How Much Do LLMs Hallucinate across Languages? On Multilingual Estimation of LLM Hallucination in the Wild

Anonymous ACL submission

Abstract

In the age of misinformation, hallucination—the tendency of Large Language Models (LLMs) to generate non-factual or unfaithful responses—represents the main risk for their global utility. Despite LLMs becoming increasingly multilingual, the vast majority of research on detecting and quantifying LLM hallucination are (a) English-centric and (b) focus on machine translation (MT) and summarization, tasks that are less common "in the wild" than open information seeking. In contrast, we aim to quantify the extent of LLM hallucination across languages in knowledge-intensive long-form question answering (LFQA). To this end, we train a multilingual hallucination detection model and conduct a large-scale study across 30 languages and 6 open-source LLM families. We start from an English hallucination detection dataset and rely on MT to translate-train a detection model. We also manually annotate gold data for five high-resource languages; we then demonstrate, for these languages, that the estimates of hallucination rates are similar between silver (LLM-generated) and gold test sets, validating the use of silver data for estimating hallucination rates for other languages. For the final rates estimation, we build open-domain QA dataset for 30 languages with LLM-generated prompts and Wikipedia articles as references. Our analysis shows that LLMs, in absolute terms, hallucinate more tokens in high-resource languages due to longer responses, but that the actual hallucination rates (i.e., normalized for length) seems uncorrelated with the sizes of languages' digital footprints. We also find that smaller LLMs hallucinate more, and significantly, LLMs with broader language support display higher hallucination rates.

1 Introduction

Generalizing seamlessly to (seemingly) arbitrary language understanding, reasoning, and generation tasks, Large Language Models (LLMs) (Kojima et al., 2022; Dubey et al., 2024; Aryabumi et al., 2024; Yang et al., 2024) have arguably become the first ubiquitously adopted language technology, with application ranging from search engines (Xiong et al., 2024), interactive agents (Teubner et al., 2023) and knowledge retrieval (Yu et al., 2023) to various content generation tasks (Liu et al., 2022b). Their utility, however, is hindered by their tendency to *hallucinate* (Maynez et al., 2020; Zhou et al., 2021; Ji et al., 2023; Zhang et al., 2023), that is, produce information that is either (i) inaccurate or factually incorrect with respect to the objective state of the world (e.g., in open-ended question answering) or (ii) unfaithful with respect to some reference (e.g., in summarization).

Consequently, a large body of work on tackling LLM hallucination has emerged, with efforts falling into the three main areas: (1) detection, i.e., identification of the hallucinated content; (2) evaluation, primarily focusing on measures for quantifying the extent and severity of hallucinations ; and (3) mitigation, focusing on mitigating hallucinative tendencies of LLMs (Ji et al., 2023). While significant progress has been made in English (Maynez et al., 2020; Liu et al., 2022a; Obaid ul Islam et al., 2023; Kasai et al., 2024; Mishra et al., 2024), hallucination evaluation efforts targeting other languages have been much sparser (Clark et al., 2023; Guerreiro et al., 2023; Herrlein et al., 2024; Shafayat et al., 2024). Moreover, these efforts have primarily targeted high-resource languages (Qiu et al., 2023; Shafayat et al., 2024) with benchmarks limited to reference-based tasks—text summarization (Clark et al., 2023; Aharoni et al., 2022) and machine translation (Dale et al., 2023; Guerreiro et al., 2023). While highly relevant, these tasks are arguably less representative of LLM usage 'in the wild' (Trippas et al., 2024), where knowledge-intensive long-form question answering (LFQA) is more prominent.

In this work, we address the above gaps in mul-

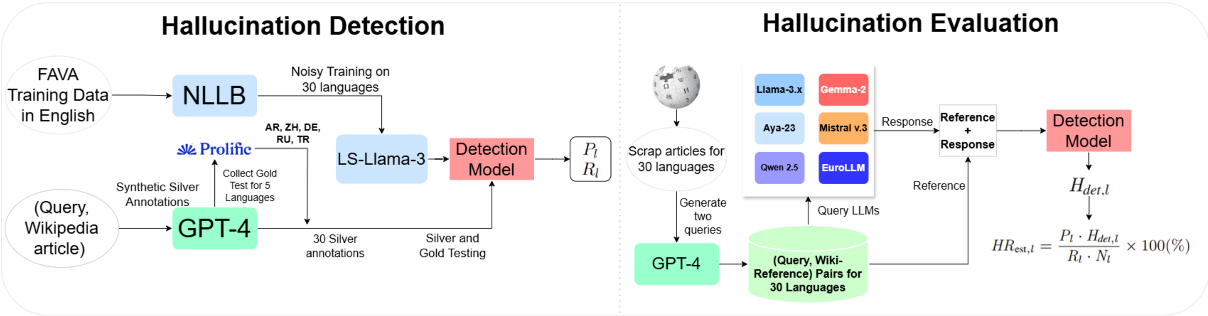


Figure 1: Illustration of our approach for estimating hallucination rates in the wild. **Hallucination Detection and Model Evaluation** (left side): (1) We automatically translate the English FAVA (Mishra et al., 2024) dataset to 30 languages and train our multilingual hallucination detection (HD) model on this (noisy) multilingual training data; (2) We synthesize a *silver* multilingual hallucination evaluation dataset by prompting a state-of-the-art LLM (GPT-4) to introduce hallucinations in its answers to knowledge-seeking questions; for a subset of five high-resource languages, we additionally collect *gold* (i.e., human) hallucination annotations; we dub this 30-language evaluation benchmark MFAVA. We use MFAVA to estimate HD model's per-language performances (precision and recall). **Hallucination Rate Estimation in the Wild** (right side): (3) We estimate the hallucination rates for all 30 languages and six different LLM families from the number of detections of the HD model and its performance.

tilingual hallucination detection and evaluation research with the ultimate goal of *estimating the "in the wild" hallucination rates of LLMs across languages*. Multilingual estimation of such hallucination rates is challenging due to the scarcity of multilingual hallucination benchmarks covering open-ended knowledge-seeking tasks that are representative of real-world LLM usage: unlike in reference-based generation tasks like summarization and machine translation, LLMs often generate long-form responses to open-ended questions, requiring more comprehensive evaluation approaches (Wei et al., 2024b). Concretely, we present a large-scale study that estimates hallucination rates for 30 languages (both high(er)- and low(er)-resource languages). Our main contributions are as follows: **(1)** We translate-train (Artetxe et al., 2023; Ebing and Glavaš, 2024) a multilingual hallucination detection (HD) model on 30 languages. **(2)** We create MFAVA HD evaluation datasets with span-level human annotations (MFAVA-GOLD) for five high-resource languages, generate synthetic (MFAVA-SILVER) HD evaluation datasets for 25 additional languages, and validate the use of MFAVA-SILVER by showing the MFAVA-SILVER and MFAVA-GOLD estimates yield similar results; **(3)** We propose a protocol for estimating "in the wild" hallucination rates of LLMs and introduce an extensive synthetic dataset (51,133 prompts across 30 languages) for estimating LLM hallucination rates in highly multilingual settings; **(4)** We offer a comprehensive hallucination rate analysis of six LLM families, validating previous findings that larger models tend to hallucinate less, and uncovering that broader LLM language coverage correlates with increased hallucination rates. This work is the first to estimate "in-the-wild" LLM hallucination rates for a wide range of languages using knowledge-intensive open-domain LFQA, reflecting real-world usage. Our comprehensive framework is illustrated in Figure 1.

2 Background and Related Work

We provide a brief overview of the body of related work on (1) hallucination detection models and (2) benchmarks for evaluating LLM hallucination.

Hallucination Detection. Coarsely, LLM hallucinations fall into two categories. *Intrinsic* hallucination are content contradicts some reference information source. The reference may be explicitly given to the LLM as part of the task (e.g., the text to be summarized in summarization or source language text in machine translation) or it may implicit (e.g., general world knowledge in question answering). In contrast, *extrinsic* hallucination refers to content that does not contradict the reference but is unnecessary or superfluous with respect to the task (e.g., additional facts in fact-based question answering) (Ji et al., 2023). Recent work introduced finer-grained taxonomies for both categories. For example, Mishra et al. (2024) distinguish between several types of intrinsic hallucinations (e.g., entity-based hallucinations or relation-based hallucinations). In a similar vein, extrinsic hallucinations are split into subtypes such as *invented*, *subjective*, and *unverifiable* content.

Unsurprisingly, most hallucination detection (and classifications) models are based on neural languages models. These are are either pre-trained encoder LM (Zhou et al., 2021; Liu et al., 2022b), discriminatively fine-tuned to classify texts as containing hallucinations or not or LLMs prompted (zero-shot or with in-context examples) to detect hallucinations (Manakul et al., 2023; Yang et al., 2023) or fine-tuned to generate hallucinated spans (Mishra et al., 2024). In this work, we cast hallucination detection as a span-detection task, formulated discriminatively, with a classifier on top of an "encoder-based" LM. However, instead of resorting to small pretrained encoder LMs, we bidirectionally (i.e., discriminatively) fine-tune a larger generative LLM, following recent advances in converting decoder LMs into encoders (Li et al., 2023b; Dukić and Šnajder, 2024; BehnamGhader et al., 2024; Schmidt et al., 2024).

