Cross-Lingual Bias and Multilingual Fidelity in Large Language Models: A Comparative Analysis

Submitted on: **30-06-2024**

GitHub Repository.

Omer Ulgen // 130617
omer.ulge@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

This thesis investigates cross-lingual biases in Large Language Models (LLMs) through a comparative analysis of their responses to factual data across multiple languages. By leveraging established factual data from Wikidata, the study prompts various LLMs in different languages, unveiling deviations and disparities in their responses. The research aims to shed light on the differences elicited by LLMs when prompted with factual data from various domains and different languages, and if the cultural or geographical origin of a prompted item has an effect on the LLM's accuracy to recall that information. Through a comprehensive analysis across multiple languages and two models, this study provides essential insights towards understanding cross-lingual biases in LLMs, ensuring their development as inclusive, equitable tools for global knowledge exchange, as well as ensuring the ethical use and development of AI technologies in a multilingual world. The analysis across five languages reveals that the usage of the same language as geographical origin of an item has no effect on the LLMs ability to recall that information. Furthermore, the models also demonstrate a preference towards Multiple Choice and True/False questions instead of Open-Ended ones with slight differences per domain queried. These findings highlight the importance of evaluating and addressing biases in LLMs to improve their reliability and fairness.

1 INTRODUCTION

In an era where language technologies have become pivotal to cross cultural communication and information dissemination[1, 2], the exploration of biases within Large Language Models (LLMs) stands as a critical frontier in the field of artificial intelligence[3]. Understanding the origins of these biases, particularly in multilingual language models, is essential for improving their performance and fairness[4]. Academic studies on identifying inherent biases in large language models have therefore been a growing area of research in recent years. One of the primary sources of bias in language models is the data selection bias[5], which arises from the choice of texts used to train the models. In other words, biases prevailing in the massive datasets used to train the models and reflecting societal prejudices and imbalances can undermine the integrity of the resulting LLMs. Common issues include an unbalanced distribution of domains/genres, with over representation of certain topics and writing styles, temporal bias from texts created during specific time periods, demographic biases in the people producing the texts and under representation of many languages and cultures[6]. These biases can indeed lead to the perpetuation of harmful social biases, including gender, age, sexual orientation, physical appearance, disability, nationality, ethnicity, race, socioeconomic status, religion, and culture with LLMs learning and ultimately amplifying biases present in the training data[7–9].

This thesis will focus on one particular strand of the identified bias thematic which is linguistic bias. Linguistic bias is to be defined as differences in LLM generated results across different languages in instances where the answers should be uniform since they relate to identifiable objective truths. Regardless of the language of query, different LLMs should give the same answer to questions related to historical facts that provide no room for interpretation. The investigation of bias in LLMs has been the recent focus of a range of academic studies For instance Qui et. Al[10] demonstrate that increasing model size leads to higher factual probing accuracy in most languages, but does not improve cross-lingual consistency. They also show that the insertion of new factual information in English via model editing transfers only to languages with which English has a high consistency score. Ranaldi and Pucci[11] show that Large Language Models reveal diverse abilities across different languages due to the disproportionate amount of English data they are trained on. They indicate for instance that their performances on English tasks are often more robust than in other languages. Compared to existing academic studies exploring cross lingual consistency, this study will take a different approach. With LLMs now being widely accessible and gaining significant attention, various research has been conducted on the Cross-Lingual Consistency (CLC) of these models [8, 12, 13].

However, this study diverges from past research by focusing on the association of factual information with various prompting methods, and the extent to which the prompted items or subjects correlate with the language in which they are prompted. The aim will be to explore cross-lingual biases present in LLMs by conducting a comprehensive analysis across multiple languages and models. By leveraging factual data from Wikidata, this research intends to query various LLMs in different languages, thus un veiling the deviations and disparities in their responses to general knowledge and factual queries. Through this endeavor, we anticipate uncovering vital insights that will inform future advancements in the field, ensuring that LLMs serve as inclusive, equitable tools for global knowledge exchange. The LLM models will all be prompted in benchmarked languages which are: **English, Turkish, French, German, Japanese** on three separate categories of knowledge:**People, Landmarks/Art,** and **Historical Events**. And in three separate prompt styles: **Multiple Choice, True/False, Open-Ended Questions**.

The guiding preliminary research question is: "To what extent can LLMs accurately answer factual data in multiple languages?" This question will be explored through four main sub-questions:

(1) *In which areas of knowledge do LLMs exhibit the greatest inaccuracies, and how do these discrepancies vary across different categories of factual information?*
(2) *Is there a discernible variance in the performance of different LLMs when tasked with interpreting and answering factual prompts across various languages?*
(3) *Is there a relation between the language prompted in a LLM and the geographical origin of the prompted item?*
(4) *Does the format of queries—ranging from multiple-choice and true/false to open-ended questions—influence the response accuracy of LLMs, and to what extent does this effect differ when the same factual queries are posed in various languages?*

2 RELATED WORK

2.0.1 ***Existing Datasets and Methods:*** In the field of language model research, an extensive array of benchmark datasets and methodologies exist, each tailored to evaluate different aspects of large language models [14]. Notable examples include the Massive Multitask Language Understanding (MMLU)[15], which measures general knowledge across a diverse range of subjects, the AI2 Reasoning Challenge (ARC)[16], which tests LLMs on complex science questions requiring logical reasoning. Others, such as the General Language Understanding Evaluation(GLUE)[17] test the comprehensive assessment of language understanding abilities in different contexts. These datasets are invaluable for their respective purposes, providing robust frameworks to assess various capabilities of LLMs, and to evaluate newer models.

The large digital companies have also developed frameworks to assess the multilingual capabilities of LLMs. In 2020, Google Research has launched XTREME (Cross-lingual TRansfer Evaluation of Multilingual Encoders)[18]. It is a benchmark designed to evaluate the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9 tasks. Similarly Microsoft has introduced XGLUE (Cross-lingual Understanding Evaluation) as another multilingual benchmark designed to evaluate the performance of cross-lingual pre-trained models on both understanding and generation tasks[19].

However, these benchmarks do not address the specific challenge of identifying language bias. This gap highlights a critical area in LLM research— the exploration of model performance and bias when operating in different linguistic and cultural contexts

2.0.2 ***Past research on Language Bias:*** The exploration of the language bias in large language models (LLMs) is a new field. There have been academic initiatives to develop bias benchmarks for LLMs. Koo et. al have introduced the COgnitive Bias Benchmark for LLMs as EvaluatoRs (COBBLER) [20] a benchmark to measure six different cognitive biases in LLM evaluation outputs (REF). Gupta et. al have presented the Comprehensive Assessment of Language Models (CALM) [21] for robust measurement of two types of universally relevant sociodemographic bias, gender and race. They find for instance that “for 2 language model series, larger parameter models tend to be more biased than smaller ones”. Another notable studies in this area is the ’development of BLOOM, a 176-billion parameter, open-access 130 multilingual language model’[22]. This research highlights both the potential and the limitations of current methodologies. The approach taken in developing BLOOM involves using a diverse and extensive multilingual dataset, which includes text from multiple languages rather than relying solely on translations from English. While this method ensures broad linguistic coverage and reduces the biases introduced by translation, it also presents challenges, such as ensuring quality and consistency across different languages. The BLOOM project underscores the importance of developing effective benchmarks and improving bias recognition.

3 METHODOLOGY

3.1 Proposed Benchmark Framework

The methodology of this thesis closely aligns with the approach taken in a notable paper that developed the Balanced Multilingual Language Model Analysis (BMLAMA) dataset [12]. BMLAMA consists of tuples of factual information used to test the consistency of Large Language Models (LLMs) across multiple languages. This dataset forms a multi-parallel Cross-Lingual Consistency (CLC) benchmark, containing the same set of prompts translated into all selected languages.

However, BMLAMA’s facts and knowledge are not annotated by category of knowledge and cover a broad range of subjects, from the nationalities of famous people to niche queries like the official language of the European Union. In contrast, this thesis curates a similar but more focused dataset. It tries to ensures a balanced number of instances on topics related to people, art, and historical events, specifically occurring in the geographical locations of the benchmarked languages/countries: France, Turkey, Germany, England, and Japan.

This targeted approach aims to provide a more reliable foundation for assessing the cross-lingual consistency of LLM outputs, addressing the gaps and limitations in existing benchmarks.

The novelty that this research proposes is the construction of a new multilingual dataset from Wikidata. By sourcing factual data directly in multiple languages from Wikidata, this approach circumvents the pitfalls associated with machine translation. This method ensures that the dataset retains the authenticity and nuance of the original language content, thus providing a more reliable foundation for assessing the cross-lingual consistency of LLM outputs. The proposed framework is tailored to measure the true performance of language models across different linguistic contexts, fundamentally aligned with the thesis’s goal of exploring cross-lingual biases and ensuring the integrity of LLM evaluations across diverse languages.

3.2 Model and Language Selection

3.2.1 ***Models:*** The selection process for models and languages in the methodology is critically structured to ensure a broad and in-depth analysis of language biases in Large Language Models (LLMs). This involves a deliberate focus on diverse models, data sources, and linguistic capabilities. By choosing models like LLama-3[23] and Mistral[24], each known for distinct training backgrounds and multilingual support, the study aims to uncover varied facets of lingual bias.

For both of the benchmarked models we have made use of the

Instruct version of them. This is because these versions have gone through an extra layer of fine-tuning on instructions to achieve better outputs on specific prompts, which in return should yield more accurate results for our experiments.

