Multilingual Fact-Checking using LLMs

Aryan Singhal*, Thomas Law*, Coby Kassner*, Ayushman Gupta*
Evan Duan, Aviral Damle, Ryan Luo Li
Association of Students for Research in Artificial Intelligence (ASTRA)
astra.ai.lab@gmail.com

Abstract

Due to the recent rise in digital misinformation, there has been great interest in using LLMs for fact-checking and claim verification. In this paper, we answer the question: *Do LLMs know multilingual facts and can they use this knowledge for effective fact-checking?* To this end, we create a benchmark by filtering multilingual claims from the X-fact dataset and evaluating the multilingual fact-checking capabilities of five LLMs across five diverse languages: Spanish, Italian, Portuguese, Turkish, and Tamil on our benchmark. We employ three different prompting techniques: Zero-Shot, English Chain-of-Thought, and Cross-Lingual Prompting, using both greedy and self-consistency decoding. We extensively analyze our results and find that GPT-4o achieves the highest accuracy, but zero-shot prompting with self-consistency was the most effective overall. We also show that techniques like Chain-of-Thought and Cross-Lingual Prompting, which are designed to improve reasoning abilities, do not necessarily improve the fact-checking abilities of LLMs. Interestingly, we find a strong negative correlation between model accuracy and the amount of internet content for a given language. This suggests that LLMs are better at fact-checking from knowledge in low-resource languages. We hope that this study will encourage more work on multilingual fact-checking using LLMs.

1 Introduction

In an era marked by the proliferation of digital misinformation, the need for fact-checking on a global scale has never been more pressing. Recent research has shown promising capabilities in large language models (LLMs) for fact-checking and claim verification (Lee et al., 2020; Hoes et al., 2023; Zhang and Gao, 2023; Choi and Ferrara, 2024). However, this research has predominantly focused on English and Chinese facts and claims, overlooking billions of people who do not speak these languages (Quelle and Bovet, 2024; Cao et al., 2023; Zhang et al., 2024). In this paper, we evaluate the multilingual fact-checking capabilities of LLMs across five languages: Spanish, Italian, Portuguese, Turkish, and Tamil, sourcing claims from the X-Fact dataset (Gupta and Srikumar, 2021). With this selection of languages, we ensure geographic and typological diversity and can probe LLMs' performance in low-resource as well as high-resource languages.

We utilize a variety of prompting techniques, including Zero-Shot (Baseline), English Chain-of-Thought (Wei et al., 2023a), Cross-Lingual Prompting (Qin et al., 2023), and Self-Consistency (Wang et al., 2023a) to evaluate the performance of LLMs. To our knowledge, this is the first work to assess the factual multilingual knowledge and inherent fact-checking capabilities of a variety of LLMs across a spectrum of languages worldwide, using a variety of prompting techniques.

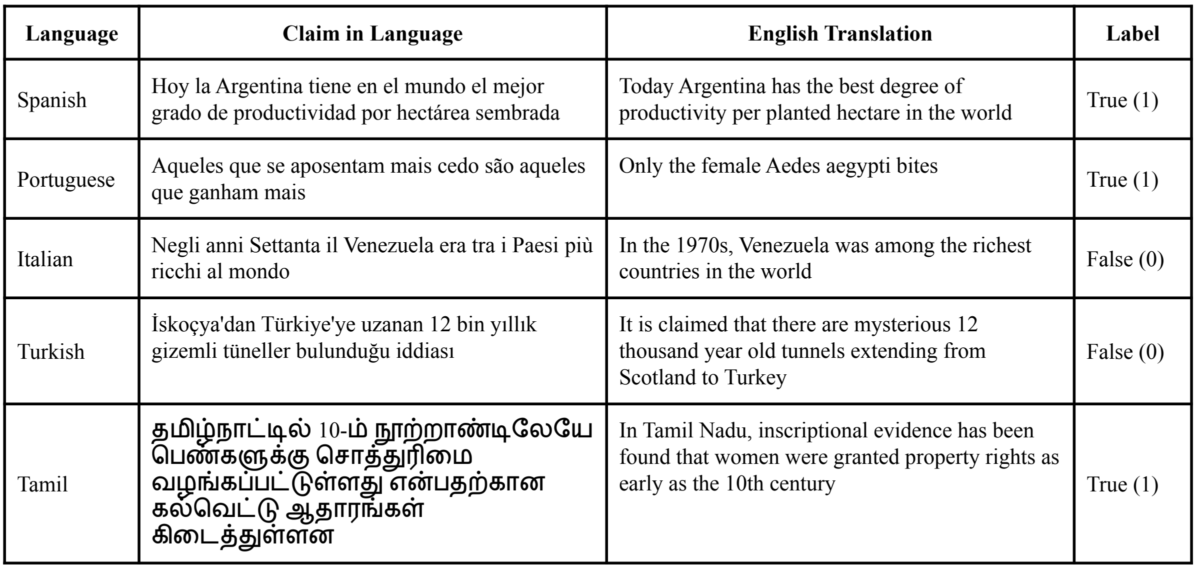
The remainder of this paper is organized as follows: In Section 2, we review related work. In Section 3, we detail the datasets, models, and evaluation method used. In Section 4, we discuss the prompting methods we use. In Section 5, we present our results. In Section 6 we analyze and interpret our findings and explore their implications. Finally, we conclude in Section 7 and suggest directions for future research.

