An Analysis of Multilingual FActScore

Vu Trong Kim[1], **Michael Krumdick**[2], **Varshini Reddy**[2]
Franck Dernoncourt[3], **Viet Dac Lai**[2,3]*
[1]KAIST [2]Kensho Technologies [3]Adobe Research
kim_vu_010801@kaist.ac.kr
{michael.krumdick,varshini.reddy,viet.lai}@kensho.com
{franck.dernoncourt,daclai}@adobe.com

Abstract

FActScore has gained popularity as a metric to estimate the factuality of long-form texts generated by Large Language Models (LLMs) in English. However, there has not been any work in studying the behavior of FActScore in other languages. This paper studies the limitations of each component in the four-component pipeline of FActScore in the multilingual setting. We introduce a new dataset for FActScore on texts generated by strong multilingual LLMs. Our evaluation shows that LLMs exhibit distinct behaviors in both fact extraction and fact scoring tasks. No LLM produces consistent and reliable FActScore across languages with varying levels of resources. We also find that the knowledge source plays an important role in the quality of the estimated FActScore. Using Wikipedia as the knowledge source may hinder the true FActScore of long-form text due to its limited coverage in medium- and low-resource languages. We also incorporate three mitigations to our knowledge source that ultimately improve FActScore estimation across all languages.

1 Introduction

Recent advancements in LLMs have demonstrated significant capabilities (Brown et al., 2020; Chowdhery et al., 2022; Anil et al., 2023; Team, 2024; OpenAI, 2024) in many applications (Zhao et al., 2023). Despite this advancement, LLMs remain prone to generate false information in response to information-seeking queries (Huang et al., 2023; Min et al., 2023). To address this critical problem, LLMs have been trained at unprecedented scales (Brown et al., 2020; Chowdhery et al., 2022) to cope with the massive world knowledge and aligned to reduce hallucination (Shi et al., 2024; Chuang et al., 2024; Dhuliawala et al., 2023). To further prevent the generation of false information, the Retrieval Augmented Generation method provides retrieved documents from trustworthy sources to the LLM (Ram et al., 2023; Yu et al., 2023b).

FActScore was introduced to estimate the factuality of generated texts automatically (Min et al., 2023) and at a low cost by combining LLM-as-a-judge scoring (Zheng et al., 2024) with existing reliable knowledge sources such as Wikipedia. FActScore has been enhanced to incorporate a larger knowledge base, like the internet, and to utilize more powerful retrieval models such as Google Search, resulting in better estimation across a larger domain coverage (Wei et al., 2024).

With the rapid development of multilingual LLMs (01.AI et al., 2024; Aryabumi et al., 2024), many more people are interacting with LLMs in an increasingly diverse set of languages. Hence, there is a crucial need to monitor and improve the factuality of texts beyond just the English language, making it helpful and safe for users across the entire world (Huang et al., 2023; Ji et al., 2023).

In this paper, we study the feasibility of the FActScore pipeline (Min et al., 2023) in a multilingual setting. The FActScore pipeline consists of multiple components: a knowledge source, a retrieval model, an LLM-based fact extractor, and an LLM-based fact scorer. We aim to scrutinize each component individually to identify bottlenecks and address these issues. However, there is no existing multilingual dataset for evaluating FActScore besides the original English-only dataset published by Min et al. (2023). To bridge this gap, we annotate a new **native** dataset of factuality in 3 non-English languages representing high-, medium-, and low-resource levels. This dataset is created on the texts generated by strong multilingual LLMs, i.e., GPT-4 and Gemini-Pro-1.0. We find that all evaluator models show decreased FActScore accuracy in lower-resource languages. We attribute this to several components. First, the performance of fact extraction, the simplest task in the FActScore pipeline,

*This work was mainly done when the author was employed by Kensho Technologies.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4309–4333
November 12-16, 2024

deteriorates with lower resource languages. To address this issue, we finetuned an open-source LLM for this task and achieved better performance than GPT-3.5. Second, the quality of the knowledge source is crucial to the overall accuracy of FActScore. Higher resource languages typically have Wikipedia pages with higher quality and coverage, leading to better FActScore estimation. Using the Internet as the knowledge source (Wei et al., 2024), therefore, has the greatest impact on improving the accuracy of FActScore estimation in medium and low-resource languages.

Our contributions are as follows:

- We annotated a new **native** dataset on the text generated by 2 strong multilingual LLMs in 3 languages for the multilingual FActScore task.
- We highlighted the importance of selecting knowledge sources in evaluating FActScore in the multilingual setting due to the variation in the quality of the knowledge sources in different languages.
- We found that increasing the quality of the knowledge source, either by utilizing the Internet or even another LLM's internal knowledge, has a great impact in improving the FActScore accuracy in all languages.

2 Related Work

With the advancement of language model development, numerous methods have been proposed to assess their factual alignment. A significant portion of these efforts involves using questions and corresponding short answers (Lin et al., 2021; Li et al., 2023), slot-filling (Cheng et al., 2023) task related to specific pre-collected factoids, however, they do not reflect practical use cases (Huang et al., 2023). Instead, directly assessing open-ended generated texts offers a clearer signal of the level of factuality in real use cases (Huang et al., 2023). Min et al. (2023) estimate the FActScore of biographies generated by LLMs by evaluating individual candidate facts in the text. Wei et al. (2024) extended topic coverage and utilized the Google API to query references for evaluation, thereby accessing a broader range of domains. Our study builds heavily on these approaches, focusing on the effectiveness of FActScore across high-, medium-, and low-resource languages. In these scenarios, both the language models' performance in each component of the evaluation pipeline and their multilingual capabilities are critical. Other approaches rely on language models' internal knowledge pools for factuality assessment (Azaria and Mitchell, 2023; Dhuliawala et al., 2023). While this approach offers simplicity, it raises concerns about the intrinsic factual alignment of these evaluators.

Considering multilingual factuality, X-FACTR (Jiang et al., 2020) and MLAMA (Kassner et al., 2021), adapted from LAMA (Petroni et al., 2019), assess models' relational knowledge through the "fill-in-the-blank" task. X-Fact (Gupta and Srikumar, 2021) releases a multilingual fact-checking benchmark, a factual correctness classification task covering various topics and 25 typologically diverse languages across 11 language families. Qi et al. (2023) introduces an extension of MLAMA and X-FACTR and a new metric to assess the cross-lingual consistency of language models. While these attempts shed light on multilingual factuality alignment, they mainly involve pre-collected sets of factual statements. Our work aims to evaluate the factuality of open-ended text generation.

Shafayat et al. (2024) adapted FActScore for a multilingual context by translating the biographies to English. Our work investigates both translation and performing the entire FActScore pipeline directly in the reference language. We also designed a comprehensive set of biographies to better capture the cultural proclivities of the target population.

3 Tasks & Resources

In this work, we evaluate the FActScore in multilingual settings using two resources: a translated annotation from previous work and a new native annotation.

3.1 Tasks

The FActScore pipeline (Min et al., 2023) consists of two main steps:

Atomic Fact Extraction that employs an extractor $\mathcal{E}$ to break a long-form biography x generated by a subject LLM $\mathcal{M}$ into atomic candidate facts $A^{\mathcal{E}}(x) = \{a_i^{\mathcal{E},x}\}$

Factuality Scoring is a binary classification task. It employs an evaluator $\mathcal{V}$ assigning a binary (*supported/not supported*) label $y_i^{\mathcal{E},x,\mathcal{V},\mathcal{C}}$ to every candidate fact a_i based on a knowledge source $\mathcal{C}$.

The final FActScore estimates the precision of

the generated biographies $\mathcal{X}$:

$$f_{\mathcal{C},\mathcal{V}}(\mathcal{E},x) = \frac{1}{|A^{\mathcal{E}}(x)|} \sum_{a_i \in A^{\mathcal{E}}(x)} \mathbb{1}(a_i)$$

$$\text{FActScore}_{\mathcal{E},\mathcal{C},\mathcal{V}}(\mathcal{M}) = \mathbb{E}_{x\in\mathcal{X}}[f_{\mathcal{C},\mathcal{V}}(\mathcal{E},x)]$$

3.2 Translated Annotation (en → X) (R1)

The original FActScore published a set of biographies $\mathcal{X}^{\mathcal{M}}$ generated by several subject LLMs $\mathcal{M}$ and their corresponding FActScore (Min et al., 2023) with full annotation of atomic fact and supporting label pairs $(a_i^{\mathcal{E},x}, y_i^{\mathcal{E},x,\mathcal{V},\mathcal{C}})$. We use Google Translate to translate each atomic fact $a_i^{\mathcal{E},x}$ in English into every other target language t to produce a newly translated annotation $(a_i^{\mathcal{E},x,t}, y_i^{\mathcal{E},x,\mathcal{V},\mathcal{C}})$. The knowledge source $\mathcal{C}$ (written in English) is also translated into corresponding target languages. We select a set of target languages (X) in 3 groups: high-resource (i.e., French (fr), Spanish (es), Chinese (zh-cn), Russian (ru), and Vietnamese (vi)), medium-resource (i.e., Arabic (ar) and Hindi (hi)), and low-resource (i.e., Bengali (bn)).

3.3 Native Annotation (R2)

The translated annotations are able to provide some insights into potential issues with FActScore in the multilingual setting. However, they provide a confounding factor: cascading errors due to issues with the translations themselves. This is especially relevant for low-resource languages. Therefore, we also annotate new FActScore data in non-English languages to better estimate FActScore and explore the issues of this task. In particular, we aim for a broad language coverage spanning high-, medium-, and low-resource languages. We investigated one language across each of these resource categories: Spanish, Arabic, and Bengali, respectively.

Following Min et al. (2023), we carefully curated a set of biographies for each language are from 4 geographical regions and 5 levels of rarity (See Appendix A). We attempted to use the same generative models as in (Min et al., 2023). However, these models are not explicitly designed to be multilingual and as a result, could not generate biographies of an acceptable quality, specifically in the low-resource language. To address this, we analyze the performance of explicitly multilingual LLMs, i.e., GPT-4 (GPT4) and Gemini Pro (GemP) to generate biographies.

We hired 2 native annotators for each language and followed the same annotation guidelines by Min et al. (2023) to evaluate the true FActScore of generated text. The Kappa agreement scores between Spanish, Arabic, and Bengali annotators are 79.8, 73.1, and 80.2, respectively. These show a substantial agreement (61-81) to close to almost perfect agreement (81-100) between native annotators.

	Subject	#Bios	R	I	A	#Facts	FActScore	
							WN-1	WN-All
es	GemP	100	62	27	11	79.4	67.20	70.77
	GPT4	100	72	1	27	81.4	82.86	86.83
ar	GemP	100	61	37	2	70.9	59.27	61.81
	GPT4	100	63	8	29	61.8	74.34	78.10
bn	GemP	100	60	40	0	58.9	58.77	60.55
	GPT4	100	68	29	3	46.4	71.95	74.48

Table 1: Statistics of the generated biographies by **GemP** and **GPT4** including the percentage of Relevant (**R**), Irrelevant (**I**), Abstain (**A**) biographies; the average number of atomic facts in relevant generated biographies; and FActScore evaluated by native speakers using one native Wikipedia page (WN-1) and the whole native Wikipedia (WN-All). Note that the FActScore is computed on the relevant generated texts only.

