language models to better understand how to answer the questions, especially during chain-of-thought few-shot prompting, which, although primarily used for logical reasoning tasks, proves to be effective for language tasks as well. The second be-

Model	Neutral, T	Neutral, F	Unknown	Contra., T	Entail., F
Llama 3.1 70B	21	5	5	9	4
Gemini-2.5	17	7	0	17	2
HyGPT 10B	6	9	0	2	7

Table 3: Comparison of NLI responses of each model with the Human Verification Annotation for **Experiment 2, Part 2**, i.e., without options but with paragraph.

Model	Neutral, T	Neutral, F	Unknown	Contra., T	Entail., F
Llama 3.1 70B	15	9	1	11	11
Gemini-2.5	17	10	1	12	5
HyGPT 10B	2	2	7	0	9

Table 4: Comparison of NLI responses of each model with the Human Verification Annotation for **Experiment 2, Part 3**, i.e., without options and without paragraph.

haviour is abnormal and requires further investigation.

3. Intuition suggests that the hallucination reported should be inversely proportional to the average semantic similarity. If LLM responses match the ground truth answers more closely at a semantic level, hallucination should be low. However, Fig. 2 empirically suggests that this relationship is not strict, but rather weak. Therefore, high semantic similarity does not necessarily imply low hallucination.

4. While SeaLLM 1.5B demonstrates a significantly low factual hallucination rate, it has been observed that this lack of factual inaccuracy comes at the cost of informativeness. The responses in 40-50 (%) of examples were very generic. Factual inaccuracy in SeaLLM 1.5B was mainly seen in short question-answers, requiring a yes or no type of output with short explanations, where the LLM made the wrong choice. In questions requiring in-depth medicine-related facts, SeaLLM was unable to respond in a detailed format. This highlights the requirement of more factual training in the form of domain-specific datasets in Burmese. The SeaLLM model demonstrates an over-reliance on a few domain-specific keywords (like disease and symptoms) across unrelated contexts, suggesting lack of understanding of word interrelationships. The hallucination patterns observed in LLM responses are shown in Table 6.

5. As defined in Eq. 1, vague responses increase N_n in the denominator, thereby mathematically reducing the hallucination rate H. Qualitative analysis from Experiment 1 shows that SeaLLM often produces neutral, low-information responses, effectively reducing hallucinations by saying less rather than by providing correct answers. This suggests that a low hallucination rate is not always indicative of stronger factual reliability, but may instead reflect a conservative strategy of avoiding errors.

5.2 Experiment 2:

1. The statistics of human verification on top of the quantitative metrics are shown in Tables 3 and 4 for Parts 2 and 3 of Experiment 2, respectively. The main conclusion drawn from these tables is that only a very small percentage of examples are labelled as <Entailment, TRUE> or <Contradiction, TRUE>, while the majority of examples are marked as Neutral by the NLI model. This indicates that the confidence of identifying hallucination through NLI is low, highlighting the need for more robust and reliable evaluation metrics.

2. For the first part of Experiment 2 (shown in Table 2), i.e., choosing one correct option out of four for a question with a background paragraph, Gemini 2.5 gives the best performance, achieving an accuracy of 76.00%. However, when the options and the paragraph are progressively removed in Parts 2 and 3 of the experiment, Llama 3.1 70B shows more stable performance. It is worth noting that there is a drastic increase in the hallucination rate for HyGPT 10B when the supporting paragraph is removed. This may imply that Llama 3.1 70B has been trained on a larger knowledge base compared to HyGPT 10B.

3. The qualitative analyses for the Armenian and Burmese datasets are shown in Tables 5 and 6 respectively.

6 Conclusion

This paper evaluates hallucination in multilingual LLMs (Llama 3.1 8B/70B and Gemini 2.5) and regional monolingual models (HyGPT 10B and SeaLLM 1.5B) using low-resource Burmese and Armenian datasets. A hybrid evaluation framework combining semantic similarity, NLI, and human verification is employed. Results across multiple prompting strategies reveal distinct hallucination patterns, with Experiment 2 showing increased hallucination as supporting context is removed, particularly for HyGPT 10B. Notably, smaller monolingual models achieve performance comparable to larger multilingual models despite limited training data, underscoring the need for more low-resource datasets to improve model reliability.