Hallucination Benchmarks. Hallucination detection models as well as evaluation datasets have largely focused on English vary in the granularity from document-level (Yang et al., 2023) of annotations/predictions, over passage- and sentence-level annotations (Zhou et al., 2021; Manakul et al., 2023), to fine-grained token- or span-level annotations (Liu et al., 2022a; Mishra et al., 2024). Notable examples include SelfCheckGPT (Manakul et al., 2023), HaluEval (Li et al., 2023a), and ScreenEval (Lattimer et al., 2023), which measure hallucination detection rates in summarization and single-fact question answering. Multilingual benchmarks for evaluating hallucination detection models remain sparse and focus on reference-based tasks like machine translation (Dale et al., 2023) and summarization (Qiu et al., 2023) which poorly represent the LLM usage in the wild.

Faithfulness in reference-based tasks is complemented by truthfulness (i.e., factuality) in question answering. Most benchmarks, e.g., TruthFulQA (Lin et al., 2022), RealtimeQA (Kasai et al., 2024), FreshQA (Vu et al., 2023), and SimpleQA (Wei et al., 2024a) here are English-centric and cover only questions that require a simple single-factoid answer. LongFact (Wei et al., 2024b), Factscore (Min et al., 2023) and mFactScore (Kim et al., 2024) do test LLMs truthfulness in generating long and free-form answers. However, LongFact is an English-only benchmark, whereas Factscore and mFactscore, albeit multilingual, cover a very specific domain of biographic questions.

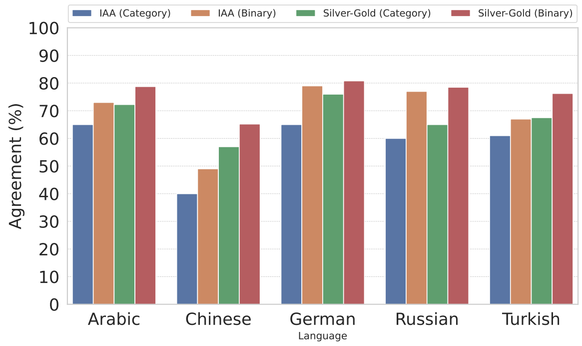


Figure 2: **1)** Inter-annotator agreement (IAA) for hallucination span detection (Binary; blue bars) and classification (Category; orange bars) for five high-resource languages; **2)** Hallucination span and class agreement between human labels and GPT-4 generated hallucinations (Silver-Gold; agreement on spans only: red bars; agreement on spans *and* hallucination type: green bars).

3 Hallucination Detection

We first describe how we obtained multilingual hallucination detection (HD) datasets (§3.1) and then report on training and evaluation of a multilingual hallucination detection model (§3.2).

3.1 mFAVA Benchmark

HD Evaluation Datasets. We start from the English FAVA (Mishra et al., 2024) dataset and its respective set of fine-grained hallucination types. FAVA's evaluation portions were created by (1) eliciting information-seeking prompts (i.e., questions) from various sources, (2) generating responses with three LLMs and (3) having human annotators label hallucinated span [1]s. We follow a similar protocol to create evaluation datasets for 30 languages.[2] We start from 300 information-seeking prompts, 150 from evaluation portion of FAVA and 150 from the Natural Questions dataset (Kwiatkowski et al., 2019). We then ask GPT-4 (Achiam et al., 2023) to (1) first create answer passages in a target language and then to (2) explicitly introduce the hallucinations of the fine-grained FAVA types into the answer. We refer to these synthetically labeled hallucination evaluation datasets, comparable across the 30 target languages, as mFAVA-Silver.

For five linguistically diverse high-resource languages—Arabic, Chinese, German, Russian,

[1] See the original paper for more details and §A for prompts for (2) and (3).

[2] §A.1 Figure 7 lists the mFAVA languages.

Very Unlikely	Unlikely	Neutral	Likely	Very Likely
21.8%	24.7%	13.0%	25.3%	15.2%

Table 1: Annotator ratings for probability of augmented text fooling the reader for the 5 gold languages.

	ENT	REL	INV	CON	UNV	SUB	Total
RU	184	65	188	287	211	153	1,088
AR	144	10	171	123	150	69	667
ZH	264	18	259	282	265	139	1,227
DE	546	25	311	324	333	238	1,777
TR	149	27	288	244	161	149	1,018
Total	1,287	145	1,217	1,260	1,120	748	5,777

Table 2: Hallucinated span counts in the gold dataset across languages. ENT (Entity), REL (Relation), INV (Invented), CON (Contradictory), UNV (Unverifiable), SUB (Subjective).

and Turkish—we also collect human hallucination annotations. To this end, we provide to the annotators the reference Wikipedia page, and the (hallucination-enriched) generation from mFAVA-Silver (of course, without the GPT-4's hallucination annotations). We source the annotations via Prolific, recruiting 5 annotators per language: all five annotators first annotated the same 50 instances, after which they were given non-overlapping sets of 50 more instances. We provide more details on the annotation process (and costs) in the §A.2. We measure the inter-annotator agreement (IAA) in terms of pairwise-averaged Cohen's kappa on token-level class decisions, both with (*IAA Category*) and without (*IAA Binary*) considering the fine-grained hallucination types. As shown in Figure 2, we observe satisfactory to good IAA for all five languages. Regarding the 50 instances labeled by all annotators, we ultimately take the annotations of the annotator that has the highest IAA with hallucination annotatios of GPT-4 from mFAVA-Silver. We denote the final human-labeled evaluation datasets for the five high-resource languages with mFAVA-Gold. Figure 2 also shows the overall IAA between human annotations from mFAVA-Gold and GPT-4's synthetic annotations from mFAVA-Silver (*Silver-Gold*): interestingly, we observe that human annotators on average agree more with GPT-4 than with one another.

Because we synthesize the hallucinated content with GPT-4 (the annotators, of course, did not know that nor which part of the generation was meant to be a hallucination according to GPT-4), there is a risk that these hallucinations may not be realistic in the sense that they can fool a human reader. Because of this, we asked our annotators to additionally indicate (on a 5-degree Likert scale from "very unlikely" to "very likely") the likelihood of hallucination fooling a human reader for each span that they labeled. Table 1 reveals that more than half of the labeled hallucinations were judged as convincing (i.e., not unlikely to fool a human). The silver test set statistics for all 30 languages are shown in §A Figure 6. Gold annotations statistics are shown in Table 2.

Training Dataset. The FAVA training set, consisting of ca. 30K instances, is fully synthetically created in the same way as the test portion, just without the human annotation step. We automatically translate the training portion of the FAVA dataset using NLLB (Costa-jussà et al., 2022) to our 30 target languages. After translation, we project the span-level annotations to token-level labels using the simple Inside-Out (I-O) scheme (Ramshaw and Marcus, 1995)[3]. Like our evaluation benchmark mFAVA, we prepare training data for two tasks: (1) detecting hallucinated spans, regardless of hallucination type (*Binary* task: tokens are classified as either part of a hallucinated span or not) and (2) detection and hallucination type classification (*Category* task: 7-way classification, tokens classified into one of 6 FAVA hallucination types or as not part of a hallucinated span).

3.2 Multilingual Hallucination Detection

Models. Using the translations of ca. 30K FAVA training instances in our 30 target languages, we train the following models: (1) Mono denotes monolingual models trained on data of one language (and evaluated for the same language on the respective mFAVA portion), i.e., we train 30 Mono models, one for each of our target languages; (2) Multi refers to a single multilingual model trained on concatenated training data of all 30 languages. We train the Mono models and the Multi model for both tasks, *Binary* and *Category*. We follow the recent body of work that successfully converts generative decoder LLMs into encoders for discriminative tasks (Li et al., 2023b; Dukić and Šnajder, 2024; BehnamGhader et al., 2024; Schmidt et al., 2024) and fine-tune Llama-3-8B-base (Dubey et al., 2024) by removing future-token masking, i.e., allowing for bidirectional contextu-

[3]In preliminary experiments, we also tested the B-I-O scheme, but I-O led to better span detection performance.