2 Related Work

English Fact-Checking using LLMs Prior research examines the capabilities of LLMs for fact-checking and claim verification in English. LLMs such as GPT-3.5 and GPT-4 excel in fact-checking when provided with sufficient contextual information, though they suffer from inconsistent accuracy (Quelle and Bovet, 2024). Tian et al. 2023 suggest enhancing LLM factuality by fine-tuning models with automatically generated factuality

*Equal contribution

Proceedings of the Third Workshop on NLP for Positive Impact, pages 13–31
November 15, 2024 ©2024 Association for Computational Linguistics



Language	Claim in Language	English Translation	Label
Spanish	Hoy la Argentina tiene en el mundo el mejor grado de productividad por hectárea sembrada	Today Argentina has the best degree of productivity per planted hectare in the world	True (1)
Portuguese	Aqueles que se aposentam mais cedo são aqueles que ganham mais	Only the female Aedes aegypti bites	True (1)
Italian	Negli anni Settanta il Venezuela era tra i Paesi più ricchi al mondo	In the 1970s, Venezuela was among the richest countries in the world	False (0)
Turkish	İskoçya'dan Türkiye'ye uzanan 12 bin yıllık gizemli tüneller bulunduğu iddiası	It is claimed that there are mysterious 12 thousand year old tunnels extending from Scotland to Turkey	False (0)
Tamil	தமிழ்நாட்டில் 10-ம் நூற்றாண்டிலேயே பெண்களுக்கு சொத்துரிமை வழங்கப்பட்டுள்ளது என்பதற்கான கல்வெட்டு ஆதாரங்கள் கிடைத்துள்ளன	In Tamil Nadu, inscriptional evidence has been found that women were granted property rights as early as the 10th century	True (1)

Figure 1: Examples of claims in the testing datasets for each language, their English translations, and respective ground-truth label

preference rankings, leading to improved factual accuracy without human labeling. Cheung and Lam 2023 incorporate external evidence-retrieval to bolster fact-checking performance for the Llama 2 model. In comparison, our work examines LLM fact-checking performance in several languages.

Multilingual Fact-Checking using LLMs Numerous studies address the linguistic divide caused by focusing solely on LLM-based fact-checking for English and Chinese. However, the detailed exploration of the multilingual capabilities of LLMs for fact-checking beyond these two languages is limited. Shafayat et al. 2024 examines the factual accuracy of GPT-3.5 and GPT-4 across nine languages and finds that the models exhibit an inherent bias towards factual political information from Western continents. Huang et al. 2022 augment mBERT (a multilingual version of the language model BERT) with cross-lingual retrieval to improve the fact-checking performance of LLMs on the X-Fact dataset. Cekinel et al. 2024 explores cross-lingual learning and low-resource fine-tuning for fact-checking in Turkish. Hu et al. 2023 benchmarks the factual knowledge possessed by ten different LLMs and their multilingual fact-checking capabilities in 27 languages. They also employ several different prompting techniques. However, their study predominantly focuses on smaller models (e.g., under 15B parameters). Moreover, their multilingual analysis only distinguishes between English and Chinese. All other languages are benchmarked together in a mixed testing set, and interlingual comparisons (besides English and Chinese) are not drawn. To the best of our knowledge, our study is the first to benchmark and closely analyze the multilingual fact-checking abilities of several LLMs across various domains, both political and non-political, using a range of different prompting techniques.