Table 1 presents the statistics of generated biographies by both subject models. Both models generate more candidate atomic facts in higher-resource languages than lower-resource languages, however, this phenomenon seems to be clearer with GPT4. GPT4 generates more relevant biographies than GemP in all three languages. GPT4 also abstains significantly more than GemP in Spanish and Arabic, whereas GemP produces many more irrelevant biographies. This shows that GPT4 has a broader knowledge and a higher awareness of its knowledge limitation. In terms of FActScore, GPT4 yields much higher FActScore(s) than GemP in all three languages, using either a single Wikipedia page or the whole Wikipedia with an average margin of approximately 14.6%. Last but not least, FActScore(s) evaluated based on the whole Wikipedia (WN-All) are higher than FActScore evaluated on a single Wikipedia page (WN-1) in all cases (on average 3%). This suggests that a larger knowledge source gives a higher FActScore. In other words, the knowledge source is the ceiling of evaluating factuality.

4 Experiments

4.1 Atomic Fact Extraction

FActScore decomposes a long-form text into multiple atomic statements, each containing a single

piece of information. The original methodology uses few-shot demonstrations to prompt InstructGPT for this task (Min et al., 2023). We examine the performance of different models and pinpoint issues of existing models for this task.

Settings: Due to the higher quality of text generated in English, prior work by Min et al. (2023) only considered if the candidate facts need to be merged or split, mainly concerning whether the facts are atomic. However, in a multilingual setting, the texts generated by LLM may contain other kinds of errors where the facts need to be merged or split, not grounded, duplicated, missing some information, and linguistic errors.

We choose GPT-3.5 (GPT3.5), GPT4, and Gemma for evaluation in this task. These models were selected for their best performance via a pilot study on a small subset of R1 (See Appendix B). We evaluate the GPT3.5 and GPT4 as few-shot In-Context Learning while Gemma is further supervised finetuned for this task. In particular, we finetune Gemma on 42k pairs of (sentence, extracted atomic facts) derived from R1. Then these three models are evaluated on a subset of 200 sentences, sampled randomly from R2 with a 1:1 ratio facts generated by GPT4 and GemP.

Results: Table 2 shows the number of errors made by 3 models (GPT3.5, Finetuned Gemma, and GPT4). Among these three models, GPT4 is the best model by a relatively large margin across all three languages. Finetuned Gemma is competitive to GPT3.5 in high-resource and better in low-resource and medium-resource languages.

However, GPT4 and GPT3.5 's performances deteriorate rapidly with low-resource language (approximately double the average error rate in Bengali, compared to Spanish and Arabic). On the other hand, the FT Gemma does not show a performance reduction in low-resource language. In fact, its error rate in Bengali is lower than those in Spanish and Arabic. This suggests that finetuning has potentially helped this model maintain a steady performance across all resource languages.

More importantly, due to the better performance of LLMs in English, Min et al. (2023) did not consider other types of errors that may happen in multilingual settings. In particular, we see a large number of grounding errors in medium and low-resource languages (Arabic and Bengali), while we don't see that in high-resource languages such as Spanish. LLMs also missed some detailed information in the given generated text in this task.

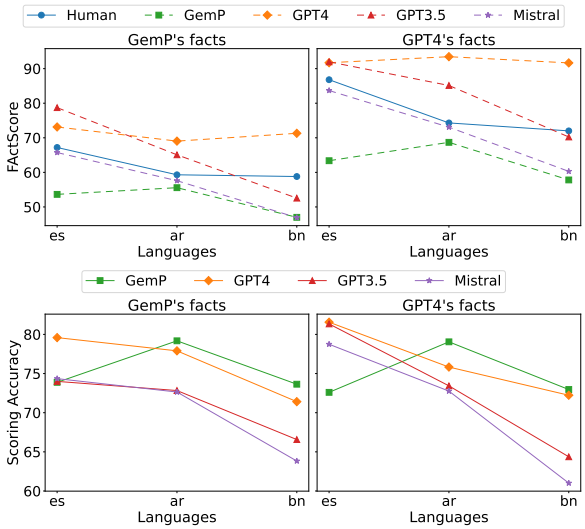

Figure 1: FActScore (upper) and Scoring Accuracy (lower) predicted by 4 scorers (GPT4, GemP, GPT3.5, Mistral) in comparison with FActScore by human (R2) on texts generated by GPT4 and GemP in native languages.

4.2 Factuality Scoring

This section investigates the feasibility of using LLMs as factuality scorers in multilingual settings.

Settings: We use GPT4 to extract facts from biographies generated by two subject models namely GPT4 and GemP to provide the same denominator for this evaluation. We evaluate 4 LLMs as fact scorers (GPT3.5, GPT4, Mistral, and GemP) on the text generated by GPT4 and GemP in native languages. The human-annotated dataset (R2) is used as the ground truth.

Results: Figure 1 (upper) shows the FActScore predicted by LLMs and by humans (R2). GemP consistently underestimates FActScore, whereas GPT4 significantly overestimates FActScore across both subject models. GPT3.5 overestimates Spanish and Arabic while closely estimating FActScore for Bengali. On the other hand, Mistral closely estimates FActScore for Spanish and Arabic while substantially underestimating the FActScore for Bengali. This experiment suggests that none of these models offers a reliable FActScore across the whole spectrum of languages, even with strong LLMs (e.g., GPT4 and GemP).

Figure 1 (lower) shows the scoring accuracy of the LLM scorers. GemP shows a steady accuracy on both GPT4 and GemP facts. Its accuracy does not show a clear dependency on the resource level.

Lang.	Extractor	#Sent	Need Merge	Need Split	Not Grounded	Duplication	Missing Information	Linguistic Error	Average ↓
es	GPT3.5	192	1	24	0	9	14	0	0.25
	FT Gemma	192	0	35	0	0	14	0	0.26
	GPT4	192	1	2	0	3	9	0	**0.08**
ar	GPT3.5	180	1	22	10	2	13	3	0.28
	FT Gemma	180	0	20	10	0	14	2	0.26
	GPT4	180	3	0	5	0	8	0	**0.09**
bn	GPT3.5	175	3	36	19	12	24	8	0.58
	FT Gemma	175	2	22	5	2	7	1	0.22
	GPT4	175	2	9	3	2	8	0	**0.14**

Table 2: Fact Extraction: Total number of errors by categories and the average number of errors per sentence on texts generated by GPT4 and GemP. The descriptions of the errors are presented in Appendix I

On the other hand, the accuracy of GPT4, GPT3.5, and Mistral decreases in turn with the level of language resources. In particular, GPT3.5 and Mistral's accuracy decreases at a steeper pace than GPT4's. Further discussion on this component will be provided in Section 5.

4.3 Knowledge Source

Since FActScore is a function of knowledge source (Min et al., 2023), the quantity and quality of the information of the knowledge source greatly affect the subsequent score (Wei et al., 2024). This section investigates the sensitivity of FActScore to changes in the underlying knowledge sources.

Settings: We collected 32 biographies of entities per language in four categories of popularity and geographical relevance: *internationally popular*, *internationally unpopular*, *locally popular*, and *locally unpopular* (See Appendix A). The annotators evaluate facts using three different sources: the native Wikipedia, the English Wikipedia, and the whole Internet. Since the Internet is a superset of knowledge sources, we considered the annotations created with access to the Internet as the golden annotations for evaluating the quality of other knowledge sources.

Results: Figure 2 shows the scoring accuracy between evaluating 4 categories of popularity in 3 languages. Using Spanish Wikipedia pages yields higher accuracy in labeling locally popular figures (L+P), whereas English Wikipedia pages are better for internationally unpopular entities (I+UP). For Arabic, the Arabic Wikipedia is better for local popular entities (L+P), while the English Wikipedia is better for international entities (I+P and I+UP). For Bengali, the Bengali Wikipedia has a much lower performance compared to the English counterpart in all four categories, especially for the international entities (I). This suggests that Bengali Wikipedia has a very low coverage, inadequate for most cases. Last but not least, even though English pages provide better coverage for local entities (L+P and L+UP) than Bengali pages, the scoring accuracies using English pages for Bengali local entities are still lower than those of international entities. These differences in performance between international and local figures highlight the importance of choosing local entities and local knowledge sources in multilingual FActScore evaluation and estimation.

4.4 Retriever

Due to the limitation of the LLM context length, a Wikipedia page of the evaluated entity is split into short passages. A retriever model retrieves k relevant passages. These passages are used as reference knowledge sources.

Settings: We examine both traditional retrieval method (i.e., BM25) and vector-based retrieval method. In particular, we use the multilingual embedding models (*distiluse-base-multilingual-cased-v2* and *paraphrase-multilingual-MiniLM-L12-v2*) (Reimers and Gurevych, 2019) to encode the texts. For each translated fact, $k = 5$ retrieved passages are retrieved out of all passages. We measure the Recall@k and the average hit rate of the top 1 and top 2 passages.

Results: Table 3 reports the retrieval performance in Recall@k of the retrieval models on 9 languages of 3 resource-level groups. For vector retrievals (Distill and Paraphrase), we see a gradual drop of performance as the resource is scarcer. While the performance of vector-based retrievals for medium languages is slightly worse than ones on high-resource languages, their performance on low-resource is significantly lower than

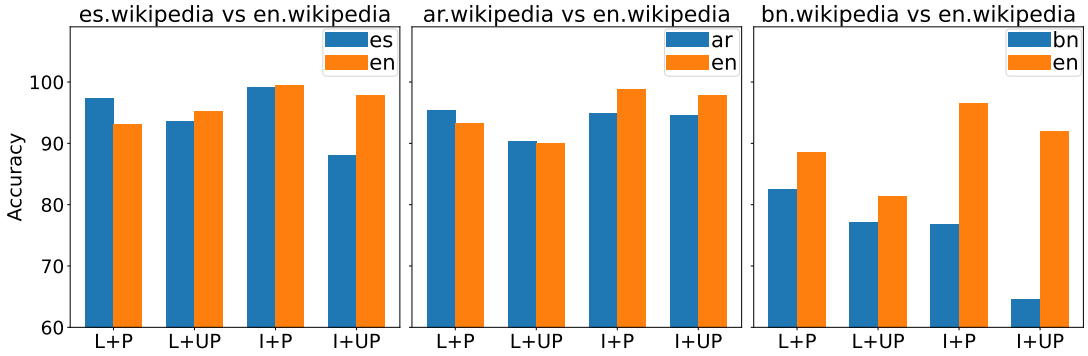

Figure 2: Accuracy of Factuality Scoring task with different knowledge sources. **L** stands for Local/Domestic, while **I** stands for International. **P** stands for Popular and **UP** stands for UnPopular.

Resource	Lang	Distill	Paraphrase	BM25
High	en	67.7	65.8	67.0
	fr	66.0	63.8	66.5
	es	66.1	63.8	66.3
	ru	66.1	63.2	62.2
	zh-cn	64.7	63.8	59.3
	vi	65.4	63.4	67.4
Medium	ar	63.4	60.9	62.8
	hi	61.5	61.4	64.7
Low	bn	51.2	54.7	61.8

Table 3: Retrieval performance of the retriever in Recall@k (%).