Task	Model	Context	German		Chinese		Arabic		Russian		Turkish	
			Silver	Gold	Silver	Gold	Silver	Gold	Silver	Gold	Silver	Gold
Binary	MONO	*Bidirect*	78.0	58.0	62.4	55.1	75.3	54.4	78.9	60.7	78.5	66.7
	MULTI	*Bidirect*	**89.5**	**65.0**	69.7	58.7	**82.5**	**61.6**	**89.1**	**65.5**	**86.4**	**72.5**
	MULTI	*Causal*	81.8	59.6	**76.3**	**62.2**	75.3	60.0	75.8	55.6	75.7	67.3
Category	MONO	*Bidirect*	53.4	38.3	35.2	22.6	14.6	7.3	63.3	36.2	49.1	30.3
	MULTI	*Bidirect*	**73.2**	**45.0**	46.5	30.1	**66.1**	**37.2**	**72.3**	**41.5**	**72.9**	**51.8**
	MULTI	*Causal*	68.7	43.4	**56.5**	**34.1**	51.8	29.4	62.6	37.9	58.6	42.4

Table 3: Token-level F1 performance of multilingual (MULTI) and monolingual (MONO) hallucination detection models for five high-resource languages with both *Silver* and *Gold* evaluation data in MFAVA. Performance reported for hallucination detection alone (*Binary*) and hallucination detection and type classification (*Category*). Models fine-tuned without (*Bidirect*) or with (*Causal*) future token masking. **Bold**: best result in each column.

alization (*Bidirect*). For comparison, for the *Multi* model, we also fine-tune the decoder as-is, using the default causal token masking (i.e., unidirectional contextualization; *Causal*).

Training. In all cases, we freeze the original model parameters and train QLora adapters (Dettmers et al., 2024), with three runs (random seeds) for each experiment, reporting mean performance. The input to the models is the reference Wikipedia article, prepended to the LLM-generated answer, with the cross-entropy loss computed exclusively over the tokens of the LLM-generated answer. We provide further training details in §A.3.

Results. Table 3 summarizes the hallucination detection performance for five high-resource languages for which we have both LLM-synthesized Silver data and human-annotated Gold portions in our MFAVA benchmark. We first observe that, expectedly, just detecting hallucinated spans (*Binary* task) is much easier than additionally correctly recognizing the type of hallucination (*Category* task). Although category labels offer finer-grained insight into the nature of LLM hallucination, we deem the models' performance on fine-grained hallucination type classification—especially on Gold, human-labeled portions of MFAVA—insufficient for reliably estimating type-specific hallucination rates "in the wild" (see §4.2). These results are in line with IAA from Figure 2, with consistently larger IAA for hallucination detection (Binary) then for type classification (Category). This renders fine-grained hallucination type classification difficult for both humans and models and warrants a broader research effort on hallucination type taxonomies as well as better hallucination type detection models. We leave this for future work.

Models' performance on the detection-only (Binary) tasks is much better across the board, but the results are much better on the Silver portions (hallucinations generated by GPT-4) of MFAVA than on the Gold (human-labeled hallucination spans). This is expected, because the hallucinated spans in our training data have also been generated by GPT-4—this means that the human-annotated Gold *mFAVA* portions introduce much more of a distribution shift w.r.t. training data than the corresponding Silver portions. At this point it is important to (re-)emphasize that we are not really interested in the absolute performance of the detection models, but rather using these detection performance estimates to produce reliable hallucination rate estimates for LLMs in the wild (§4.2).

We next observe that the 30-language multilingual model (MULTI) is consistently better than language-specific monolingual models (MONO), with gaps being particularly wide in the Category task (e.g., +30 F1 points for Arabic on the Gold MFAVA portion). Albeit smaller, the differences are also substantial in the Binary hallucination detection (e.g., +7 F1 points for Arabic and German, on respective Gold MFAVA portions). Finally, bidirectional contextualization in fine-tuning (*Bidirect*) seems to be generally more effective than fine-tuning with future-token masking (*Causal*), with Chinese performance as the only exception. This is in line with findings from other token-classification tasks (Li et al., 2023b; Dukić and Šnajder, 2024).

4 Estimating Hallucination in the Wild

We next propose a protocol for estimating hallucination rates of LLMs (for a wide range of languages) in the wild, based (1) on the number of hallucinated tokens detected by a hallucination detection (HD) model in the wild and (2) estimates of HD model's performance (precision and recall).

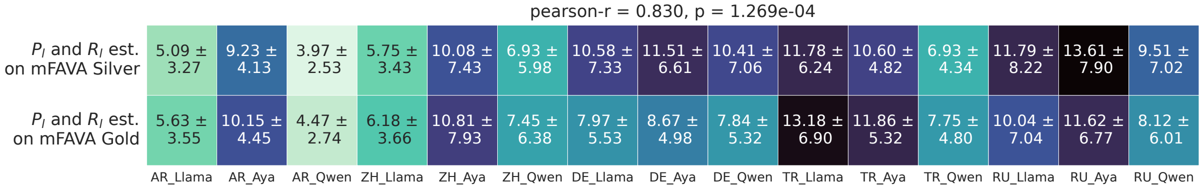


Figure 3: Comparison of hallucination rate estimates $HR_{\text{est},l}$ (mean $\pm$ std over five LLM runs) for Arabic (AR), Chinese (ZH), German (DE), Russian (RU), and Turkish (TR) for 3 LLMs based on the estimates of P_l and R_l of the MULTI (*Bidirect*) model on (1) MFAVA-Silver (top row) and (2) MFAVA-Gold (bottom row). The two sets of estimates are highly correlated $(r = 0.83, p = 1.26e-04)$.

4.1 From Model Performance to Hallucination Rates Estimates

Estimating Hallucination Rates in the Wild. Let P_l and R_l be the estimates of token-level precision and recall of a HD model for some language l and let $H_{det,l}$ be the number of hallucination tokens that the HD model detected (i.e., predicted) on some corpus C_l of LLM generations in language l, which serves as an approximation of the LLM outputs in the wild. We then posit that the estimate of the true hallucination rate of the LLM in the wild for language l, $HR_{\text{est},l}$, is given as follows:

$$HR_{\text{est},l} = \frac{P_l \cdot H_{det,l}}{R_l \cdot N_l} \times 100(\%) \tag{1}$$

where N_l is the total number of tokens in C_l, i.e., the total number of tokens generated by the LLM across answers to all user prompts. Intuitively, multiplying the number of model's detections $H_{det,l}$ with its estimated precision P_l discounts $H_{det,l}$ by the number of tokens falsely detected as hallucinated by the model—while we do not know exactly which token predictions are false positives, the expected rate of false positives is, by definition, exactly captured by P_l. Analogously, dividing $H_{det,l}$ with R_l accounts for the tokens that are hallucinated, but will (falsely) not be detected by the model—and R_l is exactly the estimate of the rate of such false negatives. We divide the estimate of the absolute number of truly hallucinated tokens (i.e., $P_l \cdot H_{det,l}/R_l$) with N_l, making $HR_{\text{est},l}$ a relative measure, that is, a rate (i.e., proportion) of all generated tokens that are hallucinated (multiplied by 100 and expressed as %). We provide a more detailed explanation/justification of Eq. (1) in §A.4.

Estimation Dataset. We next create corpora C_l (one corpus for each of our 30 target languages) of free-text LLM answers to knowledge-intensive queries, as approximations of the LLM usage in the wild. We start by randomly selecting articles from the language-specific Wikipedia, to serve as ground truth reference text. To ensure quality of reference text, we choose only from Wikipedia articles that are at least 2,000 characters long and have the collaborative Wikipedia depth (Alshahrani et al., 2023) of at least 5.[4] We then prompt GPT-4 to generate two knowledge-intensive queries for each selected article, ensuring that the information required to answer to the query is fully contained in the article text (see Table 11 in the §A.5 for the exact prompt). As a sanity check, we manually checked for 50 synthesized queries and five languages from Table 3—by translating the query and reference article to English—whether the answers to queries are indeed contained in the article, establishing that this is indeed so in 98% of cases. Our final dataset for multilingual hallucination rate estimation consists of 25,685 Wikipedia articles (spanning over 15,940 unique Wikipedia categories) and 51,133 queries. Table 9 in §A provides per-language statistics. We provide details on constructing the datasets in §A.5.

Finally, we collected responses to all queries from a total of 11 instruction-tuned open-source LLMs from 6 families (ranging in parameter count from 2 to 9 billion): Llama-3.x (Dubey et al., 2024), Aya-23 (Aryabumi et al., 2024), Euro-LLM (Martins et al., 2024), Gemma-2 (Team, 2024) Qwen-2.5 (Yang et al., 2024), and Mistral v3 (Jiang et al., 2023). We divided the queries into five subsets: for each subset the LLMs generated responses with a different random seed (see Table 8 for details on the generation configurations).

Estimates from MFAVA-Silver Performance. On the one hand, creating Gold datasets for hallucination detection evaluation is prohibitively expensive (see §A.2)—this is why we obtained such annotations for only five of 30 MFAVA languages. On the other hand, the estimates of HD model's performance are much higher on MFAVA-Silver (see

[4]The depth indicates the number of collaborative edits and correlates with the quality/factuality of the content.