3 Experimental Setup

3.1 Datasets

We source 500 random claims (250 false and 250 true) for each selected language, i.e. Spanish, Portuguese, Italian, Turkish, and Tamil, from the X-Fact dataset (Gupta and Srikumar, 2021). The claims in our final datasets encompass a diverse range of topics that are both political and non-political.

In some cases, the X-fact dataset did not contain enough fully true or false claims for a given language, and we included claims labeled as 'mostly true,' 'mostly false,' and 'partly true/misleading' by mapping them to 'true,' 'false,' and 'false,' respectively. While we acknowledge that there are distinctions between the labels given for the claims, they can still be mapped to a binary of 'true' and 'false.' For instance, the Portuguese claim "*O desmatamento ilegal subiu de 2012 pra cá em torno de 37%*" ("Illegal deforestation has increased by around 37% since 2012") is labeled as 'mostly true.'

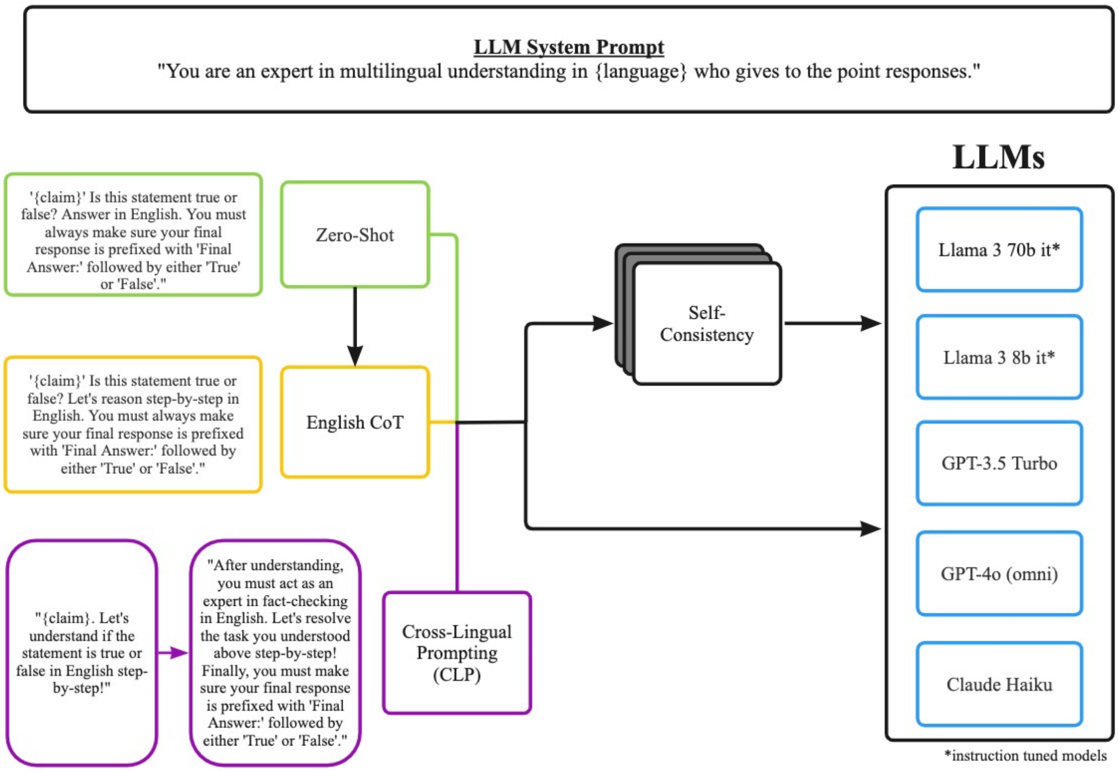

LLM System Prompt
"You are an expert in multilingual understanding in {language} who gives to the point responses."
LLMs
"'{claim}' Is this statement true or false? Answer in English. You must always make sure your final response is prefixed with 'Final Answer:' followed by either 'True' or 'False'."
Zero-Shot
"'{claim}' Is this statement true or false? Let's reason step-by-step in English. You must always make sure your final response is prefixed with 'Final Answer:' followed by either 'True' or 'False'."
English CoT
"{claim}. Let's understand if the statement is true or false in English step-by-step!"
"After understanding, you must act as an expert in fact-checking in English. Let's resolve the task you understood above step-by-step! Finally, you must make sure your final response is prefixed with 'Final Answer:' followed by either 'True' or 'False'."
Cross-Lingual Prompting (CLP)
Self-Consistency
Llama 3 70b it*
Llama 3 8b it*
GPT-3.5 Turbo
GPT-4o (omni)
Claude Haiku
*instruction tuned models

Figure 2: Prompting Methods: Zero-Shot, English Chain-of-Thought, Cross-Lingual Prompting, and Self-Consistency for multilingual fact-checking using LLMs

Although there is a minor inaccuracy in the quoted year among the five articles of evidence used by X-Fact to verify the claim, the core assertion is true. Therefore, we can reasonably map the claim to 'true.' We follow a similar line of reasoning for claims labeled as 'mostly false.' Additionally, consider an instance of a Spanish claim "*[El proyecto de Cambiemos] establece una quita de entre el 30% y el 60% para los jubilados que tienen juicio*" ("[The Cambiemos project] establishes a reduction of between 30% and 60% for retirees who have lawsuits") which is labeled as 'partly true/misleading.' While the claim contains a factual element (the reduction percentage), the primary assertion about the voluntary payment proposal applying to all retirees with lawsuits is misleading[1]. This misleading information outweighs the partly true aspect. Therefore, we can reasonably map the claim to 'false.' We follow a similar line of reasoning for the other claims labeled as 'partly true/misleading.'