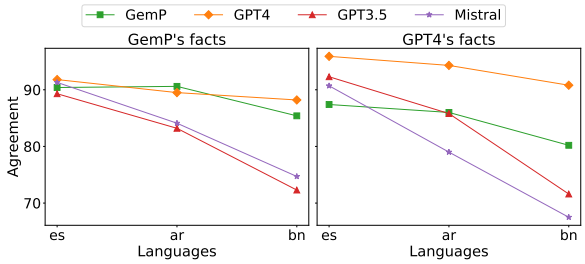

Figure 3: Prediction agreement between two variants of facts (in target language and in translated English).

the medium- and high-resource languages. On the other hand, even though BM25 also offers lower performance for lower-resource languages, it does not see a significant drop as the neural-based models.

5 Discussion

5.1 Would translation help?

A simple method for a multilingual FActScore is first translating non-English long-form text and knowledge sources into English (**X→en**) and estimating the FActScore on these proxy translated English texts (Shafayat et al., 2024). This is a promising method given that the quality of machine translation has improved significantly in the last decade. To do this, we translated the Native Annotation (R2) into English to get a translated English annotation (R3).

Figure 3 shows the prediction matching for the Factuality Scoring task on texts in the target language and in translated English. GemP and GPT4 are the two strong scorers with consistently high matching, GPT3.5 and Mistral have significantly lower matching scores in lower-resource languages. Additionally, GPT4 and GemP see a slighter decline in matching scores for lower-resource languages than GPT3.5 and Mistral. This matching variation across different languages for this task among even the most advanced LLMs may lead to unreliable FActScore estimation in lower-resource languages.

Figure 4 (lower) compares the scoring accuracy between using R3 and using R2. We see a significant improvement in scoring accuracy for Mistral and GPT3.5 in Arabic and Bengali and GemP in Bengali, all on both GPT4 and GemP texts. We attributed this to both better reading comprehension and retrieval performance in English compared to non-English languages, especially Bengali. Appendix D explores the impact of translation on retrieval performance in more detail. On the other hand, we see a significant decline in the accuracy of the scorer GPT4 on GemP's texts for all three languages while a slight increase in the accuracy in Arabic on GPT4 texts.

Figure 4 (upper) shows the FActScore predicted by these models in native texts and translated English texts. The translation contributes to the overestimation of FActScore by GPT3.5 and Mistral in medium and low-resource languages. On the

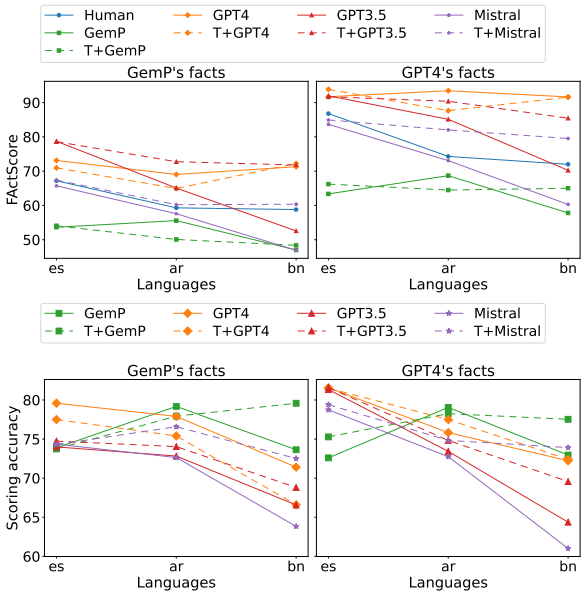

Figure 4: FActScore (upper) and Scoring accuracy (lower) by fact scorers with and without translation in comparison with FActScore by human (R2) on texts generated by GPT4 and GemP. Dash lines denote the translation being used, along with corresponding scorers.

other hand, translation has little effect on stronger scorers such as GPT4 and GemP. This suggests that these models are more consistent in understanding both English and non-English texts.

5.2 Error analysis

Figure 4 shows significant differences in the factuality-scoring task remain between the most advanced model evaluators, i.e., GPT4 and GemP, and native speakers. We conducted an error analysis to investigate the categories of these disagreements. For each language and each subject model, we randomly select 60 disagreement samples between LLMs and humans. We manually inspect this to identify the primary disagreement.

Table 4 reports the raw number of errors. The primary cause of errors by the scorer GPT4 is contextual unfaithfulness, accounting for 73% of the errors across 3 languages and 2 subject models. This issue is more severe in lower-resource languages. However, many contextually unfaithful samples are factually correct according to other knowledge sources beyond the given Wikipedia page. This suggests that GPT4 uses its internal knowledge in the Factuality Scoring task. Appendix F further discusses the behaviors of GPT4 as a scorer. The scorer GemP has a much lower contextual unfaithfulness error (especially factually correct)

	Subject	Context Unfaithful			Others			
		FC	Hal.	RD	Ret.	Tab.	Deb.	Misc.
(a) Fact Scorer: GPT4								
es	GemP	12	4	2	1	2	4	5
	GPT4	17	6	0	1	0	2	4
ar	GemP	14	5	3	4	1	0	3
	GPT4	17	4	2	0	1	3	3
bn	GemP	18	1	0	3	2	0	6
	GPT4	22	3	2	0	2	0	1
(b) Fact Scorer: GemP								
es	GemP	8	0	5	10	2	3	2
	GPT4	4	0	4	8	6	4	4
ar	GemP	4	1	7	4	6	1	7
	GPT4	7	1	6	3	6	3	4
bn	GemP	5	2	5	4	7	3	4
	GPT4	7	2	6	4	5	4	2

Table 4: Error analysis: Factually Correct (FC), Hallucination (Hal.), Reading Deficiency (RD)), Retrieval Error (Ret.), Tabular Data (Tab.), Debatable (Deb.), and miscellaneous error (Misc.))

compared to GPT4. However, GemP makes more errors due to retrieval errors and tabular data. This shows that GemP is more context-dependent and less internal-knowledge-dependent for the Factuality Scoring task.

6 Mitigations

The previous sections have shown evidence of a correlation between lower resource languages, lower retrieval performance (See Table 3), lower coverage of the native knowledge source (See Figure 2) and subsequently lower fact scoring accuracy (See Figures 1 and 3). To mitigate this problem, we empirically examine three techniques including: improving retrieval performance by (1) increasing the number of retrieved passages, (2) employing language models as Internet search agents and evaluators (Wei et al., 2024), and (3) using language models as a knowledge generator (Yu et al., 2023a; Chen et al., 2023)).

Settings: We use GemP as the fact scorer for all proposed techniques. GemP is more persistent to the change in languages (as shown in Figure 1). It is more sensitive to external knowledge than its internal knowledge (Section 5.2), making it more suitable for evaluating these mitigations than GPT4. The baseline is the original pipeline (Min et al., 2023) with GemP as the scorer and Wikipedia pages in native languages as the knowledge sources.

We use the 32 generated biographies in the three studied languages that we used to assess knowledge

sources in section 4.3. We consider the facts annotated by native speakers using the whole internet as the golden data. We evaluate these techniques by measuring their scoring accuracy with the golden labels. Table 5 illustrates the performance of the proposed methods.

6.1 Expanding Retrieved Passages

This method increases the number of retrieved passages from 8 to 20, aiming to extend the amount of information given to the scorer. This mitigation should alleviate the impact of poor recall in retrieval. Although the mildest of the three mitigations, this led to a considerable increase in performance across all three languages. The performance gap is particularly large in Bengali, correlating with observations in Section 4.4 regarding the retriever's deteriorating performance in this language. This retrieval problem might be further mitigated thanks to the increase in context length of recent language models (Xiong et al., 2023) allowing feeding more information to the LLM-based scorer.

6.2 Internet as a knowledge source

Adapted from Wei et al. (2024), GemP is prompted to send queries to the Google Search API on a given fact and determine the fact's factual accuracy from the query results. We see a clear improvement in fact-scoring accuracy and higher FActScore (closer to the golden) across the subject models and languages. For example, the accuracy on Bengali improved from 60.6 to 86.8. This shows the benefit of accessing a larger pool of information results in substantial improvement, much greater than merely increasing the number of passages from Wikipedia.

6.3 LLM as a knowledge source

Since previous experiments suggested that GPT4 heavily relies on its internal knowledge to assess factuality, we experiment with allowing GPT4 to directly augment the low-coverage knowledge source. We prompt GPT4 to create a question based on a given fact and then generate related information to answer that question (Yu et al., 2023a). This generated knowledge is combined with retrieved passages, as suggested by (Yu et al., 2023a), and used with a separate evaluator, GemP, for factual labeling. It is worth noting that this text is entirely unverified and likely contains some amount of factual errors.

This approach results in a substantial improvement, larger than that of simply increasing the number of Wikipedia passages across all languages. Compared to using the Google Search API, the GPT4 augmented knowledge base shows higher gains in high- and medium-resource languages. This suggests the reliability of GPT4's internal knowledge and its effectiveness as a knowledge generator. However, in Bengali, querying evaluation references via Google API yields significantly better factual labeling. The improvement from using GPT4's internal knowledge is attributed to the additional relevant information that it provides.

	Method	FActScore		Accuracy	
		GemP	GPT4	GemP	GPT4
es	GemP+Wiki (k=5)	58.8	68.4	70.3	69.8
	GemP+Wiki (k=8)	63.3	72.8	74.3	74.0
	GemP+Wiki (k=20)	67.0	77.3	77.3	78.5
	GemP+Google API	81.5	90.3	83.2	89.9
	GemP+GPT4's IK	78.8	93.5	84.6	91.9
	Human+Wiki	75.3	88.4	91.9	90.1
	Human+Internet	82.9	97.3	-	-
ar	GemP+Wiki (k=8)	56.4	73.7	77.9	78.3
	GemP+Wiki (k=20)	60.6	76.7	79.9	81.1
	GemP+Google API	72.3	87.7	80.3	86.3
	GemP+GPT4's IK	64.4	82.4	83.6	84.6
	Human+Wiki	60.9	81.9	90.8	89.7
	Human+Internet	69.2	90.5	-	-
bn	GemP+Wiki (k=8)	43.8	53.0	60.6	55.1
	GemP+Wiki (k=20)	52.6	63.5	69.1	65.7
	GemP+Google API	75.6	88.0	86.8	87.8
	GemP+GPT4's IK	59.4	70.0	74.8	71.1
	Human+Wiki	57.8	61.2	74.1	62.7
	Human+Internet	82.0	97.5	-	-

Table 5: FActScore and accuracy of introduced evaluation methods on GemP and GPT4's generated facts. We use GemP as the LLM scorer. +Wiki (k=x) denotes using x passages from 1 Wikipedia page as references. +Google API denotes using GemP as the Internet search agent and evaluator (evaluation is based on query results). +GPT4's IK denotes using GPT-4's generated Internal Knowledge (IK) and retrieved passages as references. Natives+Wiki/Internet denotes natives, using 1 Wikipedia page or the entire Internet as references for annotations. Natives+Internet is considered as golden labeling to conclude accuracy.