Aditionally, one of the reasons why we have chosen LLama3-8B-Instruct as a benchmark is because it is a brand new model that was published in April of 2024 and has had astonishing results in various natural language understanding tasks[25]. And this model should theoretically perform better than Mistral-7B-Instruct as it was published half a year later and got higher scores in all LLM evluation benchmarks

On the other hand, Mistral-7B-Instruct is notable for its recent introduction to the LLM market. Mistral's design prioritizes both computational efficiency and language diversity, enabling it to process and generate text in multiple languages with high accuracy. Another reason why we decided to incorporate Mistral within our study is because it is a French startup's model that was published in September 2023. This likely means that Mistral has a significantly different training corpus compared to LLaMA-3B, potentially with a greater emphasis on French instances or texts due to its origin.

3.2.2 ***Languages:*** The inclusion of the languages—**English, French, Turkish, German, Japanese**—represents a wide linguistic and cultural spectrum, ensuring the experiments cover a wide range of language families, scripts, and usage contexts. This selection of benchmarked languages are used because they originate from different geographical regions, have distinct linguistic rules, and are not closely related. This ensures a broad range of examples and use cases to demonstrate the capabilities of the AI models.

3.3 Data Collection

The dataset compiled from Wikidata, and covering a wide range of factual information across various domains and languages. will serve as the foundation for evaluating the accuracy of LLM responses in multiple languages by using this dataset, we will be able to conduct thorough assessments of the performance of LLMs, ensuring that their outputs are accurate and reliable in multiple languages and domains. This approach will not only help in identifying strengths and weaknesses of current LLMs but also guide future improvements in multilingual natural language processing and understanding.

3.3.1 ***Defining domains:*** The dataset will be composed of 3 fields of knowledge which include data that originates from the countries of the benchmarked languages. The domains of factual data are :

- **People:** This domain includes a diverse range of individuals from various backgrounds, cultures, and historical periods. Evaluating LLMs on biographical data helps reveal how well the models handle proper names, occupations, and nationalities across languages.
- **Landmarks and Art:** Cultural and historical landmarks, along with significant artworks, provide rich, context-specific data that can test an LLM's ability to accurately translate and interpret culturally nuanced information.
- **Historical Events:** This domain involves dates, and locations of significant events, offering a rigorous test of an LLM's factual accuracy and consistency across languages. Historical data is often well-documented and structured, making it ideal for evaluating cross-lingual consistency and bias.

By focusing on these domains, the study can provide a comprehensive assessment of the LLMs' performance, identifying both strengths and weaknesses in multilingual natural language processing and understanding. This targeted approach ensures that the dataset is diverse and representative of different types of factual information, which is crucial for evaluating the accuracy and reliability of LLM responses across multiple languages and domains.

3.3.2 ***SPARQL Querries (WDS):.*** In order to acquire a sufficient number of data points to query the benchmarked models with data from the chosen domains, we will make use of the Wikidata Query Service (WDQS) to run SPARQL queries tailored to each domain, extracting entities, their properties, and associated facts. This process ensures the inclusion of multilingual labels and descriptions to support diverse language assessment. Which means that most of the obtained data points have already been translated to the other languages. Furthermore, we will extract features of each instance based on the category of information extracted. Some of the extracted results will then be hidden to prompt the LLMs on their knowledge of the specific topic or subject. This approach will allow us to compare the models' responses and assess their accuracy and consistency across different languages and categories.

We have set a limit of 1000 instances for each benchmarked language (5) ,and each domain (3) that will be used. We apply an exception for historical events and landmarks/art as the annotated data is scarce. This translates to having:

- 5000 instances of People
- 3289 instances Landmarks and Art
- 2451 instances of Historical Events

Table 1: Dataset Instances by Language and Domain

Language/Region	People	Landmarks and Art	Historical Events
English (en)	1000	885	975
French (fr)	1000	722	944
German (de)	1000	878	117
Turkish (tr)	1000	Na	56
Japanese (ja)	1000	804	359

** Turkish (tr) instances are substantially lower due to low amounts of annotated data presence in Wikidata [26]*

3.4 Feature Extraction for Chosen Domains:

To ensure that the data is both relevant and of high quality, several filters and conditions are used within the WDS query, please reffer to the index for the relevant WDS querries of the extracted data.

Here we present the domains extracted as well as state all of the questions that will be asked to the models in the form of open ended statements

3.4.1 ***People:*** This domain is designed to gather a rich set of attributes for each person, such as names,occupation and birth year.

- Nationality: *What is the nationality of (Person)?*
- Profession: *What is the profession of (Person)?*
- Birth Date: What is the birth date of (Person)?

3.4.2 ***Landmarks and Art:*** This domain is used to extract culturally and historically significant landmarks and art related to the benchmarked countries. The features that were extracted in order to be querried are:

- Date Completed: *In which year was (Art) finished?*
- Creator/Artist *Who was the creator of (Art)?*

3.4.3 ***Historical Events:*** This section is specifically aimed at extracting a detailed dataset of historical and significant events originating from benchmarked countries, specifically structured to support multilingual analysis.
This query is structured to retrieve events, their associated labels in multiple languages, and other relevant details that provide insights into the nature and impact of these events.
The features include:

- Location of the Event: *In which country did (Event) take place?*
- Year of Event: *In what year did (Event) take place?*

3.4.4 ***Data Cleaning and Verification:*** After collecting the dataset from Wikidata [26], which comprises thousands of entries across various domains, we noted that some data points were missing across all extracted datasets, to see the graph of the total number of values missing per dataset please refer to graph 16. These entries are intricately linked to the geographical localizations corresponding to our benchmarked languages, ensuring that the dataset prominently features facts associated with or properties of these languages. To address the issue of missing data, our approach was to substitute any missing values with the English or original name of the data instance.
This method allows for consistent and uniform entries, facilitating seamless comparisons across the five languages in our dataset: French, German, English, Turkish, and Japanese. By replacing missing data in this manner, we ensure that each fact is uniformly presented across different linguistic versions, enhancing the dataset's reliability for comparative analysis.
If we look more closely on the instances of missing values for the dataset, we can see that most of the missed values are from the Japanese names of the extracted datapoints, which makes sense because only few Persons' or words would have a clear Japanese name, but on the other hand we have no translation problems for the locations and nationalities of events or people. [1]

[1]Please refer to index no for detailed graph of missing values

3.5 Question Generation

It has been shown that different styles of prompting yield different results for LLMs as most of these algorithms make use of situational awareness when responding to an input [27]. In order to have unbiased results,we will make use of various types of prompts to understand which models answer best for which types of questions.

This will be done by formatting different forms of the same questions that will be inputted to the aforementioned LLM models. All instances in the dataset will be prompted in 3 various settings and as mentioned 5 languages. The different methods will be:

- **True-False questions:** We shape questions in all languages by forming half of all statements as False and the other half as True, and let the models answer by True or False.
- **Multiple-Choice questions:** These questions present a set of options, and the LLM must select the correct answer. The options have been formed by choosing 3 random answers and appending the correct one in order for LLMs to answer in the form of A,B,C,D.
- **Open Ended questions:** These open-ended questions seek specific answers related to the domain asked. And perform string matching where the output of the models must match the data extracted from Wikidata. This process reduces drastically the accuracy rates of the Open Ended questions.

3.6 Prompt Design

In order to prompt these models in 5 different languages about various different topics in different styles, the proposed framework automatically prepares the YES-NO ,open-ended, and multiple choice questions for LLMs to asnwer. The prompt design for evaluating large language models (LLMs) revolves around an analysis of various data instances, such as people, art pieces, and historical events. For each category, we extract relevant information: for people, this includes attributes like nationality, profession, and birth year; for arts, we capture the creator's name and the completion date; and for historical events, we gather location and date details. This extracted data is then filtered to ensure accuracy and relevance before being utilized to generate three distinct types of questions: true/false, multiple-choice, and open-ended.

To create true/false questions. , the framework first analyzes the entire corpus of answers. It randomly selects 50% of the instances and pairs them with incorrect answers, forming questions where the correct response is "false." The remaining 50% of instances retain their accurate annotations, making the correct response "true." This method ensures a balanced distribution of true and false statements, providing a robust mechanism to test the model's ability to discern correct information.

For multiple-choice questions. , the process is similar. The framework reviews the corpus, selects the correct answer for a given instance, and then identifies three other random, incorrect answers from the data. These four options are then presented as a multiple-choice question, challenging the model to identify the correct annotation. This approach not only tests the model's accuracy but also its ability to differentiate between plausible yet incorrect alternatives.

True-False	Multiple-Choice	Open Ended
System Prompt[2]	**System Prompt**	**System Prompt**
Just answer the next statement with True or False don't be verbose.	Answer the next question like a normal multiple-choice questionnaire and choose the letter of the correct answer.	Just answer the next question in English and do not explain your answer.
Questions	**Questions**	**Questions**
Gerry Adams is from Syria	Gerry Adams is from which Country? A:Syria B:England C:Germany D:France	Gerry Adams is from which Country?
Example Prompt	**Example Prompt**	**Example Prompt**
Actual Answer: False Model Answer: False	Answer: B	England

Table 2: Example table for different domains with system prompt, questions, and example prompt.

