Do LLM hallucination detectors suffer from low-resource effect?

Debtanu Datta,[1] Mohan Kishore Chilukuri,[1] Yash Kumar,[1]
Saptarshi Ghosh,[1] Muhammad Bilal Zafar[2,3]

[1]Indian Institute of Technology Kharagpur, India
[2]Ruhr University Bochum, Germany,
[3]UAR Research Center for Trustworthy Data Science and Security, Germany
Correspondence: debtanudatta04@gmail.com

Abstract

LLMs, while outperforming humans in a wide range of tasks, can still fail in unanticipated ways. We focus on two pervasive failure modes: (i) hallucinations, where models produce incorrect information about the world, and (ii) the low-resource effect, where the models show impressive performance in high-resource languages like English but the performance degrades significantly in low-resource languages like Bengali. We study the intersection of these issues and ask: do hallucination detectors suffer from the low-resource effect? We conduct experiments on *five tasks* across *three domains* (factual recall, STEM, and Humanities). Experiments with *four LLMs* and *three hallucination detectors* reveal a curious finding: As expected, the task accuracies in low-resource languages experience large drops (compared to English). However, *the drop in detectors' accuracy is often several times smaller than the drop in task accuracy*. Our findings suggest that even in low-resource languages, the internal mechanisms of LLMs might encode signals about their uncertainty. Further, the detectors are robust within language (even for non-English) and in multilingual setups, but not in cross-lingual settings without in-language supervision.

https://github.com/aisoc-lab/low-resource-hallucination-detection

1 Introduction

LLMs have demonstrated remarkable capabilities across a wide range of tasks like web search (Gasca, 2023), coding (Peng et al., 2023) and scientific research (Gottweis et al., 2025; Shao et al., 2024). In these applications, the correctness of answers and the ability of models to cater to various languages is of paramount importance. However, a long line of work suggests that LLMs face issues along both dimensions.

First, although LLMs generate highly fluent and coherent text, this fluency often masks a critical problem: LLMs can confidently generate inaccurate information, a phenomenon commonly referred to as *hallucination* (Ji et al., 2023; Xiao and Wang, 2021). This behavior severely undermines user trust and the reliability of LLMs, and has prompted a flurry of research on detecting and mitigating hallucinations (Simhi et al., 2025; Snyder et al., 2024; Farquhar et al., 2024).

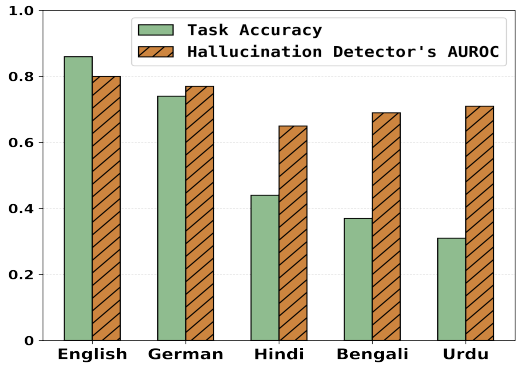


Figure 1: Comparison of task accuracy (1 denotes a correct answer, 0 denotes a hallucination) of a model (`LAM-70B`) with hallucination detector's (`MAM`) performance across languages on `mTREx-Capitals` dataset. When going from English to lower-resource languages, the task accuracy drops significantly. But the hallucination detector's performance remains relatively stable.

Second, the performance of LLMs varies significantly across languages. Models that perform well in high-resource languages like English often degrade substantially in low-resource ones (Pava et al., 2025), a phenomenon often referred to as the *low-resource effect*. Despite the advancement of multilingual pretraining and alignment, significant performance gaps persist. A number of prior studies have focused on addressing this issue (Song et al., 2025; Jiang et al., 2020).

In this work, we study the intersection of these two issues. Specifically, we ask:

If LLMs hallucinate more in low-resource languages, do hallucination detectors perform equally worse in these languages?

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2959–2985
March 24-29, 2026

We conduct extensive experiments on *five question-answer (QA) tasks* over *five languages* – English (EN), German (DE), Hindi (HI), Bengali (BN), and Urdu (UR) – with *four LLMs* of varying sizes (ranging from 7B to 70B parameter models). Due to the lack of multilingual datasets, we translate the three factual recall tasks from English into the four other languages. These five languages span *different resource levels* and *geographic regions*, and are widely spoken across the world. Moreover, they use a *typologically diverse set of four scripts*: Latin, Devanagari, Perso-Arabic, and Bengali. We tested *three hallucination detectors* (HD methods) from two popular families: (i) methods that examine the model's internal artifacts during generation, and (ii) sampling-based black-box methods that utilize model responses only.

We find that although the task accuracy of LLMs drops sharply for low-resource languages, the degradation in HD performance is often much smaller (Figure 1). In fact, in some cases, the HD performance is even better than in English. To further quantify this phenomenon, we design a metric, TPHR, that compares the drop in task accuracy with the drop in detection performance across languages relative to the EN baseline. To summarize, our key contributions are:

1. Due to the scarcity of factual QA datasets in low-resource languages, we developed a novel multilingual factual QA benchmark, **mTREx** (detailed in §3.1), by translating the original English text into DE, HI, BN, and UR. Figure 2 shows an example in all languages.
2. We perform a comprehensive evaluation of three HD methods using four LLMs over five tasks across five languages.
3. We further conduct extensive cross-lingual and multilingual analyses with ablation studies. This provides new insights into HD methods' behavior, demonstrating that detectors perform poorly in pure cross-lingual transfer, but the multilingual training with in-language supervision mitigates the performance gap.
4. We introduce the TPHR metric to compare task accuracy with HD performance across languages, which further validates that HD is relatively robust even when task accuracy drops significantly in low-resource languages.
5. We find that hallucination detectors that utilize the model's internal artifacts outperform sampling-based black-box methods, even in multilingual and low-resource settings.

Question (EN): What is the capital of Irion County?
Ref. Ans. (EN): Mertzon
Model Ans.: Mertzon (**no hallucination**)

Question (DE): Was ist die Hauptstadt von Irion County?
Ref. Ans. (DE): Mertzon
Model Ans.: Mertzon (**no hallucination**)

Question (HI): इरियन काउंटी की राजधानी क्या है?
Ref. Ans. (HI): मर्टज़ोन
Model Ans.: अल्बेमार्ल (**hallucination**)

Question (BN): আইরিয়ন কাউন্টি এর রাজধানী কি ?
Ref. Ans. (BN): মার্টজন
Model Ans.: হার্বার বিচ (**hallucination**)

Question (UR): ہے؟ کیا دارالحکومت کا کاؤنٹی ایریون
Ref. Ans. (UR): میرٹزون
Model Ans.: پنسلوانیا ایری، (**hallucination**)

Figure 2: Example from mTREx-Capitals along with the response from LAM-70B across five languages. The model answers correctly in English (EN) and German (DE) but hallucinates in the low-resource languages.

2 Related Work

Performance disparities for under-represented groups or languages are a well-known problem in broader ML (Barocas et al., 2023) as well as for language models (Liang et al., 2023; Atari et al., 2023; Moayeri et al., 2024). Causes for these disparities have been attributed to a variety of factors, including lack of training data, lack of sufficient coverage at a linguistic and cultural level (Singh et al., 2025; Bender et al., 2021), and modeling choices like capacity (Chen et al., 2018) and tokenization (Schwöbel et al., 2023; Zhou et al., 2022; Neitemeier et al., 2025). Prior work has also found that LLMs can struggle in answering the same question in different languages (Jiang et al., 2020; ul Islam et al., 2025; Wang et al., 2025), which is the focus of our paper.

Benchmarking model performance across languages can lead to a number of challenges, such as translation quality issues, lack of cultural and domain-specific context, and evaluation issues; see (Singh et al., 2025; Mahapatra et al., 2025; Datta et al., 2023) and references therein. In this work, we translate a well-known factual recall benchmark called TREx (Elsahar et al., 2018). Inspection by native speakers reveals similar challenges; see §3.

With the emergence of generative models, hallucinations have come to the fore as an important issue. Hallucinations can manifest in a variety of language modeling tasks such as summa-

rization, translation, and question-answering (Lin et al., 2021; Rawte et al., 2023; Ji et al., 2023). In this work, we focus on the question answering (QA) setting.

Plenty of recent work has looked into detecting and mitigating hallucinations. Most popular Hallucination Detection (HD) approaches operate either by repeatedly looking inside the model (Snyder et al., 2024; Azaria and Mitchell, 2023; Ferrando et al., 2024) or by inspecting the model output (Manakul et al., 2023; Farquhar et al., 2024). In our analysis, we consider HD techniques from both types and analyze their performance in low-resource scenarios. In the multilingual context, ul Islam et al. (2025) introduced a multilingual dataset to study hallucination across languages, showing that models covering more languages tend to hallucinate more. Recently, Vazquez et al. (2025) and Abdaljalil et al. (2025) also addressed multilingual hallucination, but their language overlap with our study is very limited, and their task is to detect spans of text corresponding to hallucinations. To our knowledge, there is no prior work on comparing HD performance across high- and low-resource languages for the factual QA task.

3 Multilingual QA Datasets

While QA datasets can span many domains, we focus on two popular categories: *knowledge of the facts about the real world* and *knowledge about academic disciplines*, such as STEM and humanities. We chose these two datasets: (1) **TREx** (Elsahar et al., 2018) and (2) **Global MMLU** (Singh et al., 2025). TREx offers structured factual triples grounded in encyclopedic knowledge, enabling an evaluation of factual recall. Global MMLU provides questions spanning a wide range of disciplines. Notably, the *response generation style* for these two datasets is also different, allowing us to study hallucination detection performance in different generation settings. TREx requires concise, fact-based short-form answers, whereas for Global MMLU, the model has to select one answer from four available options.

3.1 Multilingual TREx (mTREx) dataset

The existing TREx dataset (Elsahar et al., 2018) is originally available only in English. **We developed a multilingual version of TREx**, which we refer to as **mTREx**, by translating parts of TREx into four additional languages: German (DE), Hindi (HI), Bengali (BN), and Urdu (UR). Note that BN and UR are *resource-poor languages* (Haddow et al., 2022). Our objective in constructing mTREx is to benchmark hallucination detectors (HDs) across typologically and resource-wise diverse languages. To the best of our knowledge, there is no directly comparable, existing multilingual QA benchmark for hallucination detection, which motivated our creation of this mTREx.

Subject Categories. TREx comprises over 11 million factual triples sourced from Wikipedia and covers more than 600 unique Wikidata predicates. Each triple encodes a factual relationship between entities in the form of subject-predicate-object. For this study, we focus on three factual relations:

(1) Capitals: The capital of X is Y.
(2) Country: The country X is located in is Y.
(3) Official Language: The official language of X is Y.

Other predicates are excluded due to data imbalance, ambiguity, or poor translation fidelity, as discussed in Appendix A. For each of the five languages, the number of samples present in Capitals, Country, and Official languages of mTREx, are 2500, 2500, and 2374, respectively.

Selection of Translation Model via Language-specific Considerations. Both the *subject* and *object* entities in TREx are translated into the four languages stated above. For this, we try a range of translation (MT) models, including open-source neural MT models (e.g., *IndicTrans2* (Gala et al., 2023)), and proprietary multilingual LLMs (e.g., *GPT-4*, *GPT-4o-mini*). During translation, we encountered two major challenges:

1. ***Proper nouns with embedded adjectives often led to mistranslations***. For example, 'Future Shop', a company name, was mistranslated by some MT models into Hindi as 'भविष्य की दुकान', meaning 'shop of the future'. Also, 'Spectre' was translated by some models to 'भूत–प्रेत' ('ghosts') in Hindi, distorting the originally intended meaning.
2. ***Models struggle with abbreviations***. For example, *IndicTrans2* rendered 'Kingston Rd' as 'किंग्स्टन आरडी', failing to transliterate 'Rd' (which is an abbreviation for 'Road') correctly, whereas *GPT-4o-mini* translated it correctly as 'किंग्स्टन रोड', preserving semantic accuracy.

We selected *GPT-4o-mini* as our translation model for its superior handling of named entities, and preservation of semantic intent.

Lang.	% of correct translation		
	Capitals	Country	Official Language
EN-to-DE	90	88	96
EN-to-HI	88	90	80
EN-to-BN	94	92	90
EN-to-UR	96	86	92
Average	92	89	90

Table 1: Assessment of translation quality in mTREx.

Evaluation of translation quality. We conduct an extensive evaluation of the translations by the author team that included *native speakers* of HI, BN and UR, and an intermediate-level speaker for DE. The manual evaluation was over 50 randomly sampled QA pairs from each of the mTREx categories for all four target languages (600 samples in total) along three dimensions: semantic fidelity, named-entity correctness, and fluency / naturalness. As reported in Table 1, our evaluation shows consistently high quality of mTREx across languages. On average, more than 90% of translations were correct. Most of the errors observed were typographical or script variations in how entity names are written in particular languages, rather than semantic mistranslations or factual errors.

3.2 The Global-MMLU (G-MMLU) dataset

The G-MMLU dataset (Singh et al., 2025) is a multilingual extension of MMLU (Hendrycks et al., 2021). It offers a broad set of domain-diverse multiple-choice questions spanning a large number of disciplines. In our study, we focus on two disciplines (*STEM* and *Humanities*) and four languages, namely, English, German, Hindi, and Bengali. Urdu is not covered in G-MMLU. We uniformly sample 2,500 examples from each category while trying to retain a balance between subcategories. In the end, we obtained 2,510 samples from *STEM* and 2,506 from *Humanities*.

4 Hallucination Detection Methods

We focus on three hallucination detection (HD) methods that represent two popular families of HD methods: (i) *methods that leverage model internal artifacts during generation*, and (ii) *sampling-based blackbox detectors utilizing only model responses*. We briefly describe the HD methods below; detailed descriptions are in Appendix E.

4.1 Model Artifacts Method (MAM)

These techniques utilize a model's internal artifacts to identify signals of hallucination (Snyder et al., 2024; Azaria and Mitchell, 2023). We consider two types of artifacts: *self-attention scores* and *fully-connected activations*. Following (Snyder et al., 2024), we consider the hidden states from the final layer at the first generated token, as they showed no gains from alternative layers or token positions. We also perform ablation studies with averaging models' artifacts across multiple generated tokens (up to the first 10 generated tokens) from the final layer. The performances remain largely similar in both setups, as discussed in Appendix F.2.

Classifier for hallucination detection. After extracting the hidden states, we train a single-layer neural network with a hidden dimension of 256 to classify whether a response was factually correct or hallucinated. For each dataset category, we train and evaluate the classifier on a random 80/20 split in three settings: (i) *same language setup* (train and test over the same language data), (ii) EN-*to-target cross-lingual setup* (train on EN and test over target language data), and (iii) *multilingual setup* (train on combined multilingual data including EN and test over each target langauge data).

4.2 SelfCheckGPT Method (SGM)

This sampling-based method (Manakul et al., 2023) utilizes the LLM itself to determine the factuality of a generated response by measuring the factual consistency across multiple responses sampled for the same question. The intuition is that factual responses tend to remain consistent across multiple generations, whereas hallucinated responses are likely to contradict one another.