Limitations

Although we considered the entire Mpox Myanmar dataset (consisting of 99 samples), we evaluated only the first 50 samples from the test split of the SynDARin dataset (out of a total of 1.2K samples) due to API limits on non-premium accounts. Owing to the lack of external funding, we restricted the sample size used for testing. This study could be extended to include more languages; however, direct question–answer (Q&A) datasets are very scarce for low-resource languages, particularly for domain-specific knowledge. Consequently, dataset pre-processing would be required to construct Q&A pairs from the original data, which is time-consuming and demands manual effort. Future work will remove the English-language scaffolding used in our prompting strategies and instead present prompts directly in the low-resource language, reflecting realistic user behavior. We will then compare model performance under both methodologies.

7 Acknowledgements

We thank Ms. Subhoshri Pal for facilitating temporary access to higher-tier API limits, which supported this work.

References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*.

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. Nitibench: Benchmarking llm frameworks on thai legal question answering capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34292–34315.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Haoyang Chen, Dian Kong, Xiangyuan Li, and Chongjun Wang. 2025. Multi-round q&a hallucination analysis of llms in geographical education. In *Advanced Intelligent Computing Technology and Applications*, pages 14–26, Singapore. Springer Nature Singapore.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Souvik Das, Sougata Saha, and Rohini K Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the association for computational linguistics: EMNLP 2022*, pages 684–699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gayane Ghazaryan, Erik Arakelyan, Isabelle Augenstein, and Pasquale Minervini. 2025. Syndarin: Synthesising datasets for automated reasoning in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6459–6466.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and 1 others. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pages 2627–2638.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *Preprint*, arXiv:2311.15548.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Eliza Mik. 2024. Caveat lector: Large language models in legal practice. *Preprint*, arXiv:2403.09163.

MinSiThu. Mpox-myanmar.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Yan Pan, Davide Cadamuro, and Georg Groh. 2024. Exploring hallucinations in task-oriented dialogue systems with narrow domains. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 609–618.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pritika Rohera, Chaitrali Ginimav, Gayatri Sawant, and Raviraj Joshi. 2025. Better to ask in english? evaluating factual accuracy of multilingual llms in english and low-resource languages. *Preprint*, arXiv:2504.20022.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos C Siontis, Zachi I Attia, Samuel J Asirvatham, and Paul A Friedman. 2023. Chatgpt hallucinating: can it get any more humanlike? *European Heart Journal*, 45(5):321–323.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovickỳ, and 1 others. 2025. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes. *arXiv preprint arXiv:2504.11975*.

Jun Xu, Hao Zhang, Haijing Zhang, Jiawei Lu, and Gang Xiao. 2024. Chattf: A knowledge graph-enhanced intelligent qa system for mitigating factuality hallucinations in traditional folklore. *IEEE Access*, 12:162638–162650.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *Preprint*, arXiv:2305.13534.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, and 1 others. 2025. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 96–105.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing-Long QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

A Qualitative Analysis: LLM Responses to Armenian Questions

Table 5 presents two examples for each of the five possible combinations of NLI labs and human annotation, namely ⟨Neutral, TRUE⟩, ⟨Neutral, FALSE⟩, ⟨DOES NOT KNOW⟩, ⟨Contradiction, TRUE⟩, and ⟨Entailment, FALSE⟩.

B Qualitative Analysis: LLM Responses to Burmese Questions

Table 6 illustrates the observed failure modes in Burmese, providing two representative examples for each of the three manually identified categories: short-answer questions, domain-specific knowledge gaps, and semantic collapse.

C Prompting Templates for Burmese Hallucination Detection

Figures 4 and 5 depict the prompting templates used for zero-shot, one-shot, few-shot, and chain-of-thought prompting strategies.

D Key Plots

Figure 1 illustrates the variation in hallucination rates across two parameters, model choice and prompting strategy, in Experiment 1. Figure 2 demonstrates the weak relationship between hallucination and semantic similarity observed in Experiment 1. Figure 3 depicts the sizes of the language models evaluated in this study.

