

# FIBER: A Multilingual Evaluation Resource for Factual Inference Bias

Evren Ayberk Munis<sup>1</sup>, Deniz Yılmaz<sup>2</sup>, Arianna Muti<sup>3</sup>, Çağrı Toraman<sup>2</sup>

<sup>1</sup>Politecnico di Torino, Italy

<sup>2</sup>Middle East Technical University, Computer Engineering Department, Turkey

<sup>3</sup>Bocconi University, Italy

<sup>1</sup>evrenayberk.munis@studenti.polito.it

<sup>2</sup>deniz.yilmaz\_12@metu.edu.tr, ctoraman@metu.edu.tr

<sup>3</sup>arianna.muti@unibocconi.it

## Abstract

Large language models are widely used across domains, yet there are concerns about their factual reliability and biases. Factual knowledge probing offers a systematic means to evaluate these aspects. Most existing benchmarks focus on single-entity facts and monolingual data. We therefore present FIBER, a multilingual benchmark for evaluating factual knowledge in single- and multi-entity settings. The dataset includes sentence completion, question-answering, and object-count prediction tasks in English, Italian, and Turkish. Using FIBER, we examine whether the prompt language induces inference bias in entity selection and how large language models perform on multi-entity versus single-entity questions. The results indicate that the language of the prompt can influence the model's generated output, particularly for entities associated with the country corresponding to that language. However, this effect varies across different topics such that 31% of the topics exhibit factual inference bias score greater than 0.5. Moreover, the level of bias differs across languages such that Turkish prompts show higher bias compared to Italian in 83% of the topics, suggesting a language-dependent pattern. Our findings also show that models face greater difficulty when handling multi-entity questions than the single-entity questions. Model performance differs across both languages and model sizes. The highest mean average precision is achieved in English, while Turkish and Italian lead to noticeably lower scores. Larger models, including Llama-3.1-8B and Qwen-2.5-7B, show consistently better performance than smaller 3B–4B models.

**Keywords:** factual knowledge probing, hallucinations, multilinguality, inference bias

## 1. Introduction

Despite the widespread adoption of large language models (LLMs), they often exhibit unreliability and a tendency to generate false or fabricated information, which is known as LLM hallucination (Huang et al., 2024). Consequently, assessing their reliability has become a crucial research objective. In this context, Factual Knowledge Probing plays a fundamental role, serving as a systematic method to evaluate whether models accurately store and retrieve factual information without bias.

One of the notable biases observed in LLMs is their tendency to favor the language of the prompt when generating responses. This phenomenon is referred as inference bias (Kim and Kim, 2025; Li et al., 2024a)). Inference bias is measured by analyzing how frequently the model generated words or entities related to the region associated with the prompt language, using human or model-based evaluators. In this study, we extend this concept to the factual knowledge probing setting by introducing a new term, **Factual Inference Bias**. While inference bias focuses on the frequency of language-aligned responses in generated text, Factual Inference Bias examines the same phenomena by asking factual questions and evaluating by the prob-

ability ranking of ground truth tokens. Specifically, it captures cases where the model assigns higher likelihoods to factual entities that are geographically aligned with the prompt language. For instance, while the official languages of Canada are English and French, the model assigns higher probabilities to answers associated with the prompt language when the input is provided in Italian or Turkish as shown in Figure 1.

Kim and Kim (2025)'s analysis is also limited to question-answering tasks with single-entity answers. However, in real-world scenarios, a single subject can be linked to multiple objects, leading to multi-entity answers (e.g. official languages of countries). Moreover, the degree of bias may vary depending on the prompt type, such as question-answering versus sentence completion. Existing factual knowledge probing benchmarks (Petroni et al., 2019; Elazar et al., 2021; Kwiatkowski et al., 2019; Lin et al., 2021) mainly focus on single-entity questions, while multi-entity datasets are rare and often monolingual. For instance, MyridLAMA (Zhao et al., 2024) explores various prompt formats; however, it focuses on different structural variations of subjects and answers rather than having subjects that have one-to-many relationships. Therefore, there is still no multilingual dataset that combines

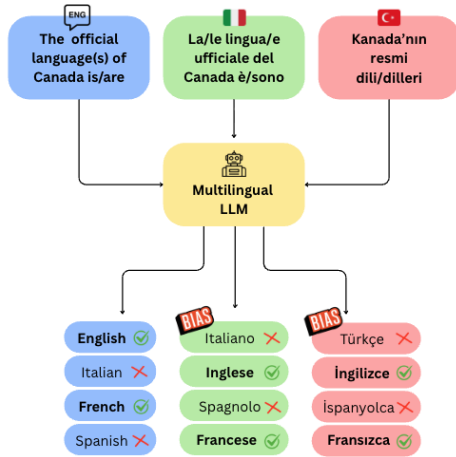


Figure 1: Multilingual prompts are provided to a multilingual language model, and candidate answers are ranked in descending order according to their cumulative log-probability scores. In the figure, gold answers are marked with a green check, incorrect predictions with a red cross, and inference-biased answers with a bias label indicating cases where the model’s prediction is influenced by the language of the prompt rather than by factual correctness.

both single- and multi-entity answers across comparable topics, making it difficult to evaluate and compare model behavior under such conditions. To address this gap, we investigate the following research questions:

**RQ1** Does the language of the prompt induce Factual Inference Bias in LLMs in different topics, such that models tend to favor entities associated with the country or region where the prompt language is mostly spoken?