6.4 Error analysis

We further conduct an error analysis for the fact-scoring task with these improvements and report in Table 6. The result shows that all three approaches reduce false negatives (and thereby increase true positives) due to their ability to provide more factual coverage.

Surprisingly, the unverified LLM augmented wikipedia articles significantly increase the true positive rate (by 12.9%, 6.8%, and 14.9% for GemP

	Method	GemP				GPT4			
		TP	FN	FP	TN	TP	FN	FP	TN
es	Wiki (k=8)	60.3	22.6	3.0	14.1	72.1	25.2	0.8	1.9
	Wiki (k=20)	63.6	19.3	3.4	13.7	76.5	20.8	0.8	1.9
	Google API	73.8	9.1	7.7	9.4	88.7	8.5	1.6	1.1
	GPT4's Internal Knowledge	73.2	9.7	5.6	11.4	91.3	6.0	2.2	0.6
ar	Wiki (k=8)	52.1	17.7	4.4	25.9	72.1	20.1	1.6	6.2
	Wiki (k=20)	55.1	14.6	5.4	24.8	75.0	17.2	1.7	6.1
	Google API	60.8	8.5	11.5	19.2	81.7	8.8	6.0	3.5
	GPT4's Internal Knowledge	58.9	10.9	5.5	24.7	79.6	12.6	2.8	5.0
bn	Wiki (k=8)	43.2	38.8	0.6	17.4	52.8	44.7	0.2	2.3
	Wiki (k=20)	51.8	30.2	0.7	17.3	63.4	34.2	0.2	2.3
	Google API	72.2	9.8	3.4	14.6	86.6	10.9	1.3	1.2
	GPT4's Internal Knowledge	58.1	23.9	1.3	16.7	69.3	28.2	0.7	1.8

Table 6: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) rates for different FActScore pipelines that use GemP as the scorers.

on es, ar, and bn respectively) without in turn significantly increasing the false positive rate (by 2.6%, 1.1% and 0.7% respectively). The increase in false positives was lower than using the Google API in all but one case. Conversely, adding additional Wikipedia data always leads to a lower rate of false positives compared to the GPT4 augmented data but also a lower rate of true positives. This implies that the benefits of increased factual coverage from using the unverified GPT4-generated data outweigh the costs of potentially false information introduced. However, these benefits diminish for lower-resource languages, while using the Google API shows more consistent gains across all languages.

7 Conclusion

This paper scrutinizes the FActScore pipeline for long-form generated texts in the multilingual setting. We generated new fact candidates and annotated a new corpus for FActScore evaluation. The most recent open-source LLMs struggle with the atomic fact extraction task. Finetuning on this task can match the performance of much larger close-source models, e.g., GPT3.5. More importantly, the Fact Scoring task is very sensitive to the coverage of the knowledge source. Although Wikipedia is reliable, it lacks coverage in lower-resource languages, which leads to a severe underestimation of the FActScore. We show that mitigation such as extending the knowledge source through increasing the amount of Wikipedia data, allowing access to the Internet, and even augmenting low-coverage Wikipedia articles with unverified text generated by an LLM improve multilingual FActScore estimation.

Acknowledgement

We thank Chris Tanner for his comments and incredible support throughout the project and Mohit Iyyer for his valuable feedback. We thank anonymous reviewers for their constructive feedback and comments.

Limitation

Even though this paper offers insights into the multilingual FActScore, the paper was not able to address more languages than the 3 examined languages and on a larger sample size due to funding limits and the extremely high cost of this task as reported in previous work (Min et al., 2023; Wei et al., 2024). As a result, the data might contain cultural biases and variations in information and knowledge exposure. Therefore, generalizing our findings to languages other than the examined ones should be considered carefully. Due to the rapid development of LLMs when the study was done, some models might be obsolete by the publication time, however, we believe this paper still provides insightful knowledge into multilingual factuality scoring.

Ethical Consideration

In this work, we hire 6 international crowd-sourced workers from 3 countries as native annotators. The annotators were paid between US$15 to US$25 per hour, adjusted to their geographical location.

While the biographies generated by the two subject models exhibit a certain level of factuality, we observed a significant amount of false information. Using these biographies as references or in real-world scenarios carries the risk of spreading misinformation and negatively impacting the individuals whose biographies are studied.

All the systems presented in this paper do not offer a perfect factual guarantee, especially with the texts and knowledge beyond the studied scope. These systems should not be used as alternate tools for traditional factual verification methods.

Given the nature of this task which involves assessing human biographies generated by LLMs, our collected data includes identifications, information, and opinions about them, including false and biased content. We only share the generated texts upon request to enhance the proper use of the data and minimize the risk of spreading false information.

References

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341, Singapore. Association for Computational Linguistics.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim,

Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection.

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

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

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

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *arXiv preprint arXiv:2402.18045*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

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

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Biography Selection

We select a set of people names from the following regions: *North America*, *Europe*, *Asia*, *Oceania*, *South America*, and *Africa*; and 5 levels of rarity based on their Wikipedia page views very frequent, frequent, medium, rare, and very rare.

In Section 4.3, four additional categories are introduced: *internationally popular*, *internationally unpopular*, *locally popular*, and *locally unpopular*. The terms *locally* and *internationally* refer to the geographical or linguistic exposure of the entities whose biographies are being factuality evaluated. Local entities might be native speakers of the language or reside in nearby regions where the language is predominantly spoken as a first language. For example, for Spanish, this includes regions such as South America and Spain. For Arabic, this includes the Arab world, and for Bengali, the Indic region. Entities deemed *popular* include those classified as *very frequent*, *frequent* or *medium* while *unpopular* encompasses *medium*, *rare*, and *very rare* entities according to rarity as introduced above according to Wikipedia page views.

B Pilot Experiments on Fact Extractor

We randomly selected 10 sentences from the original work (Min et al., 2023) and then translated them into target languages. Tested models were prompted (few-shot) to break down those sentences into individual facts. These were translated back to English for assessment based on metrics from Section 4.1.

Tables 14, 15, 16, 17, 18, 19 represents extractions of GPT4, GemP, GPT3.5, Mistral-7B-Instruct (Mistral), Llama-7B-Chat (Llama2) and Gemma-7B-Instruct respectively. All closed models are decent at the task across all studied languages. Among open models, Mistral, Llama2, and Gemma could understand the instruction and perform fact extraction, whereas Aya and BLOOMZ were lost in this task (Aya simply returns the original sentence, whereas BLOOMZ does not produce any outputs). However, in non-English languages, Llama2 shows errors even in a high-resource language like Spanish, while Gemma and Mistral begin to show errors in medium- and low-resource languages.

For native annotations with R2, we chose two closed models, GPT3.5, and GPT4, and finetuned an open-source model for the extraction task in 3 studied languages. Gemma-7B is chosen considering its large vocab size, thus saving inference costs in the multilingual context. Table 20 illustrates that the finetuned model consistently shows proper extractions across studied languages.

C Open-Source Models Performance as Scorers on More Languages

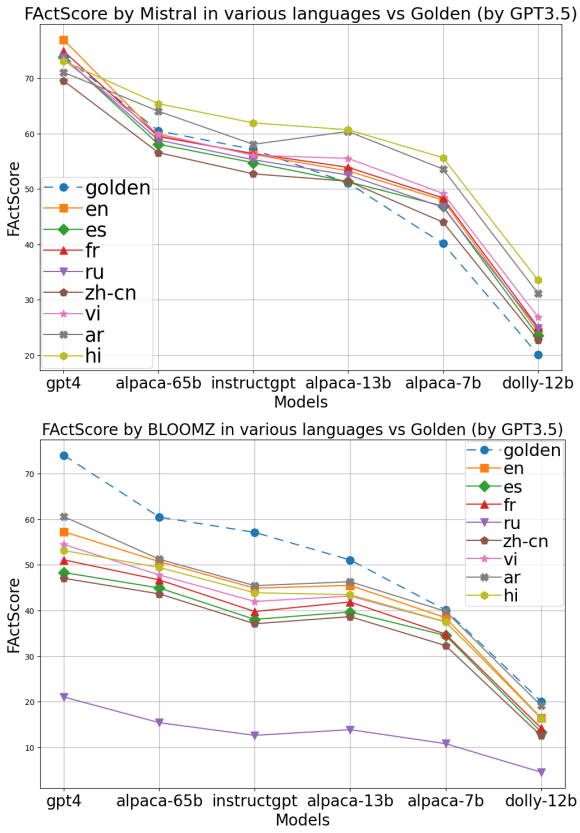

Figure 5: FActScore by Mistral and BLOOMZ on translated facts generated by studied subject models from (Min et al., 2023) (R1), compared to golden scoring by GPT3.5, as suggested by Min et al. (2023).

Figure 5 depicts the scoring of subject models by two open-source models, Mistral-7B-Instr (Mistral) and BLOOMZ-7b1 (BLOOMZ) on subject models from Min et al. (2023). Both models demonstrate

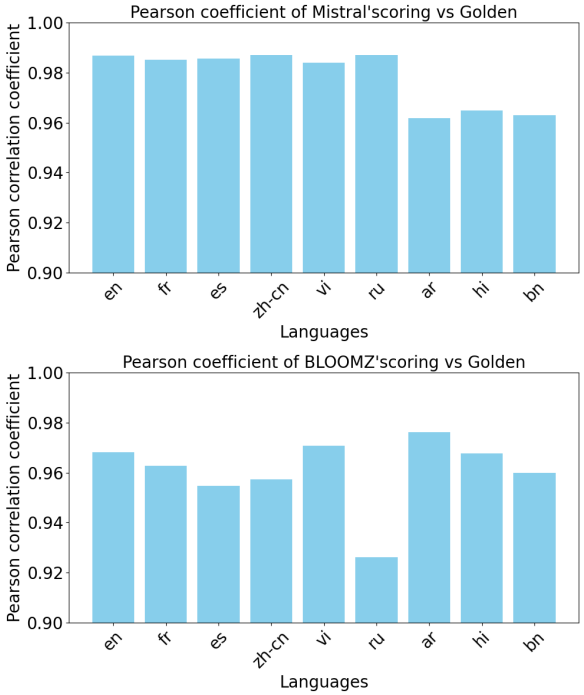

Figure 6: Pearson correlation coefficient between Mistral (up) and BLOOMZ (down) scoring on subject models from (Min et al., 2023) with that by GPT3.5 (golden labeling proposed by (Min et al., 2023)).

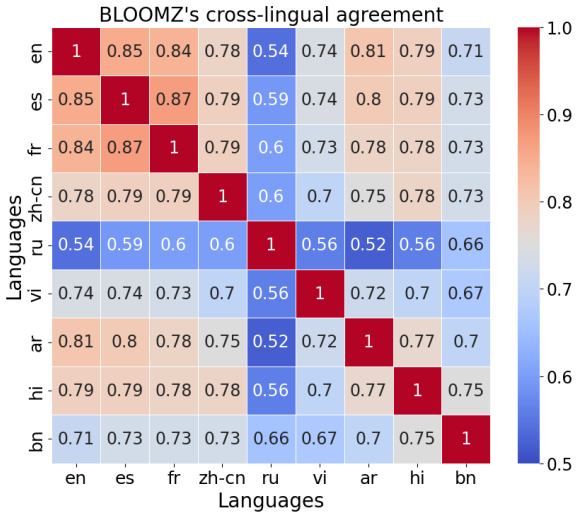

Figure 7: Cross-lingual agreement of Mistral (up) and BLOOMZ (down) when scoring different language versions of the same fact.

significant agreement in the ranking of subject models when compared to the golden labels provided in the original study (Min et al., 2023). It is important to note that the ranking order among evaluated models is the primary concern of Min et al. (2023). This is further supported by Figure 6, representing relatively high Pearson correlation coefficients of scoring by two scorers in different languages with golden labeling.