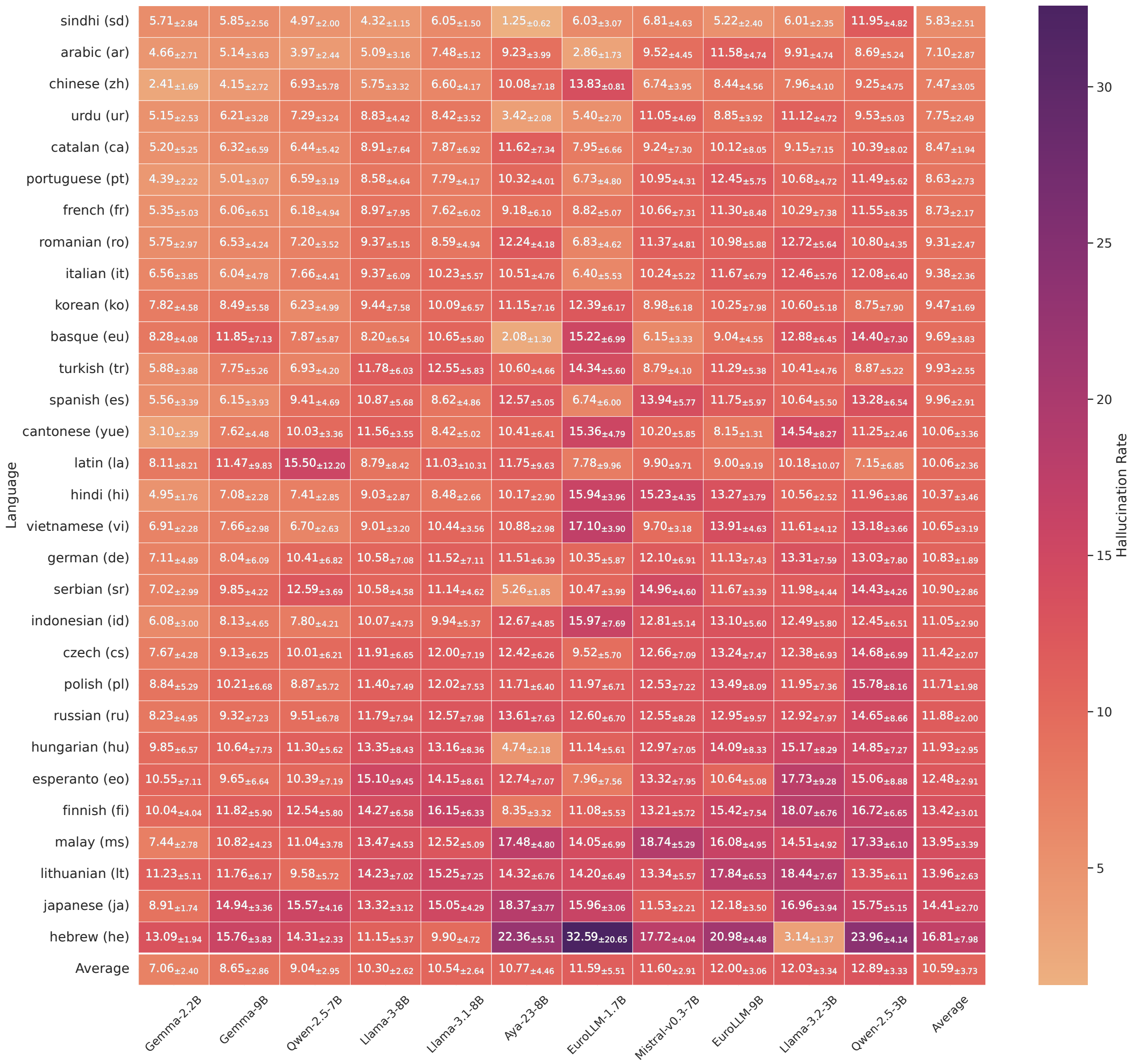
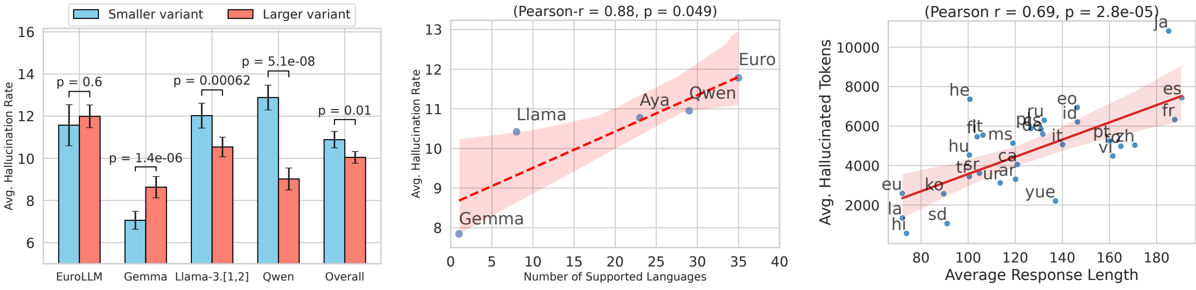


Figure 4: Mean estimates of in-the-wild hallucination rates (± std) for 30 languages and 11 LLMs. Each mean score is an average of 15 $HR_{\text{est},l}$ estimates, (3 different HD model instances applied to 5 different LLM responses). Average rates increase from top to bottom (over languages) and from left to right (over LLMs).

Table 3), with GPT-4-labeled hallucinations: this, at first glance, questions the validity of estimating 'in the wild' hallucination rates based on P_l and R_l estimated on Silver data, for the 25 languages for which we do not have MFAVA-Gold portions. Recall, however, that we do not care about HD model's absolute P_l and R_l, but whether the P_l and R_l estimates can produce reliable hallucination rate estimates $HR_{\text{est},l}$. Looking at Eq. 1, $HR_{\text{est},l}$ depends on the ratio P_l/R_l and not absolute values of P_l and R_l. We thus next test, for the five languages with both Silver and Gold portions in MFAVA, whether the $HR_{\text{est},l}$ estimates based on the Silver P_l and R_l (roughly) match those based on Gold P_l and R_l. Figure 3 shows $HR_{\text{est},l}$ estimates, computed from the performance of our MULTI (*Bidirect*) model on Silver and Gold portions, respectively, and number of its hallucination detections $H_{det,l}$ on outputs of three LLMs: Llama-3-8B, Qwen-2.5-7B, and Aya-8B. We observe very strong Pearson correlation ($r = 0.83, p = 1.26e-04$) between the Gold-based and Silver-based $HR_{\text{est},l}$ estimates, which, we argue, justifies the usage of Silver MFAVA datasets for estimating $HR_{\text{est},l}$ for the 25 languages without the Gold MFAVA portions.

4.2 Final Estimates

Figure 4 shows our in-the-wild hallucination rate estimates $HR_{\text{est},l}$ for all 30 MFAVA languages and 11 LLMs. The average rate across all languages varies between 7% and 12%, with both Gemma models offering the lowest rates. Smaller Qwen-2.5



(a) (b) (c)

Figure 5: 5a Larger models hallucinate significantly less than smaller ones. Bars are labeled with p-values from t-test. 5b Correlation between hallucination rates (averaged over all 30 languages) and the officially declared number of supported languages. 5c On average, as response length increases, so do the absolute hallucinations $H_{\text{detected},l}$.

(3B) model hallucinates the most and, interestingly, significantly more than its larger counterpart (9B).

More Parameters, Less Hallucination? For each LLM, we have 15 estimates (3 HD model instances × 5 generations by the LLM) of *HR* (averaged across all languages): we apply the Student's t-test to determine if the differences between models (smaller and larger) significantly differ. Figure 5a summarizes the results. The difference between the two EuroLLM variants is not significant; larger Gemma model hallucinates more (significantly), but the HR are low for both variants; for Llama and Qwen, the smaller models hallucinate significantly more. Finally, we aggregate the estimates across all "small" models (1.7-3B) and all "large" models (7-9B) and see that, overall (column "Overall" in Figure 5a), smaller LLMs hallucinate significantly more ($p = 0.01$). This agrees with Wei et al. (2024b) who report larger models to be more truthful in long-form answer generation.

More Languages, More Hallucination? Figure 5b compares LLMs' hallucination rates against their declared number of supported languages. here we a surprising trend see that LLMs that support more languages tend to hallucinate more (e.g., EuroLLM supports 35 languages, whereas Gemma is declared to support English only)—the correlation is strong and significant ($r = 0.88, p = 0.049$).

Say Less, Hallucinate Less? Intuitively, one would expect LLMs' hallucination rates to be larger for languages in which they are less competent (i.e., seen the least in pretraining and instruction-tuning). Surprisingly, however, we do not find this to be the case for any of the models. E.g., we observe the lowest hallucination rate for Sindhi (5.83% of tokens are hallucinated), a language with merely 18,000 Wikipedia articles and largest hallucination rate for Hebrew (16.81%). Across all 30 languages, however, we find *no* correlation between the hallucination rates and measures of language "resourceness": (i) proportion of language-specific data in Common Crawl and (ii) number of articles in the language-specific Wikipedia. As illustrated in Figure 5c, we do observe that LLMs generate longer responses for languages in which they are more competent—this entails a larger number of hallucinated tokens for longer responses, but not (necessarily) a larger (per-token) hallucination rate (recall that we account for the response length in Eq. 1). Indeed, we observe no correlation whatsoever between the response length and hallucination *rates* across languages ($r = -0.05$). This suggests that a trade-off between the answer length and the amount (not rate!) of hallucinations is a largely language-independent property of LLMs.