For a given question, a set of responses is generated with high temperature, and a 'base response' is generated by setting the temperature to zero to obtain the deterministic output from the model. Next, the method computes the average negative log-likelihood (NLL) of the base response, to evaluate how much the base response is varied with respect to the distribution defined by the sampled responses. The NLL is then transformed into the *final SelfCheck n-gram score* $1.0 - \exp(-avg_{NLL})$, ranging between 0 and 1. Higher scores correspond to lower NLL and thus higher consistency. For each question, the computed SelfCheck n-gram score serves as the detector's output, predicting the likelihood of the base response being factual or hallucinated.

4.3 Semantic Entropy Method (SEM)

SEM is also a sampling-based method (Farquhar et al., 2024) where the key idea is that when an LLM is uncertain, it is more likely to generate responses that diverge significantly in their semantic content, leading to higher entropy. Conversely, when a model is confident, it generates semantically consistent responses that are expected to cluster around a single or a few closely related meanings, resulting in lower *semantic entropy*.

For a given question, a set of diverse responses is generated with high temperature, and a 'base response' is generated with zero temperature to serve as the reference for determining correctness. To assess the semantic equivalence of sampled responses, we utilize the language-agnostic *LaBSE* model (Feng et al., 2022) to encode each response into a fixed-dimensional semantic embedding. The sampled responses are grouped into semantic clusters if the mutual cosine similarity exceeds a threshold ($\tau = 0.75$, chosen based on the internal evaluation by the native speakers). The probability of each cluster C_k is computed by summing the probabilities of the individual responses belonging to the cluster. If C_k contains I_k responses, then the aggregated probability is given by: $P(C_k) = \sum_{i \in I_k} P(\text{response}_i)$. The *semantic entropy* is then computed over the distribution of $P(C_k)$. Finally, the entropy values are compared against binary ground-truth labels.

4.4 Defining Hallucinations

We consider a response to be a hallucination if it does not match the reference answer (Snyder et al., 2024; Ji et al., 2023). However, the nature of the expected model responses and evaluation criteria differ across the datasets.

mTREx. The questions from mTREx require short-form factual answers. But, due to the potential verbosity and variation in LLM responses, exact string match is *not* a reliable metric for correctness (Adlakha et al., 2024). For instance, for the question *'What is the official language of Italy?'*, the reference answer is *'Italian'*. However, responses such as *'Italian.'* or *'Italian is the official language of Italy.'* are both correct. Hence, we adopt the same heuristic as suggested in (Snyder et al., 2024; Liang et al., 2023; Adlakha et al., 2024), where a generated response A is marked correct if the reference answer R is a substring of A. We perform all comparisons in lowercase.

We manually evaluated this heuristic over a set of 50 model responses in all five languages (750 samples in total) and observed close alignment with human judgment – **on average, this heuristic was found to be correct in more than 95% cases for EN and more than 88% cases for non-EN languages**. Details of this evaluation are in Appendix D.

G-MMLU. This dataset consists of multiple-choice questions where the model has to select one correct option from four. So, if the model's prediction does not match the correct option, the response is labeled as hallucinated.

5 Experimental Setup

We now describe our models, prompts and evaluation metrics.

5.1 Models

We considered four popular instruction-tuned LLMs, representing a range of sizes, to analyse the effect of model size on hallucinations across both low and high-resource languages – Mistral-7B-Instruct, LLaMA-8B-Instruct, Mistral-24B-Instruct and LLaMA-70B-Instruct. More details about the LLM versions, hyperparameters and infrastructure are given in Appendix B.

5.2 Prompts

We observed that LLMs often generate responses in English, even when the question is in a different language. To mitigate this issue we designed dedicated, language-specific prompt templates guiding the models to answer in the target language only.

Prompts used for mTREx. To generate responses from LLMs for mTREx, we adopted a structured prompt template consisting of a language-specific general instruction (*system prompt*) followed by a question template (*user prompt*) with a designated answer placeholder (*assistant prompt*), explicitly asking for a short answer in the same language. For the *Country* relation, the prompt in English is shown in Figure 3. Details of prompts for other languages are in Appendix C.1.

Prompts used for G-MMLU. For the experiments with G-MMLU, we designed the prompt that incorporates few-shot learning with the two smallest language-specific examples from the same category. Here, the prompt follows a structured format consisting of an instruction (*system prompt*), a small set of two QA pairs, and a test question

System Prompt: Suppose your job is to answer given questions in English. When you are asked a question in English, please give a brief answer in English only. Do not write anything else except the answer.

User Prompt: Question: In which country is the <X> located?

Assistant prompt: Answer (in English):

Figure 3: Example prompt for mTREx (Country relationship) in English.

System Prompt: You are a helpful assistant trained to answer objective questions from <subject-category>. Each question comes with 4 options (A, B, C, and D). Provide your answer in the format of a single letter (A, B, C, or D) followed by an explanation in 20 words. Use the given examples to guide your answers. The examples do not have an explanation, but your response should have.

User Prompt: Q1: <example-question-1>

Assistant Prompt: Answer: <answer-example-question-1>

User Prompt: Q2: <example-question-2>

Assistant Prompt: Answer: <answer-example-question-2>

User Prompt: Question: <actual-test-question-with-multiple-choices>

Assistant Prompt: Answer:

Figure 4: Prompt for G-MMLU in English.

(*user prompt*). The model is expected to output the correct option (in A/B/C/D format) followed by a brief explanation. The English prompt is provided in Figures 4. Full prompt templates for each language are detailed in Appendix C.2.

5.3 Evaluation Metrics

Task performance. We report percentage accuracy for the LLMs, where correct / hallucination is decided as described in §4.4.

Hallucination detection performance. To evaluate hallucination detectors, we choose the *Area Under the Receiver Operating Characteristic* (**AUROC**) curve due to its robustness against class imbalance and its threshold-independent nature. Since the ratio of correct vs. hallucinated responses is often highly imbalanced (Table 8), usage of binary classification accuracy could be misleading. AUROC ranges from 0 to 1, where 1 indicates perfect classification, 0.5 corresponds to random guessing, and values below 0.5 suggest worse-than-random performance. For easier readability, we reported AUROC scores multiplied by 100.

Task Performance to Hallucination Ratio. To investigate the alignment between disparities in task performance and Hallucination Detector's (HD) performance across languages, we introduce a novel metric: *Task Performance to Hallucination Ratio (TPHR)*. It quantifies how the difference in the model's task accuracy for a given language L compares to the difference in its HD's performance for the same language relative to the corresponding English (EN) baseline, thus serving as a *useful metric for examining the low-resource effect in multilingual hallucination detection*. Formally, for a specific LLM and a specific HD method, the TPHR for a target language L is defined as:

$$\texttt{TPHR}(L) = \log_{10}\left(\frac{|\text{Accuracy(EN)} - \text{Accuracy}(L)|}{|\text{AUROC}_{\text{HD}}(\text{EN}) - \text{AUROC}_{\text{HD}}(L)|}\right)$$

TPHR provides interpretable signals:

- TPHR ≈ 0: The disparity in HD's performance aligns with the disparity in task accuracy, i.e., both task performance and HD performance change equally w.r.t. the English baseline.
- TPHR $= +k$ ($k \in \mathbb{R}^+$): The difference in task accuracy is 10^k times larger than the difference in HD's performance. In particular, TPHR > 1 implies the degradation in LLM's task accuracy is more than 10 times the degradation in HD's performance.
- TPHR $= -k$ ($k \in \mathbb{R}^+$): Difference in HD's performance is 10^k times larger than the difference in LLM's task accuracy.

In edge cases, TPHR is assigned special values, namely UN and NA. UN arises when the task accuracy delta is zero, making the ratio 0, and thus rendering the logarithm undefined; in such cases, no meaningful conclusion can be drawn regarding the HD's behavior. In contrast, NA occurs when HD's AUROC delta is zero for a language compared to the EN baseline, which directly indicates strong detector robustness, i.e., HD's performance for the other language is equivalent to that in EN.

6 Results and Observations

We now present our observations on task performance and hallucination detection performance.

6.1 Task Performance

Table 2 presents the percentage of increment or decrement of task accuracy in answering factual questions for MST-7B, LAM-8B, MST-24B, and LAM-70B relative to their respective EN baselines. The exact accuracy values are shown in Table 8 of Appendix F. **We observe a substantial drop in performance for low-resource languages such as BN and UR.** On average, accuracies decrease

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	69	↓17	↓71	↓72	↓80	73	↓32	↓90	↓93	↓93	82	↓13	↓85	↓87	↓83	43	↓16	↓44	↓47	53	↓19	↓45	↓51
LAM-8B	74	↓11	↓57	↓62	↓62	77	↓36	↓52	↓60	↓58	70	↓0	↓37	↓51	↓43	56	↓12	↓25	↓38	61	↓16	↓28	↓39
MST-24B	80	↓25	↓62	↓64	↓64	77	↓36	↓51	↓60	↓62	85	↓15	↓46	↓55	↓48	70	↓09	↓30	↓36	72	↓17	↓38	↓44
LAM-70B	86	↓14	↓49	↓57	↓64	84	↓32	↓33	↓46	↓49	87	↓07	↓29	↓54	↓39	74	↓07	↓27	↓14	80	↓12	↓18	↓21
Average	77	↓17	↓58	↓64	↓66	78	↓35	↓56	↓64	↓65	81	↓09	↓49	↓62	↓53	61	↓11	↓31	↓31	66	↓15	↓30	↓36

Table 2: The left column for each dataset shows the task accuracy for English (EN). The following columns show % increase ↑ or decrease ↓ in task accuracy for the corresponding language w.r.t. EN. Darker color shades indicate larger differences. The task performance decreases for all languages when compared with EN.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	82	↓12	↓09	↓10	↓13	79	↑08	↓16	↑05	↓03	81	↑04	↓07	↓05	↑06	76	↓12	↓25	↓32	73	↓08	↓22	↓30
LAM-8B	79	↓06	↓09	↓06	↓11	83	↑05	↓07	↑01	↓07	88	↓03	↓05	↓08	↓05	78	↓05	↓15	↓15	78	↓05	↓19	↓19
MST-24B	76	↑01	↓05	↑01	↓08	86	↑09	↓05	↓01	↓05	82	↑09	↑05	↑05	↑05	82	↓04	↓11	↓16	78	↑04	↓17	↓18
LAM-70B	80	↓04	↓19	↓14	↓11	87	↑08	↓10	↓10	↓08	82	↑07	↑07	↓0	↑05	84	↓05	↓01	↓11	86	↓08	↓06	↓16
Average	79	↓05	↓10	↓06	↓11	84	↑07	↓10	↓02	↓06	83	↑04	↓0	↓01	↑04	80	↓06	↓12	↓18	79	↓05	↓16	↓22

Table 3: Increase ↑ or decrease ↓ in hallucination detector AUROC scores (shown in percentages) w.r.t. the EN baseline. The table show the results for the MAM (fully connected activations) method. Results of other hallucination detection methods are reported in Appendix F.2. The HD performance often drops when compared with EN but the drops are usually smaller than corresponding drops in the task accuracy (Table 8).

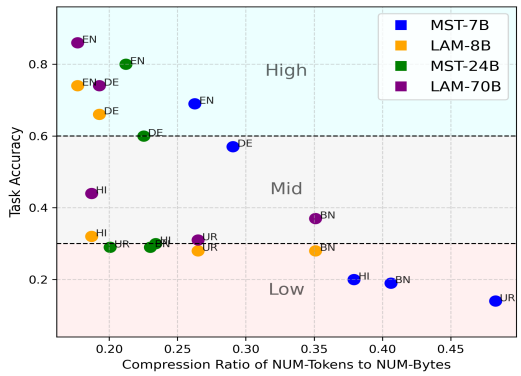


Figure 5: Task accuracy vs. token compression ratio for the Capitals dataset. Low-resource languages have higher compression ratios, showing inefficient tokenization, along with lower task performance.

by more than 62% for BN, 53% for UR, and 49% for HI relative to the EN baseline, in the mTREx categories. For DE, the performance degradations are comparatively much lower, mostly around 10%–20%. The performance drop is consistently more pronounced in mTREx than in G-MMLU. *These observations highlight a severe performance disparity in LLMs when evaluated beyond high-resource languages and motivate the need for multilingual hallucination mitigation strategies.*

Tokenization capabilities across languages. Inspired by Zhou et al. (2022); Schwöbel et al. (2023); Neitemeier et al. (2025), we investigate how tokenization efficiency relates to task performance. Specifically, we analyze the *token compression ratio*, defined as the ratio of the number of tokens produced by the model's tokenizer to the number of bytes in the input string. Figure 5 illustrates the relation between this ratio and the task accuracy over the Capitals dataset, across all LLMs and languages. Similar figures for other categories are presented in Appendix F.7. We find **an inverse relation between token compression ratio and model accuracy**. In particular, *low-resource languages such as* BN *and* UR*, show high compression ratios – indicating inefficient tokenization – while simultaneously yielding very low task accuracy*. In contrast, for resource-rich languages like EN and DE, models show more optimized tokenization and higher performance. This analysis supports the finding that *current LLM tokenizers are poorly adapted to non-English scripts, and tokenization inefficiency may be a key factor to reduced performance in multilingual scenarios.*

6.2 Hallucination Detectors Performance

Next, we analyze the performances of HDs across languages by comparing the AUROC scores in non-EN languages against their respective EN baselines. Tables 3 present the percentage of increment or decrement in AUROC for MAM (fully connected activations). Results for other HD methods: MAM (self-attention), SEM, and SGM, along with ablation and statistical studies for examining the robustness of HDs are reported in detail in Appendix F.2. We observe a consistent trend across datasets: **although the task accuracy for resource-poor languages (BN and UR) drops significantly compared to EN** (discussed in §6.1), **the degradation in HD's performance is often much smaller**. For example, in Capitals and Coun-