Table 5: Examples of LLM Responses in Armenian with English Translations in **Experiment 2**

Label	Armenian	English Translation
1. Neutral, TRUE	**Question:** Ո՞ր մարզաձևում է Էստոնիան նվաճել իր օլիմպիական մեդալների մեծ մասը:	**Question:** In which sport has Estonia won the majority of its Olympic medals?
	Ground Truth: Ըմբշամարտ	**Ground Truth:** Wrestling
	LLM's Response: Էստոնիան իր օլիմպիական մեդալների մեծ մասը նվաճել է ըմբշամարտում:	**LLM's Response:** Estonia has won most of its Olympic medals in wrestling
	Question: Ինչպես է Վիկտորիա ջրվեժը դասակարգվում աշխարհի ջրվեժների մեջ չափերով:	**Question:** How is Victoria Falls classified among the world's waterfalls by size?
	Ground Truth: Այն աշխարհի ամենամեծ ջրվեժներից մեկն է	**Ground Truth:** It is one of the world's largest waterfalls
	LLM's Response: Ըստ չափերի աշխարհի ջրվեժների մեջ Վիկտորիա ջրվեժը դասակարգվում է...	**LLM's Response:** According to size Victoria Falls is classified among the world's largest waterfalls
2. Neutral, FALSE	**Question:** Ե՞րբ է ստեղծվել Էստոնիայի ազգային օլիմպիական կոմիտեն:	**Question:** When was the Estonian National Olympic Committee founded?
	Ground Truth: 1933 թ	**Ground Truth:** 1933
	LLM's Response: Այն ստեղծվել է 1923 թվականի դեկտեմբերի 8-ին:	**LLM's Response:** It was established on December 8, 1923.
	Question: Ինչպիսի՞ էկոհամակարգ է հայտնաբերվել Վիկտորիա ջրվեժի շրջակայքում:	**Question:** What kind of ecosystem was found around Victoria Falls?
	Ground Truth: Անտառային անտառ	**Ground Truth:** Rainforest
	LLM's Response: Էկոհամակարգը հայտնաբերվել է Վիկտորիա ջրվեժի շրջակայքում:	**LLM's Response:** An ecosystem was found around Victoria Falls.
3. DOES NOT KNOW	**Question:** Ու՞մ է նվիրված «Անցյալ գիշեր Սոհոյում»-ը:	**Question:** To whom is 'Last Night in Soho' dedicated?
	Ground Truth: Դիանա Ռիգ և Մարգարեթ Նոլան	**Ground Truth:** Diana Rigg and Margaret Nolan
	LLM's Response: Տեքստում չի նշվում, թե ում է նվիրված «Անցյալ գիշեր Սոհոյում» ֆիլմը:	**LLM's Response:** The text does not mention to whom the film 'Last Night in Soho' is dedicated.
	Question: Ի՞նչ է նշանակում Ապոլլոն հունարեն Koinē-ում:	**Question:** What does Apollon mean in Greek Koin?
	Ground Truth: Կործանիչ	**Ground Truth:** Destroyer
	LLM's Response: Թվում է տեղեկատվություն չկա Ապոլլոնի մասին, այդ պատճառով չեմ կարող պատասխանել այս հարցին տրված տեքստի հիման վրա:	**LLM's Response:** It seems there is no information about Apollonius, therefore I cannot answer this question based on the provided text.
4. Contradiction, TRUE	**Question:** Ո՞րն է Ռիչարդսոն-Դուշման էֆեկտի այլ անվանումը:	**Question:** What is another name for the Richardson-Dushman effect?
	Ground Truth: Թերմիոնային արտանետում	**Ground Truth:** Thermionic Emission
	LLM's Response: Ռիչարդսոն-Դուշման էֆեկտի այլ անվանումն է ջերմաէլեկտրոնային էմիսիա:	**LLM's Response:** Another name for the Richardson-Dushman effect is thermionic emission
	Question: Որտե՞ղ է անցկացվել 1980 թվականի ամառային օլիմպիական խաղերը:	**Question:** Where were the 1980 Summer Olympics held?
	Ground Truth: Տալլին	**Ground Truth:** Tallinn
	LLM's Response: Այն անցկացվել է Տալլինում:	**LLM's Response:** It was held in Tallinn.
5. Entailment, FALSE	**Question:** Ե՞րբ ավարտվեցին ռասիզմ-թուրանիզմի դատավարությունները:	**Question:** When did the racism-turanism trials end?
	Ground Truth: 1947 թվականի մարտ	**Ground Truth:** March 1947
	LLM's Response: Դրանք ավարտվեցին 1945 թվականի մարտին:	**LLM's Response:** They ended in March 1945
	Question: Ո՞ր մարզաձևում է Էստոնիան նվաճել իր օլիմպիական մեդալների մեծ մասը:	**Question:** In which sport has Estonia won most of its Olympic medals?
	Ground Truth: Ըմբշամարտ	**Ground Truth:** Wrestling
	LLM's Response: Ծանրամարտում	**LLM's Response:** In weightlifting