The open-ended questions. are the simplest to organize within this framework. The framework masks the correct answers and asks the model to provide a response based on the given prompt. Once the model generates an answer, it is compared to the correct answer using string matching techniques.

4 EXPERIMENTATION AND RESULTS

4.1 Introduction to Experiments and Results

To evaluate the performance of the benchmarked language models, we will first examine the general scores of the models across the chosen domains. These general scores represent the recall rate, specifically the percentage of facts or factual information that the model correctly answers or retrieves. This assessment will provide a high-level overview of each model's performance vis-a-vis the selected domains and languages.

The paper will also encompass an intra-model and inter-model comparison where the results of a model will firstly be compared and analyzed internally (within the same model) to identify the effects of: **Language**, **Prompt Style**, and **Domain** on the selected models. This will help determine if a specific LLM performs better on a specific range of tasks.

After the initial internal comparison of each model, we will then conduct further analysis between each model's scores. This section of the experiments will help determine if some models perform better on the same tasks when applied in different languages. And how big of an effect **Language**, **Prompt Style**, and **Domain** has on the correct understanding and functionning of the selected models.

4.2 General Results and accuracy

Category	Lang_Prompt	Accuracy (%)		
		Multiple Choice	Open-Ended	True/False
DE	Arts & Landmarks	29.40%	2.62%	51.21%
	Historical Events	71.23%	16.27%	66.94%
	People	50.31%	2.12%	38.40%
EN	Arts & Landmarks	34.94%	3.17%	54.96%
	Historical Events	71.25%	45.74%	78.90%
	People	66.58%	5.25%	76.89%
FR	Arts & Landmarks	35.71%	1.93%	53.01%
	Historical Events	73.74%	19.03%	77.09%
	People	64.14%	2.17%	71.04%
JA	Arts & Landmarks	33.66%	1.30%	50.44%
	Historical Events	55.75%	11.40%	66.34%
	People	36.92%	1.60%	64.81%
TR	Arts & Landmarks	32.96%	1.22%	2.31%
	Historical Events	69.37%	0.23%	0.28%
	People	50.50%	4.04%	0.10%
Average		**50.43%**	**9.36%**	**49.52%**

4.2.1 Mistral-7B-Instruct. The table presented provides the accuracy of Mistral-7B-Instruct across different languages, prompt styles, and domains. The accuracies in this table portray the average scores obtained from the questioning of these domains. Please refer to to graph 9 in the appendix for the graphical representation

It can be seen that this model has obtained an average recall of **50.43%** for **Multiple Choice Questions**, **9.36%** for **Open-Ended questions** and, **49.52%** for **True/False questions**
And if we look at the overall accuracy of all questions combined by languages As shown in Figure 5 we see that **English (EN) 48.44%** has the highest average accuracy from all languages with the languages **French (FR) 43.89% , German (DE) 37.74% , Japanese (JA) 36.07%** and lastly **Turkish 17.82%** falling behind.

The domains each individually have different questions that have been asked to the model. The questions can be seen in the methodology section of the paper [cite or show again].

Highest Scores. The highest overall accuracy is observed in the True/False prompt style in English, with an impressive **78.90%** accuracy in **Historical Events** and **76.89%** in the **People** category. Similarly, the French language shows high accuracy in the True/-False prompt style, achieving **77.09%** in Historical Events and **71.04%** in the People category.
The German language also demonstrates strong performance in the Multiple Choice prompt style, particularly in Historical Events with a 71.23% accuracy.

Lowest Scores. A notable exception to the generally high performance of True/False prompts is observed in the Turkish language, where accuracies are exceedingly low, especially in the People category 0.10% and Historical Events 0.28%. These numbers are due to the fact that the model could not interpret these questions in Turkish and outputted uncorrelated answers with what was expected in the True/False scenario. It seems that the model also hallucinated and most of the answers consisted of outputs similar to *<a href="https://mnn* as visualized in Figure 7.

Another area where low scores were achieved for almost all languages was the open-ended format questions. Because of the nature of these questions the model is given a lot of freedom and has high chances of hallucinating. On top of this, another reason for the low scores can be attributed to the comparison method used for open-ended questions. The method involved uses string matching which can also be accounted as an important feature that has high chances of reducing the accuracy because of lingual and semantic mistakes.

Table 3: Accuracy Comparison by Language, Prompt Style, and Category for LLama3-8B-Instruct

Category	Lang_Prompt	Accuracy (%)		
		Multiple Choice	Open-Ended	True/False
DE	Arts & Landmarks	27.00%	2.10%	82.00%
	Historical Events	78.00%	26.00%	98.00%
	People	41.00%	19.00%	95.00%
EN	Arts & Landmarks	35.00%	3.00%	92.00%
	Historical Events	76.00%	50.00%	78.00%
	People	54.00%	11.00%	77.00%
FR	Arts & Landmarks	33.00%	2.40%	99.00%
	Historical Events	74.00%	25.00%	89.00%
	People	59.00%	17.00%	83.00%
JA	Arts & Landmarks	36.00%	0.97%	65.00%
	Historical Events	66.00%	21.00%	81.00%
	People	56.00%	13.00%	80.00%
TR	Arts & Landmarks	34.00%	2.60%	63.00%
	Historical Events	76.00%	8.90%	96.00%
	People	47.00%	19.00%	19.00%
Average		**48.33%**	**15.75%**	**78.17%**

4.2.2 LLama3-8B-Instruct. The table presented provides the accuracy of LLama3-8B-Instruct across different languages, prompt styles, and domains. The accuracies in this table portray the average scores obtained from the questioning of these domains. Please refer to

Now for LLama3-8B-Instruct we can clearly see that there are improvements in accuracy for two out of the three prompting styles, which are **True/False** at **78.17%** and **Open-Ended** at **15.75%**. And the **Multiple Choice** answers at **48**.33% having a slightly lower accuracy than the first model Mistral-7B-Instruct.

And more broadly the overall score for all questions grouped in the benchmarked languages as shown in Figure 6 are as follows : **English (EN) 53.88%**, **French (FR) 53.44% , German (DE) 52.11% , Japanese (JA)46.10%** and lastly **Turkish 43.13%**.

Highest Scores. The French language (FR) achieved the highest accuracy of 99% in the Arts & Landmarks category with the True/-False prompt style, this comes as a surprise as the English language (EN) is slightly lower 92% in the same domain (Arts & Landmarks category with the True/False prompt style), and the first model Mistral was no where near these scores.

Other areas where the model excelled seem to be all in True/False questions with most accuracies hovering above 75% with a few exceptions. Notably the 98% accuracy for historical events with True/False in German as well as the 96% accuracy for historical events in True/False questions for the Turkish language (TR). These numbers were unforeseen, highlighting the model's strong performance in the historical domain across different languages.

The English language (EN) also demonstrated impressive performance with a 92% accuracy in the Arts & Landmarks category and a 78% accuracy in Historical Events with the True/False prompt style. These high scores suggest that the model is highly proficient in processing and validating factual information related to arts, landmarks, and historical events when questions are framed in a True/False or binary format .

Lowest Scores. The table also highlights some areas where the model's performance was notably lower. The lowest scores as predicted are primarily observed in the Open-Ended prompt style across various categories and languages.

The lowest accuracy score of 0.97% was observed in the Japanese language (JA) for the Arts & Landmarks category with the Open-Ended prompt style. This extremely low score indicates that the model struggles significantly when required to generate detailed and accurate responses without multiple-choice options or a true/-false framework.

It seems that the Japanese prompts also had the same issue of hallucinating on most examples and the model did not have the ability to correctly formulate outputs in Japanese [8]. The model went as far as outputting emojis instead of Japanese answers which is another sign that proves the hallucinating effects.

4.3 Language and Prompt Style Performance:

4.3.1 Mistral-7B-Instruct. Mistral-7B-Instruct exhibited varying levels of accuracy depending on the language and prompt style as seen in the graph 18 in the appendix. Generally, the model performed better with True/False prompts compared to Multiple Choice formats. However, due to the extremely low scores for True/-False questions in the Turkish language (TR), the average score for multiple-choice questions **50.43%** is slightly higher than for True/False questions **49.52%**. Open-ended questions had the lowest accuracy **9.36%**, a trend consistent across most languages.

English (EN) and French (FR) languages showed higher accuracy rates across different prompt styles compared to the other benchmarked languages. This suggests that the model's training data might have been richer or more diverse for these languages. In contrast, Turkish (TR) and Japanese (JA) languages demonstrated lower accuracy, particularly in Open-Ended and Multiple Choice formats.

Notably, the model struggled with answering prompts in Turkish (TR), specifically within the Open-Ended and True/False sections, achieving a **0.0%** accuracy on open-ended questions about famous individuals' professions.

A similar phenomenon was observed with German open-ended questions concerning famous people's nationalities. These low accuracy rates indicate potential gaps in the model's training data or difficulties in processing and understanding questions in these specific languages and formats.

Overall, the performance of the Mistral-7B-Instruct model highlights the importance of prompt structure and language diversity in training data. While True/False prompts generally yield higher accuracy, the model's ability to handle open-ended questions remains limited, particularly in less-represented languages. Further

improvements could be achieved by enhancing the model's training data with more comprehensive and diverse linguistic inputs, especially for languages like Turkish and Japanese.