However, there are notable variations in FActScore across languages. This indicates that while the pipeline effectively operates in multilingual environments for comparing factuality alignment among language models in a particular language, it is not suitable for assessing model performances across different languages.

Figure 7 displays the cross-lingual agreement heatmap between texts written in two languages of two open-source models, i.e., Mistral-7B-Instr (Mistral) and BLOOMZ-7b1 (BLOOMZ). The first row of the heat map illustrates the labeling agreement of both models when evaluating facts in English and non-English languages. The agreement for both models decreases in correlation with the resource levels of the non-English languages. This decline is clearly observable in Mistral's heat map, but only partially in BLOOMZ's heat map. Specifically, BLOOMZ's agreement in Russian and Vietnamese is consistently lower than expected, given their high-resource status in the Common Crawl corpus. This issue is attributed to BLOOMZ's alignment training dataset, namely xP3. The xP3 dataset does not include any Russian data and contains a limited amount of Vietnamese data (2.11% in xP3), less than that for Arabic (2.72% in xP3), a lower-resource language.

Figure 8 further illustrates the cross-lingual agreement of two proprietary models, GemP and GPT3.5, with a subset of three out of the nine studied languages. The leading open-source model, Mistral, slightly trails behind GPT-3.5, with average scores of 0.83 and 0.85 respectively. However, Mistral's performance is significantly lower than that of GemP, which achieves an average score of 0.88.

D Impact of Translation on Retriever

Section 4.1 discusses the impact of translation on scoring accuracy by different scorers with clear positive effects on GPT3.5, Mistral, and GemP. However, this phenomenon might be attributed to translation's contribution to addressing the multilingual deficiency of the retriever (illustrated in Section 4.4) as well. This section explores that hypothesis by comparing the effect of translation if it is performed before (T+R) and after retrieval (R+T).

As shown in Table 7, while the difference is

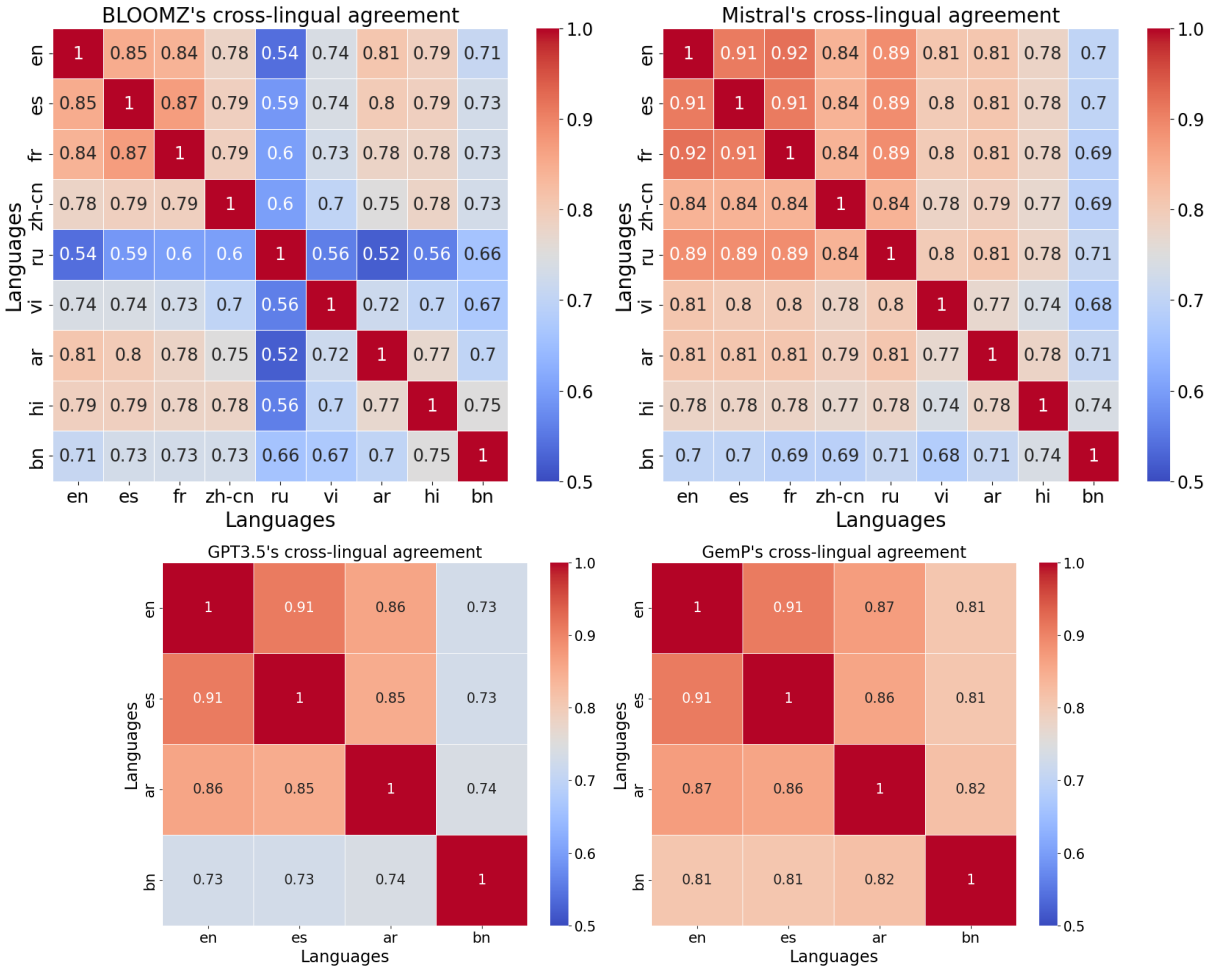

Figure 8: Cross-lingual agreement of BLOOMZ (a), Mistral (b), GPT3.5 (c), and GemP (d) when evaluating different language versions of the same fact.

	Method	FActScore		Accuracy	
		GemP	GPT4	GemP	GPT4
es	GemP	58.81	68.39	72.72	73.85
	GemP (T->R)	59.97	72.92	72.18	77.66
	GemP (R->T)	59.74	71.41	72.10	76.01
ar	GemP	56.40	73.71	80.75	78.76
	GemP (T->R)	52.22	69.42	80.25	78.33
	GemP (R->T)	50.79	70.06	80.00	77.47
bn	GemP	43.78	52.97	71.89	70.63
	GemP (T->R)	50.22	62.05	79.50	70.13
	GemP (R->T)	39.53	49.34	70.28	65.02

Table 7: FActScore and accuracy of performing translation before and after retrieval regarding regarding two metrics. Golden labels are human annotations with 1 Wikipedia page as the knowledge source.

not significant in high- and medium-resource languages, for Bengali, performing translation after retrieval (retrieval is in Bengali) significantly diminishes the benefits of translation. Consequently, using translation even results in lower accuracy compared to not using translation at all.

E GPT4's Behaviors as a Scorer

Concurrent with the discussion in Section 5.2, among context-unfaithful samples, there are also factually incorrect ones, including hallucinations and reading deficiencies. A significant portion (72%) of these factually incorrect samples contains information not found in the knowledge source, hallucination.

This category, similar to the discussed factually correct samples, lacks grounded information within the provided context, highlighting an interesting behavior of GPT-4 as a scorer. The model heavily relies on its internal knowledge during the scoring process.

This reliance may partially explain the decreasing accuracy of GPT-4's scoring in lower-resource languages, as demonstrated in Figure 1 (lower).

Specifically, Table 8 shows that the information available in the Wikipedia versions of the studied languages diminishes in correlation with their resource levels. This might result in their growing distances with the GPT4's internal knowledge. Consequently, it contributes to lower accuracy (see Figure 1 (lower)) when GPT-4 is the scorer.

Correspondingly, error analysis in Section 5.2 reveals a higher number of context-unfaithful samples in lower-resource languages. This indicates GPT-4's increased tendency to rely on its internal knowledge in more limited-resource circumstances.

Table 4 illustrates that GPT4 as a scorer is factually correct in about half of the disparity samples with native annotators. However, as shown in Table 9, excluding the retriever and knowledge source from the pipeline and relying solely on GPT4's internal knowledge leads to a decrease in factually correct evaluations overall. This implies that despite their limitations, external knowledge sources are essential for maintaining the reliability of the evaluation process.

Lang	#Facts	#Passages
es	391.6	15.5
ar	317.5	13.0
bn	277.3	12.8

Table 8: Average number of facts and passages in a Wikipedia page in three languages.

F Error Analysis Setup

For each language, we collected 60 disagreement samples, proportionally distributed according to false positives and false negatives by these model scorers against golden labels by human.

To categorize disagreement cases, we do the following steps:

- Thoroughly read the entire Wikipedia article to identify relevant text (sentences, paragraphs) for evaluating the fact and checking for annotator errors.
- If no text within the Wikipedia page relates to the fact, it should be labeled as "not supported" by annotators (or it would be a mistake from the annotator) and "supported" by the model scorer. We then proceed to evaluate the fact based on external sources and determine whether the labeling should be classified as "context unfaithful but factually correct" (if supported by external sources) or "context unfaithful and hallucinated" (if not supported by external sources).
- If related information is found within the Wikipedia page, classify the labeling disagreements as follows:
  - Tabular data: The information is in a table and has not been processed by Wikipedia's HTML conversion to text.
  - Retriever error: The information is not in the passages retrieved.
  - The information is in the retrieved passages but missed by scorers.
  - Cannot Deduct from Context: Correct evaluation of the fact, while not being explicitly specified, but deductible from the provided context, but the modeling evaluator fails to do so.
  - Subjective opinion: The labeling is hugely influenced by the annotator's subjective opinion.
- Other cases to consider:
  - Assistant Generation: If the sentence is part of the model's service generated content.

G Experimental Settings

We utilized and assessed the following models to study components of the FActScore pipeline.