Models	mTREx – Capitals				mTREx – Country				mTREx – Official Language				G-MMLU – STEM			G-MMLU – Humanities		
	DE	HI	BN	UR	DE	HI	BN	UR	DE	HI	BN	UR	DE	HI	BN	DE	HI	BN
MAM (fully connected activations)																		
MST-7B	0.079	0.845	0.796	0.699	0.584	0.706	1.230	1.531	0.564	1.067	1.249	1.134	-0.109	0.000	-0.079	0.222	0.176	0.089
LAM-8B	0.204	0.778	0.964	0.709	0.845	0.824	1.663	0.875	UN	0.813	0.711	0.875	0.243	0.067	0.243	0.398	0.054	0.204
MST-24B	1.301	1.097	1.708	0.929	0.544	0.989	1.663	1.079	0.269	0.989	1.070	1.011	0.301	0.368	0.284	0.602	0.317	0.359
LAM-70B	0.602	0.447	0.649	0.786	0.586	0.493	0.637	0.768	0.000	0.620	NA	0.929	0.097	1.301	0.046	0.155	0.447	0.084
Average	0.512	0.750	0.991	0.753	0.653	0.740	1.398	1.009	0.368	NA	1.699	1.156	0.146	0.279	0.133	0.398	0.187	0.150
MAM (self-attention)																		
MST-7B	0.125	0.991	1.000	0.837	0.760	1.217	0.987	1.134	0.342	1.544	1.249	1.230	0.000	0.000	-0.041	0.222	0.150	0.130
LAM-8B	0.204	0.582	0.663	0.621	1.447	0.648	1.186	0.750	UN	0.716	0.711	1.000	0.368	0.105	0.281	0.301	0.117	0.234
MST-24B	1.000	0.854	1.009	0.804	0.447	1.114	1.362	1.681	0.637	1.290	1.672	NA	0.477	1.021	0.319	0.301	0.653	0.602
LAM-70B	0.234	0.419	0.576	0.661	0.586	0.845	0.746	1.011	-0.067	0.699	1.672	0.686	0.222	NA	0.097	0.097	0.368	0.084
Average	0.336	0.653	0.736	0.708	0.732	0.944	NA	1.009	0.544	NA	1.097	1.332	0.243	0.376	0.165	0.523	0.301	0.234
SelfCheckGPT																		
MST-7B	0.778	0.991	1.000	0.786	0.760	0.865	0.577	0.833	0.439	1.368	1.249	0.833	0.544	0.677	0.699	1.000	0.681	0.653
LAM-8B	0.903	1.146	1.186	1.362	1.447	0.561	0.362	0.539	UN	0.336	0.711	0.477	0.845	0.368	0.419	0.398	0.385	0.477
MST-24B	0.347	0.585	0.708	0.753	0.669	0.591	1.061	0.426	0.000	0.688	0.769	1.136	NA	0.368	0.444	1.079	NA	0.903
LAM-70B	0.301	0.544	0.649	1.041	0.317	0.125	0.688	0.768	-0.176	1.097	0.827	1.054	0.398	1.301	1.000	1.000	0.845	1.230
Average	0.512	0.699	0.787	0.862	0.829	0.530	0.620	0.804	0.146	1.301	1.699	1.156	0.845	0.580	0.580	0.699	0.699	0.681
Semantic Entropy																		
MST-7B	0.778	0.611	0.745	0.594	0.517	1.121	0.878	0.753	0.564	1.067	0.810	0.686	0.544	0.434	0.398	0.523	0.426	0.431
LAM-8B	0.125	0.544	0.517	0.459	0.333	0.648	0.885	0.653	UN	0.938	1.556	1.000	0.544	0.192	0.208	1.000	0.385	0.380
MST-24B	0.301	0.585	0.666	0.561	0.669	0.813	0.709	0.640	0.415	0.591	0.973	0.613	0.079	0.720	0.921	0.380	1.431	1.505
LAM-70B	0.234	0.669	1.088	1.439	0.477	0.493	1.591	0.835	0.301	0.921	1.672	0.490	NA	1.301	0.699	NA	1.146	NA
Average	0.269	0.574	0.690	0.628	0.477	0.798	0.854	0.862	0.544	1.602	1.398	1.633	0.845	0.580	0.501	1.000	0.699	0.681

Table 4: TPHR scores ($\log_{10}$ of ratio of accuracy delta to AUROC delta w.r.t EN) for **MAM (fully connected activations)**, **MAM (self-attention)**, **SelfCheckGPT**, and **Semantic Entropy** methods. Cells are marked as NA when the AUROC delta is zero, and as UN when the task accuracy delta is zero. Darker color shades indicate a higher TPHR value, signifying that although the task accuracy for these languages drops drastically, the HD's performance remains much more stable.

try, the drop in AUROC for BN and UR is limited to around 5–10% for most LLMs, whereas the degradation in task performance is as high as 45–90% for these low-resource languages. In some cases, the HD performance even increases. For instance, for MST-24B and LAM-70B, around 5% gains are observed in HD in UR, BN, HI on the Official Language dataset. Overall, the findings highlight that HDs remain relatively robust across languages, especially compared to the significant performance drop in task accuracy.

Insights from TPHR values. Table 4 shows that TPHR values are notably high (> 1) in low-resource languages like BN and UR, particularly for the mTREx datasets. For instance, in MAM (fully connected activations), MST-24B shows a high TPHR value of 1.71 in BN for Capitals, while for Country, MST-7B shows TPHR = 1.53 in UR. Note that TPHR > 1 indicates the degradation in the model's task accuracy is more than 10 times the degradation in HD's performance w.r.t. EN baseline, i.e., **although the task accuracy for these languages drops drastically, the HD's performance remains much more stable**. In comparison, DE tend to exhibit lower TPHR values for most cases, implying a more balanced degradation in both task and HD performance. Thus, the TPHR metric provides a quantitative lens to interpret these disparities and *validates the stability of HDs in multilingual settings, particularly for low-resource languages where task accuracy suffers the most.*

Entropy analyses between generators and detectors. To further investigate why detectors' performance remains stable across languages even when generators' performance (i.e., LLMs' task accuracy) degrades, we analyze the *entropy of softmax distributions* for both generators and detectors. Low-resource languages are typically tokenized into a larger number of tokens, which are also seen far less frequently during pretraining. Consequently, at each generation step, the model must choose from a larger and less familiar token set, yielding a more spread-out softmax distribution in low-resource languages. Whereas high-resource languages exhibit more concentrated softmax distributions. On the other hand, detectors make a binary decision (hallu vs. non-hallu) with the same training setup across languages, leading to more comparable softmax distributions irrespective of language. Building on this intuition, our analysis also reveals that *generators exhibit much higher entropy in low-resource settings* (e.g., MST-24B on Capitals: 0.325 in EN vs. 1.762 in UR), *while detector entropies remain largely stable across languages* (e.g., MAM (fully connected activations) on Country with MST-24B: 0.3 in EN vs. 0.336 in UR). *This contrast suggests that low-resource effects primarily impact the high-dimensional generation space, whereas detection operates over a simpler, language-agnostic binary decision space (hallu vs. non-hallu), leading to more stable detectors' performance.* Detailed anal-

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	↓28	↓32	↓35	↓48	77	↓18	↓23	↓35	↓22	78	↓19	↓24	↓37	↓41	73	↓14	↓27	↓26	72	↓08	↓25	↓25
LAM-8B	79	↓15	↓33	↓29	↓28	86	↓31	↓27	↓24	↓37	88	↓22	↓47	↓25	↓44	77	↓06	↓13	↓16	78	↓17	↓21	↓19
MST-24B	78	↓21	↓29	↓24	↓28	83	↓27	↓24	↓24	↓33	86	↓12	↓36	↓24	↓31	74	↓09	↓16	↓24	70	↑03	↓21	↓20
LAM-70B	82	↓21	↓34	↓30	↓33	86	↓21	↓35	↓17	↓34	82	↓23	↓33	↓26	↓34	83	↓05	↓16	↓12	86	↓12	↓19	↓21
Average	80	↓22	↓32	↓30	↓35	83	↓25	↓28	↓25	↓33	84	↓20	↓36	↓29	↓38	77	↓09	↓19	↓19	76	↓09	↓21	↓21

Table 5: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(self-attention)** method w.r.t. the corresponding EN baseline for the model and dataset in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	78	↓12	↓08	↓04	↓22	79	↓01	↓06	↓0	↓11	74	↑09	↓04	↓05	↑05	74	↓11	↓20	↓28	68	↓03	↓19	↓24
LAM-8B	80	↓11	↓19	↓12	↓16	83	↑01	↓06	↓0	↓07	88	↓05	↓06	↓10	↓07	77	↓04	↓10	↓10	76	↓03	↓14	↓17
MST-24B	72	↑03	↓07	↑01	↓04	82	↑12	↓05	↑02	↓02	84	↑05	↓01	↓0	↓0	77	↓06	↓09	↓19	71	↑08	↓13	↓13
LAM-70B	82	↓09	↓17	↓16	↓11	86	↑07	↓05	↓10	↓05	78	↑12	↑09	↓0	↑10	85	↓05	↓0	↓12	85	↓08	↓06	↓15
Average	78	↓08	↓13	↓09	↓14	82	↑05	↓05	↓02	↓06	80	↑05	80	↓04	↑02	77	↓05	↓09	↓17	75	↓03	↓13	↓17

Table 6: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(self-attention)** method w.r.t. the corresponding EN baseline for the model and dataset in the multilingual setting (train over data across all languages and test over target language data).

yses and results are provided in Appendix F.3.

Comparison among HD methods. Another key finding in our study is that both MAM methods (fully connected activations and self-attention) outperformed sampling-based blackbox methods SEM and SGM across all languages, demonstrating that LLMs' internal signals – captured through artifacts – remain informative for detecting hallucination even in a multilingual scenario.

Cross-lingual and multilingual analyses for MAM detectors. So far, we evaluated the MAM classifiers in the same-language settings. We now conduct cross-lingual and multilingual experiments to observe language transfer effects.

Cross-lingual setting: In this setup, classifiers are trained on 80% of EN data and evaluated on 20% of each non-EN language separately, across all datasets and LLMs. We **observe substantial performance drops when transferring across languages**. For instance, on the Capitals dataset for MAM (self-attention), average AUROC across all LLMs decreases by more than 19% in DE, 29% in HI, 26% in BN, and 34% in UR (see Table 9 & Table 25).

Multilingual setting: Here, classifiers are trained on 80% of the combined multilingual data (all languages including EN) and evaluated on 20% of each language individually. In contrast to the cross-lingual case, **performance in this multilingual setting remains close to the same-language baseline**. For instance, for MAM (self-attention) over Capitals, average AUROC across all LLMs is nearly unchanged: 74 vs. 72 for DE, 70 vs. 68 for HI, 71 vs. 71 for BN, and 70 vs. 67 for UR (see Table 9 & Table 29).

Table 5 & Table 6 present the percentage of increment or decrement in AUROC w.r.t. the EN baseline for MAM (self-attention) for cross-lingual and multilingual setups, respectively. Detailed cross-lingual and multilingual results are discussed in Appendix F.4. *These experiments demonstrate that zero-shot cross-lingual transfer is challenging, but multilingual training with in-language supervision mitigates the performance gap.*

7 Conclusion

We present a comprehensive study of LLM performance and hallucination detection in a multilingual QA setting, especially with resource-poor languages. Our findings reveal that although there is a consistent drop in task accuracy of LLMs for low-resource languages, the performance of hallucination detectors (HDs) remains relatively stable across languages. To quantify this phenomenon, we introduce a novel metric (TPHR) which reveals that HDs suffer significantly less from the low-resource effect than the underlying LLMs. We hypothesize that this robustness may be since hallucination detection is an easier binary classification problem than generation that requires predicting the correct token sequence from a large vocabulary at each step. Our findings also show that HDs that leverage models' internal artifacts (while potentially benefiting from explicit supervision over these internal states) serve as more stable indicators of hallucination across languages than black-box approaches that rely solely on generated responses.

Acknowledgments

This research was partly conducted during Debtanu Datta's visit to Ruhr-Universität Bochum, Germany. The visit was supported by the UAR Research Center for Trustworthy Data Science and Security (RC Trust), Germany. Debtanu Datta is also supported by the Prime Minister's Research Fellowship (PMRF) from the Government of India.

Limitations

While our study provides valuable insights regarding hallucination detectors' performance in low-resource languages, several limitations remain. We focus only on two specific types of question-answering tasks, where model responses contain a single true answer. Extending our analyses on tasks such as dialogue generation—where multiple hallucinations can be observed in a single generation—can be considered as a direction for future work. Extending the analysis to a broader range of QA tasks, domains, and generation types is also an important future direction.

Also, hallucinations can arise from diverse underlying sources, such as reasoning failures, lack of knowledge, or the generation of fabricated information, and these factors may play distinct roles in model behavior. However, the scope of this study is limited to a comparative evaluation of hallucination detector performance across languages, rather than an analysis or classification of the root causes of hallucinations. We therefore leave a systematic investigation of hallucination sources as an interesting direction for future work.

In this study, we evaluate five languages (`EN`, `DE`, `HI`, `BN`, and `UR`) that use four distinct scripts (Latin, Devanagari, Perso-Arabic, Bengali) and represent a range of resource levels. This diversity suggests that our framework should extend to other languages, but of course, a dedicated analysis is needed. Incorporating more languages would further generalize our findings. It can be noted that the detectors we study are largely language-agnostic and domain-agnostic, so they can be extended beyond the five languages and factual-QA tasks. Specifically, MAM detectors require access to model artifacts (self-attention, activations, etc.) and therefore could extend readily to any language that an LLM supports. Sampling-based detectors (SEM and SGM) are black-box and thus easily applied to new languages and domains.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. ArXiv, abs/2503.07833.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. Transactions of the Association for Computational Linguistics, 12:681–699.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. Which humans?

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 967–976, Singapore. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 610–623.

Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? Advances in neural information processing systems, 31.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5291–5302, Singapore. Association for Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. Nature, 630:625 – 630.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. arXiv preprint arXiv:2411.14257.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. Preprint, arXiv:2305.16307.

David Gasca. 2023. Help us build the future of search with search labs.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an ai co-scientist. Preprint, arXiv:2502.18864.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. Computational Linguistics, 48(3):673–732.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. Preprint, arXiv:2009.03300.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM computing surveys, 55(12):1–38.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5943–5959, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. Transactions on Machine Learning Research. Featured Certification, Expert Certification.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

Sayan Mahapatra, Debtanu Datta, Shubham Soni, Adrijit Goswami, and Saptarshi Ghosh. 2025. Milpac: A novel benchmark for evaluating translation of legal text to indian languages. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 24(8).

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9004–9017, Singapore. Association for Computational Linguistics.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 1211–1228.

Pit Neitemeier, Björn Deiseroth, Constantin Eichenberg, and Lukas Balles. 2025. Hierarchical autoregressive transformers: Combining byte- and word-level processing for robust, adaptable language models. In The Thirteenth International Conference on Learning Representations.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. Last accessed on 2025-07-28.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot. ArXiv, abs/2302.06590.

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

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cedric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12310–12324, Singapore. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i'm wrong: High-certainty hallucinations in llms. Preprint, arXiv:2502.12964.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. Preprint, arXiv:2412.03304.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, page 2721–2732, New York, NY, USA. Association for Computing Machinery.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State, Tegawendé F. Bissyandé, and Jacques Klein. 2025. Is llm the silver bullet to low-resource languages machine translation? Preprint, arXiv:2503.24102.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How much do llms hallucinate across languages? on multilingual estimation of llm hallucination in the wild. Preprint, arXiv:2502.12769.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2734–2744, Online. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. Richer countries and richer representations. arXiv preprint arXiv:2205.05093.

Appendix

A Cleaning, Predicate Selection and Filtering Criteria for TREx

First, we cleaned the TREx dataset by removing entries where the subject or object is missing and by excluding entries where the subject or object is a pronoun (e.g., 'he', 'she', 'it', 'we', 'they'), as these lack the specificity needed for factual relationships. Next, several predicates from the original TREx dataset are excluded based on qualitative and quantitative considerations. The *Continent* predicate is discarded due to class imbalance, with 'Antarctica' appearing as the answer for more than half of the instances. The *Discoverer or Inventor* predicate exhibited ambiguity in subject references, creating difficulty in ensuring consistent interpretation. The *Head of Government* predicate was outdated in many entries, reducing its relevance for current factual evaluation of LLMs. The *Award Received* predicate included numerous acronyms and abbreviations that the translation models failed to translate accurately, and are prone to being accepted universally without proper understanding in the case of multilingual settings. Additionally, several other predicates – such as *Basic Form of Government*, *Currency*, *Location of Discovery*, and *Symptoms and Signs* – are excluded due to insufficient data coverage (fewer than 500 instances). By filtering predicates with high semantic clarity, sufficient scale, and multilingual relevance, we aim to ensure a robust and fair benchmark for hallucination detection across languages.