**RQ2** Given that most existing benchmarks are limited to datasets containing only single-entity answers, how do LLMs perform when faced with multi-entity questions?

Factual Knowledge Probing is an experimental framework designed to evaluate how effectively a model recalls and utilizes real-world factual information when answering queries. To assess factuality in LLMs, researchers commonly compare performance across two task types: open-ended generation Youssef et al. (2023), such as sentence completion or continuation prompts, and knowledge-seeking question-answering (QA) tasks Youssef et al. (2023). The performance of a model may vary significantly between these formats and across different languages. Therefore, evaluating both task types and multiple linguistic settings is important for obtaining a comprehensive understanding of a model’s overall factual knowledge capabilities.

We examine these two research questions using a new multilingual dataset called FIBER<sup>1</sup>, which stands for Factual Inference Bias Evaluation Resource. FIBER includes both question-answering and sentence completion tasks for single- and multi-entity answers in English, Italian, and Turkish. As shown in Figure 1, multilingual prompts are provided to the model, and the log probabilities of a predefined surface set—containing all possible answers across subjects—are computed. Instead of generating new tokens, we analyze the model’s assigned probabilities to these candidate answers. The ranked list of probabilities is then evaluated using the Average Precision (AP)(Zhu, 2004) metric, which measures how well the model ranks the ground-truth answers among all possible candidates. This ranking also enables the analysis of Factual Inference Bias, by observing whether the model systematically favors entities related to the region of language of the prompt.

In this work, we introduce a new multilingual dataset that includes English, Italian, and Turkish prompts in both sentence completion and question-answering formats. The dataset contains single-entity and multi-entity answers, allowing us to examine a new type of factual inference bias. We share our evaluation scripts and the dataset online<sup>1</sup>. Our findings show that language contributes to factual inference bias; however, this effect depends on the topic. Questions related to language-specific topics have a higher tendency to trigger factual inference bias. In addition, models perform better on questions with single-entity answers than on those requiring multiple entities. Language also plays an important role in overall performance, with English prompting higher accuracy compared to Italian and Turkish.

## 2. Related Work

### 2.1. Datasets for Probing Factual Knowledge

In the question-answering (QA) dataset format, the model answers direct questions about factual details of a topic. The most fundamental QA datasets are Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). Natural Questions contains real search queries from Google, and it has two types of answers: one-word or two-word short answers, and a short paragraph that explains the short answer. On the other hand, TriviaQA is a reading comprehension dataset that has triplets of question, answer, and evidence.

Other datasets focus on the cloze-style format

<sup>1</sup>The code and dataset to replicate the experiments can be found at <https://github.com/metunlp/fiber>

with sentence completion tasks. Language Model Analysis (LAMA) (Petroni et al., 2019) utilizes a sentence completion dataset from various data resources to assess the extent to which knowledge is preserved in the model. The data consists of subject, relation, and object triplets, and the model's goal is to predict the object from incomplete sentences. ParaREL (Elazar et al., 2021) employs a similar subject-relation-object structure while testing its robustness by expressing each relation in several paraphrased variations. WIKI-UNI (Cao et al., 2021) maintains a similar structure to previous datasets. These datasets are used in various studies related to LLM hallucination. For instance, LAMA is employed to signify rank-based factual knowledge probing in the LAMA paper, while ParaREL is utilized for mechanistic factual knowledge analysis (Yu et al., 2024).

Existing studies offer limited resources for factual knowledge probing with multi-entity answers. MultiSpanQA (Li et al., 2022) features questions with answers spanning multiple discontinuous text segments. In contrast, MyriadLAMA (Zhao et al., 2024) broadens factual probing by varying subjects and objects, and incorporating both cloze and question–answering formats.