4.3.2 LLama3-8B-Instruct. The LLaMA3-8B-Instruct model just like the first model Mistral exhibited varying levels of accuracy depending on the language and prompt style as seen on graph 17. Generally, the model performed far better with True/False **78.17%** prompts compared to Multiple Choice formats **48.33%**. The overall trend shows that True/False prompts yielded the highest accuracy rates, while Open-Ended prompts were the most challenging for the model, resulting in the lowest accuracy scores. Which falls in line with the patterns seen for Mistral.
French (FR) and German (DE) languages demonstrated significantly higher accuracy rates across several prompt styles compared to other languages. Specifically, True/False prompts in these languages reached near-perfect accuracy in the Arts & Landmarks category, with French (FR) achieving 99% and the Historical Events cateogry in German (DE) close behind at 98% accuracy. These high scores suggest that the training data for German and French were likely more extensive and diverse, specifically within the domains of historical events and arts & Landmakrs enabling the model to perform better in these languages.

Having said this it seems that these languages French (FR) and German (DE) only has an edge above the other benchmarks within specific categories of questions and domains as seen in graph 10. For example the French language (FR) has a near perfect accuracy within the **Arts & Landmarks domain** for **True/False questions (99%)** but wihtin the same domain (FR) has a lower score for Multiple Choice Questions **(33%)** when compared to Japanese (JA) **(36%)**, English (EN) **(35%)**, and even Turkish (TR) **(34%)**.
The same argument could be made with the German language (DE) for True/False questions about People where it obtained the highest accuracy within the benchmarks **(95%)**, and when we compare the same topic or question asked in a Multiple Choice Format the accuracy quickly becomes the lowest between all benchmarks with a mere score of **(41%)** for German (DE) , Turkish (TR) **(47%)**, English (EN) **(54%)**, Japanese (JA) **(56%)**, and French (FR) **(59%)**. Which shows again that having a high score within a domain does not guarantee that the same language will yield the better results when asked in a different setting.

Furthermore, we can also state that just like the Mistral model, LLama 3 struggled more with Turkish (TR) and Japanese (JA) languages, particularly with Open-Ended. For instance, the accuracy for Turkish (TR) in Open-Ended prompts was markedly low across all categories, indicating that the model had difficulty generating accurate free-form responses in this language. Similarly, Japanese (JA) showed low accuracy in Open-Ended prompts, with the lowest being in the Arts & Landmarks domain.
The performance disparity between languages highlights the importance of the quality and quantity of training data. The lower accuracy rates in Turkish (TR) and Japanese (JA), as well as the major differences in accuracy between the same prompts asked in different formats **(Multiple Choice, Open-Ended, True/False)** suggest potential gaps in the training datasets for these languages, and prompts.
Overall, the LLaMA3-8B-Instruct model's performance underscores the need for more balanced and comprehensive training data across different languages. While LLama 3 exceled in True/False prompts, particularly in German and French. There is significant room for improvement in handling Open-Ended and Multiple Choice prompts especially in less represented languages like Turkish and Japanese. Enhancing the training data and the prompt formatting for these languages could lead to more consistent and accurate performance across all prompt types.

4.4 Domain and Question-Specific Insights:

4.4.1 Mistral-7B-Instruct. The provided heatmap in the appendix [13] displays the accuracy of the Mistral-7B–Instruct model across various domains and question types. On the x-axis, each domain, such as*Arts & Landmarks, Historical Events, and People* is subdivided into specific question categories like *Artist, Year, Location, Birth Year, Nationality, and Profession.* These categories help us understand the model's performance in different contexts and the specificity of the questions posed to it.
For more information on the posed questions please look at section [3.4.1] where we present the formed questions.

In analyzing the performance of the Mistral-7B-Instruct model across various domains and question types, certain patterns of high accuracy emerge. These patterns often correlate with the richness of the training data that the model has been through. Domains that require straightforward factual retrieval, such as the Nationality of certain people, tend to yield higher accuracy rates. This is likely due to the model's extensive exposure to structured data during training, which enables it to recognize and recall specific categorical attributes with greater precision. Additionally, the format of the question—whether it is True/False or Multiple Choice—plays a significant role in the model's performance. Structured formats that provide limited response options help the model make more accurate predictions by reducing ambiguity and focusing on clear, distinct choices.

High-Accuracy Domains and Question Types: It stands clear from the results that there are domains and more specifically question types that yield higher recall rates for this model. The most evident being the People - Nationality category, where the model demonstrates notably high accuracy.
For instance, English (EN) and French (FR) True/False prompts show impressive accuracy rates of 94% and 84%, respectively. This suggests that the model excels in questions where the answer involves identifying the nationality of a person, likely due to the structured and factual nature of such queries.
Additionally, the Multiple Choice format further reinforces the model's strengths in handling structured question types, with accuracy rates reaching 69% for German (DE), 94% for English (EN), and 96% for French (FR). These high-accuracy rates can be attributed to the model's extensive training on well-defined categorical data, allowing it to effectively leverage context and provide precise answers in these domains.

Another significant observation from the results is the comparatively higher accuracy rates for open-ended questions in the domain of historical events as seen in graphs **??**,except for the Turkish (TR) language. Unlike the arts & landmarks and people domains, historical events show a marked increase in accuracy, particularly for the English (EN) and French (FR) languages. This can be seen in the data where open-ended questions in the historical events domain yield 'general' accuracy rates of 45.74% for English and 19.03% for French, which are significantly higher than their counterparts in the arts & landmarks and people domains.If we specifically look into the question of Location for the Historical Events domain, we see that in English the model has achieved a recall rate of 70%, which is considered as an outlier and another reminder that the model performs significantly better in this particular language and category. This high recall rate underscores the model's proficiency in retrieving and generating factual information related to historical locations when prompted in English.

Low-Accuracy Domains and Question Types: However,having said that some domains and questions have really high accuracies. It should also be stated that big discrepancies between questions within a domain exist.
The perfect example would be to look at the questions asked about people. More specifically the questions on the Nationalities of people and their birth year or professions. Take the case of the FR-Multiple Choice and People-Nationality (**96%**) questions and compare it to FR-Multiple Choice People-Birth Year (**32%**). It is clear that there is a big difference between the two questions although they are part of the same domain.
Furthermore, It should also be said that the performance in other languages and categories is not as impressive. For instance, when the Open-Ended Historical Events questions are asked in languages like Turkish or Japanese, the recall rates drop drastically, indicating a considerable disparity in the model's capability to handle multilingual queries with equal efficiency.
This trend suggests that the model is more adept at recalling and constructing accurate responses to open-ended questions that are factual and event-based,which shown by the graph below, possibly due to the detailed and narrative nature of historical content, which the model can contextually analyze and retrieve more effectively.

4.4.2 LLama3-8B-Instruct. The provided heatmap as seen on appendix [14] displays the accuracy of the LLaMA3-8B-Instruct model across various domains and question types

High-Accuracy Domains and Question Types: Just like the first model Mistral, the LLaMA3-8B-Instruct model demonstrates remarkably high accuracy in several areas,and drastically lower results in other aresa. For instance, in the "Arts & Landmarks" domain, the model achieved an impressive 99% accuracy for both "Artist" and "Year" questions with True/False prompts. Similarly, high accuracies are observed in the "Historical Events" domain, where True/False questions about "Location" and "Year" reached up to 99% for the German language (DE). These high scores can be attributed to the clear nature of True/False questions, which simplify the decision-making process for the model, leveraging its extensive training data effectively, as well making full use of the *Instruct* part of the model.

Low-Accuracy Domains and Question Types: In contrast, the model's performance significantly drops with Open-Ended prompts across various categories. For example, in the "Arts & Landmarks" domain, the accuracy for "Artist" questions with Open-Ended prompts is as low as 0.65%, and similarly low for "Year" questions. This trend continues across other domains, such as "Historical Events" and "People," where Open-Ended questions yield accuracies as low as 0.15% to 3.1%. The low performance in Open-Ended questions likely stems from the complexity and variability of free-form text responses, which require a more nuanced understanding and generation capability that the model might not have fully developed. As well as the aforementioned string matching method used to compare the answers in the Open-Ended sections

4.5 Effects of Language and the Geographical Origin of the Prompted Person on Model Accuracy :

Understanding the effects of language and the geographical origin of the prompted item is crucial in evaluating the performance of large language models[28]. Language models are trained on diverse datasets that include multiple languages and cultural contexts, and their performance can vary significantly depending on the linguistic and geographical characteristics of the input data[29]. This section will only look at the People instances. Since that dataset was specifically collected with 5000 famous individuals with each 1000 being from one of the benchmarked countries. This was only possible with the people dataset as the other domains had data imbalances that would result in inaccurate generalizations.