B LLM versions, Hyperparameters & Infrastructure

LLM versions: In this study, we use the following LLMs whose sizes (number of parameters) vary from 7 billion to 70 billion:

1. Mistral-7B-Instruct (`MST-7B`)[1]
2. LLaMA-8B-Instruct (`LAM-8B`)[2]
3. Mistral-24B-Instruct (`MST-24B`)[3]
4. LLaMA-70B-Instruct (`LAM-70B`)[4]

Hyperparameters & Infrastructure: All experiments were conducted on a server equipped with 8x NVIDIA H200 GPUs (shared with several other researchers). For all three HD methods, the base LLM responses are generated by sampling the most likely token according to the Softmax probability, i.e., the greedy decoding with *temperature* as 0, continuing until an <end of text> token is reached or the output length reaches 50 tokens. For `SEM` and `SGM` methods, we generate 20 sample responses for an input with a high temperature of 1.0 to compare with the base response at zero temperature. The classifiers for the Model Artifacts method are trained using the Adam optimizer with a learning rate of 10^{-4}, weight decay of 10^{-2}, a batch size of 128, and for 1000 iterations.

C Prompts

C.1 Prompt Templates for mTREx

In this section, we describe the exact prompt templates used for all five languages and three relation types (Capitals, Country, and Official Language) for the **mTREx** Dataset. For each language, the prompt comprises a general instruction (*system prompt*) followed by a language-specific question (*user prompt*) with a designated answer placeholder (*assistant prompt*). Figures 6, 7, 8, 9, and 10 present the detailed prompts for the **mTREx** dataset in English, German, Hindi, Bengali, and Urdu, respectively.

System Prompt: Suppose your job is to answer given questions in English. When you are asked a question in English, please give a brief answer in English only. Do not write anything else except the answer.

User Prompt (Capitals): Question: What is the capital of <X>?

User Prompt (Country): Question: In which country is the <X> located?

User Prompt (Official Language): Question: What is the official language of the <X>?

Assistant prompt: Answer (in English):

Figure 6: Prompt for **mTREx** in English.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[4] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

System Prompt: Angenommen, Ihre Aufgabe ist es, Fragen auf Deutsch zu beantworten. Wenn Ihnen eine Frage auf Deutsch gestellt wird, geben Sie bitte eine kurze Antwort auf Deutsch. Nichts außer der Antwort.

User Prompt (Capitals): Frage: Was ist die Hauptstadt von <X>?

User Prompt (Country): Frage: In welchem Land liegt die <X>?

User Prompt (Official Language): Frage: Was ist die Amtssprache von <X>?

Assistant prompt: Antwort (auf Deutsch):

Figure 7: Prompt for **mTREx** in German.

System Prompt: मान लीजिए कि आपका काम हिंदी में सवालों के जवाब देना है। जब आपसे हिंदी में कोई सवाल पूछा जाए, तो कृपया उसका संक्षिप्त जवाब हिंदी में ही दें। जवाब के अलावा कुछ और न लिखें।

User Prompt (Capitals): प्रश्न: <X> की राजधानी क्या है?

User Prompt (Country): प्रश्न: <X> किस देश में स्थित है?

User Prompt (Official Language): प्रश्न: <X> की आधिकारिक भाषा क्या है?

Assistant prompt: उत्तर (हिंदी में):

Figure 8: Prompt for **mTREx** in Hindi.

System Prompt: ধরুন আপনার কাজ হল বাংলায় প্রশ্নের উত্তর দেওয়া। যখন আপনাকে বাংলায় প্রশ্ন করা হবে, অনুগ্রহ করে শুধুমাত্র বাংলায় একটি সংক্ষিপ্ত উত্তর দিন। উত্তর ছাড়া অন্য কিছু লিখবেন না।

User Prompt (Capitals): প্রশ্ন: <X> এর রাজধানী কি ?

User Prompt (Country): প্রশ্ন: কোন দেশে <X> অবস্থিত ?

User Prompt (Official Language): প্রশ্ন: <X> এর সরকারী ভাষা কী ?

Assistant prompt: উত্তর (বাংলায়):

Figure 9: Prompt for **mTREx** in Bengali.

System Prompt: جب ہے۔ دینا جواب کے سوالوں میں اردو کام کا آپ کریں فرض دیں۔ میں اردو صرف جواب مختصر مہربانی برائے تو جائے پوچھا سوال کوئی میں اردو سے آپ لکھیں۔ نہ کچھ علاوہ کے جواب

User Prompt (Capitals): سوال: <X> کا دارالحکومت کیا ہے؟

User Prompt (Country): سوال: <X> واقع میں ملک کس ہے؟

User Prompt (Official Language): سوال: <X> کی زبان سرکاری کیا ہے؟

Assistant prompt: جواب (اردو میں):

Figure 10: Prompt for **mTREx** in Urdu.

C.2 Prompt Templates for G-MMLU

We describe below the prompt templates used for all languages (English, German, Hindi, and Bengali) for the G-MMLU dataset. Each prompt includes a general instruction (*system prompt*) followed by two language-specific in-context examples and the actual multiple-choice question (*user prompt*). Figures 11, 12, 13, and 14 provide the full prompt templates for English, German, Hindi, and Bengali, respectively.

System Prompt: You are a helpful assistant trained to answer objective questions from <subject-category>. Each question comes with 4 options (A, B, C, and D). Provide your answer in the format of a single letter (A, B, C, or D) followed by an explanation in 20 words. Use the given examples to guide your answers. The examples do not have an explanation, but your response should have.

User Prompt: Q1: <example-question-1>

Assistant Prompt: Answer: <answer-example-question-1>

User Prompt: Q2: <example-question-2>

Assistant Prompt: Answer: <answer-example-question-2>

User Prompt: Question: <actual-test-question-with-multiple-choices>

Assistant Prompt: Answer:

Figure 11: Prompt for **G-MMLU** in English.

System Prompt: Sie sind ein hilfreicher Assistent, der darauf trainiert ist, objektive Fragen von <subject-category> auf Deutsch zu beantworten. Jede Frage hat vier Antwortmöglichkeiten (A, B, C, und D). Geben Sie Ihre Antwort in Form eines einzelnen Buchstabens (A, B, C, oder D) an, gefolgt von einer Erklärung in 20 Wörtern. Orientieren Sie sich bei Ihren Antworten an den Beispielen. Die Beispiele enthalten keine Erklärungen, Ihre Antwort sollte jedoch welche enthalten.

User Prompt: Q1: <example-question-1>

Assistant Prompt: Answer: <answer-example-question-1>

User Prompt: Q2: <example-question-2>

Assistant Prompt: Answer: <answer-example-question-2>

User Prompt: Question: <actual-test-question-with-multiple-choices>

Assistant Prompt: Answer:

Figure 12: Prompt for **G-MMLU** in German.

D Human Evaluation of the Heuristic for Identifying Hallucinations

In this study, we adopt a substring-based heuristic to identify factual hallucinations in LLM responses, which has been applied in several prior works in the generative question answering task (Snyder et al., 2024; Liang et al., 2023; Adlakha et al., 2024). According to this heuristic, an LLM-generated response A is marked correct if the reference answer R (for the same question) is a substring of A; otherwise, A is considered a hallucination. All comparisons are performed in lowercase.

To quantify the reliability of this heuristic, we conducted a detailed human evaluation. We manually evaluated responses from the `LAM-70B` on

System Prompt: आप <subject-category> से हिन्दी में वस्तुनिष्ठ प्रश्नों का उत्तर देने के लिए प्रशिक्षित एक सहायक सहायक हैं. प्रत्येक प्रश्न में 4 विकल्प (A, B, C, और D) दिए गए हैं। अपना उत्तर एक अक्षर प्रारूप में (A, B, C, या D) दर्ज करें और उसके बाद 20 शब्दों में स्पष्टीकरण दें। अपने उत्तरों को निर्देशित करने के लिए दिए गए उदाहरणों का उपयोग करें। उदाहरणों में स्पष्टीकरण नहीं है, लेकिन आपके उत्तर में स्पष्टीकरण होना चाहिए।

User Prompt: Q1: <example-question-1>

Assistant Prompt: Answer: <answer-example-question-1>

User Prompt: Q2: <example-question-2>

Assistant Prompt: Answer: <answer-example-question-2>

User Prompt: Question: <actual-test-question-with-multiple-choices>

Assistant Prompt: Answer:

Figure 13: Prompt for **G-MMLU** in Hindi.

System Prompt: আপনি <subject-category> থেকে বাংলায় বস্তুনিষ্ঠ প্রশ্নের উত্তর দেওয়ার জন্য প্রশিক্ষিত একজন সহায়ক সহকারী। প্রতিটি প্রশ্নের জন্য ৪টি বিকল্প (A, B, C, এবং D) দেওয়া আছে। আপনার উত্তর একটি অক্ষরের (A, B, C, অথবা D) আকারে দিন এবং তারপর ২০ শব্দে ব্যাখ্যা দিন। আপনার উত্তরের দিকনির্দেশনা দিতে প্রদত্ত উদাহরণগুলি ব্যবহার করুন। উদাহরণগুলিতে ব্যাখ্যা নেই কিন্তু আপনার উত্তরে ব্যাখ্যা থাকতে হবে।

User Prompt: Q1: <example-question-1>

Assistant Prompt: Answer: <answer-example-question-1>

User Prompt: Q2: <example-question-2>

Assistant Prompt: Answer: <answer-example-question-2>

User Prompt: Question: <actual-test-question-with-multiple-choices>

Assistant Prompt: Answer:

Figure 14: Prompt for **G-MMLU** in Bengali.

50 samples from each mTREx category across all five languages: EN, DE, HI, BN, and UR (750 samples in total). A human annotator was shown the question, the LLM-generated response A, the reference answer R, and then asked to judge if A is correct or a hallucination. The human judgment was matched with the decision according to the above-mentioned heuristic.

Results (percentage of cases where the human judgement matched with the decision according to the heuristic) across all languages are presented in Table 7. The results show that the heuristic aligns closely with human judgment: **for EN, it is correct for more than 95% of cases on average**, with errors typically arising in scenarios where the gold answer is part of the question itself, as shown in Figure 15.

For non-EN languages, the heuristic achieves more than 88% accuracy on average. As shown in Figure 16, most of the errors in these non-EN languages are False Positives (the LLM is actually correct, but the heuristic marks the answer as wrong), primarily due to typographical or script variations in how the same entity name can be written in a particular language. Overall, this analysis **demonstrates high agreement between the substring-match heuristic and human evaluation**.

Question (EN): What is the capital of Houghton County?
Ref. Ans. (EN): Houghton, MI <OR> Houghton <OR> Houghton, Michigan
Model Res.: Houghton County is a county in the U.S. state of Michigan, and Hancock is its county seat.
Heuristic Evaluation: Correct (Not a hallucination)
Human Evaluation: Wrong (Hallucination)

Figure 15: Example in English (EN) language where hallucination heuristic evaluation is incorrect. Here, the LLM response is wrong, but the heuristic marks the response as correct.

Lang.	% of correct detection using heuristic		
	Capitals	Country	Official Language
EN	90	96	100
DE	80	90	94
HI	80	96	92
BN	82	96	78
UR	80	94	96

Table 7: Assessment of substring-based heuristic for hallucination detection.

E Hallucination Detection Methods

In our study, we focus on three Hallucination Detection (HD) methods that represent popular families of HD methods: (i) *methods utilizing model internal artifacts during generation*, (ii) *sampling-based black-box detectors that leverage only responses generated by the model*.

E.1 Model Artifacts Method (MAM)

Inspired by the work of Snyder et al. (2024), *we investigate whether the model's internal artifacts associated with the generation can provide signals on hallucination across languages*. This method is model-agnostic, enabling the probing of factual reliability directly from the model's internal states.

Artifacts for Detecting Hallucinations. We focus on the following key artifacts from the final decoder layer (say L) of the Transformer architecture, specifically after processing the input prompt and generating the *first token* of the response.

- **Self-Attention Scores (S_ℓ):** These represent the outputs of the final linear projection within the multi-head self-attention module at layer $\ell \in \{1, \ldots, L\}$, encoding contextual dependencies

Question (DE)**:** Was ist die Hauptstadt von Schaki-Khanat?
Ref. Ans. (DE)**:** Schach <OR> Scheqi, Aserbaidschan <OR> Scheki
Model Res.: Nukha (heute Şəki)
Heuristic Evaluation: Wrong (Hallucination)
Human Evaluation: Correct (Not a hallucination)

Question (HI)**:** फेयट काउंटी, इलिनॉयस की राजधानी क्या है?
Ref. Ans. (HI)**:** वंडालिया
Model Res.: फेयट काउंटी, इलिनॉयस की राजधानी वांडलिया है।
Heuristic Evaluation: Wrong (Hallucination)
Human Evaluation: Correct (Not a hallucination)

Question (BN)**:** সেন্ট জর্জ প্যারিশ, সেন্ট ভিনসেন্ট এবং গ্রেনাডাইনস এর রাজধানী কি ?
Ref. Ans. (BN)**:** কিংস্টাউন
Model Res.: কিংসটাউন
Heuristic Evaluation: Wrong (Hallucination)
Human Evaluation: Correct (Not a hallucination)

Question (UR)**:** ہے؟ کیا دارالحکومت کا ماناواتو
Ref. Ans. (UR)**:** نارتھ پالمسٹون
Model Ans.: ہے۔ نارتھ پالمرسٹن دارالحکومت کا ماناواتو
Heuristic Evaluation: Wrong (Hallucination)
Human Evaluation: Correct (Not a hallucination)

Figure 16: Examples in non-EN languages, where hallucination heuristic evaluation is incorrect. Here, the LLM responses are actually correct, but the heuristic marks the responses as wrong.

among tokens. We extract the self-attention scores for the final decoder layer L associated with the interaction between the final token of the input prompt (say, I_N for N tokens in input) and the first generated response token (say, O_1). This is denoted as $\mathbf{S}_L(I_N, O_1)$.

- **Fully connected activations ($\mathbf{FC}_\ell$):** These are the fully connected hidden representation at layer ℓ. We similarly extract $\mathbf{FC}_L(I_N, O_1)$ to represent the activations linking the final input token I_N and the first answer token O_1.

Classifier for hallucination detection. After extracting these features, we train a single-layer neural network with a hidden dimension of 256 to classify whether a response was factually correct or hallucinated. For each dataset category, we train and evaluate the classifier on a random 80/20 split.