Each claim has a binary ground truth label, i.e., '0' for false and '1' for true. As such, for a given language l, our dataset $\delta_l = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$,

A sample claim for each language from their respective datasets is presented in Figure 1. Appendix A contains a detailed breakdown of the test data for each language. It should be noted that all the claims were sourced from 2021 and earlier.

3.2 Models

We conduct our experiments on the instruction-tuned Llama 3 8B (8 billion parameters) and Llama 3 70B (70 billion parameters) (MetaAI, 2024), GPT-3.5-turbo[2], GPT-4o (OpenAI, 2024), and Claude 3 Haiku (Anthropic, 2023), all of which are pre-trained on multilingual corpora. For each model, we set the temperature to 0.7. The maximum possible token length for the model's outputs was set according to their respective context lengths. We provide the following system prompt to each LLM: "You are an expert in multilingual understanding in `{language}` who gives to-the-point responses," where "`{language}`" is the language

[1] A majority (3/4) of the articles used by X-Fact to verify the claim explicitly clarify that the reduction applies specifically to the 300,000 retirees with lawsuits against the National Social Security Administration (Anses), and not to all retirees with lawsuits.

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

in which the claim is written.

3.3 Evaluation

For each experiment, we record the number of correct, incorrect, and inconclusive responses returned by the LLM. We express the accuracy score of the LLM as the percentage of correct answers.

4 Experiments

Figure 2 displays the various prompting techniques we explore in this study.

Zero-Shot We use zero-shot prompting to create a baseline for each LLM's performance. We add the instruction "Answer in English" to our zero-shot prompts to ensure that the LLM's response is in English, as in preliminary tests the LLM would, in some cases, generate outputs in the language specified in the system prompt. This issue is specific to the zero-shot setting.

English Chain-of-Thought Chain-of-Thought (CoT) prompting performs significantly better than zero-shot prompting on a variety of reasoning tasks (Wei et al., 2023b) including fact-checking and claim verification (Hu et al., 2023). In CoT prompting, models are encouraged through k-shot examples to reason explicitly, in written-out steps.

We employ English CoT (EN-CoT) (Shi et al., 2022) by adding the instruction "Let's reason step-by-step *in English*" to the original instruction.