## 2.2. Multilinguality

In multilingual and cross-lingual settings, hallucination becomes even more problematic. Qiu et al. (2023) show that multilingual LLMs hallucinate more frequently in non-English languages. This finding is also supported by Chataigner et al. (2024), who show how hallucination rates differ by language in free-form generation: they generate biographies across 19 languages and compare to Wikipedia pages, finding systematic variation in hallucination frequencies especially between high- and low-resource languages. Several datasets focus on evaluating text generation models within a multilingual framework. Some of these datasets include XQUAD (Artetxe et al., 2020), a fundamental benchmark for multilingual reading comprehension and knowledge extraction based on Wikipedia passages with questions related to the passage; MLQA (Lewis et al., 2020), which has a similar structure to XQUAD, but it has seven languages and contains real passages from Wikipedia rather than translations; MPARAREL (Fierro and Søgaard, 2022) and Multilingual LAMA (Kassner et al., 2021), two examples of multilingual cloze-style datasets for knowledge probing, which are the direct translation of the PARAREL and LAMA datasets into various languages. To the best of our knowledge, limited work explore hallucinations in Turkish and Italian. For Turkish, Turk-LettuceDetect (Taş et al., 2025) proposes token-level hallucination detection models in retrieval-augmented generation (RAG) settings, trained on a Turkish-translated version

of the RAGTruth benchmark (Taş et al., 2025). In the multilingual benchmark space, HalluVerse25 (Abdaljalil et al., 2025) includes a Turkish portion, allowing fine-grained categorization of hallucination types (entity, relation, sentence) and comparisons of model performance on Turkish vs. English and Arabic. In Italian, the only work addressing hallucinations applies a combined detection and mitigation framework in the Italian healthcare domain Priola (2025). Moreover, Muti et al. (2025) note that LLMs occasionally generate lexical hallucinations (invented or distorted words) in sentence completion when describing feminists.

## 2.3. Factual Probing Evaluation Methods

Recent research has extended factual evaluation beyond traditional static probing, incorporating retrieval-augmented and verification-based approaches that assess whether model-generated outputs are grounded in accessible evidence. Evaluation methods in this domain commonly include a variety of quantitative and qualitative metrics, such as ROUGE-L, precision@1 and precision@k metrics. Together, these metrics provide a more comprehensive assessment of both the factual reliability and robustness of large language models. In multilingual settings, knowledge-grounded metrics such as mFACT and retrieval-enhanced datasets like HaluEval 2.0 demonstrates that integrating retrieval and claim-verification modules substantially improve factual probing across languages (Qiu et al., 2023; Li et al., 2024b).

## 2.4. Inference Bias

Kim and Kim (2025) explore this phenomenon by evaluating inference bias within factual knowledge probing tasks. They posed questions such as “What is your country’s name?” across multiple languages and analyze how frequently the models’ answers align with either the dominant language of their training data or the query language. Their analysis reveal that LLMs tend to adapt their responses to the language of the query, indicating that the linguistic context substantially influences the retrieval of facts.

In a related line of research, Li et al. (2024a) examined geopolitical bias, defined as the inconsistency in how LLMs recall geographical or political knowledge when queried in different languages. They ask questions concerning territorial disputes—for example, “Is Ceuta a territory of Spain or Morocco?”—in the native languages of the countries referenced. Their findings show that multilingual models exhibit marked geopolitical bias and encode geographical knowledge unevenly across languages.

### 3. Dataset

To construct a benchmark that provides broad insights into the Factual Inference Bias behavior of the LLMs, we divide the dataset into two main classes: multi-entity and single-entity. The single-entity data represent one-to-one relationships between subjects and entities, while the multi-entity data represent one-to-many relationships among subjects and entities, containing multiple valid answers per query. The subject refers to the source entity that holds a specific relationship with one or more target entities. For example, in the language relation, “Canada” serves as the subject for the entities “English” and “French.”

The dataset consists of data in three languages: English, Italian, and Turkish. Single-entity data are taken from Elazar et al. (2021) for English and CoAStAL NLP Group (2025) for Turkish, after manually correcting grammatical inconsistencies<sup>2</sup>. The remaining queries are manually constructed following the ParaREL format. As the sources are mainly in English, we translate subjects and entities into Italian and Turkish with Gemini and verify them with native speakers by following the guidance reported in Umutlu et al. (2025). For each topic in the multi-entity class, we create sentence completion and question-answering queries for both the entities themselves, and the number of entities related to the subjects. In Turkish, however, sentence completion can occur in two grammatical modes (with and without a suffix in the answer), so we include both variants. Example queries are given in Table 1.

We examine the effects of Factual Inference Bias in LLMs, focusing on the following topics: UNESCO World Site Heritages, Neighbors, Car Brands, Top 500 Universities, Founding Locations of Companies, Mobile Network Operators, Country Code Top-Level Domain, Locations of Sites, Time Zones, Capital Cities, Polyglot Celebrities, Original Languages of Books, Official Languages. Those help assess whether the prompt’s language affects the model’s factual predictions. For each qualifying topic, we identify candidate entities that may exhibit bias when the prompt is in Italian or Turkish. The primary reason for excluding English from the Factual Inference Bias score calculations is that English is associated with multiple regions, making it less suitable for isolating language-specific effects. We use English only for RQ2. For example, {Bari, Bologna, Italia, Roma, Sardegna, Sicilia, Torino, Venezia} and {Ankara, İstanbul, Türkiye} represent entities that could be biased toward the Italian and Turkish versions of the prompt, respectively, under the topic Locations of Sites.