5 Conclusion

We presented the first effort towards understanding how much multilingual LLMs hallucinate "in the wild". To this end, we proposed a novel framework for hallucination rate estimation, which adjusts the number of detected hallucinations based on the detector's performance resulting in more reliable rate estimates. We trained a series of multilingual detection models, and measured their precision and recall scores on our newly created MFAVA datasets across 30 languages. To estimate hallucinations, we build a novel synthetic open-domain knowledge-intensive QA dataset for which we collected answers from eleven open-source LLMs. Our findings indicate that smaller models and models that cover more languages hallucinate significantly more, and that model response-length does not correlate with hallucination rate.

Limitations

We acknowledge that our method of using GPT-4 to insert synthetic hallucinations may not perfectly replicate natural model errors. However, this approach was chosen due to the immense difficulty and expense of manually curating such a dataset in 30 languages (detailed in §3.1). Crucially, our findings in Table 1 indicate that more than 50% of these synthetic hallucinations were still perceived as convincing and realistic.

We adopted the common translation-train approach and thus used MT to translate the original FAVA into our 30 target languages. While one may argue that we thus add some noise to the training process resulting in unreliable detectors, recall that we are not opting for the highest possible detection performances, but rather interested in obtaining reliable performance estimates.

We only have gold annotations for 5 languages. Here, one might argue that, thus, our performance estimates might be unreliable. This is why in §4.1, we compare estimates obtained on mFAVA-Silver with ones obtained on mFAVA-Gold and show that silver annotations can serve as a reliable proxy.

For our hallucination evaluation, we only manually check a subset of the Arabic, Chinese, German, Russian, and Turkish queries to ensure that the answers to the synthetic prompts are present in the Wikipedia references. The high rate of 98% we observed makes us confident that the potential error we introduce via such "non-grounded" questions for other languages is negligible, especially for high-resource languages. We still acknowledge, however, that the Wikipedia articles we use might be limited in terms of the knowledge they cover (Kim et al., 2024), this is why we carefully filter via minimum length and collaborative Wikipedia depth towards higher-quality articles with high coverage.

Finally, we deliberately limited the scope of this work to assessing factual correctness and we do not cover factual coverage. We decided to do so as quantifying hallucinations in long-form generation is already difficult for English Xu et al. (2023); Min et al. (2023); Wei et al. (2024b) and more so in non-english languages (Kim et al., 2024), and currently, resources for assessing multilingual factual coverage are still lacking.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. *arXiv preprint arXiv:2212.10622*.

Saied Alshahrani, Norah Alshahrani, and Jeanna Matthews. 2023. DEPTH+: An enhanced depth metric for Wikipedia corpora quality. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 175–189, Toronto, Canada. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta R. Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only llms for sequence labeling. In *Annual Meeting of the Association for Computational Linguistics*.

Benedikt Ebing and Goran Glavaš. 2024. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. 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 5325–5344.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Janek Herrlein, Chia-Chien Hung, and Goran Glavaš. 2024. Anhalten: Cross-lingual transfer for german token-level reference-free hallucination detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 92–100.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Vu Trong Kim, Michael Krumdick, Varshini Reddy, Franck Dernoncourt, and Viet Dac Lai. 2024. An analysis of multilingual FActScore. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4309–4333, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Barrett Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1703.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023b. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022a. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6723–6737.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022b. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337.

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.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Sebastian Nordhoff and Harald Hammarström. 2012. Glottolog/langdoc:increasing the visibility of grey literature for low-density languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3289–3294, Istanbul, Turkey. European Language Resources Association (ELRA).

Saad Obaid ul Islam, Iza Škrjanec, Ondrej Dusek, and Vera Demberg. 2023. Tackling hallucinations in neural chart summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 414–423.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-FAct: Assessing factuality of multilingual LLMs using FActscore. In *First Conference on Language Modeling*.

Gemma Team. 2024. Gemma.

Timm Teubner, Christoph Flath, Christof Weinhardt, Wil Aalst, and Oliver Hinz. 2023. Welcome to the era of chatgpt et al.: The prospects of large language models. *Business & Information Systems Engineering*.

Johanne R Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2703–2707.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024b. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3225–3245.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. *arXiv preprint arXiv:2310.06498*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

A Appendix

A.1 Choice of languages

Initially, we wanted to cover all 14 language families based on Glottolog 5.0 (Nordhoff and Hammarström, 2012)), however, as we progressed through the languages, we found that even the best closed-source LLMs like GPT-4 and Gemini are bad at generating text in low-resource languages (e.g. Amharic, Aymara, Hausa and Tamil) and we could not employ LLMs to generate and annotate a silver hallucination detection dataset in these languages. See Table 7 for 30 languages.

A.2 Annotation Process

We provide the FAVA seed passage generations and hallucination insertion prompts in Tables 12, 13, 14, 15, 16, 17.

Cost of Silver Annotations The total cost for generating silver data for 30 languages using GPT-4 was ∼\$2,310 with ∼\$77 per language. Distribution of categories across 30 languages is provided in Table 4 and and per language label distribution is provided in Figure 6.

	ENT	REL	INV	CON	UNV	SUB
Count	11143	9036	5649	4024	5670	6396

Table 4: Distribution of categories across 30 languages in silver set.

Gold Annotations: The annotators were sourced through prolific platform. Each annotator was screened on 10 samples and if they met the threshold of 40% agreement with the silver annotation, they were invited to participate in the full study.

It is worth noting that as the hallucination annotation task for longform QA is very cognitively demanding, it took us a long time to find annotators who could do the task correctly with high-effort. Most of the time, annotators who passed the screening test, decided to leave the study because of the high effort requirement of the task even though our study was paying above minimum wage (14 \$/hr) for the full study. Moreover, Table 2 reveals that Inter-Annotator Agreement (IAA) and Silver-Gold agreement for category annotations are both below 80%, underscoring the inherent difficulty of this task. This challenge is further reflected in our token-level agreement scores, which are impacted by minor inconsistencies in annotator decisions regarding minimal span selection.

The total cost of the gold annotations was (including platform and annotation fees) **\$4581** where each annotator was paid 14 \$/hr. All the annotators were at least bachelor's level and bilingual because in addition to understanding their own language (e.g. Arabic) they also needed to understand the task instructions and Wikipedia content (in English). The task instructions are given in Figure 8. Each annotator was asked if they consent to storing their prolific IDs during manual and automatic assessment stage. Following the assessment, their prolific IDs were deleted.

We will release MFAVA data under an open scientific licensing.

A.3 Training Details

All the classifiers were trained utilizing the Bi-LLM (Li et al., 2023b) and transformers (Wolf, 2019) library. The models were trained with three seeds (42, 47, 49) on 4xH100 until convergence. Seeds are set for *torch.manual_seed()* and *random.seed()*. The exact hyper-parameters are given in the Table 5. Total GPU hours: 1134.

Parameter	Value
Translate Train-Val Split	70:30
Seeds	[42, 47, 49]
Quantization	4-bit BF16
Model	Llama-3-8B (base)
GPUs	4× H100
LoRA r	32
LoRA α	32
LoRA Dropout	0.05
LoRA Target Modules	All
Epochs	∼2 (until convergence)
Input Length	4096
Learning Rate	1×10^{-4}
Weight Decay	0.01
Batch Size	8
Gradient Accumulation	8

Table 5: Training Details

A.4 Adjusting for P_l and R_l

The hallucination rate $HR_{est,l}$ for a given language l, is defined as the ratio of hallucinated tokens detected by the model ($H_{\text{detected},l}$) to the total number of generated tokens (N_l):

$$HR_l = \frac{H_{\text{det},l}}{N_l}\,. \tag{2}$$

To refine this rate, we adjust for the detection model's precision (P_l) and recall (R_l). Precision is defined as:

$$P_l = \frac{TP_l}{TP_l + FP_l} \tag{3}$$

where (TP_l) FP_l denote true and false positives respectively. Rearranging this equation gives the number of true positives:

$$TP_l = P_l \cdot HR_{\text{det},l} \tag{4}$$

Recall is defined as:

$$R_l = \frac{TP_l}{TP_l + FN_l} \tag{5}$$

where FN_l denotes false negatives. The total number of corrected hallucinations ($HR_{\text{est},l}$) can thus be expressed as:

$$HR_{\text{est},l} = TP_l + FN_l = \frac{TP_l}{R_l} \tag{6}$$

Substituting Equations 4 in 6, we derive the $H_{\text{est},l}$ as:

$$H_{\text{est},l} = \frac{P_l \cdot H_{\text{det},l}}{R_l} \tag{7}$$

By incorporating the model's P_l and R_l, our estimation framework effectively corrects for the imperfections of a hallucination detector. When estimating the hallucination rate HR_l on a large corpus (see §4), EQ 7 provides a reliable measure of the true number of hallucinations. This accounts for the detector erroneously flagging $1 - P_l\%$ of its identified instances and failing to capture $1 - R_l\%$ of genuine hallucinations.