Table 6: Examples of LLM Failure Modes in with English Translations in **Experiment 1**

Failure Observed	Burmese	English Translation
1. Hallucination in Short Question-Answer	**Question:** ကလေးငယ်များ မျောက်ကျောက်ရောဂါ ကူးစက်ခံရနိုင်လား?	**Question:** Can children get Monkeypox?
	Correct Answer: ကလေးငယ်များသည် လူငယ်လူရွယ်များနှင့် လူကြီးများထက် ရောဂါလက္ခဏာ ပိုမိုပြင်းထန်စွာ ခံစားရနိုင်ပါသည်။ ရောဂါပိုးသည် သန္ဓေသား သို့မဟုတ် မွေးကင်းစ ကလေးငယ်မွေးဖွားသည့်အချိန် သို့မဟုတ် မွေးကင်းစကလေးငယ်များအား ကနဦး ကိုင်တွယ်ထိတွေ့ခြင်းစသည်တို့မှ ကူးစက်နိုင်ပါသည်။	**Correct Answer:** Yes, and they may experience more severe symptoms than adults. Transmission can occur to a fetus in the womb, during birth, or through initial close contact after birth.
	LLM Output: ကလေးငယ်များသည် မျောက်ကျောက်ရောဂါ ကူးစက်ခံရနိုင်ခြင်း မရှိပါ။	**LLM Output:** Children cannot get infected by Monkeypox
2. Domain-Specific Knowledge Gap	**Question:** မျောက်ကျောက်ရောဂါ လက္ခဏာတွေက ဘာတွေလဲ?	**Question:** What are the symptoms of Monkeypox?
	Correct Answer: ရောဂါလက္ခဏာအနေဖြင့် ဖျားခြင်း၊ ခေါင်းအရမ်းကိုက်ခြင်း၊ ကြွက်သားနာကျင်ခြင်း၊ ခါးနာခြင်း၊ အားအင်ကုန်ခမ်းခြင်း၊ ပြန်ရည်အိပ်များ ရောင်ရမ်းခြင်းနှင့် အရေပြားပေါ်တွင် အနီကွက်များ ထွက်ခြင်း သို့မဟုတ် အဖုအပိန့်များ၊ အရည်ကြည်ဖုများထွက်ခြင်းတို့ကို အများစုတွေ့ရတတ်ပါသည်။	**Correct Answer:** Symptoms include fever, severe headache, muscle aches, back pain, low energy, swollen lymph nodes, and skin rashes or lesions. Rashes usually appear 1–3 days after fever starts, progressing from flat spots to fluid-filled or pus-filled blisters before crusting over and falling off.
	LLM Output: မျောက်ကျောက်ရောဂါတွင် များစွာသော ရောဂါလက္ခဏာများ ရှိသည်။ အဓိကအားဖြင့် ဖျားနာခြင်းနှင့် အရေပြားဆိုင်ရာ လက္ခဏာများ ဖြစ်ပေါ်တတ်သည်။ ထိတွေ့မှုများရှိပါက ရောဂါလက္ခဏာများကို ဂရုစိုက်သင့်ပြီး ကျန်းမာရေးအတွက် လိုအပ်သည်များကို လုပ်ဆောင်သင့်သည်။	**LLM Output:** Monkeypox disease has many symptoms. Mainly, fever and skin-related symptoms usually occur. If there are contacts, one should be careful of the symptoms and perform what is necessary for health.
3. Semantic Collapse	**Question:** မျောက်ကျောက်ရောဂါ ကူးစက်နိုင်ခြေ အများဆုံးဖြစ်သူတွေက ဘယ်သူတွေလဲ?	**Question:** Who are most at risk of catching Monkeypox?
	Correct Answer: ကျန်းမာရေးဝန်ထမ်းများ၊ အတူနေမိသားစုဝင်များနှင့် လိင်ဆက်ဆံဖက်တို့သည် ကူးစက်ခံရနိုင်ခြေ ပိုများသည်။	**Correct Answer:** Health workers, household members, and sexual partners are at higher risk of infection.
	LLM Output: ကူးစက်ခံရနိုင်ခြေ အရှိဆုံးသူများမှာ ရောဂါများ၊ ရောဂါလက္ခဏာများနှင့် အခြားသော ရောဂါထိတွေ့မှုများသာ ဖြစ်ပါသည်။	**LLM Output:** The people most at risk are diseases, symptoms, and other disease contacts only...