<sup>2</sup>Grammatical inconsistencies include incorrect use or omission of suffixes in the sentences.

The distribution of the topics, along with the sources, is shown in Table 2.

## 4. Experiments

### 4.1. Methodology

#### 4.1.1. Evaluation Metric

**RQ1.** To assess Factual Inference Bias, we measure how strongly a model’s outputs are influenced by language-specific associations rather than actual facts. For a specific topic  $t$ , we select a corresponding set of subjects  $S(t)$  and prompt the model in language  $l$  to retrieve the top- $n$  answers, denoted as  $P(n, m, s, t)$ . We evaluate these answers against two distinct groups: the truly related entities,  $R(s)$ , and the language-specific entities,  $Q(l, t)$  (a set representing incorrect or biased associations unique to that language-topic pair). The bias is computed by calculating the proportion of these language-specific entities ( $Q$ ) that appear in the model’s output, showing when the model defaults to linguistic stereotypes instead of retrieving the true entities ( $R$ ). Finally, these scores are averaged across all subjects in the topic to derive a single value representing the overall degree of Factual Inference Bias.

**RQ2** To evaluate model performance for single- and multi-entity answers, we employ a rank-based metric: Average Precision (Zhu, 2004), computed over the model’s logarithmic probabilities. We also utilize log-probabilities within an information retrieval framework to evaluate model performance, following the approach used in LAMA. However, while LAMA (Petroni et al., 2019) employs Precision@ $k$  as its evaluation metric, we replace it with Average Precision (AP) to better account for multi-entity answers and to more accurately assess how these multiple correct entities are ranked across the possible answers.

Precision is defined as the fraction of retrieved entities that are relevant, the intersection between the true entities of subject ( $G$ ) and the model’s entity predictions  $|S|$ , divided by the size of model’s entity predictions.

$$\text{precision} = \frac{|\{\text{true entities}\} \cap \{\text{model predictions}\}|}{|\{\text{model predictions}\}|}$$

*Precision@ $k$*  extends this notion by considering only the top- $k$  retrieved candidates instead of the full set for both gold and surface set.

$$\text{precision@k} = \frac{|\{\text{true entities@k}\} \cap \{\text{model predictions@k}\}|}{|\{\text{model predictions@k}\}|}$$

Average Precision measures how well relevant items are retrieved and ranked. It is computed as

Dataset	Language	Query Type	Query Target	Query	Answer
Multi-Entity	English	SC	Entities	The official language(s) of Canada is/are	English, French
		QA	Entities	What is/are the official language(s) of Canada?	English, French
		SC	Entity Count	The number of official languages of Canada is	2
		QA	Entity Count	How many official languages does Canada have?	2
	Italian	SC	Entities	La/Le lingua/e ufficiale/i del Canada è/sono	Inglese, Francese
		QA	Entities	Qual è/quali sono la/le lingua/e ufficiale/i del Canada?	Inglese, Francese
		SC	Entity Count	Il numero di lingue ufficiali del Canada è	2
		QA	Entity Count	Quante lingue ufficiali ha il Canada?	2
	Turkish	SC	Entities	Kanada'nın resmi dili/dilleri	İngilizce, Fransızca'dır
		SC	Entities	Kanada'nın resmi dili/dilleri şudur/şunlardır:	İngilizce, Fransızca
QA		Entities	Kanada'nın resmi dili/dilleri nedir/nelerdir?	İngilizce, Fransızca'dır	
SC		Entity Count	Kanada'nın resmi dil sayısı:	2	
QA		Entity Count	Kanada'nın kaç tane resmi dili vardır	2	
Single-Entity		English	SC	Entities	macOS is product of
	QA		Entities	Which company does macOS belong to?	Apple Inc.
	Italian	SC	Entities	macOS è prodotto da	Apple Inc.
		QA	Entities	A quale azienda appartiene macOS?	Apple Inc.
	Turkish	SC	Entities	macOS'un üreticisi	Apple Inc.'dir
		SC	Entities	macOS'un üreticisi şudur:	Apple Inc.
		QA	Entities	macOS'un üreticisi kimdir?	Apple Inc.

Table 1: Sample query–answer pairs of **Multi-Entity** (Official Languages) and **Single-Entity** (Product Makers) from FIBER. SC : Sentence Completion, QA : Question Answering.