A.5 Hallucination Evaluation Dataset

To construct the hallucination evaluation dataset, we aimed to scrap $\sim$ 1000 articles per language with more than 2000 characters. However, problem with non-English languages (especially moderate-low resource) is that $\geq$ 2000 character articles can be scarce. Furthermore, sometimes Wikipedia has articles tagged as *Unreferenced, Failed Verification,* or *Under Construction* which flag the article as unfinished or not factually verified. Such tags are very prominent in languages other than English and we do not include such articles in our dataset.

We use Wikipedia article summary (text before the first heading) as references and prompt gpt-4 to generate 2 knowledge-intensive queries per article. Sometimes, it generated only one query even though we explicitly state to generate two queries. We did not prompt the GPT-4 again to generate the second query due to budget constraints. The total cost to generate prompts for 31 languages is $192. Per language statistics can be found in Table 9. We will release hallucination evaluation data under an open scientific licensing.

Given the following reference in language <*language name*>: Generate **two knowledge-intensive queries** in `<language name>`. Ensure the questions are concise but knowledge-intensive. The questions should require thorough reading of the reference text to answer. Separate the questions with a newline.

Table 6: Prompt for generating knowledge-intensive queries.

A.6 Response Collection for Hallucination Evaluation Dataset

We collect LLM responses on 5 seeds: 42, 43, 44, 47, 49 for 6 LLM model[5]. The generation configurations that we used are provided in the *generation_config.json* in model repositories on huggingface. Seeds are set for *torch.manual_seed()* and *random.seed()*.

A.7 Manual Analysis

To further check the quality and informativeness (See §A.8 for definitions of informativeness) of the responses, we manually analyze 60 responses from Aya-23-8B for German and Arabic, two languages for which we have gold annotations. Overall, 10% of the responses had repetitive words and sentences and 5% of the responses were *I don't know* responses. For the remainder of the samples, the responses were fluent and long and were relevant to the input prompt.

A.8 Informativeness

Currently, there is no agreed-upon definition of informativeness. Lin et al. (2022) considers a response to be informative if it is potentially relevant to the question and Wei et al. (2024b) considers a response to be informative if it has a certain number of supporting facts from the reference text.

[5]We comply with licensing agreement for each of the LLMs we use.

Language	Language Family	Script	Test-Set
Arabic	Afro-Asiatic (Semitic)	Arabic	Gold
Chinese	Sino-Tibetan (Sinitic)	Chinese (Han)	Gold
German	Indo-European (Germanic)	Latin	Gold
Russian	Indo-European (Slavic)	Cyrillic	Gold
Turkish	Turkic (Common Turkic)	Latin	Gold
Basque	Language Isolate	Latin	Silver
Cantonese	Sino-Tibetan (Sinitic)	Chinese (Han)	Silver
Catalan	Indo-European (Romance)	Latin	Silver
Czech	Indo-European (Slavic)	Latin	Silver
Esperanto	Constructed	Latin	Silver
Finnish	Uralic (Finnic)	Latin	Silver
French	Indo-European (Romance)	Latin	Silver
Hebrew	Afro-Asiatic (Semitic)	Hebrew	Silver
Hindi	Indo-Aryan	Devanagari	Silver
Hungarian	Uralic (Ugric)	Latin	Silver
Indonesian	Austronesian (Malayo-Polynesian)	Latin	Silver
Italian	Indo-European (Romance)	Latin	Silver
Japanese	Japonic	Kanji	Silver
Korean	Koreanic	Hangul	Silver
Latin	Indo-European (Italic)	Latin	Silver
Lithuanian	Indo-European (Slavic)	Latin	Silver
Malay	Austronesian (Malayo-Polynesian)	Latin	Silver
Polish	Indo-European (Slavic)	Latin	Silver
Portuguese	Indo-European (Romance)	Latin	Silver
Romanian	Indo-European (Romance)	Latin	Silver
Serbian	Indo-European (Slavic)	Cyrillic	Silver
Sindhi	Indo-Aryan	Arabic	Silver
Spanish	Indo-European (Romance)	Latin	Silver
Urdu	Indo-Aryan	Arabic	Silver
Vietnamese	Austroasiatic (Vietic)	Latin	Silver

Table 7: Classification of languages by language family (based on Glottolog 5.0), script, and test-set status. Gold test sets are available for 5 languages, while the rest have silver test sets.

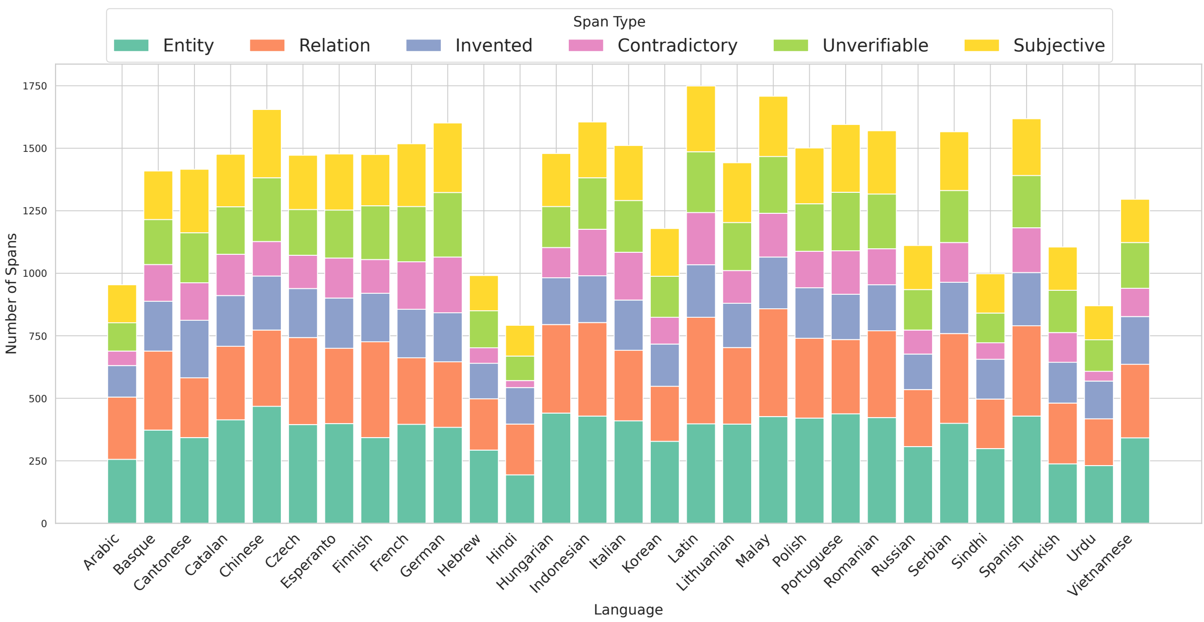


Figure 6: Distribution of 6 labels across 30 languages in MFAVA-SILVER dataset.

Model	max_new_tokens	temperature	top_p	top_k	repetition_penalty	do_sample
Llama-3.x	1024	0.6	0.9	–	–	True
Aya	1024	–	0.3	–	–	True
Qwen-2.5	1024	0.7	0.9	20	1.05	True
Mistral	1024	–	–	50	–	True
Gemma-2	1024	–	–	–	–	True
EuroLLM	1024	–	–	–	–	True

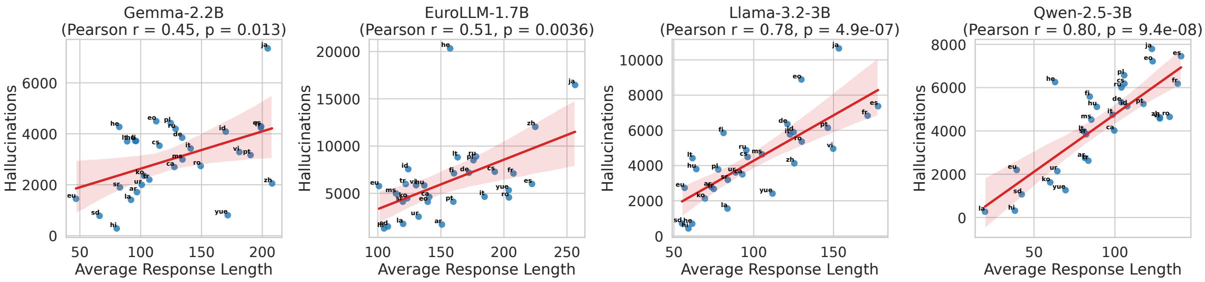
Table 8: Huggingface MODEL.GENERATE() parameters for each model family. – indicate default is used. Generation configurations are provided in model's respective HuggingFace (Wolf, 2019) repositories