E.2 SelfCheckGPT Method (SGM)

Along with analyzing models' internal artifacts, we have also benchmarked sampling-based black-box methods such as **SelfCheckGPT** method (SGM) (Manakul et al., 2023). It leverages the LLM itself to determine the factuality of a generated response by measuring the factual consistency across multiple responses sampled for the same question. The intuition is that factual responses tend to remain consistent across multiple stochastic generations, whereas hallucinated responses are likely to diverge and contradict one another. We adopt the n-gram based variant of SGM to assess lexical consistency across sampled responses. The key steps are outlined below:

(i) Stochastic Sampling: For a given input, a set of diverse responses has been generated with high temperature to encourage variability in the outputs. In contrast, the base response is generated by setting the temperature to zero to obtain the deterministic output from the model.

(ii) Consistency Scoring: To evaluate how much the base response is varied with respect to the distribution defined by the sampled responses, the average negative log-likelihood (NLL) of the base response has been calculated. Which is then transformed into the final SelfCheck n-gram score. The specific transformation used in our implementation, following common practice, is $1.0 - \exp(-avg_{NLL})$. This maps the NLL (which ranges from 0 to ∞) to a score between 0 and 1, where higher scores correspond to lower NLL and thus higher consistency.

Finally, for each question, the computed SelfCheck n-gram score serves as the detector's output, predicting the likelihood of the base response being factual or hallucinated.

E.3 Semantic Entropy Method (SEM)

In contrast to the previous n-gram-based lexical approach, we evaluate a semantic approach, which is also a sampling-based method, called **Semantic Entropy** (SEM) (Farquhar et al., 2024). It assesses hallucination by estimating *semantic uncertainty* across sampled responses generated by the model. The key idea is that *when an LLM is uncertain, it is more likely to generate responses that diverge significantly in their semantic content*, leading to higher entropy. Conversely, when a model is confident, it generates semantically consistent responses. This diversity is quantified as *semantic entropy*. The confident and factual responses are expected to cluster around a single or a few closely related meanings, resulting in lower *semantic entropy*. The method proceeds in the following steps:

(i) Stochastic Sampling: For a given question, sample responses have been generated with high temperature to encourage output diversity. For each response, the sequence-level log-likelihood has been computed by averaging the token-level

log-probabilities. The base response has been generated at zero temperature to serve as the reference for determining correctness.

(ii) Semantic Clustering: To assess the semantic equivalence of sampled responses, we utilize the language-agnostic *LaBSE* model (Feng et al., 2022), which encodes each response into a fixed-dimensional semantic embedding. Next, the sampled responses are grouped into semantic clusters if the mutual cosine similarity exceeds a threshold ($\tau = 0.75$)[5]. This enables the clustering of responses that convey the same underlying meaning, even if phrased differently.

(iii) Cluster Probability and Entropy: The probability of each cluster C_k is computed by summing the probabilities of the individual responses that belong to it. If C_k contains I_k responses, then the aggregated probability is given by:

$$P(C_k) = \sum_{i \in I_k} P(\text{response}_i)$$

The *semantic entropy* is then computed over the distribution of $P(C_k)$. Finally, the entropy values are compared against binary ground-truth labels.

F Additional Results

F.1 Task Accuracy Details

Table 8 presents the task accuracies of `MST-7B`, `LAM-8B`, `MST-24B`, and `LAM-70B` across all datasets in answering factual questions for all five languages: English (`EN`), German (`DE`), Hindi (`HI`), Bengali (`BN`), and Urdu (`UR`).

F.2 Results of Hallucination Detection Methods

Tables 9, 10, 11, and 12 present exact AUROC scores of **`MAM (self-attention)`**, **`MAM (fully connected activations)`**, **`SGM`**, and **`SEM`** methods for `MST-7B`, `LAM-8B`, `MST-24B`, and `LAM-70B` across datasets and languages, respectively. Tables 13, 14, and 15 present the percentage of increment or decrement in AUROC scores for `MST-7B`, `LAM-8B`, `MST-24B`, and `LAM-70B` across languages for **`MAM (self-attention)`**, **`SGM`**, and **`SEM`** methods w.r.t. the corresponding `EN` baseline for the model and dataset, respectively.

Ablation Studies for Validating Robustness of Hallucination Detectors. To further validate the robustness of hallucination detectors (HDs), we conducted the following detailed ablation studies covering both families of detector methods:

- For **`MAM` detectors** (both **`self-attention`** and **`fully connected activations`** variants by considering artifacts from the final layer at the first generated token), we trained the classifiers 20 times in the same language setup using different random seeds to create diverse train-test splits. Across all four LLMs and five languages, we **observe that the standard errors of AUROC scores across runs remain very small (only around 0.02)**, demonstrating stable detector performance. Detailed results for **`MAM`** (`self-attention` and `fully connected activations`) variants are reported in Tables 18 and 19, respectively.
- Additionally, we further evaluate **`MAM` detectors** (both **`self-attention`** and **`fully connected activations`** by averaging artifacts across multiple generated tokens (up to the first 10 tokens) from the final layer. Results for the **`MAM`** (`self-attention` and `fully connected activations`) variants are reported in Tables 16 and 17, respectively. We observe that, *on average, the performance remains largely similar to the previous setup, where only the 'first generated token' was considered*. For instance, **`MAM`** (`self-attention`) over the Capitals dataset in `DE` with the `MST-24B` model yields 77 vs. 76 AUROC (see Tables 16 and 9). Also, in several cases, performance slightly improves upon considering multiple generated tokens, e.g., for **`MAM`** (`self-attention`) over the Capitals in `HI` with `LAM-8B`: 72 vs. 68 AUROC (see Tables 16 and 9).
  Furthermore, as reported in Tables 33–40, we have also conducted the cross-lingual and multilingual analyses (discussed in Appendix F.4) with this new 'multiple generated token' variant of **`MAM`** detectors, and it exhibits the same qualitative trends as the 'first generated token' setup: (i) EN-to-target cross-lingual transfer remains challenging, and (ii) multilingual training mitigates this gap. Tables 33 & 35 present exact AUROC scores, and Tables 34 & 36 present the percentage of increment or decrement in AUROC w.r.t. the `EN` baseline in the cross-lingual setting. For the multilingual setting, Tables 37 & 39 show exact AUROC scores, and Tables 38 & 40 show the percentage of increment or decrement in AUROC w.r.t. the `EN` baseline.

[5]The threshold has been chosen based on the internal evaluation by the native speakers.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	69	57	20	19	14	73	50	07	05	05	82	71	12	11	14	43	36	24	23	53	43	29	26
LAM-8B	74	66	32	28	28	77	49	37	31	32	70	70	44	34	40	56	49	42	35	61	51	44	37
MST-24B	80	60	30	29	29	77	49	38	31	29	85	72	46	38	44	70	64	49	45	72	60	45	40
LAM-70B	86	74	44	37	31	84	57	56	45	43	87	81	62	40	53	74	69	54	64	80	70	66	63
Average	77	64	32	28	26	78	51	34	28	27	81	74	41	31	38	61	54	42	42	66	56	46	42

Table 8: Task accuracy of MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets in answering factual questions in English (EN), German (DE), Hindi (HI), Bengali (BN), and Urdu (UR) languages.

- For **sampling-based detectors** – Semantic Entropy (SEM) and SelfCheckGPT (SGM) – we performed ablation experiments by varying the number of sampled responses $N \in \{10, 15, 20\}$ at high temperature settings. Experiments were conducted across all five datasets for MST-7B, LAM-8B, and MST-24B models in all five languages. **Results show that absolute AUROC values change only minimally as N varies, and the standard errors in AUROC remain extremely small (less than 0.01) across all models and languages**. This further strengthens the stability of our findings. Detailed results for SGM and SEM detectors are reported in Tables 20 and 21, respectively.

F.3 Entropy Analyses Between Generators and Detectors

To further investigate the stability of detectors' performance across languages compared to the performance of generators (i.e., LLMs' task accuracy), we analyze the *entropy of the softmax distributions* produced by both LLM generators and hallucination detectors across high- and low-resource languages.

It is true that inefficient tokenization of LLMs for low-resource languages primarily affects the generation task at each step. More specifically, low-resource languages usually get tokenized into a larger number of tokens than high-resource languages, and these tokens are observed far less frequently during pretraining due to the low-resource nature of the language. Consequently, *at each generation step, the model must select from a larger and less familiar token set, leading to a more spread out softmax distribution in low-resource languages*, that is, the token probabilities are closer to each other. In contrast, high-resource languages tend to exhibit more concentrated softmax distributions.

On the other hand, the hallucination detectors have to pick between the same set of options (exactly two: hallucinated vs. non-hallucinated) regardless of the language, and are trained on the same amount of data, potentially leading to more comparable softmax distributions across languages.

To quantify this phenomenon, we compute the entropy (a measure of concentration, or lack thereof) of the softmax distribution for the first generated token for all LLMs across all five languages. As reported in Table 22, generator entropies are consistently low for high-resource languages but substantially higher for low-resource ones, indicating greater uncertainty during generation in low-resource settings. For instance, on the Capitals dataset for MST-24B: 0.325 for EN vs. 1.762 for UR.

In contrast, the detector entropies (as reported in Tables 23 & 24 for MAM (self-attention) and MAM (fully connected activations), respectively) remain consistent across languages, for both high- and low-resource languages. For instance, MAM (fully connected activations) on Country dataset with MST-24B: 0.3 for EN vs. 0.336 for UR. *These results demonstrate that generator entropies vary significantly between high- and low-resource languages, whereas detector entropies do not*, as they operate on a fundamentally simpler prediction space (hallucination vs. non-hallucination).

F.4 Cross-lingual and Multilingual Analyses for MAM Detectors

We conduct additional experiments to analyze the behavior of MAM detectors in cross-lingual and multilingual settings. These experiments complement the same-language setup reported in the main paper (where training and testing had been done over the same language data) and provide deeper insights into the cross-lingual transfer effects.

Cross-lingual Setting. In the cross-lingual setting, we train MAM detectors on 80% of the EN data and test them on 20% of the target-language data, across all dataset categories and for all four LLMs. **We observe significant**

performance drops in this cross-lingual setting compared to the same-language setup. For instance, on `mTREx-Capitals` using `MAM (self-attention)`, average AUROC across all LLMs decreases from 74 → 62 in `DE`, 70 → 54 in `HI`, 71 → 56 in `BN`, and 70 → 52 in `UR`. Similar drops are observed across other datasets and both `MAM` variants (`self-attention` and `fully connected activations`). Detailed results across all datasets are reported in Tables 25, 26, 27, and 28, where Tables 25 and 27 present exact AUROC scores and Tables 26 and 28 present the percentage of increment or decrement in AUROC w.r.t. the `EN` baseline.

Multilingual Setting. In this setup, `MAM` detectors had been trained on 80% of the data across all five languages (EN, DE, HI, BN, UR) and tested on 20% of each language separately. **Unlike the cross-lingual setting, here the performance remains close to the same-language setup, showing that multilingual supervision mitigates the cross-lingual transfer gap.** For example, on `mTREx-Capitals` with `MAM (self-attention)`, average AUROC across all LLMs is nearly unchanged: 74 vs. 72 in `DE`, 70 vs. 68 in `HI`, 71 vs. 71 in `BN`, and 70 vs. 67 in `UR`. Detailed results across all datasets and LLMs are provided in Tables 29, 30, 31, and 32, where Tables 29 and 31 present exact AUROC scores and Tables 30 and 32 present the percentage of increment or decrement in AUROC w.r.t. the `EN` baseline.

These experiments together show two key findings: (i) **zero-shot `EN`-to-target cross-lingual transfer for hallucination detection is challenging**, and (ii) **combined multilingual training mitigates this gap**. A brief theoretical intuition helps explain these effects. Consider the very simple setting where the last layer last token embeddings for each class (correct vs. hallucinated) follow a 1D Gaussian distribution: In EN, hallucinated responses follow $N(-5, 1)$, that is, a mean of 5 and a variance of 1, and non-hallucinated responses follow $N(5, 1)$. Clearly, both can be separated with almost 100% accuracy. Now, consider another language, say `UR`, where hallucinated responses follow $N(10, 1)$ and non-hallucinated responses $N(-10, 1)$. These are also well separated, but the `EN`-trained threshold would give almost zero accuracy in the case of `UR`. The key point is that while the correct and hallucinated responses in both languages are almost perfectly separable, they are not separable via the same classifier and **need a language-specific classifier**. Hence, zero-shot cross-lingual transfer struggles, but modest in-language supervision or pooled multilingual training restores performance.

F.5 Accuracy of Classifiers for `MAM` method

Tables 41 and 42 present the binary classification accuracy of the Model Artifacts method across all four LLMs, covering all the dataset categories in mTREx and G-MMLU datasets and all languages.

F.6 Interpretation of $10^{\texttt{TPHR}}$ Values

To illustrate the magnitude of the differences in the model's task accuracy compared to the differences in HD's performance across languages with respect to the `EN` baseline, we also present the values of $10^{\texttt{TPHR}}$ in Table 43. For a given language L, $10^{\texttt{TPHR}(L)}$ directly represents the *ratio of disparities*, that is, how many times larger the change in task accuracy is compared to the change in HD's performance. Here,

- $10^{\texttt{TPHR}} \approx 1$ indicates the disparities in task accuracy and HD's performance are of equal magnitude relative to the `EN` baseline. In other words, both the task performance and the HD performance change equally with respect to the `EN` baseline.
- $10^{\texttt{TPHR}} \gg 1$ means the difference in task accuracy is much higher than the difference in HD's performance.
- $10^{\texttt{TPHR}} \approx 0$ represents the difference in HD's performance is significantly greater than the difference in LLM's task accuracy.

As reported in the Table 43, we observe for low-resource languages like `BN` and `UR`, the $10^{\texttt{TPHR}}$ values are very high, even many times in the range of 30 to 50, particularly for the mTREX datasets.