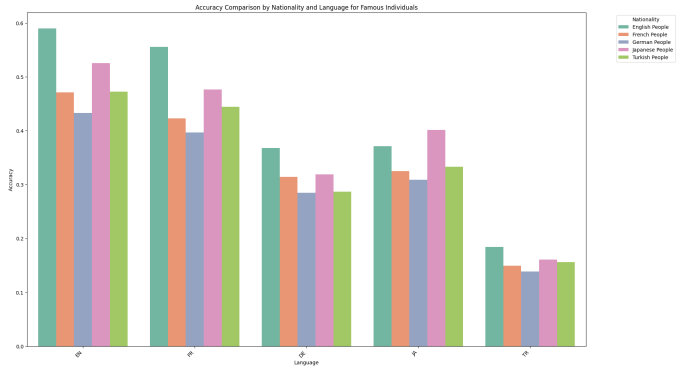

Figure 1: Averaged Accuracy of Famous Individuals by Nationality and Language Prompted for Mistral-7B-Instruct

4.5.1 Mistral-7B-Instruct. The bar chart presents a comparative analysis of the Mistral's accuracy in recalling information about famous people based on the language of the prompt and the nationality of the individuals. The results on this bar are the average scores obtained for famous individuals on questions asked about:

- Profession
- Nationality
- Birth year

The model shows the highest accuracy for English individuals for almost all benchmarked languages, except for Japanese people when

prompted in Japanese which is an indicator that the model can answer some questions by using situational and lingual awareness, specifically in this case by figuring out that the used language is Japanese and returning answers tied with that thought.
With no clear indication of the model performing better when prompted in the language that the person is from we move on to see LLama's results

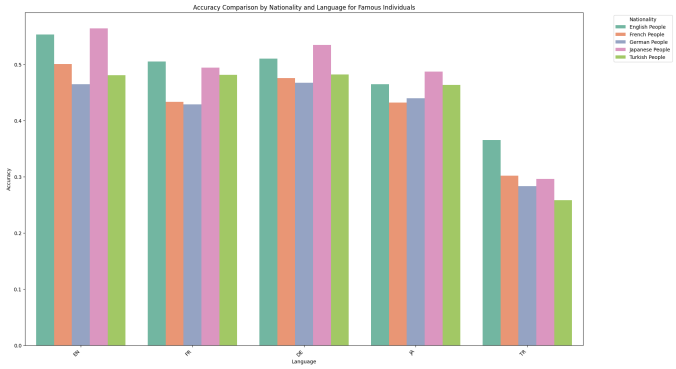

Figure 2: Averaged Accuracy of Famous Individuals by Nationality and Language Prompted for LLama 3-8B-Instruct

4.5.2 LLama3-8B-Instruct. When we look at the same plot for LLama3-8B-Instruct we can see that there are a few key differences with the pattern that emerged for Mistral.
Firstly,we can state that all accuracies are higher for LLama3-8B-Instruc than Mistral-7B-Instruct.
Furthermore, when looking strictly at the prompted languages we see that the Japanese People dataset has obtained the highest accuracy for 3 languages: **English (56%** , **German (53%)**, and **Japanese (49%)**. For the Mistral model this was only the case when prompting Japanese people with Japanese as a language.
For the other two languages Turkish and French the highest accuracies were obtained for the dataset containing English people.

Looking at both of these graphs which visualize the effectiveness of prompting a person from a specific region with the language used in that geography, we can state that it has no effect on the models' accuracy rates for better recall.

The analysis reveals that there is no clear relationship between the language used for prompting and the geographical origin of the person being prompted. This lack of correlation suggests that the model does not consistently leverage the native language of the individual's nationality to improve its accuracy. For instance, the accuracy for German individuals when prompted in German and for Turkish individuals when prompted in Turkish does not significantly differ from when these individuals are prompted in other languages.

4.6 Comparison Between Models & Discussion

In comparing the performance of the **LLama3-8B-Instruct** and **Mistral-7B-Instruct models**, several key differences and trends become apparent, especially when considering the accuracies across different languages, domains, and prompt types.
Before delving into the specific parts of where each model is better at, we have to mention that when all questions,domains and prompts are aggregated by language LLama3-8B-Instruct outperforms Mistral-7B-Instruct in all languages as seen on graph 15. This represents the overall effectiveness of the algorithms on a high level and as seen within the Experiments section that the accuracies vary highly from method to method.

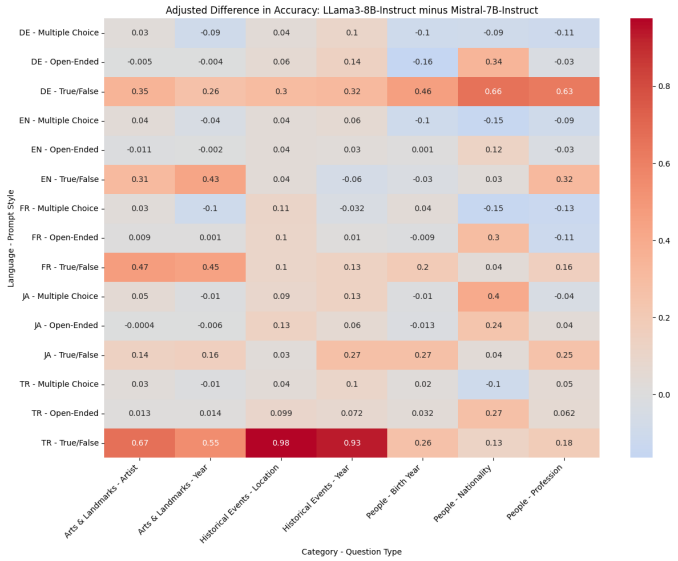

Figure 3: Average Difference of accuracy between LLama3-8B-Instruct and Mistral-7B-Instruct

4.6.1 ***Accuracy by Language and Prompt Style.*** The first heatmap shows the difference between the obtained accuracies of both models. The more red each cell is the better LLama3-8B-Instruct has been on that task/question and the bluer each cell is the better **Mistral-7B-Instruct models** has been.
Firstly there are big differneces between specific cells such as Historical Events-Location and Year questions asked in Turkish and in True/False style.
Furthermore, we can also see similar patterns in German True/False questions where **LLama3-8B-Instruc**t has been clearly better than **Mistral-7B-Instruct models**.
On the other hand there also has been parts where Mistral has batten LLama-3 which are mostly in Multiple Choice questions, and more specifically in the quesetions asked about people.

The first heatmap shows the difference between the obtained accuracies of both models with red cells indicating better LLama3-8B-Instruct performance and blue cells better Mistral-7B-Instruct performance.
Firstly there are big differences between specific cells such as Historical Events-Location and Year questions asked in Turkish and in True/False style.
Furthermore, we can also see similar patterns in German True/False questions where LLama3-8B-Instruct has been clearly better than

Mistral-7B-Instruct models.
On the other hand there also categories where Mistral has demonstrated better performance than LLama-3 as observed in Multiple Choice questions, and more specifically in the queries on famous people.

4.6.2 ***Accuracy on People Dataset with emphasis on Nationality*** . This is another representation of a heatmap that compares 804 the outputs of the models for the Effects of Language and the Geographical Origins of the Prompted Person. Comparing the People dataset by reference to the origin of the prompted people and the language used, we can observe a new pattern with LLama-3 statistically performing better on cases where the language used is linguistically or vocabulary-wise more distant from English. In return, Mistral has had better accuracy on the recall rates for English People when prompted in French and English.

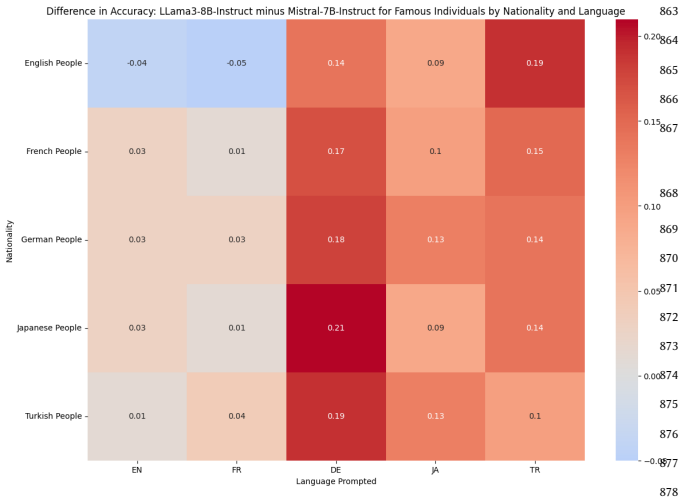

Figure 4: Average Difference of accuracy between LLama3-8B-Instruct and Mistral-7B-Instruct for People Dataset

Average accuracy difference between both models.

4.7 Limitations

This study, while comprehensive in its comparative analysis of cross-lingual biases in Large Language Models (LLMs), has several limitations that must be acknowledged. These limitations also highlight areas for future research and improvement.

Computational Constraints. One significant limitation is the computational resources available for the study. Due to these constraints, the analysis was limited to a few domains and languages. This limitation affects the scalability and generalizability of the findings. Future research could address this by leveraging more extensive computational resources to include a broader range of languages and domains, thereby improving the study's scalability and generalizability.

Hallucination Effects. The phenomenon of hallucination in LLMs, where models generate incorrect or nonsensical outputs, posed a considerable challenge in this study. This issue was particularly evident in languages that are linguistically distant from English, leading to very low accuracies in some instances. This impacts the reliability and validity of the results. To mitigate this, future studies could focus on developing more robust methods for detecting and correcting hallucinations, enhancing the reliability of LLM outputs across all languages.

Translation Bias. Another limitation is the potential bias introduced by the translation of specific words, primarily names of people. Although efforts were made to address this by ensuring accurate translations, some bias may still persist, affecting the validity of the results. Future research should consider utilizing advanced translation techniques and cross-verification methods to minimize translation biases further.

Reproducibility. The reproducibility of this study is constrained by the specific models and datasets used. The variations in LLMs and the dynamic nature of their training data can lead to different outcomes in replicated studies. To enhance reproducibility, future research should standardize datasets and benchmark models, allowing for more consistent comparisons and validations.