Cross-lingual Prompting Cross-lingual Prompting (CLP) (Qin et al., 2023) builds on EN-CoT prompting and exhibits substantial performance improvements on multilingual reasoning tasks. A CLP prompt includes a Cross-Lingual Alignment Prompt and a Task-Specific Solver prompt. The output of the Cross-Lingual Alignment prompt is included as context with the task-specific solver prompt, which generates the final output.

In this work, as presented in Figure 2, the Cross-Lingual Alignment Prompt involves prompting the LLM to "understand if the statement is true or false". The language model's prediction is generated through the Task-Specific Solver Prompt.

Self-Consistency In self-consistency, models are given an identical prompt multiple times and the most frequent answer is selected as the solution (Wang et al., 2023b). We explore a variant of each prompting method, i.e. zero-shot, EN-CoT, and CLP, modified with self-consistency. For our self-consistency experiments, we feed each prompt to the model three times and select the prediction that occurs the most frequently as the final answer. Note that if the three outputs for a given claim are all distinct, i.e. 'true', 'false' and 'inconclusive', we take the final output as 'inconclusive'.

5 Results

5.1 Zero-Shot

Accuracy As presented in Table 1, GPT-3.5-turbo has an average accuracy of 50%, GPT-4o stands out with the highest zero-shot accuracy at 55%, Llama 3 70B has an average accuracy of 54%, Llama 3 8B showcases an accuracy of 49%, and Claude 3 Haiku has an accuracy of 47%. These results more or less correspond with model size; larger models achieve a higher accuracy.

Inconclusive Responses We note that GPT-3.5-turbo, GPT-4o, Llama 3 70B, Llama 3 8B, and Claude 3 Haiku give an average of 74, 47, 48, 60, and 114 inconclusive responses respectively. Again, this more or less corresponds with model size; smaller models tend to have a higher number of inconclusive responses, and larger models tend to have fewer inconclusive responses.

5.2 English Chain-Of-Thought

Accuracy As presented in Table 1, GPT-3.5-Turbo, Llama 3 70B, and Llama 3 8B experience a significant decrease in average accuracy upon applying the English CoT method, with reductions of 9%, 7%, and 9% respectively. Conversely, GPT-4o and Claude 3 Haiku experience a slight increase in accuracy with increases of 2% and 3% respectively.

Inconclusive Responses We observe that GPT-3.5-Turbo, Llama 3 70B, and Llama 3 8B all experience a significant increase in average inconclusive responses with increases of 72, 45, and 41 respectively. Conversely, GPT-4o and Claude 3 Haiku experience a decrease in inconclusive responses, with reductions of 17 and 77 respectively. The increase in inconclusive responses alongside the decrease in accuracy suggests that models like GPT-3.5-Turbo, Llama 3 70B, and Llama 3 8B may struggle to provide the correct answer and follow simple instructions when faced with the structured reasoning demands of the English CoT method. The decrease in inconclusive responses and the slight increase in accuracy for GPT-4o and Claude 3 Haiku suggest