Language	Unique Categories	Total Articles	Total Queries
Arabic	537	959	1907
Basque	486	938	1872
Cantonese	261	401	793
Catalan	359	989	1976
Chinese	712	977	1939
Czech	720	988	1975
Esperanto	608	956	1912
French	332	987	1973
Finnish	549	995	1972
German	797	984	1967
Hebrew	660	999	1991
Hindi	153	186	367
Hungarian	745	992	1964
Indonesian	457	958	1913
Italian	678	988	1974
Japanese	667	999	1991
Korean	539	747	1488
Latin	334	465	916
Lithuanian	711	946	1888
Malay	442	778	1556
Polish	889	1000	1998
Portuguese	390	955	1909
Romanian	351	811	1618
Russian	462	999	1996
Spanish	938	977	1952
Serbian	386	798	1587
Sindhi	224	519	1029
Turkish	660	856	1650
Urdu	567	878	1749
Vietnamese	326	660	1311
Total	**15,940**	**25,685**	**51,133**

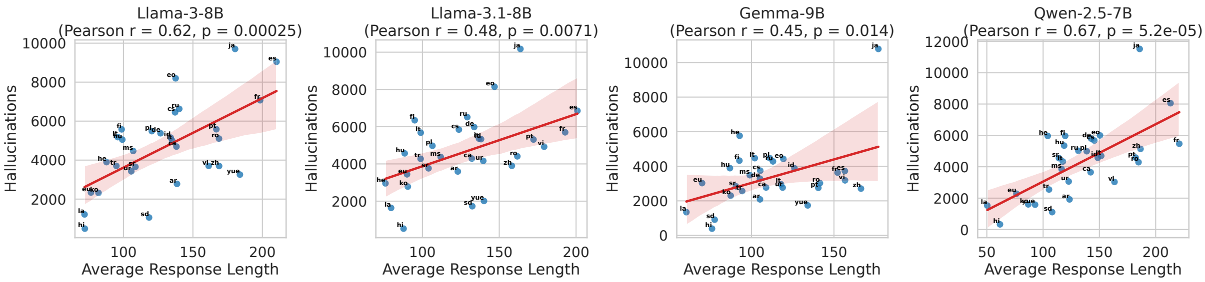
Table 9: Per language statistics for hallucination evaluation dataset.

Language	Precision (%)	Recall (%)	F1 Score (%)
GOLD			
Arabic (Gold)	73.98	53.40	61.63
Chinese (Gold)	70.73	53.93	58.79
German (Gold)	58.19	74.06	65.05
Turkish (Gold)	79.67	66.95	72.57
Russian (Gold)	63.18	68.46	65.53
Average	69.15	63.36	64.71
SILVER			
Arabic	93.28	74.81	82.59
Chinese	80.33	66.28	69.77
German	91.64	87.77	89.50
Turkish	89.58	83.92	86.43
Russian	93.05	86.04	89.15
Basque	87.22	74.46	79.80
Cantonese	78.49	49.40	56.12
Catalan	94.70	87.46	90.85
Czech	93.99	84.75	89.00
Esperanto	94.28	86.53	90.05
French	91.58	89.37	90.31
Finnish	86.67	84.26	85.15
Hebrew	82.75	32.97	44.19
Hindi	68.01	68.48	66.77
Hungarian	92.35	74.29	81.93
Indonesian	92.12	85.75	88.72
Italian	93.76	87.26	90.28
Korean	86.39	79.11	82.31
Japanese	77.06	61.03	67.15
Lithuanian	90.48	75.39	81.81
Malay	86.15	68.96	75.73
Portuguese	95.80	86.77	90.94
Serbian	86.16	76.75	79.91
Sindhi	82.00	69.38	74.36
Spanish	95.86	85.34	90.14
Vietnamese	89.35	84.57	86.71
Urdu	88.82	72.32	79.39
Average	88.22	76.42	80.71

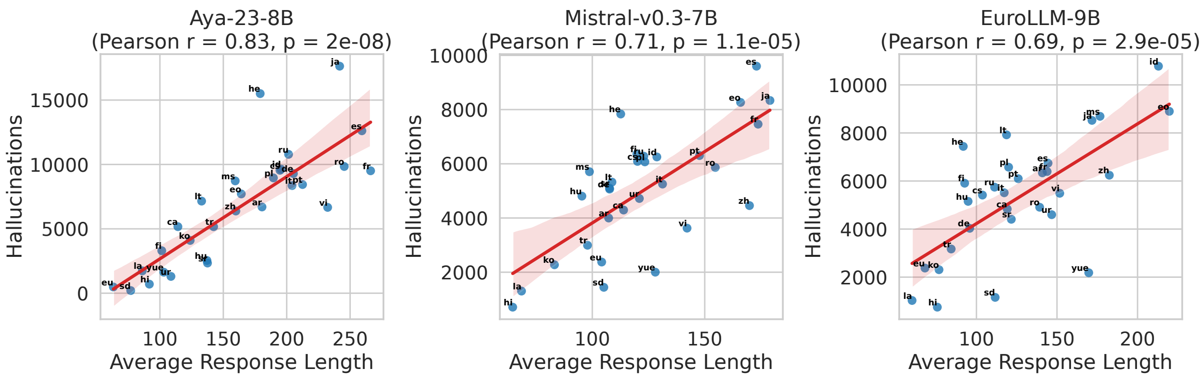
Table 10: Precision, Recall, and F1 scores for all languages, including GOLD scores for five languages.



(a) Hallucinations vs response length correlation of smaller models.



(b) Hallucinations vs response length correlation of bigger models.



(c) Hallucinations vs response length correlation of bigger models.

Figure 7: Per model correlations between hallucinations and response length.

READ THE INSTRUCTION CAREFULLY.

A hallucination in natural language generation is text generated by a large language model (ChatGPT) that is not grounded in a reference text or is nonsensical.

Read the passage carefully and label the hallucinatory text using the following hallucination taxonomy. You SHOULD use the reference Wikipedia article in English below. DO NOT MODIFY THE EXISTING TEXT. ONLY ADD ANNOTATION SPANS. to verify the information present in the text.

1. **Entity:** Refers to errors where a specific named entity (e.g., location, person) in a statement is incorrect, but changing that entity makes the sentence factually correct.
   Example: "The Eiffel Tower is in Berlin" should be labeled as an entity error because replacing "Berlin" with "Paris" corrects the error.
2. **Relation**: Involves errors where a semantic relationship in a statement is incorrect, affecting the factual accuracy of the relationship described.
   Example: "Barack Obama is the wife of Michelle Obama" should be labeled as a relation error because the relationship "wife" is incorrect and should be "husband."
3. **Contradictory**: This applies to statements that completely contradict verifiable evidence or facts.
   Example: "The sun rises in the west" is a contradictory error because it goes against the well-known fact that the sun rises in the east.
4. **Invented**: Related to statements that include concepts, entities, or events that do not exist or are entirely fabricated.
   Example: "The annual conference of dragons" should be labeled as invented because dragons do not exist in reality.
5. **Subjective:** Relates to expressions or propositions that reflect personal beliefs or opinions rather than universal facts and cannot be universally validated as true or false.
   Example: "Chocolate is the best flavor" is subjective because it is based on personal preference.
6. **Unverifiable:** Concerns statements that, although fact-based, cannot be confirmed or denied with the provided reference.
   Example: "There is an undiscovered painting by Van Gogh in my basement" is unverifiable as it cannot be easily proven or disproven without further evidence.

Annotation Guidelines:

Step 1: Review the Passage

- Read the provided text carefully to understand the context and the information presented.

Step 2: Label the text given the wikipedia reference

- You should label the text with the above-provided taxonomy.
- Entity and Relation hallucination usually consists of one word or two. This could be a numeric as well.
- Invented, subjective, and unverifiable hallucinations can be a whole statement or just one word or two.
- Contradictory can be one word or statement(s).

To label the text you should enclose it in an angle bracket like **<entity>some-text</entity>**. Always follow a closing tag with the opening tag. Do not use any other bracket types or labels.

Rating Task:

Rate the task between Very Unlikely to Very Likely.
Here, you should **only consider the text you labeled** as a hallucination.