F.7 Comparison of Task Accuracy with Tokenization capability

To gain deeper insights into the language capabilities of LLMs, we examine the relationship between tokenization efficiency and task performance. Specifically, we analyze the token compression ratio, defined as the number of tokens produced by the model's tokenizer divided by the number of bytes in the input string. Figures 17, 18 illustrate how this ratio correlates with task accuracy across all LLMs and languages for mTREx and G-MMLU, respectively.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	70	74	74	71	77	81	73	84	72	78	83	80	74	82	73	66	54	51	72	66	55	52
LAM-8B	79	74	68	69	68	86	87	77	83	78	88	84	83	81	85	77	74	66	66	78	73	65	64
MST-24B	78	76	71	73	70	83	93	80	85	82	86	89	84	85	86	74	72	72	62	70	76	64	62
LAM-70B	82	75	66	69	70	86	93	82	79	82	82	89	87	81	89	83	80	83	75	86	78	80	72
Average	80	74	70	71	70	83	88	78	83	78	84	86	84	80	86	77	73	69	64	76	73	66	62

Table 9: AUROC scores of MAM **(self-attention)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	82	72	75	74	71	79	85	66	83	77	81	84	75	77	86	76	67	57	52	73	67	57	51
LAM-8B	79	74	72	74	70	83	87	77	84	77	88	85	84	81	84	78	74	66	66	78	74	63	63
MST-24B	76	77	72	77	70	86	94	82	85	82	82	89	86	86	86	82	79	73	69	78	81	65	64
LAM-70B	80	77	65	69	71	87	94	78	78	80	82	88	88	82	86	84	80	83	75	86	79	81	72
Average	79	75	71	74	70	84	90	76	82	79	83	86	83	82	86	80	75	70	66	79	75	66	62

Table 10: AUROC scores of MAM **(fully connected activations)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	71	69	66	66	62	76	72	67	58	66	68	72	65	64	58	57	59	53	53	58	59	53	52
LAM-8B	71	72	68	68	69	78	77	67	58	65	75	69	63	68	65	63	64	57	55	64	68	57	56
MST-24B	78	69	65	68	69	81	87	71	77	63	61	74	69	69	64	63	63	54	54	57	58	57	53
LAM-70B	74	68	62	63	69	69	82	48	61	76	56	65	54	63	59	54	52	53	55	53	54	51	52
Average	74	70	65	66	67	76	80	63	64	68	65	70	63	66	62	59	60	54	54	58	60	54	53

Table 11: AUROC scores of **SelfCheckGPT** method (SGM) for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	67	65	55	58	53	76	69	71	67	64	71	68	65	60	57	61	63	54	53	64	61	55	54
LAM-8B	74	68	62	60	58	82	69	73	76	72	73	62	70	74	70	68	66	59	55	67	68	60	57
MST-24B	71	61	58	60	57	79	73	73	70	68	63	68	73	68	73	59	64	55	56	58	63	57	57
LAM-70B	71	64	62	67	69	70	61	61	69	76	64	67	67	63	75	60	60	61	58	55	55	56	55
Average	71	64	59	61	59	77	68	70	70	70	68	66	69	66	69	62	63	57	56	61	62	57	56

Table 12: AUROC scores of **Semantic Entropy** method (SEM) for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	↓11	↓06	↓06	↓10	77	↑05	↓05	↑09	↓06	78	↑06	↑03	↓05	↑05	73	↓10	↓26	↓30	72	↓08	↓24	↓28
LAM-8B	79	↓06	↓14	↓13	↓14	86	↑01	↓10	↓03	↓09	88	↓05	↓06	↓08	↓03	77	↓04	↓14	↓14	78	↓06	↓17	↓18
MST-24B	78	↓03	↓09	↓06	↓10	83	↑12	↓04	↑02	↓01	86	↑03	↓02	↓01	↓0	74	↓03	↓03	↓16	70	↑09	↓09	↓11
LAM-70B	82	↓09	↓20	↓16	↓15	86	↑08	↓05	↓08	↓05	82	↑09	↑06	↓01	↑09	83	↓04	↓0	↓10	86	↓09	↓07	↓16
Average	80	↓08	↓12	↓11	↓12	83	↑06	↓06	↓0	↓06	84	↑02	↓0	↓05	↑02	77	↓05	↓10	↓17	76	↓04	↓13	↓18

Table 13: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(self-attention)** method w.r.t. the corresponding EN baseline for the model and dataset. Darker color shades indicate a larger magnitude of percentage difference.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	71	↓03	↓07	↓07	↓13	76	↓05	↓12	↓24	↓13	68	↑06	↓04	↓06	↓15	57	↑04	↓07	↓07	58	↑02	↓09	↓10
LAM-8B	71	↑01	↓04	↓04	↓03	78	↓01	↓14	↓26	↓17	75	↓08	↓16	↓09	↓13	63	↑02	↓10	↓13	64	↑06	↓11	↓12
MST-24B	78	↓12	↓17	↓13	↓12	81	↑07	↓12	↓05	↓22	61	↑21	↑13	↑13	↑05	63	↓0	↓14	↓14	57	↑02	↓0	↓07
LAM-70B	74	↓08	↓16	↓15	↓07	69	↑19	↓30	↓12	↑10	56	↑16	↓04	↑12	↑05	54	↓04	↓02	↑02	53	↑02	↓04	↓02
Average	74	↓05	↓12	↓11	↓09	76	↑05	↓17	↓16	↓11	65	↑08	↓03	↑02	↓05	59	↑02	↓08	↓08	58	↑03	↓07	↓09

Table 14: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for **SelfCheckGPT** method (SGM) w.r.t. the corresponding EN baseline for the model and dataset.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	67	↓03	↓18	↓13	↓21	76	↓09	↓07	↓12	↓16	71	↓04	↓08	↓15	↓20	61	↑03	↓11	↓13	64	↓05	↓14	↓16
LAM-8B	74	↓08	↓16	↓19	↓22	82	↓16	↓11	↓07	↓12	73	↓15	↓04	↑01	↓04	68	↓03	↓13	↓19	67	↑01	↓10	↓15
MST-24B	71	↓14	↓18	↓15	↓20	79	↓08	↓08	↓11	↓14	63	↑08	↑16	↑08	↑16	59	↑08	↓07	↓05	58	↑09	↓02	↓02
LAM-70B	71	↓10	↓13	↓06	↓03	70	↓13	↓13	↓01	↑09	64	↑05	↑05	↓02	↑17	60	↓0	↑02	↓03	55	↓0	↑02	↓0
Average	71	↓10	↓17	↓14	↓17	77	↓12	↓09	↓09	↓09	68	↓03	↑01	↓03	↑01	62	↑02	↓08	↓10	61	↑02	↓07	↓08

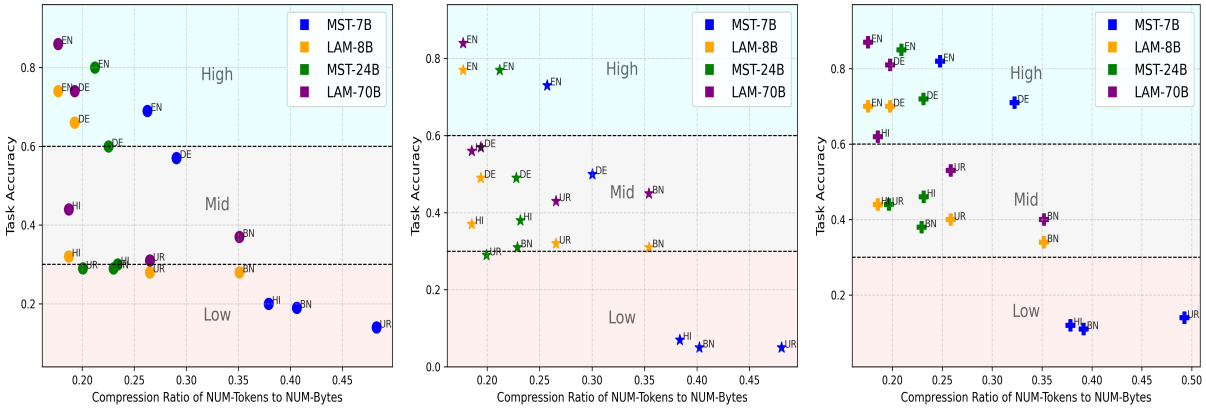
Table 15: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for **Semantic Entropy** method (SEM) w.r.t. the corresponding EN baseline for the model and dataset.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	79	74	77	75	75	82	86	76	88	83	83	84	77	75	80
LAM-8B	86	74	72	76	75	84	88	83	85	85	88	83	85	83	86
MST-24B	81	77	74	78	72	81	94	82	86	86	84	90	85	88	87
Average	82	75	74	76	74	82	89	80	86	84	85	85	82	82	84

Table 16: AUROC scores of MAM **(self-attention)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	82	77	78	75	76	85	90	76	89	89	84	85	80	77	83
LAM-8B	82	76	74	77	76	83	88	84	87	85	89	85	86	85	87
MST-24B	83	79	76	80	73	85	95	83	89	88	86	90	88	86	89
Average	82	77	75	77	74	84	90	81	88	87	86	86	84	82	86

Table 17: AUROC scores of MAM **(fully connected activations)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages.



(a) For mTREx – Capitals (b) For mTREx – Country (c) For mTREx – Official Languages

Figure 17: Comparison of task accuracy with token compression ratio (number of tokens produced by the model's tokenizer divided by the number of bytes in the input prompt) for mTREx across all LLMs and languages.

Dataset	Model	EN	DE	HI	BN	UR
mTREx – Capitals	MST-7B	0.765 ± 0.019	0.717 ± 0.022	0.763 ± 0.022	0.722 ± 0.024	0.718 ± 0.036
	LAM-8B	0.767 ± 0.019	0.739 ± 0.021	0.690 ± 0.022	0.711 ± 0.026	0.677 ± 0.022
	MST-24B	0.760 ± 0.024	0.777 ± 0.018	0.684 ± 0.031	0.713 ± 0.022	0.683 ± 0.023
	LAM-70B	0.785 ± 0.021	0.739 ± 0.021	0.669 ± 0.020	0.679 ± 0.018	0.717 ± 0.022
mTREx – Country	MST-7B	0.768 ± 0.015	0.803 ± 0.021	0.718 ± 0.035	0.809 ± 0.066	0.750 ± 0.049
	LAM-8B	0.854 ± 0.020	0.879 ± 0.012	0.763 ± 0.020	0.830 ± 0.021	0.767 ± 0.017
	MST-24B	0.849 ± 0.016	0.931 ± 0.006	0.793 ± 0.019	0.845 ± 0.015	0.806 ± 0.015
	LAM-70B	0.855 ± 0.015	0.923 ± 0.012	0.808 ± 0.020	0.804 ± 0.015	0.820 ± 0.018
mTREx – Official Language	MST-7B	0.786 ± 0.021	0.800 ± 0.021	0.784 ± 0.027	0.717 ± 0.028	0.754 ± 0.041
	LAM-8B	0.878 ± 0.018	0.837 ± 0.017	0.826 ± 0.013	0.834 ± 0.019	0.837 ± 0.016
	MST-24B	0.831 ± 0.019	0.891 ± 0.014	0.825 ± 0.015	0.842 ± 0.013	0.835 ± 0.015
	LAM-70B	0.832 ± 0.022	0.883 ± 0.022	0.854 ± 0.014	0.806 ± 0.021	0.866 ± 0.014
G-MMLU – STEM	MST-7B	0.734 ± 0.016	0.692 ± 0.020	0.556 ± 0.024	0.519 ± 0.017	–
	LAM-8B	0.737 ± 0.020	0.721 ± 0.017	0.677 ± 0.026	0.638 ± 0.024	–
	MST-24B	0.756 ± 0.022	0.719 ± 0.023	0.688 ± 0.021	0.647 ± 0.025	–
	LAM-70B	0.812 ± 0.018	0.813 ± 0.014	0.856 ± 0.010	0.758 ± 0.015	–
G-MMLU – Humanities	MST-7B	0.735 ± 0.016	0.671 ± 0.018	0.546 ± 0.024	0.528 ± 0.026	–
	LAM-8B	0.749 ± 0.019	0.736 ± 0.016	0.645 ± 0.022	0.636 ± 0.026	–
	MST-24B	0.685 ± 0.020	0.731 ± 0.020	0.626 ± 0.027	0.598 ± 0.018	–
	LAM-70B	0.798 ± 0.027	0.805 ± 0.021	0.783 ± 0.012	0.735 ± 0.015	–

Table 18: AUROC scores (mean ± standard error) [without multiplying by 100] for MAM **(self-attention)** method where the classifiers are trained and tested across 20 runs in the same language setup for all LLMs, languages, and datasets. Note that Urdu (UR) is not covered in G-MMLU.

Dataset	Model	EN	DE	HI	BN	UR
mTREx – Capitals	MST-7B	0.789 ± 0.020	0.753 ± 0.020	0.771 ± 0.021	0.740 ± 0.024	0.728 ± 0.028
	LAM-8B	0.751 ± 0.021	0.731 ± 0.020	0.691 ± 0.026	0.728 ± 0.025	0.682 ± 0.024
	MST-24B	0.742 ± 0.031	0.770 ± 0.022	0.694 ± 0.025	0.737 ± 0.021	0.708 ± 0.022
	LAM-70B	0.795 ± 0.022	0.743 ± 0.018	0.662 ± 0.026	0.697 ± 0.016	0.720 ± 0.024
mTREx – Country	MST-7B	0.793 ± 0.014	0.857 ± 0.017	0.705 ± 0.044	0.830 ± 0.054	0.746 ± 0.049
	LAM-8B	0.835 ± 0.021	0.875 ± 0.012	0.764 ± 0.020	0.837 ± 0.021	0.772 ± 0.017
	MST-24B	0.852 ± 0.015	0.931 ± 0.009	0.819 ± 0.016	0.856 ± 0.016	0.827 ± 0.015
	LAM-70B	0.851 ± 0.017	0.923 ± 0.010	0.771 ± 0.017	0.800 ± 0.016	0.797 ± 0.018
mTREx – Official Language	MST-7B	0.834 ± 0.020	0.816 ± 0.019	0.783 ± 0.030	0.729 ± 0.034	0.782 ± 0.034
	LAM-8B	0.878 ± 0.012	0.842 ± 0.018	0.836 ± 0.013	0.839 ± 0.018	0.837 ± 0.017
	MST-24B	0.834 ± 0.023	0.891 ± 0.015	0.855 ± 0.016	0.866 ± 0.014	0.852 ± 0.016
	LAM-70B	0.827 ± 0.025	0.865 ± 0.028	0.849 ± 0.013	0.826 ± 0.011	0.862 ± 0.012
G-MMLU – STEM	MST-7B	0.753 ± 0.014	0.705 ± 0.023	0.558 ± 0.024	0.547 ± 0.018	–
	LAM-8B	0.748 ± 0.019	0.709 ± 0.017	0.670 ± 0.027	0.643 ± 0.023	–
	MST-24B	0.806 ± 0.018	0.794 ± 0.012	0.721 ± 0.055	0.679 ± 0.048	–
	LAM-70B	0.814 ± 0.016	0.817 ± 0.013	0.848 ± 0.012	0.751 ± 0.013	–
G-MMLU – Humanities	MST-7B	0.754 ± 0.015	0.697 ± 0.021	0.561 ± 0.031	0.545 ± 0.034	–
	LAM-8B	0.750 ± 0.016	0.729 ± 0.013	0.641 ± 0.026	0.638 ± 0.024	–
	MST-24B	0.777 ± 0.020	0.780 ± 0.028	0.638 ± 0.073	0.631 ± 0.035	–
	LAM-70B	0.798 ± 0.027	0.807 ± 0.022	0.777 ± 0.014	0.734 ± 0.017	–

Table 19: AUROC scores (mean ± standard error) [without multiplying by 100] for MAM **(fully connected activations)** method where the classifiers are trained and tested across 20 runs in the same language setup for all LLMs, languages, and datasets.