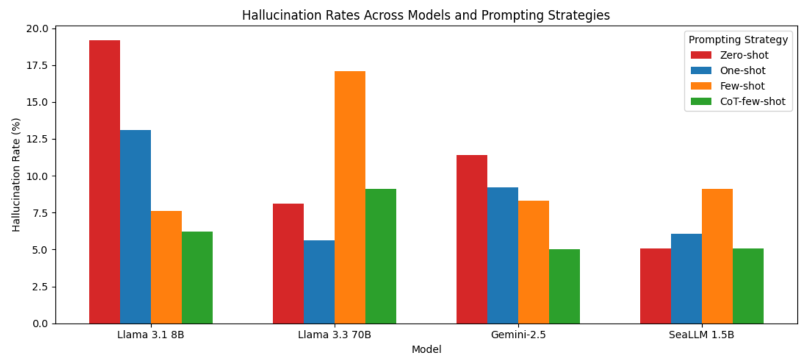

Figure 1: Bar chart for hallucination rate comparison across all models and prompting conditions in **Experiment 1**.

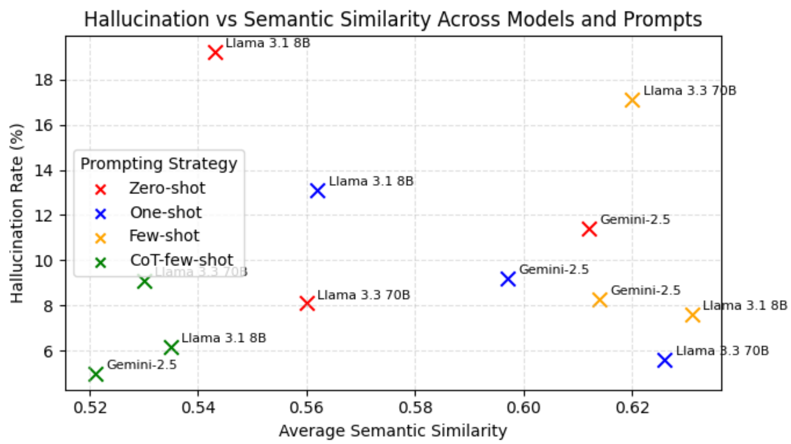

Figure 2: Scatter plot for hallucination vs. semantic similarity trade-off in **Experiment 1**.