Subject Models:

- GemP (gemini-1.0-pro)
- GPT4 (gpt-4-0125-preview)

Factuality Scorers:

- GemP (gemini-1.0-pro)
- GPT4 (gpt-4-0125-preview)
- GPT3.5 (gpt-35-turbo-0125)
- Mistral (mistralai/Mistral-7B-Instruct-v0.2)
- BLOOMZ (bigscience/bloomz-7b1)

Fact Extractors:

- GemP (gemini-1.0-pro)

Scorer	Reference	FActScore						Accuracy					
		es		ar		bn		es		ar		bn	
		GemP	GPT4	GemP	GPT4	GemP	GPT4	GemP	GPT4	GemP	GPT4	GemP	GPT4
Human	Internet	82.7	97.3	70.4	92.2	81.8	97.5	-	-	-	-	-	-
	Wiki (1 page)	75.4	88.4	61.6	81.5	57.5	60.2	92.1	91.1	91.7	91.1	89.3	90.4
	Wiki (All)	78.3	91.7	64.4	86.5	60.1	60.6	95.0	94.2	93.7	94.3	76.4	72.5
GPT4	w/ Wiki	78.2	91.7	68.4	93.5	74.0	91.7	86.2	91.0	85.9	93.3	86.0	90.9
	w/o Wiki	87.6	97.1	87.3	98.2	85.4	93.7	86.0	94.7	75.3	91.5	79.9	91.6
	T + w/ Wiki	78.2	94.4	65.7	91.6	73.4	86.3	84.9	93.7	84.1	92.0	84.0	86.4
	T + w/o Wiki	95.9	95.3	79.3	95.5	81.1	82.9	85.0	93.3	79.0	89.9	82.8	83.1
GemP	w/ Wiki	63.3	72.8	56.4	73.7	43.8	53.0	74.3	74.0	77.9	78.3	60.6	55.1
	w/o Wiki	79.4	84.4	74.1	88.4	77.3	86.2	76.2	82.7	68.5	82.1	72.7	84.1
	T + w/ Wiki	60.4	73.0	53.4	69.7	49.9	61.9	69.0	73.9	76.3	74.9	67.0	63.4
	T + w/o Wiki	75.6	80.0	67.9	84.6	71.5	73.6	73.1	78.7	68.7	80.3	72.5	73.1
GPT3.5	w/ Wiki	81.7	90.9	69.2	88.9	59.1	71.1	81.3	89.8	79.5	88.9	65.2	70.0
	w/o Wiki	84.3	91.5	82.3	93.6	74.6	80.9	79.3	89.8	67.6	87.0	71.5	78.8
	T + w/ Wiki	84.8	93.4	74.3	92.8	73.0	83.3	81.7	91.8	82.6	91.9	79.6	83.1
	T + w/o Wiki	84.7	92.2	77.9	93.4	82.8	84.7	78.0	90.4	77.1	88.0	80.2	83.9
Mistral	w/ Wiki	72.1	84.4	57.5	78.5	45.0	59.5	77.9	84.0	73.4	80.6	55.6	59.0
	w/o Wiki	58.6	71.0	43.9	65.2	60.9	72.1	61.2	69.9	54.5	61.1	57.2	71.6
	T + w/ Wiki	74.6	86.9	61.0	85.8	61.2	74.5	78.2	86.2	81.6	88.2	73.3	76.3
	T + w/o Wiki	67.8	77.7	61.0	82.5	65.2	67.7	69.3	76.9	68.6	80.0	68.7	68.2

Table 9: FActScore and accuracy by different scorers with (w/) or without (w/o) Wiki and whether translation (T) is used on generated facts and knowledge source (Wiki page). Accuracy is measured against native labeling using the Internet to find references. Human with the Internet is considered as golden.

- GPT4 (gpt-4-0125-preview)
- GPT3.5 (gpt-35-turbo-0125)
- Gemma (google/gemma-7b-it)
- Mistral (mistralai/Mistral-7B-Instruct-v0.2)
- Aya (CohereForAI/aya-101)
- BLOOMZ (bigscience/bloomz-7b1)
- Llama2 (meta-llama/Llama-2-7b-chat-hf)

Retrievers:

- sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
- sentence-transformers/distiluse-base-multilingual-cased-v2

Knowledge Generator:

- GPT4 (gpt-4-0125-preview)

Translator:

- Google Translate (Cloud Translation - Basic (v2), used from January 2024 to June 2024)

Running Trials of Experiments: All results were obtained from data conducted or collected from a single trial.

H Hyper-Parameters

All experiments are conducted from January to June 2024. The following hyper-parameters are specified, while all others are set to their default values.

Generation Temperature: All studied models' temperatures are set to 0.7.

Context, max generation length: For open-source models, the maximum output length is set to 512 tokens, and the maximum sequence length is set to 4096 tokens for high-resource languages, and 1024 and 6024 tokens for medium- and low-resource languages, respectively. For closed models accessed via API, the maximum token limit is uniformly set to 4096 tokens for all use cases.

I Annotation Guidelines

9.1 Factuality Labeling

To collect R2, the original pipeline from Min et al. (2023) is fully replicated to studied languages. Along with it, we had the qualification task to assess annotators and provided a 1-hour training session.

9.2 Fact Extraction Error Definitions

We categorized the fact extraction errors into the following groups:

- **Need Merge**: The extracted sentence is excessively extracted into fragments that are too vague or unverifiable. E.g. *"He played", "He did"*.

- **Need Split**: The extracted sentence is not atomic enough and can be further divided into meaningful atomic sentences when the list doesn't include those atomic sentences, e.g., *"He played for Real Madrid in 2010" without prior "He played for Real Madrid" or "He played in 2010" in the list.*

- **Duplicated**: The extracted sentence offers the same information as another one in the list, e.g., *"He played for Real Madrid" and "He played soccer for Real Madrid"*

- **Missing Information**: The extracted sentences don't include critical information from the original sentence, e.g., the extracted fact only contains a single sentence "He played for Real Madrid" while the original text "He played for Real Madrid in 2010". In this case, the "in 2010" information is missed.

- **Not Grounded**: The extracted sentence is incorrectly modified, e.g., the extracted fact only contains a single sentence "He played for Real Betis" while the original text "He played for Real Madrid in 2010".

- **Linguistic Error**: The extracted sentence contains grammar, spelling, or coherence errors.

- **Misc**: Any other errors that do not fall into the above errors.

Category	Candidate Fact	Comment
Context unfaithful - Factually correct	*"Naipaul fue criticado por su visión a menudo pesimista"* **Translated:** *"Naipaul was criticized for his often pessimistic vision"*	Native label: F, Model label: T, Ground truth: T. **Comment:** No related information within the provided Wikipedia page. But there is supporting evidence from en.wikipedia. **Evidence:** *"Yet he has been accused of being a"neo-colonialist" , and in this novel post-colonial Africa is depicted as spiraling into a kind of Hell...Naipaul's fiction and especially his travel writing have been criticised for their allegedly unsympathetic portrayal of the Third World. The novelist Robert Harris has called Naipaul's portrayal of Africa racist and"repulsive,"reminiscent of Oswald Mosley's fascism."*
Context unfaithful - Reading Deficiency	*"Rodrygo Goes de Souza nació el 9 de enero de 2001."* **Translated:** *"Rodrygo Goes de Souza was born on January 9, 2001."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The evaluator misses the related information (that supports the fact) within retrieved passages. **Evidence:** *"Rodrygo Silva de Goes (; Osasco, São Paulo, 9 de enero de 2001), conocido simplemente como Rodrygo, es un futbolista brasileño que juega como delantero en el Real Madrid C. F. de la Primera División de España."* **Translated:** *"Rodrygo Silva de Goes (; Osasco, São Paulo, January 9, 2001), known simply as Rodrygo, is a Brazilian footballer who plays as a forward for Real Madrid C.F. of the Spanish First Division."*
Retrieval error	*"Ingresó al seminario de Villa Devoto en Buenos Aires."* **Translated:** *"He entered the Villa Devoto seminar in Buenos Aires."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The retriever fails to retrieve the needed information passage for evaluation. **Evidence:** *"Ingresó al seminario del barrio Villa Devoto y al noviciado de la Compañía de Jesús."* **Translated:** *"He entered the seminary in the Villa Devoto neighborhood and the novitiate of the Society of Jesus."*
Tabular data	*"El primer torneo importante que Court ganó fue el campeonato australiano de tenis."* **Translated:** *"The first major tournament Court won was the Australian Tennis Championships."*	Native label: T, Model label: F, Ground truth: T. **Comment:** Related information is embedded in the table which is not processed by Wikipedia's HTML conversion to text, thus not being contained as passages to retrieve. **Evidence:** Information lies within"Victorias (24)"table at the first row
Subjective opinion	*"Es considerado uno de los trompetistas más destacados de su generación."* **Translated:** *"He is considered one of the most prominent trumpeters of his generation."*	Native label: T, Model label: F, Ground truth: T/F. **Comment:** The statement/fact is subjective, thus debatable. Chuck Mangione had a song, being recognized as the number one jazz song of all time by a radio channel, but there is no explicit mention that he is a prominent trumpeter of his generation. **Evidence:** *"Recientemente las estaciones de radio que transmiten jazz en los Estados Unidos han reconocido a Feels So Good de Mangione como la canción número uno de todos los tiempos."*(**en:** *"Recently, jazz radio stations in the United States have recognized Mangione's Feels So Good as the number one song of all time."*
Annotation error	*"Tekke dio el salto al fútbol europeo en 2006."* **Translated:** *"Tekke made the leap to European football in 2006."*	Native label: F, Model label: T, Ground truth: T. **Comment:** The annotator misses details within the Wikipedia page. **Evidence:** *"Esa temporada Tekke se convirtió en el máximo goleador de la Superliga de Turquía al anotar 31 goles. El 31 de julio de 2006 firma un contrato con su actual club, el Zenit de San Petersburgo ruso, equipo que realizó un desembolso económico de 10 millones de euros para poder hacerse con sus servicios."* **Translated:** *"That season Tekke became the top scorer in the Turkish Super League by scoring 31 goals. On July 31, 2006, he signed a contract with his current club, the Russian Zenit Saint Petersburg, a team that made a financial outlay of 10 million euros to be able to acquire his services."*
Assistant generation	*"La información podría haberse modificado."* **Translated:** *"The information may have been modified."*	Native label: F, Model label: T, Ground truth: F. **Comment:** The assistant/service generation by subject models is often labeled as *"supported"* by modeling evaluators.

Table 10: Examples from each disagreement category between natives and Gemini in Spanish.