Entity Class	Subjects	Entities	Total				Average		Resources	
			Subjects	English	Italian	Turkish	Total	Entity Count		
Single-Entity	Elements	Atomic Numbers	30	60	60	90	210	-	(Bowserinator, 2025)	
	Countries	ccTLD	62	124	124	186	434	-	(Wikipedia contributors, 2025)	
	Companies	Founding Locations	234	468	468	702	1,638	-	(CoAStAL NLP Group, 2025)	
	Books	Original Languages	214	428	428	642	1,498	-	(CoAStAL NLP Group, 2025)	
	Countries	Capital Cities	58	116	116	174	406	-	(CoAStAL NLP Group, 2025)	
	Elements	Chemical Symbols	30	60	60	90	210	-	(Bowserinator, 2025)	
	Sites	Locations	239	478	478	717	1,673	-	(CoAStAL NLP Group, 2025)	
	Products	Makers	148	296	296	444	1,036	-	(CoAStAL NLP Group, 2025)	
	Multi-Entity	Countries	Car Brands	40	160	160	200	520	9.85	(Wikipedia contributors, 2025)
		Countries	Time Zones	125	500	500	625	1,625	1.24	(Wikipedia contributors, 2025)
Polyglot Celebrities		Languages	71	284	284	355	923	3.65	(Wikipedia contributors, 2025)	
Countries		UNESCO World Site Heritages	85	340	340	425	1,105	2.44	(UNESCO World Heritage Centre, 2025)	
Countries		Universities	32	128	128	160	416	9.53	(QS Top Universities, 2025)	
Countries		Neighbors	78	312	312	390	1,014	3.86	(Wikipedia contributors, 2025)	
Countries		Official Languages	97	388	388	485	1,261	1.82	(Wikipedia contributors, 2025)	
Countries		Mobile Network Operators	24	96	96	120	312	3.17	(Wikipedia contributors, 2025)	
<b>Total (Single-Entity)</b>			1,015	2,030	2,030	3,045	7,105	-		
<b>Total (Multi-Entity)</b>			552	2,208	2,208	2,760	7,176	4.44 (Mean)		
<b>Total</b>			1,567	4,238	4,238	5,805	14,281	-		

Table 2: Main statistics of FIBER including the number of subjects, entries per language (English, Italian, and Turkish), total entries, and average entity counts (only for multi-entity) for all single-entity and multi-entity topics, along with their corresponding data sources.

the mean of the  $Precision@k$  values at the ranks where relevant answers appear. Formally:

$$AP = \frac{1}{|G|} \sum_{k=1}^{|S|} Precision@k \cdot \mathbf{1}[c_{(k)} \in G],$$

where  $|G|$  is the size of the gold set and  $|S|$  is the size of surface set.  $Precision@k$  is the precision at rank  $k$ ,  $c_k$  is the answer at rank  $k$ , and  $\mathbf{1}[c_k \in G]$  is an indicator function that equals 1 if the answer at rank  $k$  is in gold set, and 0 otherwise.

Mean Average Precision extends the concept of Average Precision to a set of multiple prompts

or subjects. It measures the overall ranking performance of a model across all evaluation instances. Formally, it is computed as the mean of the Average Precision scores over all  $N$  prompts (or subjects). It is used to evaluate the model's overall performance across different languages, models, topics, query types, and query targets:

$$MAP = \frac{1}{N} \sum_{i=1}^N AP_i,$$

where  $AP_i$  denotes the Average Precision of the  $i$ -th prompt, and  $N$  is the total number of evaluated prompts.

For each prompt, the set of candidate answers is divided into two subsets: the gold set, which includes the ground-truth answers, and the surface set, which includes all possible answers across all subjects. For each candidate in the surface set, the model’s logits are computed, and the logarithmic probability of the first token is obtained and appended to the original prompt. Then, the prompt is iteratively updated by concatenating the previously scored tokens, and the probability of the next token is calculated with respect to this new prompt. This process continues until all the tokens of the candidate are evaluated. The cumulative sum of the token-level log probabilities is taken as the overall score for that candidate. After repeating this procedure for all members of the surface set, a list of scores is obtained and sorted in descending order. The Average Precision (AP) is then computed on this ranking according to the defined formula, and the entire procedure is repeated for all subjects, topics, query types, and query targets.

#### 4.1.2. Models

In this study, we employ four different language models: Gemma-3-4B<sup>3</sup>, Qwen-2.5-3B<sup>4</sup>, Qwen2.5-7B<sup>5</sup>, and Llama-3.1-8B<sup>6</sup>. Our selection aims to encompass three distinct model families—Gemma, Qwen, and LLaMA—in order to examine how performance varies across different architectural types. Additionally, to investigate the impact of model scale within the same family, we include two variants of Qwen-2.5 with different parameter sizes (3B and 7B). Overall, our experimental setup consists of two lower-parameter models (Gemma-3-4B and Qwen-2.5-3B) and two medium-parameter models (Qwen-2.5-7B and Llama-3.1-8B), allowing us to analyze performance differences both across model families and with respect to model size.