	% Accuracy						Inconclusive Responses					
Model	Spanish	Italian	Portuguese	Turkish	Tamil	Average	Spanish	Italian	Portuguese	Turkish	Tamil	Average
GPT-3.5-Turbo												
`0-shot`	49.00	49.40	42.60	53.80	56.40	50.00	82	70	138	64	17	74
`SC 0-shot`	56.20	45.80	41.60	53.40	60.80	52.00	44	96	161	69	89	92
`EN-CoT`	32.60	38.60	41.20	45.60	45.20	41.00	212	154	175	115	74	146
`SC EN-CoT`	32.00	37.80	37.00	44.60	52.60	41.00	246	146	174	146	38	150
`CLP`	35.40	37.00	38.40	54.80	56.20	44.00	177	181	189	55	41	129
`SC CLP`	31.00	34.00	36.60	52.40	56.20	42.00	220	204	189	79	17	142
GPT-4o												
`0-shot`	42.00	48.60	56.00	58.20	67.80	55.00	99	32	21	60	24	47
`SC 0-shot`	39.00	51.60	57.20	60.20	75.00	57.00	153	26	20	48	7	51
`EN-CoT`	53.00	51.60	57.40	58.20	64.80	57.00	75	10	16	36	11	30
`SC EN-CoT`	49.00	51.60	56.60	59.80	69.20	57.00	73	16	25	41	58	43
`CLP`	54.00	54.00	63.00	64.40	61.40	59.00	48	11	32	28	57	35
`SC CLP`	53.60	51.20	59.20	63.20	62.00	58.00	64	30	52	40	12	40
Llama 3 70B												
`0-shot`	41.80	52.40	49.00	58.80	66.00	54.00	108	36	65	29	2	48
`SC 0-shot`	45.00	50.00	49.00	58.20	64.40	53.00	88	29	79	23	26	49
`EN-CoT`	38.40	46.80	41.00	52.20	57.00	47.00	157	66	143	62	36	93
`SC EN-CoT`	36.20	43.20	37.40	51.40	56.80	45.00	183	90	203	91	8	115
`CLP`	50.20	52.00	51.40	51.80	58.40	53.00	8	4	2	6	49	14
`SC CLP`	43.40	47.80	46.80	54.40	51.60	49.00	74	46	128	40	10	60
Llama 3 8B												
`0-shot`	42.00	50.40	39.00	53.40	59.80	49.00	123	34	107	24	13	60
`SC 0-shot`	50.80	51.00	52.40	52.40	57.20	53.00	26	40	25	16	54	32
`EN-CoT`	34.40	39.00	39.20	45.20	50.40	42.00	183	89	118	89	26	101
`SC EN-CoT`	40.20	41.40	42.80	45.00	53.60	45.00	149	110	95	105	10	94
`CLP`	49.80	46.20	49.00	52.40	53.80	50.00	7	12	5	8	68	20
`SC CLP`	40.00	42.00	41.00	46.40	45.20	43.00	118	78	114	58	7	75
Claude 3 Haiku												
`0-shot`	36.80	45.80	40.20	51.00	62.80	47.00	185	94	162	88	40	114
`SC 0-shot`	39.40	48.20	49.40	55.40	63.80	51.00	162	63	104	58	36	85
`EN-CoT`	45.00	45.60	47.80	54.00	58.20	50.00	96	76	81	53	27	67
`SC EN-CoT`	45.60	44.40	48.40	55.40	59.20	51.00	118	71	74	62	70	79
`CLP`	38.20	41.00	38.60	47.80	58.20	45.00	183	135	150	128	66	132
`SC CLP`	35.80	39.20	41.40	45.20	61.80	45.00	207	141	148	139	17	130

Table 1: Percent accuracy and inconclusive responses per method, model, and language

that these models benefit from the structured reasoning of the English CoT method, enabling them to provide more precise and definitive answers.

5.3 Cross-Lingual Prompting

Accuracy As presented in Table 1, GPT-3.5-Turbo, Llama 3 70B, and Claude 3 Haiku experience a slight decrease in average accuracy upon applying the Cross-Lingual Prompting method, with reductions of 4%, 1%, and 3% respectively. Conversely, GPT-4o and Llama 3 8B experience minor increases in accuracy, with increases of 1% and 2% respectively.

Inconclusive Responses We note that Llama 3 70B, Llama 3 8B, and GPT-4o experience a drastic decrease in average inconclusive responses, with reductions of 34, 40, and 12 respectively. Interestingly, we also observe that Claude 3 Haiku and GPT-3.5-Turbo experience a significant increase in inconclusive responses with increases of 18 and 55 respectively.

5.4 Self-Consistency

Accuracy We show that Self-Consistency has varying impacts on average model accuracies given the prompting method it works with. In a 0-shot setting, we observe consistent increases in accuracy across the board for all models except GPT-3.5-Turbo. Specifically, Llama 3 70B, Llama 3 8B, GPT-4o, and Claude 3 Haiku show increases of 1%, 2%, 2%, and 1% respectively. For EN-CoT and CLP, applying self-consistency proves to be more effective for GPT-3.5-Turbo and GPT-4o, with accuracy increases of 1% and 2%, respectively. However, Llama 3 70B and Claude 3 Haiku experience insignificant changes in accuracy.