Category	Fact	Comment
Context unfaithful - Factually correct	"ناجوه كلوه" **Translated:** *"He had a big mustache"*	Native label: F, Model label: T, Ground truth: T. No related information within the provided Wikipedia page. But there is supporting evidence from en.wikipedia. **Evidence:** *"Hogan grew a beard alongside his famous mustache and dyed it black, traded his red and yellow garb in for black and white clothing, often detailed with lightning bolts, and renamed himself "Hollywood" Hulk Hogan (often shortened to Hollywood Hogan"*
Context unfaithful - Hallucination	"ةينطولا ةطبارلا قوقحل ةأرملا." **Translated:** *"Stone helped found the National Women's Suffrage Association."*	Native label: F, Model label: T, Ground truth: F. The fact is false. Lucy Stone helped found the American Woman Suffrage Association (AWSA) in 1869, which was a rival organization to the National Woman Suffrage Association (NWSA) founded by Elizabeth Cady Stanton and Susan B. Anthony
Context unfaithful - Reading Deficiency	"ةيثحبلا تاماتهلاا دعب ام ةيرظن لمشت رامعتسلاا" **Translated:** *"Research interests include postcolonial theory."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The evaluator misses the related information (that supports the fact) within retrieved passages. **Evidence:** "دعب ام ـل ةسسؤملا صوصنلا نم دعي ١٩٨٥ ماع روشنملا ةرثؤملا ةيملاعلا تايصخشلا مهأ نم ًاليلحا كافيبس دعتو ،ةيلايئولوكلا نونفلل وتويك قزئاج ىلع كافيبس تلصح .بدلأاو يراضحلا دقنلا يف ةملعمو)قدقاذ قرظنم(قدقاذ تايرظن ةملاع» اهنوكل ٢٠١٢ ماعل ةفسلفلاو م لوعملا م لاعلاب قلعتي اميف يركفلا رامعتسلاا دض ةيناسذلإا مولعلا نع عفادتل" **Translated:** *"The 1985 publication is considered one of the founding texts of postcolonialism, and Spivak is currently considered one of the most important international figures influencing cultural criticism and literature. Spivak was awarded the 2012 Kyoto Prize for Arts and Philosophy for being a "critical theorist (critical theorist) and educator who defends the humanities against intellectual colonialism in relation to the globalized world.*"
Retrieval error	"لاقتذا ةقفص تناك نويلم ٢٢٢ ةميقب راميذ ورويل." **Translated:** *"Neymar's transfer was worth 222 million euros."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The retriever fails to retrieve needed information passage for evaluation. **Evidence:** "ةقفص يف نامريج ناسن سيراب ىلإ راميذ لقتذا ،٢٠١٧ ماع يف ةيسايق ورويل نويلم ٢٢٢ اهتميق تغلب ثيح ،مدقلا قرك خيرات يف مخضلأا يه" (**en:** "In 2017, Neymar moved to Paris Saint-Germain in a record deal, the largest in football history, worth 222 million euros.")
Tabular data	"نم جوزتم اديثوس ةكلملا نروكجنولاريجاف ايدوياً اذ." **Translated:** *"Married to Queen Suthida Vajiralongkorn na Ayodhya."*	Native label: T, Model label: F, Ground truth: T. **Comment:** Related information is embedded in the table which is not processed by Wikipedia's HTML conversion to text, thus not being contained as passages to retrieve. **Evidence:** Information locates on "ةجوزلا" (Wife) section of the side infobox.
Subjective opinion	"ًالاثم ربتعتي ربصدنه ةبيدنتحي." **Translated:** *"Hend Sabry is a role model."*	Native label: T, Model label: F, Ground truth: T/F. **Comment:** The statement/fact is subjective, thus debatable.
Assistant generation	"تاملعم رفوتت مل دبعب قافو لوح ةددحم ماع ىتح يواشلا ردادقلا ٢٠٢٣." **Translated:** *"No specific information about Abdelkader Chaoui's death is available until 2023."*	Native label: F, Model label: T, Ground truth: F. **Comment:** The assistant/service generation by subject models is often labeled as "supported" by modeling evaluators.
Inconsistent Wikipage	"ةيبهذلا ةركلا ةزئاج تارم ٧ اهب زاف." **Translated:** *"The Ballon d'Or he won 7 times."*	Native label: F, Model label: T, Ground truth: F. **Comment:** The Wikipedia page has conflicting information on the number of Ballon d'Or that Lionel Messi won.

Table 11: Examples from each disagreement category between natives and Gemini in Arabic.

Category	Fact	Comment
Context unfaithful - Factually correct	*Trygve Lie intentó promover la paz."* **Translated:** *"Trygve Lie tried to promote peace."*	Native label: F, Model label: T, Ground truth: T. No related information within the provided Wikipedia page. But there is supporting evidence from en.wikipedia. **Evidence:** *"He sent 50 members of the United Nations guard force from Lake Success to assist the Mediator in supervising the Truce in the former British Mandate of Palestine in 1948 and the "UNTSO", the first peacekeeping operation was established by the United Nations."*
Context unfaithful - Hallucination	*Algunas de sus películas han recibido críticas positivas a nivel internacional."* **Translated:** *"Some of his films have received positive reviews internationally."*	Native label: F, Model label: T, Ground truth: F. The fact is false. Besides several users on IMDB and Rotten Tomatoes, there is no concrete evidence that supports Surya Saputra's films are recognized internationally.
Context unfaithful - Reading Deficiency	*"Murió de un ataque al corazón."* **Translated:** *"He died of a heart attack."*	Native label: F, Model label: T, Ground truth: F. **Comment:** He died of pneumonia, not a heart attack. **Evidence:** *"...Arruinado, físicamente débil y con la mente deteriorada, Capone se retiró a una propiedad ubicada en Palm Island, en Miami Beach, Florida, donde se recluyó con su esposa del mundo exterior. El 21 de enero de 1947, sufrió un derrame cerebral, y murió cuatro días después de neumonía: Al Capone fue encontrado muerto en la bañera"* (**en:** *"...Ruined, physically weak and mentally deteriorating, Capone retired to a property located on Palm Island in Miami Beach, Florida, where he and his wife secluded themselves from the outside world. On January 21, 1947, he suffered a stroke, and died four days later of pneumonia: Al Capone was found dead in the bathtub.*
Retrieval error	*"Incluyó su papel en Ğuardianes de la Galaxia Vol. 2014"* **Translated:** *"Included his role in Ğuardians of the Galaxy Vol. 2014.""*	Native label: T, Model label: F, Ground truth: T. **Comment:** The retriever fails to retrieve the needed information passage for evaluation. **Evidence:** *"En 2014, logró el reconocimiento a nivel mundial al protagonizar la película Guardianes de la Galaxia (2014) con el papel de Peter Quill / Star-Lord.23 El filme recibió elogios de la crítica por su humor y fue un éxito comercial tras recaudar 773 millones de dólares, además de convertirse en la cuarta película más taquillera de 2014"* **Translated:** *"In 2014, he achieved worldwide recognition by starring in the film Guardians of the Galaxy (2014) with the role of Peter Quill / Star-Lord.23 The film received critical praise for its humor and was a commercial success after grossing $773. million dollars, in addition to becoming the fourth highest-grossing film of 2014"*
Tabular data	*"Drummond promedió 17.5 puntos por partido en la temporada 2020-21."* **Translated:** *"Drummond averaged 17.5 points per game in the 2020-21 season."*	Native label: T, Model label: F, Ground truth: T. **Comment:** Related information is embedded in the table which is not processed by Wikipedia's HTML conversion to text, thus not being contained as passages to retrieve.
Subjective opinion	*"Sarr es hábil."* **Translated:** *"Sarr is skillful"*	Native label: F, Model label: T, Ground truth: T/F. **Comment:** The statement/fact is subjective, thus debatable.
Assistant generation	*"Nuevos proyectos y logros pueden haberse agregado a la biografía de Surya Saputra después de 2023."* **Translated:** *"New projects and achievements may have been added to Surya Saputra's biography after 2023."*	Native label: F, Model label: T, Ground truth: F. **Comment:** The assistant/service generation by subject models is often labeled as "supported" by modeling evaluators.

Table 12: Examples from each disagreement category between natives and GPT-4 in Spanish.

Category	Fact	Comment
Context unfaithful - Factually correct	"لًافزاع ناك هدلاو." **Translated:** *"His father was a musician."*	Native label: F, Model label: T, Ground truth: T. No related information within the provided Wikipedia page. But there is supporting evidence from en.wikipedia. Evidence: His mother is dancer Kine Gueye Thiam (née Gueye), and his father is percussionist Mor Thiam. Mor Thiam was born to a Toucouleur family of Quranic scholars in Kaolack, Senegal.
Context unfaithful - Hallucination	"ربكلأا نبلا وه." **Translated:** *"He is the eldest son."*	Native label: F, Model label: T, Ground truth: F. The fact is false. The eldest son of his father is Abdelaziz bin Khalifa Al Thani
Context unfaithful - Reading Deficiency	"ربمتبس ٨ يف تيفوت ٢٠٢٢." **Translated:** *"She died on September 8, 2022."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The evaluator misses the related information (that supports the fact) within retrieved passages. **Evidence:** "ثيبازيلإ ةكلملا قافو ماهغنكاب رصق نلعأ ،٢٠٢٢ ربمتبس ٨ يف روهدت لوح قرتاوتم عابنأ عم كلذ نمازت ،اماع ٩٦ زهانيز رمع نع ةينانثلا ةيحصلا اهتلاح." (**Translated:** *"On September 8, 2022, Buckingham Palace announced the death of Queen Elizabeth II at the age of 96, coinciding with frequent reports about the deterioration of her health."*
Retrieval error	"وه يناثلا لفطلا مسا سنتيج نوج يرور." **Translated:** *"The name of the second child is Rory John Gates."*	Native label: T, Model label: F, Ground truth: T. **Comment:** The retriever fails to retrieve needed information passage for evaluation. **Evidence:** "ةثلاث ابجنأو م١٩٩٤ ماع يف شنيرف ادنيليم نم سنتيغ ليبد جوزت م)٢٠٠٢(ليدأ يبيف ،م)١٩٩٩(نوج يرور ،م)١٩٩٦(نيرثاك رفنيج :مهل لافطأ ةمصاعلا يف قريحبد ىلع لطي فلكمو مخضي رصع لزنم يف ةلئاعلا شيعتو يف لجر ىنغأ» ببقل سنتيغ ليبد لمحم ٢٠٠٦ ىتحو م١٩٩٦ ماع ذنم .نطنشاو عبرتدقو يكيرمأ رلاود رايلم ١٠٠-ب م١٩٩٩ ماع يف ةورث تردق دقف ،«مملاعلا ىلع ٢٠٠٧م ماع ىرخأ قرم شرعلا." (**Translated:** *"Bill Gates married Melinda French in 1994 and they have three children: Jennifer Katherine (1996), Rory John (1999), and Phoebe Adele (2002). The family lives in a huge, expensive modern house overlooking a lake in Washington, DC. From 1996 AD until 2006 AD, Bill Gates held the title of "the richest man in the world." His wealth was estimated in 1999 at 100 billion US dollars, and he ascended to the throne again in 2007 AD."*
Tabular data	"ىلع وهاينتذ لصحي في سويروللكبلا ةجرد ملعلا." **Translated:** *"Netanyahu received a Bachelor of Science degree."*	Native label: T, Model label: F, Ground truth: T. **Comment:** Related information is embedded in the table (infobox) which is not processed by Wikipedia's HTML conversion to text, thus not being contained as passages to retrieve. **Evidence:** Information locates on "ملأا ةسردملا " (Alma mater) section of the side info box.
Subjective opinion	"لًافورعم ناك تنيارب ةبعلا يف هينافتب." **Translated:** *"Bryant was known for his dedication to the game."*	Native label: F, Model label: T, Ground truth: T/F. **Comment:** The statement/fact is subjective, thus debatable. While there is no explicit information that Kobe Bryant was known for his dedication to the game, it could be deductible.