#### 4.1.3. Experimental Setup

The experiments for Gemma-3-4B and Qwen-2.5-3B are conducted using an L4 22.5 GB GPU, while the Llama-3.1-8B model is evaluated with an A100 80 GB GPU. The Qwen-2.5-7B model is tested with two A5000 25 GB GPUs. All models are configured with a sampling temperature of 0.1, following the approach of similar studies (Farquhar et al., 2024; Kim and Kim, 2025) to obtain a best-generation estimation. This low temperature setting increases the likelihood of sampling the most probable tokens, thereby providing a more reliable measure of model

<sup>3</sup><https://huggingface.co/google/gemma-3-4b-it>

<sup>4</sup><https://huggingface.co/Qwen/Qwen2.5-3B-Instruct>

<sup>5</sup><https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

<sup>6</sup><https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

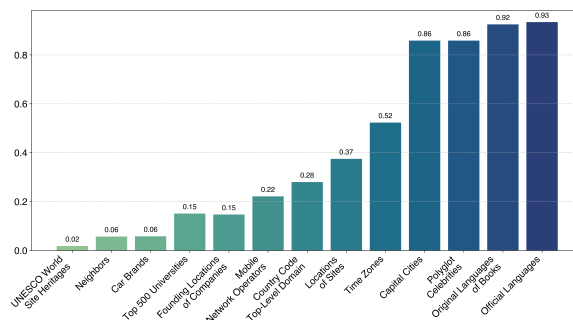


Figure 2: Factual Inference Bias by topic, averaged over model, languages, query types, and query targets.

accuracy.

## 4.2. Experimental Results

### 4.2.1. Factual Inference Bias Evaluation

Figure 2 highlights that the inference bias is much higher for topics whose entities involve countries, cities, places, and languages. The remaining experiments are conducted using the top-10 ranked positions. Polyglot Celebrities, Official Languages, Capital Cities, and Original Languages of the Books are the top four topics with the highest Factual Inference Bias, ranging between 0.86 and 0.93. Factual Inference Bias of other topics range between 0.02 and 0.52.

Figure 3 shows the comparison of the Factual Inference Bias of the models across topics. We can see that Gemma-3-4B has less bias compared to (with an average Factual Inference Bias score of 0.278). Llama-3.1-8B, Qwen-2.5-3B, and Qwen-2.5-7B, score 0.351, 0.353, and 0.364, respectively. However, despite the model shows a lower tendency toward Factual Inference Bias, Gemma-3-4B achieves the lowest performance in terms of mean average precision, as represented in the next subsection. Therefore, we argue that the outcomes of Gemma-3-4B need further investigation to understand its tendency to bias behavior. Finally, Figure 3 shows that the model size has not significant effect to the factual inference bias.

Figure 4 shows the comparison of the Factual Inference Bias of the models between Italian and Turkish across topics. When we consider the performance of the models based on their overall Factual Inference Bias scores, we observe that overall bias for Italian is (with an average Factual Inference Bias of 0.312) less than Turkish (with an average Factual Inference Bias of 0.333).

To address our first research question (RQ1), we observe that 31% of the topics exhibit a Factual Inference Bias score greater than 0.5. This finding suggests that the language in which a prompt is formulated can influence the Factual Inference Bias, although the extent of this influence appears to



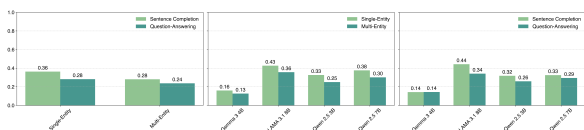


Figure 5: Subplot (Left): Mean average precision by entity count type and query type, averaged over topics, models, languages, query targets. Subplot (Middle): Mean average precision by models and entity count type, averaged over languages, query type, topics, query targets. Subplot (Right): Mean average precision by models and query type, averaged over languages, entity count type, topics, query targets.

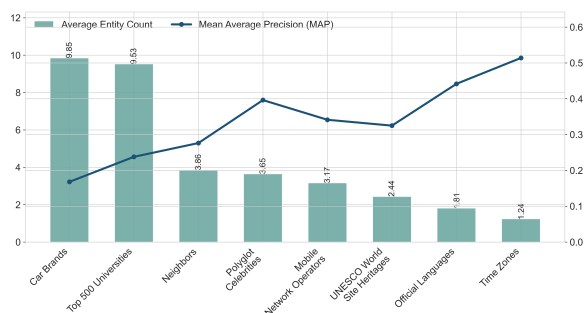


Figure 6: Average entity count (left) vs mean average precision (right) by topics, averaged over models, languages, query types, and query targets.

mains consistent across all languages for both query types.