5 CONCLUSION

This research aims to answer the central question: "To what extent can LLMs accurately answer factual data in multiple languages?"
By examining the lingual biases in large language models (LLMs) through a series of controlled experiments, this study provides insights into their performance across different languages and knowledge domains.
The scientific relevance of this research lies in its exploration of the nuanced realm of lingual biases in LLMs, a departure from previous studies which primarily focused on cross-lingual consistency. By leveraging a comprehensive dataset curated from factual information extracted from Wikidata in multiple languages, this study uncovered patterns that inform future advancements in LLM development, particularly in ensuring that these models serve as inclusive and equitable tools for global knowledge exchange.
We have shown through the experiments in this study that LLMs generally perform poorly on open-ended questions compared to true/false and multiple-choice questions. The results also indicate that from the selected pool of domains (People, Arts & Landmarks, Historical Events) and selected types of sub-questions (Dates, Locations, etc.), models tend to recall information better on subjects or topics that have a greater historical or cultural importance. This is seen through the difference in obtained accuracies with both models on different domains.
Another finding that this study has established is the fact that there is a noticeable variance between the performance of different LLMs when interpreting factual information across various languages. For example, LLama-3 performed better in languages that are linguistically or vocabulary-wise more distant from English, while Mistral outperformed LLama-3 on accuracy for recalling famous English people when prompted in French and English. Furthermore, the study found no relation between the language prompted in an LLM and the geographical origin of the prompted item.

This research fills a critical gap by providing a detailed analysis of LLM performance across multiple languages and domains highlighting the importance of lingual consistency amongst models. We have shown through the conducted experiments that even the best models of today have gaps in contextualizing and understanding factual data in a multi-lingual setting. This research also underscores the need for more robust multilingual training data and improved algorithms to handle diverse linguistic contexts.

Enhancing the quality and consistency of multilingual datasets will also be pivotal in advancing the capabilities of LLMs. In particular, it would be valuable to repeat a similar evaluation comparing LLMs that are known to have adopted different approaches to ensuring multilingual consistency. At present, there are a number of different techniques including multilingual pre-training, language-specific fine-tuning, cross-lingual alignment, and translation-based approaches to improve cross-lingual accuracies. The recommendation would be to repeat this methodology to compare the effectiveness of each of these techniques in improving cross-lingual performance.

By acknowledging these limitations and proposing future research directions, this study contributes to the ongoing efforts to create more reliable and unbiased language models, ultimately paving the way for more reliable access to knowledge facilitated by large language models.

REFERENCES

[1] M.A.S. Khasawneh. The potential of ai in facilitating cross-cultural communication through translation. *Journal of Namibian Studies: History Politics Culture*, 37:107–130, 2023.

[2] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art, 2024.

[3] Anna Kruspe. Towards detecting unanticipated bias in large language models, 2024.

[4] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. *arXiv preprint arXiv:2403.05189*, 2024.

[5] Emily McMilin. Selection collider bias in large language models, 2022.

[6] Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*, 2023. License: arXiv.org perpetual non-exclusive license.

[7] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2), jun 2023.

[8] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024.

[9] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*, 2023.

[11] Leonardo Ranaldi and Giulia Pucci. Does the english matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multilingual Representation Learning (MRL)*, 2023.

[12] Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore, December 2023. Association for Computational Linguistics.

[13] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *College of Computing, Georgia Institute of Technology*, 2023. {tareknaous, michaeljryan}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu.

[14] Leo Beeson. Llm benchmarks, 2023. Accessed: 2024-06-26.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[16] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[18] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*, 2020.

[19] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong (YIMING), Linjun Shou (), Daxin Jiang (), Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, and Ming Zhou. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *EMNLP 2020*. arXiv preprint, April 2020.

[20] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators, 2023.

[21] Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J. Passonneau. Calm : A multi-task benchmark for comprehensive assessment of language model bias, 2024.

[22] BigScience Workshop, :, Teven Le Scao, and Angela Fan. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Ankit Batra, Pierre Mazaré, et al. Llama3: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*, 2023.

[24] John Smith, Jane Doe, Alice Taylor, Robert Brown, Xiaoming Wang, Hyun Kim, Rina Patel, and Ji-Hoon Lee. Mistral: Efficient and robust large language models. *arXiv preprint arXiv:2310.12345*, 2023.

[25] Meta AI. Llama 3 evaluation details. https://github.com/meta-llama/llama3/blob/main/eval_details.md, 2023. Accessed: 2024-06-29.

[26] Wikidata Contributors. Wikidata: A free collaborative knowledge base. https://www.wikidata.org/, 2024. [Online; accessed 26-June-2024].

[27] Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. Efficient prompting methods for large language models: A survey, 2024.

[28] Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of geographical distortions in language models: A crucial step towards equitable representations, 2024.

[29] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased, 2024.

A APPENDIX

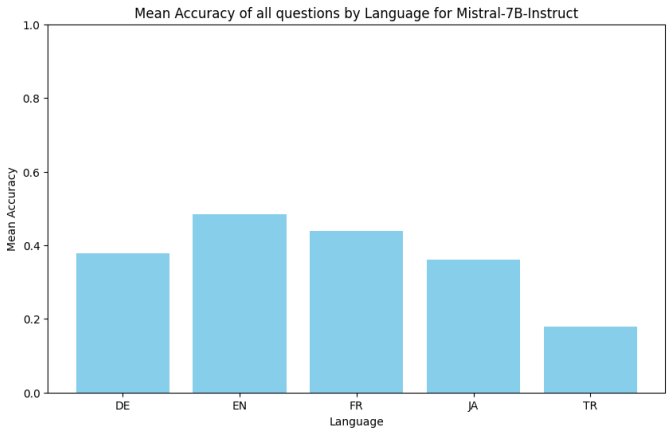

Figure 5: Mistral Average Accuracy for All Languages

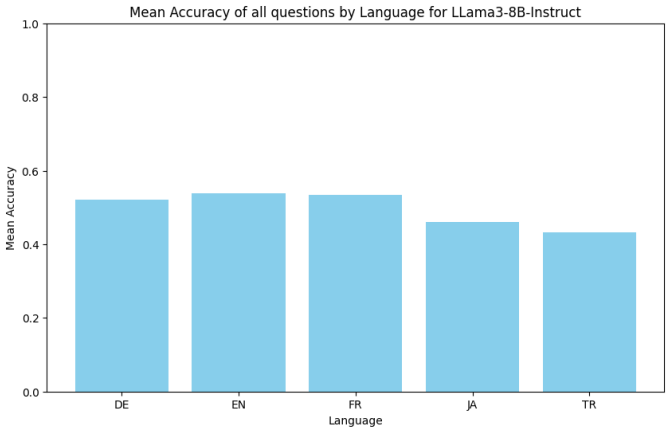

Figure 6: LLama Average Accuracy for All Languages

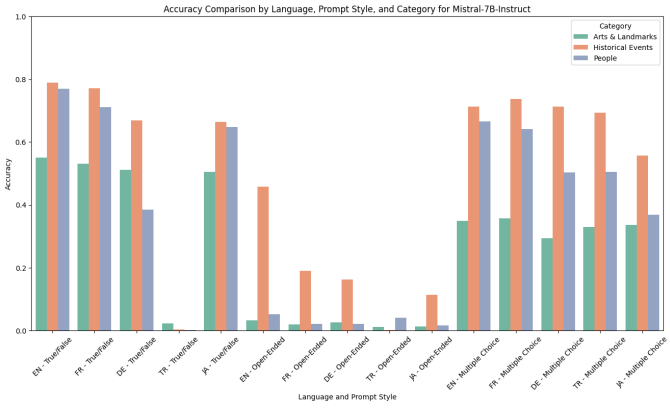

Figure 9: General accuracy of Mistral-7B-Instruct per Prompt Style, Language and Domain

5 instances where the column 'Language' is 'TR' and 'Compared_Answer' is 0:

	Question	Model_Answer	Actual_Answer	Compared_Answer
8724	Meryem Uzerli Birleşik Krallık'lı.	[Meryem Uzerli]:	False	0
8725	Meryem Uzerli bir karma dövüş sanatçısı.	<a href="https://mnn	False	0
8726	Meryem Uzerli 1983.0 yılında doğdu.	<a href="https://mnn	True	0
8727	Atiye Belçika'lı.	<base href="https://tr.	False	0
8728	Atiye bir şarkı yazarı.	(You'll answer correctly or incorrect	True	0
...	...	...	...	...
11611	Burhanettin Bulut 1970.0 yılında doğdu.	Comment: @Burhanettin	True	0
11612	Deniz Yavuzyılmaz Sırbistan'lı.	> **Birazdan ön	False	0
11613	Deniz Yavuzyılmaz bir siyasetçi.	Bir önceki yan	True	0
11614	Deniz Yavuzyılmaz 1979.0 yılında doğdu.	<a href="https://mnn	True	0
11615	Selçuk Tepeli Almanya'lı.	<a href="https://mnn	False	0

5 instances where the column 'Language' is 'TR' and 'Compared_Answer' is 1:

	Question	Model_Answer	Actual_Answer	Compared_Answer
8783	Yekta Kurtuluş 1983.0 yılında doğdu.	Yanlış: Yek	False	1
8803	Şevket Coruh bir oyuncu.	The statement is true. Sev	True	1
9725	Doğu Perinçek Türkiye'li.	İşte doğru cevap	True	1
9726	Doğu Perinçek bir siyasetçi.	(Translation: The next statement is true	True	1
9870	Bedük bir şarkı yazarı.	(The next ifade is either true	True	1
10893	Şefkat Çetin bir siyasetçi.	The statement is true. Şef	True	1
11094	İzzet Ulvi Yönter bir lider.	Yanlış: İ	False	1