Dataset	Model	EN	DE	HI	BN	UR
mTREx – Capitals	MST-7B	0.5785 ± 0.1443	0.5644 ± 0.1297	0.5377 ± 0.0979	0.5399 ± 0.0987	0.5234 ± 0.0710
	LAM-8B	0.7102 ± 0.0047	0.7159 ± 0.0027	0.6768 ± 0.0033	0.6803 ± 0.0072	0.6907 ± 0.0038
	MST-24B	0.5904 ± 0.1854	0.5667 ± 0.1307	0.5439 ± 0.0975	0.5521 ± 0.1165	0.5579 ± 0.1242
mTREx – Country	MST-7B	0.5875 ± 0.1724	0.5791 ± 0.1521	0.5641 ± 0.1172	0.5223 ± 0.0531	0.5437 ± 0.1027
	LAM-8B	0.7792 ± 0.0037	0.7680 ± 0.0016	0.6656 ± 0.0020	0.5777 ± 0.0013	0.6414 ± 0.0065
	MST-24B	0.5981 ± 0.2050	0.6219 ± 0.2464	0.5587 ± 0.1327	0.5751 ± 0.1710	0.5308 ± 0.0828
mTREx – Official Language	MST-7B	0.5575 ± 0.1180	0.5754 ± 0.1470	0.5299 ± 0.0894	0.5285 ± 0.0843	0.5295 ± 0.0585
	LAM-8B	0.7467 ± 0.0015	0.6847 ± 0.0034	0.6254 ± 0.0048	0.6762 ± 0.0045	0.6490 ± 0.0069
	MST-24B	0.5404 ± 0.0770	0.5834 ± 0.1637	0.5531 ± 0.1207	0.5511 ± 0.1227	0.5399 ± 0.0890
GMMLU – Humanities	MST-7B	0.5305 ± 0.0946	0.5300 ± 0.1034	0.5164 ± 0.0367	0.5082 ± 0.0262	–
	LAM-8B	0.5558 ± 0.1723	0.5590 ± 0.2113	0.5306 ± 0.0907	0.5158 ± 0.0668	–
	MST-24B	0.5223 ± 0.0816	0.5223 ± 0.0866	0.5169 ± 0.0835	0.5204 ± 0.0420	–
GMMLU – STEM	MST-7B	0.5217 ± 0.0833	0.5219 ± 0.0943	0.5211 ± 0.0492	0.5114 ± 0.0353	–
	LAM-8B	0.5503 ± 0.1559	0.5372 ± 0.1508	0.5236 ± 0.0854	0.5158 ± 0.0540	–
	MST-24B	0.5341 ± 0.1402	0.5385 ± 0.1425	0.5114 ± 0.0465	0.5122 ± 0.0449	–

Table 20: Average AUROC scores (mean ± standard error) [without multiplying by 100] for **SelfCheckGPT** method (SGM) by varying the number of sample responses ($N \in \{10, 15, 20\}$) in high temperature settings for MST-7B, LAM-8B, and MST-24B across all datasets and languages.

Dataset	Model	EN	DE	HI	BN	UR
mTREx – Capitals	MST-7B	0.6608 ± 0.0089	0.6596 ± 0.0019	0.5552 ± 0.0046	0.5812 ± 0.0058	0.5190 ± 0.0056
	LAM-8B	0.7434 ± 0.0040	0.6684 ± 0.0078	0.6202 ± 0.0014	0.6095 ± 0.0032	0.5880 ± 0.0058
	MST-24B	0.7144 ± 0.0014	0.6148 ± 0.0037	0.5824 ± 0.0033	0.5928 ± 0.0043	0.5738 ± 0.0039
mTREx – Country	MST-7B	0.7592 ± 0.0022	0.6981 ± 0.0041	0.7150 ± 0.0018	0.6595 ± 0.0077	0.6434 ± 0.0162
	LAM-8B	0.8090 ± 0.0036	0.6839 ± 0.0017	0.7265 ± 0.0036	0.7504 ± 0.0094	0.7101 ± 0.0036
	MST-24B	0.7945 ± 0.0019	0.7406 ± 0.0070	0.7251 ± 0.0046	0.7047 ± 0.0018	0.6637 ± 0.0086
mTREx – Official Language	MST-7B	0.6986 ± 0.0096	0.6836 ± 0.0029	0.6443 ± 0.0078	0.5952 ± 0.0117	0.5695 ± 0.0041
	LAM-8B	0.7242 ± 0.0022	0.6277 ± 0.0020	0.7014 ± 0.0014	0.7198 ± 0.0103	0.6861 ± 0.0049
	MST-24B	0.6302 ± 0.0022	0.6720 ± 0.0030	0.7232 ± 0.0031	0.6691 ± 0.0085	0.7176 ± 0.0097
GMMLU – Humanities	MST-7B	0.6335 ± 0.0098	0.6188 ± 0.0082	0.5425 ± 0.0064	0.5345 ± 0.0095	–
	LAM-8B	0.6813 ± 0.0070	0.6875 ± 0.0020	0.6006 ± 0.0054	0.5630 ± 0.0138	–
	MST-24B	0.5860 ± 0.0023	0.6075 ± 0.0181	0.5710 ± 0.0105	0.5583 ± 0.0127	–
GMMLU – STEM	MST-7B	0.6046 ± 0.0129	0.6219 ± 0.0079	0.5477 ± 0.0051	0.5305 ± 0.0056	–
	LAM-8B	0.6769 ± 0.0080	0.6642 ± 0.0019	0.5978 ± 0.0053	0.5484 ± 0.0074	–
	MST-24B	0.5851 ± 0.0064	0.6333 ± 0.0083	0.5596 ± 0.0052	0.5594 ± 0.0105	–

Table 21: Average AUROC scores (mean $\pm$ standard error) [without multiplying by 100] for **Semantic Entropy** method (SEM) by varying the number of sample responses ($N \in \{10, 15, 20\}$) in high temperature settings for MST-7B, LAM-8B, and MST-24B across all datasets and languages.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	0.306	0.381	1.052	0.968	1.773	0.154	0.625	1.310	1.344	1.205	0.247	0.713	0.896	1.001	1.039
LAM-8B	0.448	0.243	0.461	1.659	0.606	0.485	0.673	0.351	2.101	0.466	0.742	0.435	0.360	1.229	0.557
MST-24B	0.325	0.874	1.518	1.785	1.762	0.304	0.722	1.066	1.521	1.050	0.386	0.423	1.080	1.145	1.067
Average	0.360	0.499	1.010	1.471	1.380	0.314	0.673	0.909	1.655	0.907	0.458	0.524	0.778	1.125	0.887

Table 22: Entropy of the softmax distribution for the first generated token for LLMs across languages over **mTREx** datasets. Generator entropies are consistently low for high-resource languages and substantially higher for low-resource languages, indicating greater uncertainty during generation in low-resource ones.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	0.449	0.543	0.400	0.433	0.294	0.342	0.442	0.203	0.124	0.183	0.318	0.461	0.261	0.282	0.351
LAM-8B	0.394	0.533	0.472	0.472	0.475	0.303	0.394	0.486	0.436	0.473	0.295	0.410	0.426	0.431	0.427
MST-24B	0.390	0.504	0.550	0.490	0.513	0.330	0.290	0.515	0.469	0.473	0.254	0.302	0.500	0.409	0.478
Average	0.411	0.527	0.474	0.465	0.427	0.325	0.375	0.401	0.343	0.376	0.289	0.391	0.396	0.374	0.419

Table 23: Entropy of the detector's softmax distribution for MAM **(self-attention)** across languages and LLMs over **mTREx** datasets. Unlike generators, detector entropies remain low and consistent across both high- and low-resource languages, indicating stable behavior that is independent of the language resource level.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	0.482	0.570	0.400	0.510	0.304	0.325	0.334	0.072	0.094	0.115	0.258	0.461	0.148	0.222	0.219
LAM-8B	0.242	0.523	0.434	0.459	0.374	0.257	0.312	0.417	0.431	0.378	0.268	0.399	0.356	0.420	0.338
MST-24B	0.212	0.541	0.469	0.484	0.455	0.300	0.240	0.412	0.401	0.336	0.214	0.287	0.375	0.372	0.396
Average	0.312	0.544	0.434	0.484	0.378	0.294	0.295	0.300	0.309	0.277	0.247	0.382	0.293	0.338	0.318

Table 24: Entropy of the detector's softmax distribution for MAM **(fully connected activations)** across languages and LLMs over **mTREx** datasets. Unlike generators, detector entropies remain low and consistent across both high- and low-resource languages, indicating stable behavior that is independent of the language resource level.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	57	54	51	41	77	63	59	50	60	78	63	59	49	46	73	63	53	54	72	66	54	54
LAM-8B	79	67	53	56	57	86	59	63	65	54	88	69	47	66	49	77	72	67	65	78	65	62	63
MST-24B	78	62	55	59	56	83	61	63	63	56	86	76	55	65	59	74	67	62	56	70	72	55	56
LAM-70B	82	65	54	57	55	86	68	56	71	57	82	63	55	61	54	83	79	70	73	86	76	70	68
Average	80	62	54	56	52	83	62	60	62	56	84	67	54	60	52	77	70	62	62	76	69	60	60

Table 25: AUROC scores of MAM **(self-attention)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	↓28	↓32	↓35	↓48	77	↓18	↓23	↓35	↓22	78	↓19	↓24	↓37	↓41	73	↓14	↓27	↓26	72	↓08	↓25	↓25
LAM-8B	79	↓15	↓33	↓29	↓28	86	↓31	↓27	↓24	↓37	88	↓22	↓47	↓25	↓44	77	↓06	↓13	↓16	78	↓17	↓21	↓19
MST-24B	78	↓21	↓29	↓24	↓28	83	↓27	↓24	↓24	↓33	86	↓12	↓36	↓24	↓31	74	↓09	↓16	↓24	70	↑03	↓21	↓20
LAM-70B	82	↓21	↓34	↓30	↓33	86	↓21	↓35	↓17	↓34	82	↓23	↓33	↓26	↓34	83	↓05	↓16	↓12	86	↓12	↓19	↓21
Average	80	↓22	↓32	↓30	↓35	83	↓25	↓28	↓25	↓33	84	↓20	↓36	↓29	↓38	77	↓09	↓19	↓19	76	↓09	↓21	↓21

Table 26: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(self-attention)** method w.r.t. the corresponding EN baseline for the model and dataset in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	82	68	53	45	53	79	57	58	58	50	81	67	49	50	48	76	66	56	54	73	69	55	55
LAM-8B	79	67	54	61	52	83	53	63	63	55	88	65	58	50	49	78	73	69	64	78	71	63	60
MST-24B	76	65	60	61	56	86	55	62	71	59	82	76	59	64	62	82	77	67	62	78	78	60	58
LAM-70B	80	66	59	46	59	87	66	61	65	52	82	61	58	55	56	84	79	79	68	86	77	71	68
Average	79	66	56	53	54	84	57	60	64	54	83	67	56	54	53	80	73	67	61	79	73	62	60

Table 27: AUROC scores of MAM **(fully connected activations)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	82	↓17	↓35	↓45	↓35	79	↓28	↓27	↓27	↓37	81	↓17	↓40	↓38	↓41	76	↓13	↓26	↓29	73	↓05	↓25	↓25
LAM-8B	79	↓15	↓32	↓23	↓34	83	↓36	↓24	↓24	↓34	88	↓26	↓34	↓43	↓44	78	↓06	↓12	↓18	78	↓09	↓19	↓23
MST-24B	76	↓14	↓21	↓20	↓26	86	↓36	↓28	↓17	↓31	82	↓07	↓28	↓22	↓24	82	↓06	↓18	↓24	78	↓0	↓23	↓26
LAM-70B	80	↓18	↓26	↓42	↓26	87	↓24	↓30	↓25	↓40	82	↓26	↓29	↓33	↓32	84	↓06	↓06	↓19	86	↓10	↓17	↓21
Average	79	↓16	↓29	↓33	↓32	84	↓32	↓29	↓24	↓36	83	↓19	↓33	↓35	↓36	80	↓09	↓16	↓24	79	↓08	↓22	↓24

Table 28: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(fully connected activations)** method w.r.t. the corresponding EN baseline for the model and dataset in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	78	69	72	75	61	79	78	74	79	70	74	81	71	70	78	74	66	59	53	68	66	55	52
LAM-8B	80	71	65	70	67	83	84	78	83	77	88	84	83	79	82	77	74	69	69	76	74	65	63
MST-24B	72	74	67	73	69	82	92	78	84	80	84	88	83	84	84	77	72	70	62	71	77	62	62
LAM-70B	82	75	68	69	73	86	92	82	77	82	78	87	85	78	86	85	81	85	75	85	78	80	72
Average	78	72	68	71	67	82	86	78	80	77	80	84	80	77	82	77	73	70	64	75	73	65	62

Table 29: AUROC scores of MAM **(self-attention)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	78	↓12	↓08	↓04	↓22	79	↓01	↓06	↓0	↓11	74	↑09	↓04	↓05	↑05	74	↓11	↓20	↓28	68	↓03	↓19	↓24
LAM-8B	80	↓11	↓19	↓12	↓16	83	↑01	↓06	↓0	↓07	88	↓05	↓06	↓10	↓07	77	↓04	↓10	↓10	76	↓03	↓14	↓17
MST-24B	72	↑03	↓07	↑01	↓04	82	↑12	↓05	↑02	↓02	84	↑05	↓01	↓0	↓0	77	↓06	↓09	↓19	71	↑08	↓13	↓13
LAM-70B	82	↓09	↓17	↓16	↓11	86	↑07	↓05	↓10	↓05	78	↑12	↑09	↓0	↑10	85	↓05	↓0	↓12	85	↓08	↓06	↓15
Average	78	↓08	↓13	↓09	↓14	82	↑05	↓05	↓02	↓06	80	↑05	80	↓04	↑02	77	↓05	↓09	↓17	75	↓03	↓13	↓17

Table 30: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(self-attention)** method w.r.t. the corresponding EN baseline for the model and dataset in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	72	72	74	67	80	85	74	79	78	79	83	71	74	80	75	66	58	55	71	68	57	51
LAM-8B	79	72	70	72	69	84	87	78	83	77	89	84	83	79	83	78	74	68	69	75	74	64	63
MST-24B	79	75	72	77	71	84	94	79	86	83	84	89	84	85	85	82	78	71	69	76	80	66	63
LAM-70B	79	75	67	70	71	86	94	79	79	82	80	87	88	82	89	84	81	85	75	85	78	80	72
Average	78	73	70	73	69	83	89	77	81	79	83	85	81	79	84	80	75	70	66	76	74	66	62

Table 31: AUROC scores of MAM **(fully connected activations)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets and languages in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	79	↓09	↓09	↓06	↓15	80	↑06	↓08	↓01	↓02	79	↑05	↓10	↓06	↑01	75	↓12	↓23	↓27	71	↓04	↓20	↓28
LAM-8B	79	↓09	↓11	↓09	↓13	84	↑04	↓07	↓01	↓08	89	↓06	↓07	↓11	↓07	78	↓05	↓13	↓12	75	↓01	↓15	↓16
MST-24B	79	↓05	↓09	↓03	↓10	84	↑12	↓06	↑02	↓01	84	↑06	↓0	↑01	↑01	82	↓05	↓13	↓16	76	↑05	↓13	↓17
LAM-70B	79	↓05	↓15	↓11	↓10	86	↑09	↓08	↓08	↓05	80	↑09	↑10	↑02	↑11	84	↓04	↑01	↓11	85	↓08	↓06	↓15
Average	78	↓06	↓10	↓06	↓12	83	↑07	↓07	↓02	↓05	83	↑02	↓02	↓05	↑01	80	↓06	↓12	↓18	76	↓03	↓13	↓18