Furthermore, it is observed that as the entity count of the subject decreases, the average precision increases across all languages and topics, as shown in Figure 6. This finding reinforces the earlier observation that model accuracy improves when the number of possible answers is smaller or limited to one—indicating that models perform more reliably on single-entity data than on multi-entity data.

To answer our RQ2, the model shows higher performance when responding to single-entity questions than to multi-entity questions. Qualitatively, the model performs worse when multiple correct answers exist, making it more difficult to maintain factual consistency across all entities. Language choice also plays a significant role in model performance: the highest mean average precision is observed for English (35%), whereas performance decreases for the low-resource languages Turkish (25%) and Italian (27%). Furthermore, models with larger parameter counts, such as Llama-3.18B (38%) and Qwen-2.5-7B (38%), consistently have better performance according to smaller models in the 3B and 4B range. Notably, Gemma-3-4B has the weakest results, achieving only 26% mean average precision.

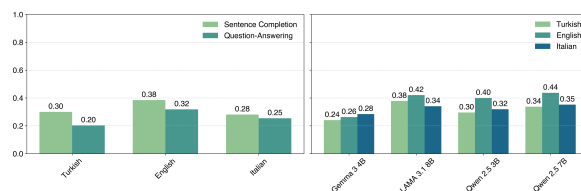


Figure 7: Subplot (Left): Mean average precision by languages and query type, averaged over topics, models, entity count type, query targets. Subplot (Right): Mean average precision by models and languages, averaged over query types, entity count type, topics, query targets.

#### 4.2.3. Multilingual Evaluation

We analyze the model’s performance across different languages and entity types to better understand the factors influencing its factual consistency.

When averaging performance across all tasks and models by language in Figure 7, English achieves the highest overall results, with a mean average precision (MAP) of 38%. Both Italian and Turkish demonstrate similar performance levels, with 32% and 31% MAP, respectively.

Performance ranking in descending order is English, Italian and Turkish across all models, except for LLaMA 3.1, which exhibits stronger performance in Turkish (37% MAP) than in Italian (34% MAP). The higher performance of Llama 3.1 on Turkish data should be investigated through further experiments to identify the reason behind this unexpected behavior.

## 5. Conclusion

This study investigates Factual Inference Bias and model performance across a novel dataset called FIBER containing answers with varying numbers of entities. We construct a dataset comprising eight single-entity and eight multi-entity topics, including question-answering, sentence completion for both entity class, and entity count-targeted for multi-entity data. To evaluate the models, we use two metrics based on the log probabilities of possible answers—one designed to measure Factual Inference Bias and the other is mean average precision, which measures the model’s ability to provide correct answers across all possible answers.

The results indicate that models exhibit Factual Inference Bias, although this tendency is topic-dependent. Topics related to language-oriented content show higher bias compared to others. Likewise, the degree of bias varies according to the prompt language, suggesting that linguistic factors play a role in shaping model responses. Furthermore, model performance differs across entity classes: models perform better on single-entity

data, as evidenced by the observation that the mean average precision increases when the average entity count decreases.

For future research, mechanistic interpretability methods (Rai et al., 2024) should be employed to uncover the underlying causes of this bias and to better understand how entity class influences model performance. Additionally, our analysis is limited to three languages, and future studies should expand this framework to include more low-resource languages. Lastly, since our study focuses on smaller models in the range of 3B to 8B, further evaluations should also be conducted on larger-scaled models and different model families to obtain more comprehensive insights.

## 6. Limitations

Our study primarily focuses on country- and language-related topics within multi-entity factual probing tasks. While this scope allows for a controlled evaluation of multilingual factuality, it also limits the generalizability of our findings to broader or more diverse knowledge domains. Similarly, our experiments include only two low-resource languages, Italian and Turkish; extending the analysis to a wider range of languages would enable a more comprehensive cross-linguistic evaluation.

Furthermore, the set of candidate answers for each subject is restricted to entities explicitly included in our dataset. Expanding this pool with semantically related entities could yield to more fine-grained and comprehensive probing results. Finally, our experiments are conducted using relatively small-scale language models (3B, 4B, 7B, and 8B parameters). Future work should extend these evaluations to larger models to better assess scalability and performance consistency across different parameter sizes.

Lastly, we make use of a possible-answer dataset and analyze the log probabilities assigned to each entity. This task can be extended to open-ended text generation, allowing us to examine how the model hallucinated and produces responses related to these concepts, rather than focusing solely on factual correctness.

## 7. Ethical and Broader Impact

Our study demonstrates that large language models can be influenced by the prompt language, exhibiting tendencies that align with the geographical or cultural context associated with that language. Such tendencies increase the likelihood of generating incorrect or biased responses when the prompt language varies. Therefore, language-induced bias should not be overlooked when evaluating the reliability and robustness of LLMs.