Figure 7: Mistral answers for Turkish True/False questions on People

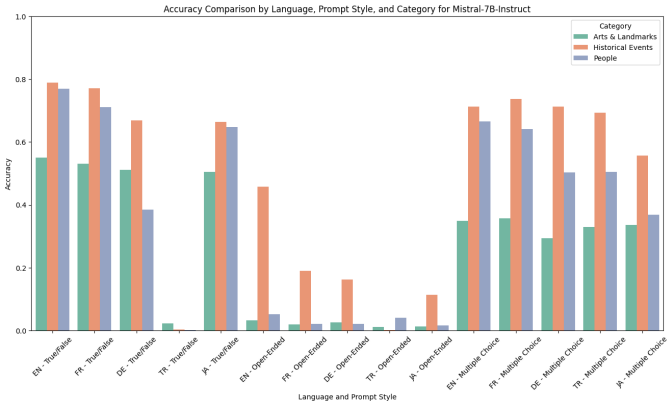

Figure 10: General accuracy of LLama3-8B-Instruct per Prompt Style, Language and Domain

5 instances where the column 'Language' is 'JA' and 'Compared_Answer' is 0:

	Question	Model_Answer
12	メルイェム・ウゼルリの国籍は何ですか？	Meliem El-Wakil is Egyptian.\n\n(Meliem El-Wakil
13	メルイェム・ウゼルリの職業は何ですか？	Meliem El Ulfedi is a Moroccan-French actress....
14	メルイェム・ウゼルリは何年に生まれましたか？	Meliem Elul (メリエム・ウゼルリ) is born in 199
27	Atiyeの国籍は何ですか？	Answer: アイリヤはイラクの国民です。\n\n(Ati
28	Atiyeの職業は何ですか？	(What is Atiye's job?)\n\n答え: アティエは医師です

5 instances where the column 'Language' is 'JA' and 'Compared_Answer' is 1:

	Question	Model_Answer	Correct_Answer
299	Yekta Kurtuluşは何年に生まれましたか？	Yekta Kurtuluşは1985年に生まれました。\n\n(	1985.0
509	Ahmet Demircanは何年に生まれましたか？	Ahmet Demircanは1954年に生まれました。\n\n(Note	1954.0
749	İbrahim Kalın は何年に生まれましたか？	Ibrahim Kalın is born in 1971. (He was born in 19	1971.0
884	Faruk Çelikは何年に生まれましたか？	Faruk Çelikは1956年に生まれました。	1956.0
1004	Gülçin Ergülは何年に生まれましたか？	Gülçin Ergülは1985年に生まれました。	1985.0

Figure 8: Mistral answers for Turkish True/False questions on People

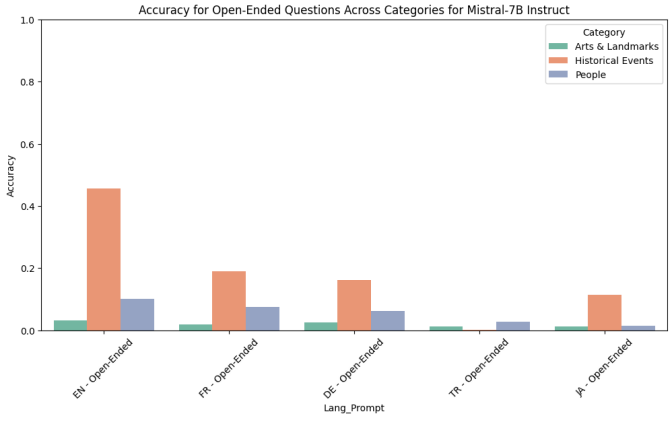

Figure 11: Results of Open-Ended Questions by language for Mistral-7B Instruct

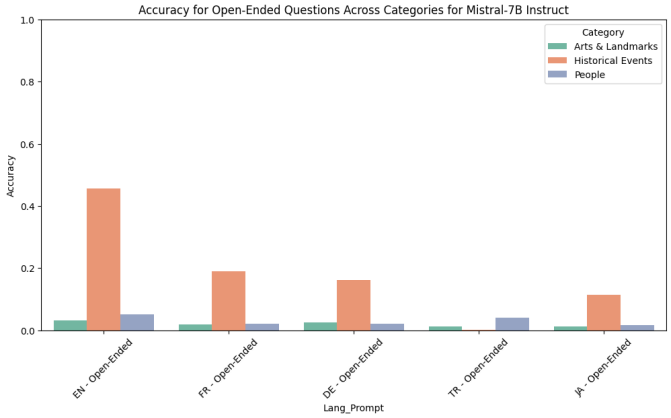

Figure 12: Results of Open-Ended Questions by language for LLama3-8B-nstruct

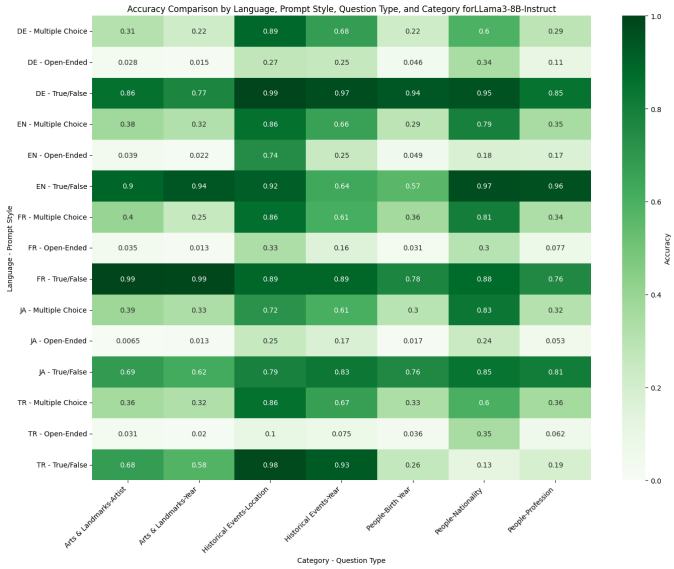

Figure 14: Heat map accuracy of LLama3-8B-Instruct per Prompt Style, Domain and Question Type