Inconclusive Responses We see there is a significant increase in average inconclusive outputs compared to the baseline. The highest number of inconclusive outputs in the Self-Consistency context comes from GPT-3.5-Turbo, with 150 inconclusive outputs. In contrast, Llama 3 70B, Llama 3 8B, GPT-4o, and Claude 3 Haiku produce 115, 75, 40, and 130 inconclusive outputs respectively.

5.5 Language-Specific Trends

Tamil consistently demonstrated higher accuracy across models when paired with any prompting method, with an average accuracy of 50%. Additionally, Tamil almost always has the lowest number of inconclusive outputs, averaging 30 inconclusive responses. Tamil was the only language in our dataset from the Dravidian language family in South Asia. In contrast, Italian and Spanish, both Romance languages, perform subpar compared to Tamil despite being higher-resourced, with average accuracies of 44% and 44% respectively, and average inconclusive outputs of 85 and 110. This disparity is discussed in more detail in Section 6.

A detailed summary of the results for each LLM's performance with every prompting method and language tested is presented in Appendix B.

6 Analysis and Discussion

6.1 Two-way ANOVA

We perform a two-way Analysis of Variance (ANOVA) to investigate the effects of two factors—the prompting techniques and the LLM model—on the observed accuracy scores. The ANOVA results reveal that both the technique ($F = 2.552$, $p = 0.03$) and model ($F = 11.633$, $p < 0.001$) factors have a statistically significant effect on the accuracy scores. To further understand the strength of the effects, we calculate the partial eta-squared η_p^2 values, which provide an estimate of the effect size for each factor.

The partial eta-squared value for the 'Model' factor is 0.2495, indicating a large effect size (Cohen, 1988). This suggests that approximately 24.95% of the variance in the accuracy score is attributable to the LLM model, after accounting for the prompting technique. In contrast, the partial eta-squared value for the 'Technique' factor is 0.0835, corresponding to a medium effect size. This suggests that approximately 8.35% of the variance in the accuracy score is attributable to the prompting technique, after accounting for the LLM model.

Given the substantial effect size associated with the LLM model factor, further analysis is needed to understand the underlying factors contributing to the statistically significant effect of prompting technique on accuracy scores. We conduct two separate two-way ANOVAs for the self-consistent (SC) and non-self-consistent (non-SC) techniques.

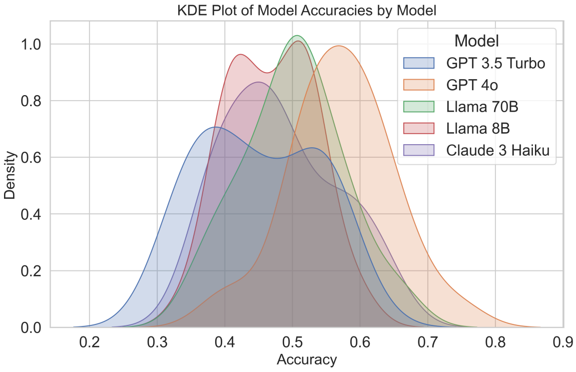


Figure 3: KDE Distribution of Accuracies by Model

6.2 Impact of Prompting Techniques

Overall, both the LLM model ($F = 5.477$, $p < 0.001$) and the SC prompting technique ($F = 4.332$, $p = 0.017$) had significant effects on the accuracy score. However, for non-SC techniques, the LLM model had a significant effect ($F = 6.149$, $p < 0.01$), but the non-SC prompting technique did not have a statistically significant impact ($F = 1.731$, $p = 0.185$) on the accuracy score. This suggests that the self-consistency decoding strategies are the primary drivers behind the significant effect of the prompting technique. EN-CoT and CLP are designed to improve reasoning capabilities in LLMs (Shi et al., 2022; Qin et al., 2023), so their negligible impact in fact-checking suggests that improvements in reasoning ability do not improve claim verification accuracy.

6.3 Visualization and Distribution Analysis

To visualize and analyze the distribution of model accuracies across various factors, we use Kernel Density Estimation (KDE) plots. KDE is a non-parametric technique that produces a smooth, continuous estimate of the probability density function for a given variable. The density curve represents the likelihood of the relative probability of observing different accuracy values for each model, technique, or language category. A higher value on the density curve indicates a higher probability of achieving that accuracy level, while a lower value on the density curve indicates a lower probability of achieving that accuracy level.