Table 32: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, MST-24B, and LAM-70B across languages for MAM **(fully connected activations)** method w.r.t. the corresponding EN baseline for the model and dataset in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	79	70	62	60	61	82	60	70	57	63	83	78	63	52	60
LAM-8B	86	67	60	53	57	84	72	74	48	68	88	71	70	57	68
MST-24B	81	67	56	61	59	81	76	66	66	64	84	78	53	48	56
Average	82	67	59	58	58	82	69	69	57	64	85	75	62	51	61

Table 33: AUROC scores of MAM **(self-attention)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	79	↓11	↓22	↓24	↓23	82	↓27	↓15	↓30	↓23	83	↓06	↓24	↓37	↓28
LAM-8B	86	↓22	↓30	↓38	↓34	84	↓14	↓12	↓43	↓19	88	↓19	↓20	↓35	↓23
MST-24B	81	↓17	↓31	↓25	↓27	81	↓06	↓19	↓19	↓21	84	↓07	↓37	↓43	↓33
Average	82	↓18	↓28	↓29	↓29	82	↓16	↓16	↓30	↓22	85	↓12	↓27	↓40	↓28

Table 34: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, and MST-24B across languages for MAM **(self-attention)** method [when averaging artifacts across multiple generated tokens] w.r.t. the corresponding EN baseline for the model and dataset in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	82	71	59	61	65	85	69	62	53	71	84	80	66	53	62
LAM-8B	82	67	51	60	52	83	71	71	69	68	89	75	58	48	57
MST-24B	83	71	57	58	59	85	72	67	62	68	86	80	75	65	73
Average	82	69	55	59	58	84	70	66	61	69	86	78	66	55	63

Table 35: AUROC scores of MAM **(fully connected activations)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	82	↓13	↓28	↓26	↓21	85	↓19	↓27	↓38	↓16	84	↓05	↓21	↓37	↓26
LAM-8B	82	↓18	↓38	↓27	↓37	83	↓14	↓14	↓17	↓18	89	↓16	↓35	↓46	↓36
MST-24B	83	↓14	↓31	↓30	↓29	85	↓15	↓21	↓27	↓20	86	↓07	↓13	↓24	↓15
Average	82	↓16	↓33	↓28	↓29	84	↓17	↓21	↓27	↓18	86	↓09	↓23	↓36	↓27

Table 36: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, and MST-24B across languages for MAM **(fully connected activations)** method [when averaging artifacts across multiple generated tokens] w.r.t. the corresponding EN baseline for the model and dataset in the cross-lingual setting (train over EN data and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	76	69	74	72	73	78	82	75	78	81	81	79	73	69	75
LAM-8B	82	72	73	74	76	84	84	84	85	83	89	81	84	80	83
MST-24B	78	75	71	77	70	80	91	81	84	83	82	87	85	86	85
Average	78	72	72	74	72	80	85	79	82	82	83	82	80	78	81

Table 37: AUROC scores of MAM **(self-attention)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	76	↓09	↓03	↓05	↓04	78	↑05	↓04	↓0	↑04	81	↓02	↓10	↓15	↓07
LAM-8B	82	↓12	↓11	↓10	↓07	84	↓0	↓0	↑01	↓01	89	↓09	↓06	↓10	↓07
MST-24B	78	↓04	↓09	↓01	↓10	80	↑14	↑01	↑05	↑04	82	↑06	↑04	↑05	↑04
Average	78	↓08	↓08	↓05	↓08	80	↑06	↓01	↑02	↑02	83	↓01	↓04	↓06	↓02

Table 38: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, and MST-24B across languages for MAM **(self-attention)** method [when averaging artifacts across multiple generated tokens] w.r.t. the corresponding EN baseline for the model and dataset in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	80	74	75	74	76	84	88	76	80	88	83	83	77	71	80
LAM-8B	82	73	75	75	76	83	87	84	86	84	89	83	86	82	85
MST-24B	82	77	74	79	70	83	92	83	87	88	84	89	86	87	86
Average	81	74	74	76	73	83	88	81	84	86	85	85	82	80	83

Table 39: AUROC scores of MAM **(fully connected activations)** method [when averaging artifacts across multiple generated tokens] for MST-7B, LAM-8B, and MST-24B over **mTREx** datasets across languages in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language				
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR
MST-7B	80	↓08	↓06	↓08	↓05	84	↑05	↓10	↓05	↑05	83	↓0	↓07	↓14	↓04
LAM-8B	82	↓11	↓09	↓09	↓07	83	↑05	↑01	↑04	↑01	89	↓07	↓03	↓08	↓04
MST-24B	82	↓06	↓10	↓04	↓15	83	↑11	↓0	↑05	↑06	84	↑06	↑02	↑04	↑02
Average	81	↓09	↓09	↓06	↓10	83	↑06	↓02	↑01	↑04	85	↓0	↓04	↓06	↓02

Table 40: Percentage of increment ↑ or decrement ↓ in AUROC scores for MST-7B, LAM-8B, and MST-24B across languages for MAM **(fully connected activations)** method [when averaging artifacts across multiple generated tokens] w.r.t. the corresponding EN baseline for the model and dataset in the multilingual setting (train over data across all languages and test over target language data).

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	74	65 ↓12	80 ↑08	80 ↑08	88 ↑19	79	72 ↓09	92 ↑16	96 ↑22	95 ↑20	86	78 ↓09	90 ↑05	91 ↑06	89 ↑03	69	67 ↓03	77 ↑12	76 ↑10	66	63 ↓05	67 ↑02	70 ↑06
LAM-8B	81	70 ↓14	67 ↓17	74 ↓09	73 ↓10	83	80 ↓04	72 ↓13	80 ↓04	74 ↓11	85	80 ↓06	75 ↓12	75 ↓12	75 ↓12	68	67 ↓01	64 ↓06	68	71	67 ↓06	59 ↓17	65 ↓08
MST-24B	80	72 ↓10	74 ↓08	72 ↓10	72 ↓10	84	87 ↑04	75 ↓11	81 ↓04	77 ↓08	89	86 ↓03	76 ↓15	77 ↓13	76 ↓15	70	68 ↓03	66 ↓06	60 ↓14	73	68 ↓07	61 ↓16	63 ↓14
LAM-70B	89	77 ↓13	62 ↓30	67 ↓25	70 ↓21	88	88 ↓0	74 ↓16	74 ↓16	76 ↓14	89	89 ↓0	81 ↓09	75 ↓16	80 ↓10	78	75 ↓04	74 ↓05	69 ↓12	83	76 ↓08	75 ↓10	65 ↓22

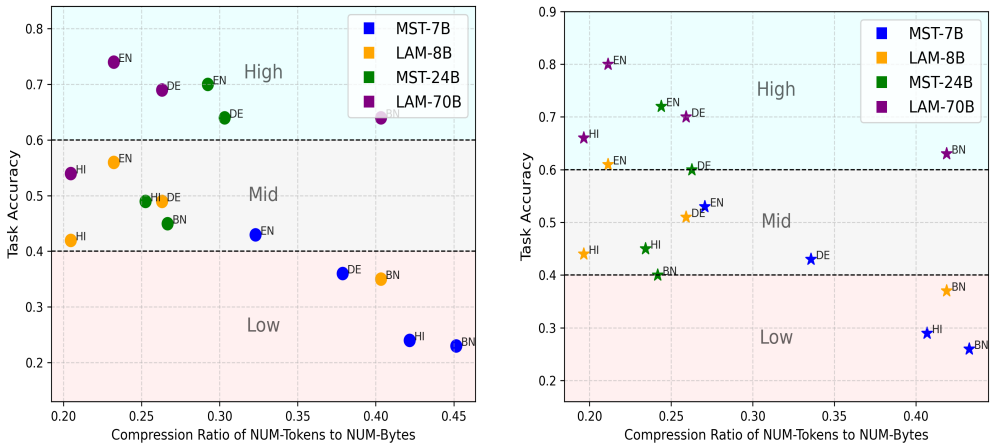
Table 41: Classifier's accuracy of the MAM **(self-attention)** method for MST-7B, LAM-8B, MST-24B, and LAM-70B across datasets in English (EN), German (DE), Hindi (HI), Bengali (BN), and Urdu (UR) languages. A value of ↑X or ↓Y represents an X% increase or Y% decrease in score w.r.t. the corresponding EN baseline for a model and dataset.

Models	mTREx – Capitals					mTREx – Country					mTREx – Official Language					G-MMLU – STEM				G-MMLU – Humanities			
	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	UR	EN	DE	HI	BN	EN	DE	HI	BN
MST-7B	74	68 ↓08	79 ↑07	79 ↑07	89 ↑20	81	78 ↓04	92 ↑14	96 ↑19	95 ↑17	87	81 ↓07	90 ↑03	91 ↑05	91 ↑05	71	67 ↓06	77 ↑08	76 ↑07	68	65 ↓04	63 ↓07	70 ↑03
LAM-8B	79	70 ↓11	70 ↓11	76 ↓04	74 ↓06	82	79 ↓04	72 ↓12	80 ↓02	75 ↓09	86	80 ↓07	78 ↓09	75 ↓13	76 ↓12	69	68 ↓01	63 ↓09	68 ↓01	70	68 ↓03	58 ↓17	65 ↓07
MST-24B	80	71 ↓11	71 ↓11	71 ↓11	73 ↓09	83	88 ↑06	75 ↓10	82 ↓01	79 ↓05	88	84 ↓05	79 ↓10	77 ↓12	79 ↓10	75	72 ↓04	67 ↓11	65 ↓13	77	74 ↓04	64 ↓17	63 ↓18
LAM-70B	88	77 ↓12	61 ↓31	67 ↓24	73 ↓17	86	87 ↑01	72 ↓16	71 ↓17	72 ↓16	88	87 ↓01	81 ↓08	77 ↓12	79 ↓10	76	75 ↓01	74 ↓03	68 ↓11	83	76 ↓08	73 ↓12	67 ↓19

Table 42: Classifier's accuracy of the `MAM` **(fully connected activations)** method for `MST-7B`, `LAM-8B`, `MST-24B`, and `LAM-70B` across datasets in English (`EN`), German (`DE`), Hindi (`HI`), Bengali (`BN`), and Urdu (`UR`) languages. A value of ↑X or ↓Y represents an X% increase or Y% decrease in score w.r.t. the corresponding `EN` baseline for a model and dataset.

Models	mTREx – Capitals				mTREx – Country				mTREx – Official Language				G-MMLU – STEM			G-MMLU – Humanities		
	DE	HI	BN	UR	DE	HI	BN	UR	DE	HI	BN	UR	DE	HI	BN	DE	HI	BN
`MAM` (fully connected activations)																		
MST-7B	1.2	7.0	6.2	5.0	3.8	5.1	17.0	34.0	3.7	11.7	17.8	13.6	0.8	1.0	0.8	1.7	1.5	1.2
LAM-8B	1.6	6.0	9.2	5.1	7.0	6.7	46.0	7.5	0.0	6.5	5.1	7.5	1.8	1.2	1.8	2.5	1.1	1.6
MST-24B	20.0	12.5	51.0	8.5	3.5	9.8	46.0	12.0	1.9	9.8	11.8	10.2	2.0	2.3	1.9	4.0	2.1	2.3
LAM-70B	4.0	2.8	4.5	6.1	3.9	3.1	4.3	5.9	1.0	4.2	NA	8.5	1.2	20.0	1.1	1.4	2.8	1.2
Average	3.2	5.6	9.8	5.7	4.5	5.5	25.0	10.2	2.3	NA	50.0	14.3	1.4	1.9	1.4	2.5	1.5	1.4
`MAM` (self-attention)																		
MST-7B	1.3	9.8	10.0	6.9	5.8	16.5	9.7	13.6	2.2	35.0	17.8	17.0	1.0	1.0	0.9	1.7	1.4	1.4
LAM-8B	1.6	3.8	4.6	4.2	28.0	4.4	15.3	5.6	0.0	5.2	5.1	10.0	2.3	1.3	1.9	2.0	1.3	1.7
MST-24B	10.0	7.1	10.2	6.4	2.8	13.0	23.0	48.0	4.3	19.5	47.0	NA	3.0	10.5	2.1	2.0	4.5	4.0
LAM-70B	1.7	2.6	3.8	4.6	3.9	7.0	5.6	10.2	0.9	5.0	47.0	4.9	1.7	NA	1.2	1.2	2.3	1.2
Average	2.2	4.5	5.4	5.1	5.4	8.8	NA	10.2	3.5	NA	12.5	21.5	1.8	2.4	1.5	3.3	2.0	1.7
SelfCheckGPT																		
MST-7B	6.0	9.8	10.0	6.1	5.8	7.3	3.8	6.8	2.8	23.3	17.8	6.8	3.5	4.8	5.0	10.0	4.8	4.5
LAM-8B	8.0	14.0	15.3	23.0	28.0	3.6	2.3	3.5	0.0	2.2	5.1	3.0	7.0	2.3	2.6	2.5	2.4	3.0
MST-24B	2.2	3.8	5.1	5.7	4.7	3.9	11.5	2.7	1.0	4.9	5.9	13.7	NA	2.3	2.8	12.0	NA	8.0
LAM-70B	2.0	3.5	4.5	11.0	2.1	1.3	4.9	5.9	0.7	12.5	6.7	11.3	2.5	20.0	10.0	10.0	7.0	17.0
Average	3.2	5.0	6.1	7.3	6.8	3.4	4.2	6.4	1.4	20.0	50.0	14.3	7.0	3.8	3.8	5.0	5.0	4.8
Semantic Entropy																		
MST-7B	6.0	4.1	5.6	3.9	3.3	13.2	7.6	5.7	3.7	11.7	6.5	4.9	3.5	2.7	2.5	3.3	2.7	2.7
LAM-8B	1.3	3.5	3.3	2.9	2.2	4.4	7.7	4.5	0.0	8.7	36.0	10.0	3.5	1.6	1.6	10.0	2.4	2.4
MST-24B	2.0	3.8	4.6	3.6	4.7	6.5	5.1	4.4	2.6	3.9	9.4	4.1	1.2	5.2	8.3	2.4	27.0	32.0
LAM-70B	1.7	4.7	12.2	27.5	3.0	3.1	39.0	6.8	2.0	8.3	47.0	3.1	NA	20.0	5.0	NA	14.0	NA
Average	1.9	3.8	4.9	4.2	3.0	6.3	7.1	7.3	3.5	40.0	25.0	43.0	7.0	3.8	3.2	10.0	5.0	4.8

Table 43: Ratio of model's task accuracy delta to HD's AUROC delta w.r.t `EN` (i.e., 10^{TPHR} values) for `MAM` **(fully connected activations)**, `MAM` **(self-attention)**, **SelfCheckGPT**, and **Semantic Entropy** methods. Values where the AUROC delta is zero are marked as NA. Darker color shades indicate a higher value, signifying that although the task accuracy for these languages drops drastically, the HD's performance remains much more stable.



(a) For G-MMLU – STEM

(b) For G-MMLU – Humanities

Figure 18: Comparison of task accuracy with token compression ratio (number of tokens produced by the model's tokenizer divided by the number of bytes in the input prompt) for G-MMLU across all LLMs and languages.