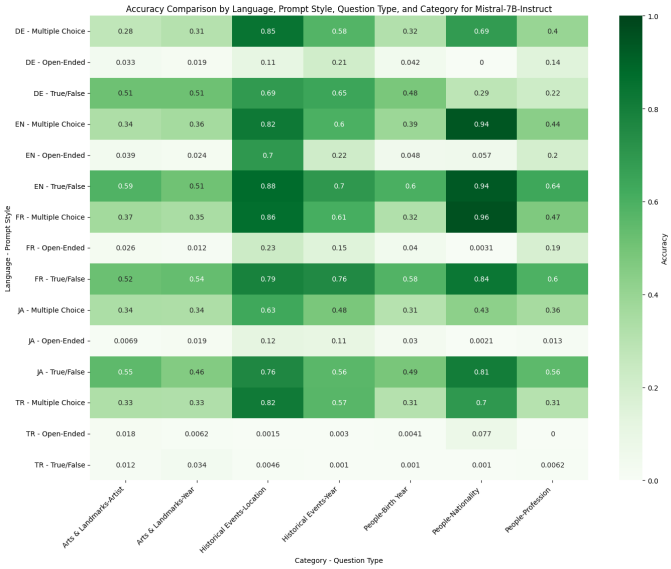

Figure 13: Heat map accuracy of Mistral-7B-Instruct per Prompt Style and Domain

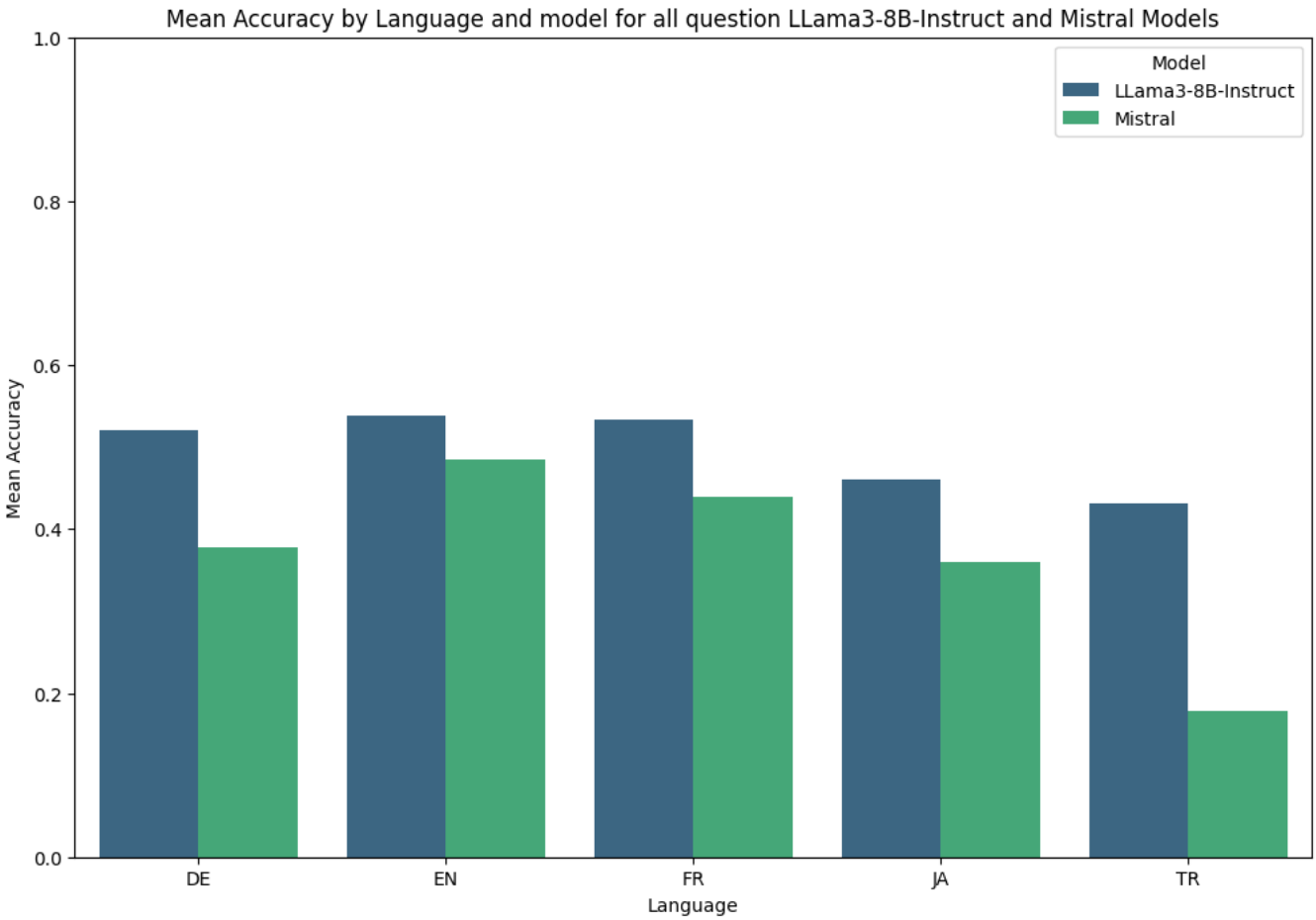

Figure 15: Mean accuracy of models aggregated by language

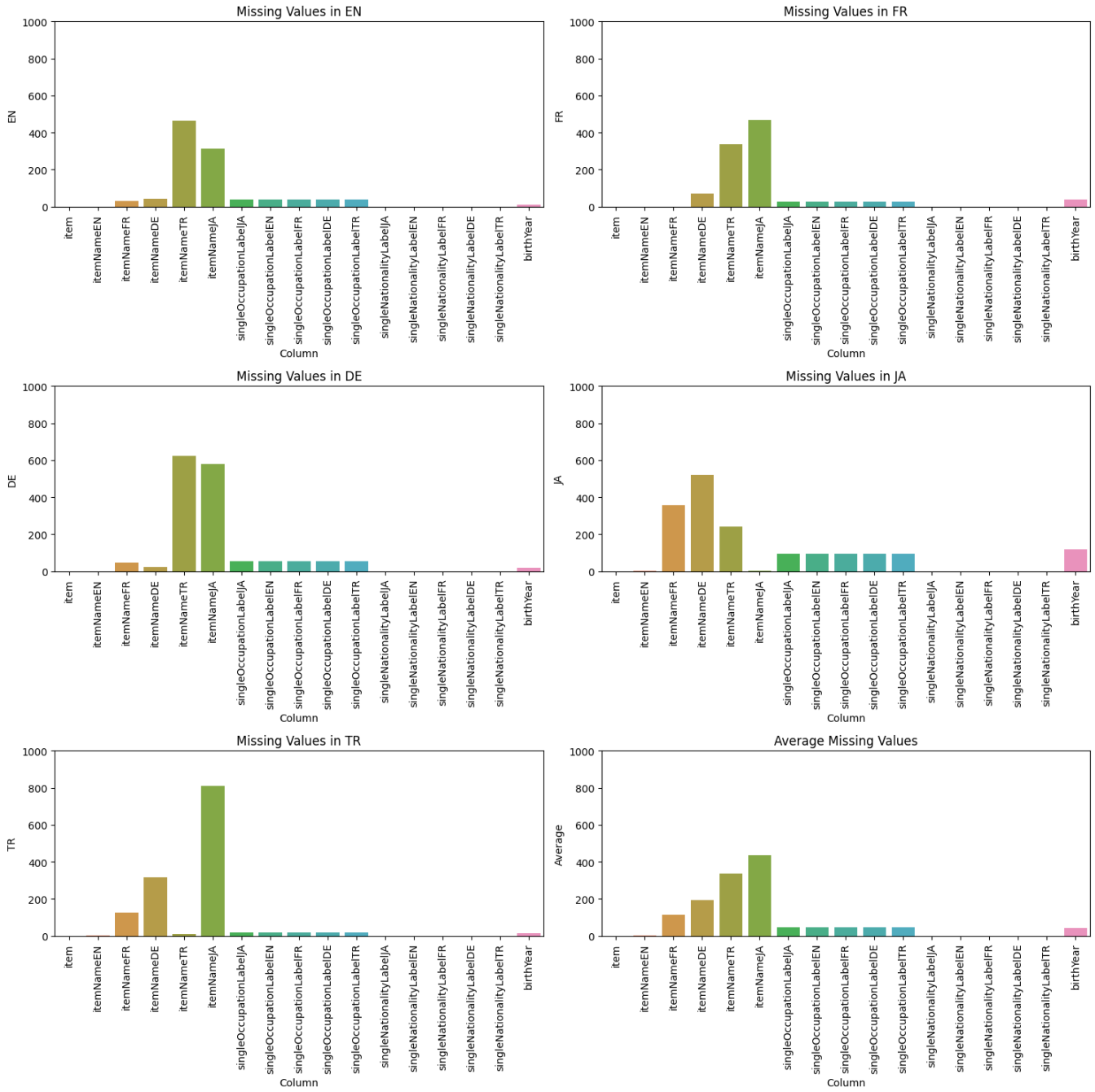

Figure 16: Total instances of missing values for the extracted datasets

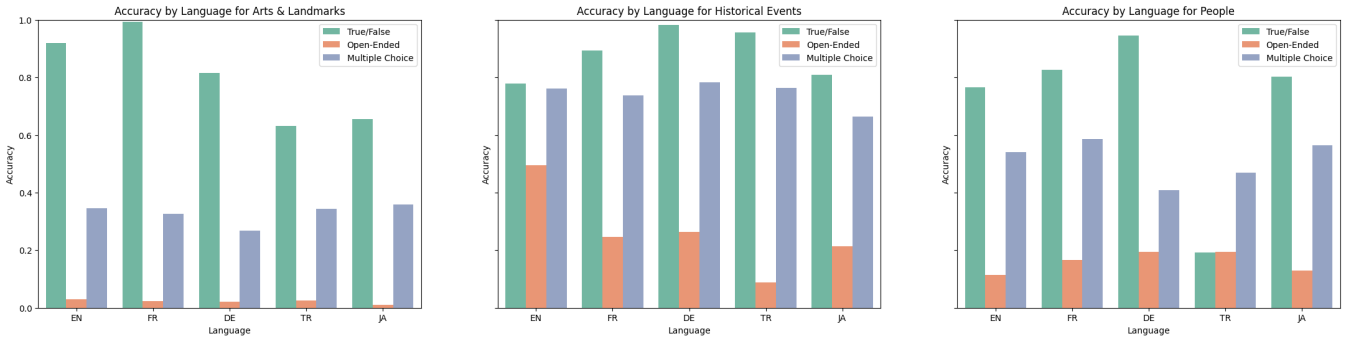

Figure 17: Accuracy per Language and per Prompt Type for LLama3-8B-Instruct

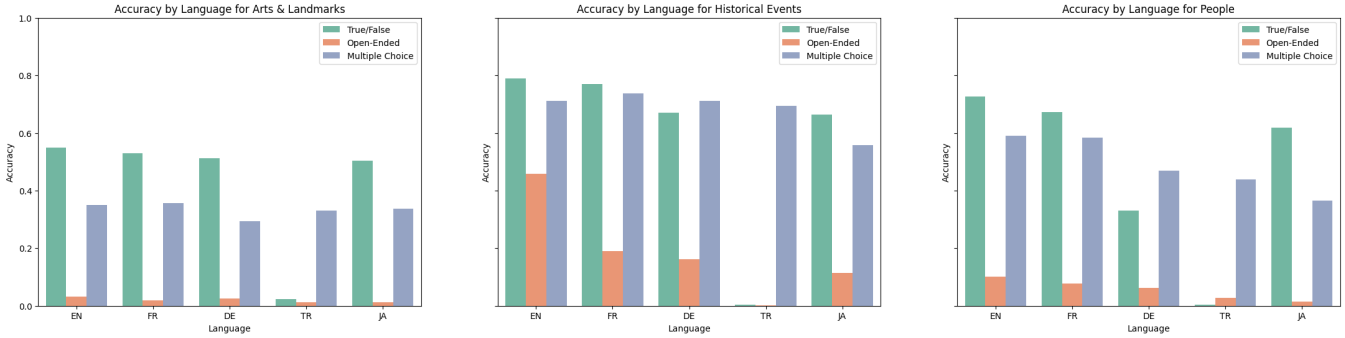

Figure 18: Accuracy per Language and per Prompt Type for Mistral-7B-Instruct