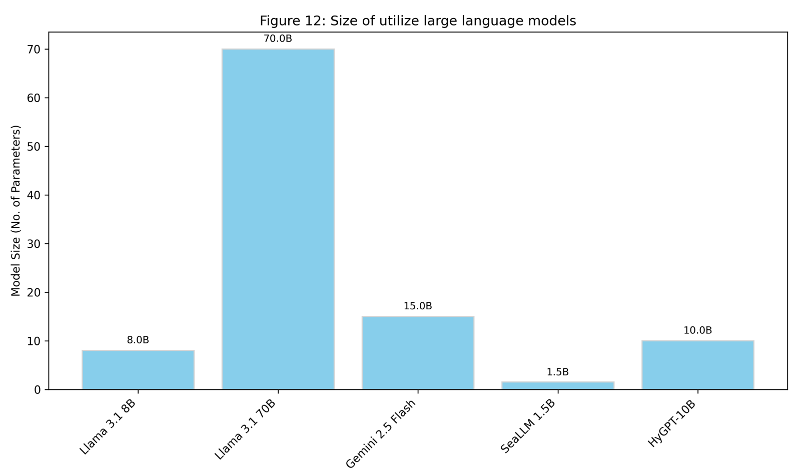

Figure 3: This bar graph shows the size (number of parameters) of the large language models we have utilized for our experiments

Zero-shot Prompting Template

Prompt:

```
Question: {question} Answer the question in Burmese language. Be accurate and concise.
```

One-shot Prompting Template

Prompt:
You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.
Important guidelines:

- For factual questions with definite answers: Be specific, detailed, and descriptive.
- For questions about preventive measures or recommendations: Only provide information you are certain about.
- Do not guess or provide uncertain information.

Example:
Question: မောက်စ်ပေါက်ရောဂါ၏ လက္ခဏာများမှာ အဘယ်နည်း။
Answer: မောက်စ်ပေါက်ရောဂါ၏ အဓိကလက္ခဏာများမှာ ဖျားနာခြင်း၊ ကိုယ်ခန္ဓာကိုက်ခဲခြင်း၊ ခေါင်းကိုက်ခြင်းနှင့် အရေပြားပေါ်တွင် အဖုများ ပေါ်လာခြင်းတို့ ဖြစ်ပါသည်။

Now answer this question:
Question: {question}
Answer:

Figure 4: Prompting templates used in **Experiment 1**.

Few-shot Prompting Template

Prompt:
You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.
Important guidelines:

- For factual questions with definite answers: Be specific, detailed, and descriptive.
- For questions about preventive measures or recommendations: Only provide information you are certain about.
- Do not guess or provide uncertain information.

Examples:
Question: မောက်စ်ပေါက်ရောဂါသည် မည်သို့ကူးစက်နိုင်သနည်း။
Answer: အနီးကပ်ထိတွေ့မှု၊ ရောဂါရှိသူ၏ အရေပြားအဖုများ သို့မဟုတ် အသုံးအဆောင်များမှတဆင့် ကူးစက်နိုင်ပါသည်။

Question: မောက်စ်ပေါက်ရောဂါကို မည်သို့ကာကွယ်နိုင်သနည်း။
Answer: လက်များကို မကြာခဏဆေးကြောခြင်း၊ ရောဂါရှိသူများနှင့် အနီးကပ်မထိတွေ့ခြင်းတို့ဖြင့် ကာကွယ်နိုင်ပါသည်။

Now answer this question:
Question: {question}
Answer:

Chain-of-Thought Template

Prompt:
You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.
Important guidelines:

- Show your reasoning process step by step before giving the final answer.
- For factual questions with definite answers: Be specific, detailed, and descriptive.
- For questions about preventive measures or recommendations: Only provide information you are certain about.
- Do not guess or provide uncertain information.

Example with reasoning:
Question: မောက်စ်ပေါက်ရောဂါသည် လူမှ လူသို့ ကူးစက်နိုင်ပါသလား။
Reasoning:

1. မောက်စ်ပေါက်ရောဂါသည် ဗိုင်းရပ်စ်ကြောင့် ဖြစ်ပွားသည်။
2. ဗိုင်းရပ်စ်များသည် အနီးကပ်ထိတွေ့မှုမှတဆင့် ကူးစက်နိုင်သည်။
3. ထို့ကြောင့် လူမှ လူသို့ ကူးစက်နိုင်ပါသည်။

Answer: ဟုတ်ပါသည်၊ မောက်စ်ပေါက်ရောဂါသည် လူမှ လူသို့ အနီးကပ်ထိတွေ့မှုမှတဆင့် ကူးစက်နိုင်ပါသည်။

Now answer the following question by first showing your reasoning and then providing the final answer:
Question: {question}
Reasoning:
Answer:

Figure 5: Prompting templates used in **Experiment 1**.