Figure 8: Annotation Instructions.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English.
Given a passage answer the question in 3-5 sentences using only the information presented in the reference passage.
Question: [QUESTION_TEXT]
Reference Passage: [REFERENCE_CONTENT]

Table 11: Prompt for seed passage generation.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English.
Given a passage with possibly already inserted error tokens wrapped in <relation>, <contradictory>, <unverifiable>, <subjective>, or <invented>, insert entity errors in the passage below, wrapped in tokens to make the passage factually incorrect. Ensure these insertions are outside these existing <> tags, and don't modify the <> tags at all. The error is defined as such:
1. entity errors (<entity>): a small part of a sentence, often an entity (e.g., location name), is incorrect (usually 1-3 words). Entity errors often involve noun phrases or nouns.
Example 1: Messi is an <entity></entity> soccer player.
Example 2: Selena Gomez was born on <entity></entity> 22.
Example 3: India's population is <entity></entity> billion people.
Now, insert entity error tokens in the given passage but make sure that you don't modify anything inside any already existing <> error tokens, only add entity errors with <entity></entity> tokens outside the already existing <> tags.
##
Paragraph: [PASSAGE_CONTENT]
Edited:

Table 12: FAVA Prompt for entity hallucination insertion.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English.
Given a passage with possibly already inserted error tokens wrapped in <entity>, <contradictory>, <unverifiable>, <subjective>, or <invented>, insert relation errors, outside the already inserted tokens without modifying the content within already existing tokens. Wrap the relational errors in tokens to make the passage factually incorrect. The error is defined as such:
1. relational error (<relation>): a sentence is partially incorrect as a small part (usually 1 - 3 words). Relational errors often involve verbs and are often the opposite of what it should be.
Example 1: FDA <relation></relation> pfizer COVID-19 Vaccine.
Example 2: Rishi Sunak <relation></relation> his role as Prime Minister in 2022.
Example 3: Millie Bobbie Brown has also starred in several popular movies, including "Godzilla vs. Kong" and "Enola Holmes" which she also <relation></relation>.
Now, insert relation error tokens in the given passage but make sure that you don't modify anything inside any already existing <> error tokens, only add relational errors with <relation></relation> tokens outside the already existing <> tags. ##
Paragraph: [PASSAGE_CONTENT]
Edited:

Table 13: FAVA prompt for relation hallucination insertion.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English. Given a reference and a passage with possibly already inserted error tokens wrapped in <entity>, <relation>, <unverifiable>, <subjective>, or <invented>, insert contradictory sentence errors in the passage outside the already inserted tokens without modifying the content within already existing tokens. Wrap the inserted errors in tokens to make the passage factually incorrect.
The contradictory error is defined as such:
1. contradictory sentence error (<contradictory>): a sentence where the entire sentence is contradicted by the given reference, meaning the sentence can be proven false due to a contradiction with information in the reference provided.
##
Example 1:
Reference: Japan participated in World War I from 1914 to 1918 in an alliance with Entente Powers (France, the United Kingdom, Russia, the United States, Italy) against the Central Powers (Germany, Austria-Hungary, the Ottoman Empire, and Bulgaria).
Contradictory Sentence: <contradictory>Japan sent its army to help Germany during World War I.</contradictory>
Explanation: The reference states that Japan was in an alliance against Germany, so Japan would not send its army to help Germany like the sentence states.
##
Example 2:
Reference: Percy Jackson & the Olympians is a series of five fantasy novels written by American author Rick Riordan.
Contradictory Sentence: The Harry Potter series was written by J.K Rowling<contradictory>, as was the Percy Jackson series</contradictory>.
Explanation: The reference states that the Percy Jackson series is written by Rick Riordan and not J.K Rowling like the sentence suggests.
##
Example 3:
Reference: As one of the busiest women in music, it'll come as no surprise that Taylor has won pretty much every award there is to win in the biz - being the proud owner of no less than 12 Grammy Awards.
Contradictory Sentence:<contradictory>Taylor Swift has never won a Grammy in her entire career since she is better known as a performer than a singer.</contradictory>
Explanation: The reference states that Taylor Swift has won 12 Grammys and is a musician while the sentence says she has won no Grammys which contradicts the reference.
Now, insert contradictory sentences with tokens in the given passage but make sure that you don't modify anything inside any already existing <> error tokens at all, keep those untouched, only insert new contradictory sentences (entire sentences) with <contradictory></contradictory> tokens outside the already existing <> tags in the passage. ##
Reference: [REFERENCE_CONTENT]
Passage: [PASSAGE_CONTENT]
Edited:

Table 14: FAVA prompt for contradictory hallucination insertion.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English. Given a subject and a passage with possibly already inserted error tokens wrapped in <entity>, <relation>, <contradictory>, <unverifiable>, or <invented>, insert subjective sentence errors outside the already inserted tokens without modifying the content within already existing tokens. Wrap the insertions in tokens to make the passage factually incorrect. The error is defined as such:
1. subjective sentence (<subjective>): an entire sentence or phrase that is subjective and cannot be verified, so it should not be included.
Example 1: <subjective>He is the greatest soccer player ever.</subjective>
Example 2: The first Harry Potter book was published in 1998 <subjective>and was a lot better than the rest in the series because of its use of a rich and evocative vocabulary.</subjective>
Example 3: <subjective>Overall, Aenir is a thrilling adventure novel that takes readers on a journey through a unique and imaginative world, filling their lives with excitement.</subjective>
Now, insert subjective sentence error tokens in the given passage but make sure that you don't modify anything inside any already existing <> error tokens at all, keep those untouched, only insert full subjective sentences or phrases with <subjective></subjective> tokens about the given subject outside the already existing <> tags in the given passage. ##
Subject: [SUBJECT_NAME]
Passage: [PASSAGE_CONTENT]
Edited:

Table 15: FAVA prompt for subjective hallucination insertion.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English. Given a reference and a passage with possibly already inserted error tokens wrapped in <entity>, <relation>, <contradictory>, <subjective>, or <invented>, insert unverifiable errors outside the already inserted tokens without modifying the content within already existing tokens. Wrap the insertions in tokens to make the passage factually incorrect. The error is defined as such: 1. unverifiable sentence (<unverifiable>): a sentence where the whole sentence or phrase is unlikely to be factually grounded although it can be true, and the sentence cannot be confirmed nor denied using the reference given or internet search, it is often something personal or private and hence cannot be confirmed. ## Unverifiable Error Example 1: <unverifiable>Apple is planning on releasing an instrument collection.</unverifiable> Explanation: Information about Apple's release plans cannot be corroborated by any information online, however could be true. ## Unverifiable Error Example 2: <unverifiable>Selena Gomez is known to love turtles.</unverifiable> Explanation: Personal information about Selena Gomez's opinion on turtle's cannot be verified online, however could be true. ## Unverifiable Error Example 3: <unverifiable>Tom Cruise wanted to act in a Bollywood film.</unverifiable> Explanation: Personal information about Tom Cruise's preference on acting in a Bollywood film could be true but cannot be found online. Now, insert unverifiable error tokens in the given passage but make sure that you don't modify anything inside any already existing <> error tokens at all, keep those untouched, only insert unverifiable sentences or phrases with <unverifiable></unverifiable> tokens outside the already existing <> tags in the given passage. Remember, unverifiable sentences seem like they are true but cannot be confirmed or denied. ## Reference: [REFERENCE_CONTENT] Passage: [PASSAGE_CONTENT] Edited:

Table 16: FAVA prompt for unverifiable hallucination insertion.

For the following instruction, generate the whole output in {lang} language even though the instruction and input is in English. Given a subject and a passage with possibly already inserted error tokens wrapped in <entity>, <relation>, <contradictory>, <unverifiable>, or <subjective>, insert invented info sentence errors outside the already inserted tokens without modifying the content within already existing tokens. Wrap the errors in tokens to make the passage factually incorrect. The error is defined as such: 1. invented info error (< invented >): these errors refer to entities that are not known or do not exist. this does not include fictional characters in books or movies. invented info errors include phrases or sentences which have unknown entities or misleading information. ## Invented Info Example 1: <invented>Kansas City has a large population of the Yuman Tribe.</invented> Explanation: Yuman tribe is not an actual tribe, they are a invented entity. ## Invented Info Example 2: Joel Embiid is a Cameroonian professional basketball player for the Philadelphia 76ers , and he was awarded the Kia NBA MVP Trophy in 2023 <invented>and received the Shaquille O'Neal trophy for being the fastest runner this season.</invented> Explanation: There is no trophy named the Shaquille O'Neal trophy in NBA and so it is not possible for Joel Embiid to have won it. Also, there is no award for being the fastest runner in the NBA. Both are invented. ## Invented Info Example 3: <invented>Andrew Ng's area of Sentiment-Infused Language Generation (SILG) which explores the influence of sentiment analysis on the generation of human-like language.</invented> Explanation: There is no field of research named Semantic compositionality analysis, so it is a invented entity. Now, insert invented information error tokens about the subject in the given passage but make sure that you don't modify anything inside any already existing <> error tokens at all, keep those untouched, only add fictional sentence errors with <invented></invented> tokens outside the already existing <> tags in the given passage. Also avoid inserting errors before the first sentence. Also make sure you tag each edit with <invented></invented> tags. ## Subject: [SUBJECT_NAME] Passage: [PASSAGE_CONTENT] Edited:

Table 17: FAVA prompt for invented hallucination insertion.