The proposed dataset enables a systematic comparison of model performance across different task types, languages, and question structures, including those with varying numbers of valid answers. It also facilitates the investigation of Factual Inference Bias and other potential bias types through knowledge probing experiments.

These findings can support model developers in identifying and analyzing the underlying sources of bias using emerging mechanistic interpretability techniques. The presence of such weaknesses and biases in LLMs cannot be disregarded; rather, they should be carefully considered during model design, training, and evaluation to ensure fair and reliable behavior across languages and contexts.

In this study, ChatGPT<sup>7</sup> and Gemini<sup>8</sup> are used to paraphrase human-written texts in order to enhance readability and improve wording.

## 8. References

- Abdaljalil, S., Kurban, H., and Serpedin, E. (2025). Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. *arXiv preprint arXiv:2503.07833*.
- Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Bowserinator (2025). Periodic-table-json: A json of the entire periodic table. <https://github.com/Bowserinator/Periodic-Table-JSON>.
- Cao, B., Lin, H., Han, X., Sun, L., Yan, L., Liao, M., Xue, T., and Xu, Z. (2021). Knowledgeable or educated guess? revisiting language models as knowledge bases. *arXiv preprint arXiv:2106.09231*.
- Chataigner, C., Taïk, A., and Farnadi, G. (2024). Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.
- CoAStAL NLP Group (2025). mpararel: Multilingual paraphrase-based relational probes for language models. GitHub repository. Retrieved October 24, 2025.
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. (2021). Measuring and improving consistency

<sup>7</sup><https://chatgpt.com/>

<sup>8</sup><https://gemini.google.com>

- in pretrained language models. *arXiv preprint arXiv:2102.01017*.
- Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Fierro, C. and Søgaard, A. (2022). Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Office Information Systems*, 43(2).
- Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Kassner, N., Duffer, P., and Schütze, H. (2021). Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.
- Kim, S. and Kim, H. J. (2025). A dual-layered evaluation of geopolitical and cultural bias in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, volume 4, pages 580–595, Bangkok, Thailand. Association for Computational Linguistics.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Lewis, P., Oguz, B., Rinott, R., Riedel, S., and Schwenk, H. (2020). Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Li, B., Haider, S., and Callison-Burch, C. (2024a). This land is {Your, My} land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.
- Li, H., Tomko, M., Vasardani, M., and Baldwin, T. (2022). Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, Washington. Association for Computational Linguistics.
- Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. (2024b). The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.
- Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Muti, A., Bassignana, E., Moscato, E., and Nozza, D. (2025). Probing feminist representations: A study of bias in LLMs and word embeddings. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLIC-it 2025)*, Cagliari, Italy. AILC / Italian Association for Computational Linguistics.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Priola, M. P. (2025). Addressing hallucinations with rag and nmiss in italian healthcare llm chatbots. *arXiv preprint arXiv:2412.04235*.
- Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., and Cohen, S. (2023). Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.
- QS Top Universities (2025). 2026 qs world university rankings [excel file]. [https://www.topuniversities.com/sites/default/files/2025-07/2026%20QS%20World%20University%20Rankings%201.2%20\(For%20qs.com\).xlsx](https://www.topuniversities.com/sites/default/files/2025-07/2026%20QS%20World%20University%20Rankings%201.2%20(For%20qs.com).xlsx).
- Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. (2024). A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.
- Taş, S., Huseyni, M. E., Özyay Ezerçeli, Bayraktar, R., and Terzioğlu, F. B. (2025). Turk-lettucedetect:

A hallucination detection model for turkish rag applications. *arXiv preprint arXiv:2509.17671*.

Umutlu, E. E., Cengiz, A. A., Sever, A. K., Erdem, S., Aytan, B., Tufan, B., Topraksoy, A., Darıcı, E., and Toraman, C. (2025). Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish. In Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlit, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Tafford, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM<sup>2</sup>)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

UNESCO World Heritage Centre (2025). World heritage list: Statistics. <https://whc.unesco.org/en/list/stat/>.

Wikipedia contributors (2025). Wikipedia: The free encyclopedia. Wikipedia. Retrieved October 23, 2025, from Wikipedia: The Free Encyclopedia.

Youssef, P., Koraş, O., Li, M., Schlötterer, J., and Seifert, C. (2023). Give me the facts! a survey on factual knowledge probing in pre-trained language models. Technical report, Findings of the Association for Computational Linguistics: EMNLP 2023. p. 15588.

Yu, L., Cao, M., Cheung, J. C. K., and Dong, Y. (2024). Mechanisms of non-factual hallucinations in language models. *arXiv preprint arXiv:2403.18167*.

Zhao, X., Yoshinaga, N., and Oba, D. (2024). What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13186–13214, Miami, Florida. Association for Computational Linguistics.

Zhu, M. (2004). Recall, precision and average precision. Technical report, Department of Statistics and Actuarial Science, University of Waterloo.