Table 13: Examples from each disagreement category between natives and GPT-4 in Arabic.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by GPT4 in **English**: + He then moved to Olympique Alès. + He then moved to Olympique Alès in 2003. + He then moved to Stade Brestois. + He then moved to Stade Brestois in 2004. + Olympique Alès is a team. + Stade Brestois is a team. + He moved to Olympique Alès and Stade Brestois in consecutive years.
- Facts extracted by GPT4 in **Spanish**: + Se mudó al Olympique Alès. (En: He then moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He then moved to Olympique Alès in 2003.) + Olympique Alès es un equipo. (En: Olympique Alès is a team.) + Luego se mudó al Stade Brestois. (En: He then moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.) + Stade Brestois es un equipo. (En: Stade Brestois is a team.)
- Facts extracted by GPT4 in **Arabic**: + “سيدأ كيبمولو أ ىلإ لقتذا”. (En: Moved to Olympique Ales) + “٢٠٠٣ ماع يف سيدأ كيبمولو أ ىلإ لقتذا.” (En: He moved to Olympique Ales in 2003) + “سيوتسير ب داتس ىلإ كلذ دعب لقتذا.” (En: He then moved to Stade Brestois.) + “٢٠٠٤ ماع يف سيوتسير ب داتس ىلإ لقتذا.” (En: He moved to Stade Brestois in 2004.)

Table 14: Example of atomic facts extracted by GPT4.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by GemP in **English**: + He moved to Olympique Alès. + He moved to Olympique Alès in 2003. + He moved to Stade Brestois. + He moved to Stade Brestois in 2004.
- Facts extracted by GemP in **Spanish**: + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He moved to Olympique Alès in 2003.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.)
- Facts extracted by GemP in **Arabic**: + “كيبمولو أ ىلإ كلذ دعب لقتذا.” (En: He moved to Olympique) + “٢٠٠٣ ماع يف سيدأ كيبمولو أ ىلإ لقتذا.” (En: He moved to Olympique in 2003) + “سيوتسير ب داتس ىلإ كلذ دعب لقتذا. ”(En: He moved to Stade Prestois.) + “٢٠٠٤ ماع يف سيوتسير ب داتس ىلإ لقتذا.” (En: He moved to Stade Prestois in 2004.)

Table 15: Example of atomic facts extracted by GemP.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by GPT3.5 in **English**: + He then moved to Olympique Alès. + He then moved to Olympique Alès in 2003. + He then moved to Stade Brestois. + He then moved to Stade Brestois in 2004.
- Facts extracted by GPT3.5 in **Spanish**: + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He moved to Olympique Alès in 2003.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.)
- Facts extracted by GPT3.5 in **Arabic**: + “كيبمولو أ ىلإ كلذ دعب لقتذا.” (En: He moved to Olympique Ales) + “٢٠٠٣ ماع يف كيبمولو أ ىلإ كلذ دعب لقتذا.” (En: He moved to Olympique Ales in 2003) + “سيوتسير ب داتس ىلإ كلذ دعب لقتذا.”(En: He moved to Stade Brestois.) + “٢٠٠٤ ماع يف سيوتسير ب داتس ىلإ لقتذا.” (En: He moved to Stade Brestois in 2004.)

Table 16: Example of atomic facts extracted by GPT3.5.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by Mistral Instruct in **English**: + He then moved to Olympique Alès. + He then moved to Olympique Alès in 2003. + He then moved to Stade Brestois. + He then moved to Stade Brestois in 2004.
- Facts extracted by Mistral in **Spanish**: + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He moved to Olympique Alès in 2003.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.)
- Facts extracted by Mistral Instruct in **Arabic**: + “ كلذ دعب لقتذا.” (En: He moved: INCOMPLETE) + “ كيبملو أ ى لإ كلذ دعب لقتذا.” (En: He moved to Olympique) + “ ٢٠٠٣ ماع يف كيبملو أ ى لإ كلذ دعب لقتذا.” (En: He moved to Olympique in 2003) + “ سيوتسير ب داتسا.” (En: Stade Prestois - INCOMPLETE) + “ ٢٠٠٤ ماع يف سيوتسير ب داتسا.” (En: Stade Prestois in 2004 - INCOMPLETE) + “ ٢٠٠٤ ماع يف سيوتسير ب داتسو ٢٠٠٣ ماع يف كيبملو أ ى لإ كلذ دعب لقتذا. ” (En: He then moved to Olympique in 2003 and Stade Prestois in 2004.)

Table 17: Example of atomic facts extracted by Mistral-Instruct.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by Llama-2 Chat in **English**: + He moved to Olympique Alès. + He moved to Olympique Alès in 2003. + He moved to Stade Brestois. + He moved to Stade Brestois in 2004.
- Facts extracted by Llama-2 Chat in **Spanish**: + Luego se mudó. (En: Then he moved.) + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) Insufficient covering information from the original sentence, years of movement
- Facts extracted by Llama-2 Chat in **Arabic**: +“ لقتذا م ث.” (En: Then move on: INCOMPLETE) +“ كلذ دعب لقتذا.” (En: Go then: INCOMPLETE) +“ سيلأ كيبملو أ ى لإ.” (En: To Olympique Alice: INCOMPLETE) +“ ٢٠٠٣ ماع يف سيلأ كيبملو أ ى لإ.” (En: To Olympique Alice in 2003: INCOMPLETE) +“ سيوتسير ب داتسا.” (En: Prestois Stadium: INCOMPLETE) +“ ٢٠٠٤ ماع يف سيوتسير ب داتسا.” (En: Prestois Stadium in 2004: INCOMPLETE)
- Facts extracted by Llama-2 Chat in **Bengali**: Does not have sufficient tokens in Bengali (text with full of UNK tokens)

Table 18: Example of atomic facts extracted by Llama-2 Chat.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by Gemma Instruct in **English**: + He moved to Olympique Alès. + He moved to Olympique Alès in 2003. + He moved to Stade Brestois. + He moved to Stade Brestois in 2004.
- Facts extracted by Gemma Instruct in **Spanish**: + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He moved to Olympique Alès in 2003.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.)
- Facts extracted by Gemma Instruct in **Arabic**: +“ كيبملو أ ى لإ لقتذا.” (En: Moved to Olympic: DUPLICATED) +“ كيبملو أ ى لإ لقتذا.” (En: Moved to Olympic: DUPLICATED) +“ كيبملو أ ى لإ لقتذا.” (En: Moved to Olympic: DUPLICATED) +“ كيبملو أ ى لإ لقتذا.” (En: Moved to Olympic: DUPLICATED)

Table 19: Example of atomic facts extracted by Gemma-7B-Instruct.

Original sentence (in English): *He then moved to Olympique Alès in 2003 and Stade Brestois in 2004.*
- Facts extracted by Finetuned Gemma in **English**: + He moved to Olympique Alès. + He moved to Olympique Alès in 2003. + He moved to Stade Brestois. + He moved to Stade Brestois in 2004.
- Facts extracted by Finetuned Gemma in **Spanish**: + Se mudó al Olympique Alès. (En: He moved to Olympique Alès.) + Se mudó al Olympique Alès en 2003. (En: He moved to Olympique Alès in 2003.) + Se mudó al Stade Brestois. (En: He moved to Stade Brestois.) + Se mudó al Stade Brestois en 2004. (En: He moved to Stade Brestois in 2004.)
- Facts extracted by Finetuned Gemma in **Arabic**: + " كييمولو أ ىلإ كلذ دعب لقتنا ." (En: He moved to Olympique) + " ٢٠٠٣ ماع يف سيلأ كييمولو أ ىلإ لقتنا ." (En: He moved to Olympique in 2003) + " سيوتسير ب داتس ىلإ كلذ دعب لقتنا ." (En: He moved to Stade Prestois.) + " ٢٠٠٤ ماع يف سيوتسير ب داتس ىلإ لقتنا ." (En: He moved to Stade Prestois in 2004.)

Table 20: Example of atomic facts extracted by Finetuned Gemma.

	Fact	**Google API Texts**
es	*"Spivak ha escrito sobre la traducción"* → *"Spivak has written about translation"*	*"Gayatri Chakravorty Spivak is an Indian scholar, literary theorist, and feminist critic. She has made a significant statement on "the politics of translation" ... Living Translation performs the invaluable service of gathering for the first time Gayatri Chakravorty Spivak's wide-ranging writings on translation."* **Comment:** Gayatri Chakravorty Spivak's Wikipedia page has no information about her research on the impact of translation. Yet, using Google Query API returns information about her two books on the topic ("The Politics of Transaltion" and "Living Translation").
ar	"نم ديدعلا صخصخ ةيموكحلا تاكرشلا." → *"He privatized many state companies."*	"دق برغملا خيراتي ةصصوخلا تايلمع ربكأ نأ ريرقتلا تايطعم نم دافتسيو قريبكً لاومأ تردأ دقو ،وطج سيرداو يفسويلا نمحرلا دبع يتموكح للاخ تمت." → *"It is clear from the report's data that the largest privatization operations in the history of Morocco took place during the governments of Abderrahmane Youssoufi and Idriss Jettou, and they generated large sums of money."* **Comment:** The additional information supports the fact about privatization initiatives under Idriss Jettou's term, whereas the Wikipedia page has no related information about it.

Table 21: Examples demonstrate that using Google Query API provides additional information to the scorer, GemP, leading to accurate fact labeling. The additional information from the examples is not present on Wikipedia pages and has been manually validated as correct.

	Fact & Labels	GPT4 Additional Texts
es	*Implementó políticas radicals.* → *"He implemented radical policies."*	**Query:** *"¿Qué tipo de políticas implementó Hafizullah Amín?"* → *"What kind of policies did Hafizullah Amin implement?"* **GPT-4 Texts:** *"Durante su tiempo en el poder, Amín se desempeñó como primer ministro y luego, brevemente, como presidente de Afganistán desde septiembre hasta diciembre de 1979. Sus políticas se centraron en la implementación de una agenda socialista radical, que incluía reformas agrarias destinadas a abolir las tradicionales tenencias feudales de tierra y redistribuir las tierras a los campesinos sin tierra."* → *"During his time in power, Amin served as prime minister and then briefly as president of Afghanistan from September to December 1979. His policies focused on the implementation of a radical socialist agenda, which included land reforms aimed at abolishing the traditional feudal land tenures and redistributing land to landless peasants."* **Comment:** The additional information generated by GPT4 supports the fact about radical reforms by Hafizullah Amin, whereas the Wikipedia page has no related information about it. The generated information is manually confirmed to be correct
ar	"يف ُتذرا هةْ ضرع أدب ماع ٢٠٠٤." → *"The Apprentice" began airing in 2004."*	**Query:** "ههدقيب مارت دلناذود ناك يذلا ُتذرا هةْ جمانرب ضرع أدب ىتم؟" → *"When did Donald Trump's show "The Apprentice" start airing?"* **GPT-4 Texts:** "نويزفلتة عوذ نم يكيرمأ يفنويزفلتة جمانرب وه ُتذرا هة جمانرب ٢٠٠٤ ريانيد ٨ يف قرم لوألا جمانربلا رهظ .ةدحتملا تايلاولا يف ه ضرع أدب عقاوا ر ةكبشلا ىلع." → *"The Apprentice" is an American reality television show that began airing in the United States. The program debuted on January 8, 2004 on NBC."* **Comment:** The additional information generated by GPT4 supports the fact about the airing time of "The Apprentice", whereas the Wikipedia page has no related information about it. The generated information is manually confirmed to be correct.

Table 22: Examples demonstrate that using GPT-4 as a knowledge generator provides additional information to the scorer, GemP, leading to accurate fact labeling. The additional information from the examples is not present on Wikipedia pages and has been manually validated as correct.