In Figure 3, we can observe that the GPT 4o model exhibits the highest accuracy density peaking at around 0.57. The relatively narrow distribution suggests that GPT 4o performs consistently close to the peak value (0.57). This consistency suggests that GPT 4o is more reliable and gener-

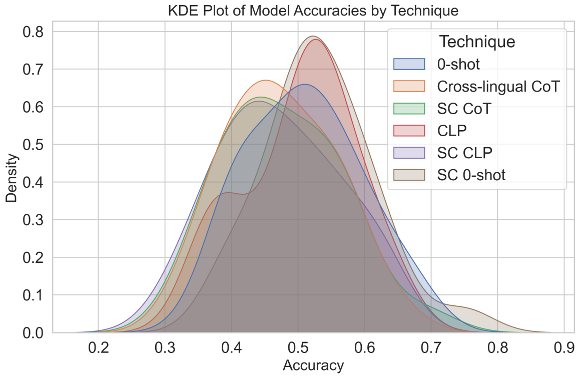


Figure 4: KDE Distribution of Accuracies by Technique

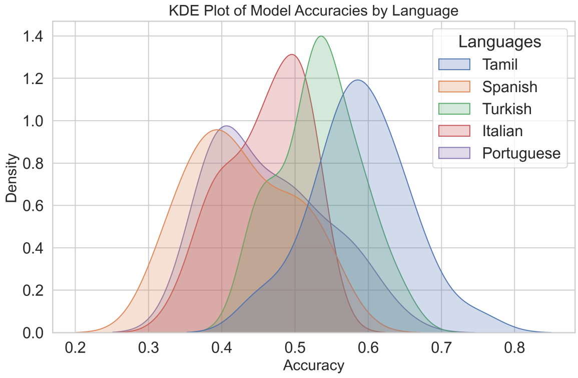


Figure 5: KDE Distribution of Accuracies by Language

ally outperforms the other models.

In Figure 4, we can observe a close performance between CLP and SC 0-shot. CLP has a slightly higher accuracy density peaking around 0.54 while SC 0-shot's highest accuracy density peaks at around 0.52. However, the distribution of SC 0-shot is broader, indicating greater variability in accuracy. This variability gives SC 0-shot the potential to achieve higher accuracy scores, approximately up to 0.85. This variability indicates that SC 0-shot is generally more likely to outperform other techniques.

In Figure 5, we can observe that Tamil, categorized as a low-resource language, exhibits the highest accuracy among these languages. This finding contradicts the conventional expectation that high-resource languages, with the abundance of data, would outperform low-resource counterparts.

6.4 Correlation Analysis

Language	Internet Content (%)
Spanish	5.8%
Portuguese	3.6%
Italian	2.6%
Turkish	1.9%
Tamil	$< 0.1\%$

Table 2: Percentage of internet content by language

Table 2 presents the percentage of internet content for each language (W3Techs, 2024). Using this data, we perform a correlation analysis where we calculate the Pearson correlation coefficient ρ[3] between the language's accuracy and its percentage of internet content. The correlation analysis reveals a strong negative correlation where $\rho = -0.924$, suggesting that languages with less internet data tend to have higher accuracy scores, while those with more internet data tend to have lower accuracy scores. We hypothesize that for languages like Tamil, which have relatively scarce internet content, the available data is likely of higher quality and reliability. Conversely, the abundance of content for high-resource languages like Spanish or Portuguese may introduce significant noise, contradictory information, and lower-quality data into the training corpus for the LLMs tested.

7 Conclusion and Future Work

In this study, we assessed the performance of five large language models (LLMs) in verifying claims in five languages (Spanish, Portuguese, Italian, Turkish, and Tamil) using the X-Fact dataset. Our findings indicate that both the choice of model and the prompting technique significantly impact fact-checking performance. Notably, GPT-4o consistently achieved higher accuracy than the other models, likely due to its advanced architecture and larger size. Interestingly, a simple self-consistency and zero-shot prompt combination outperformed all other prompting and decoding strategies, suggesting that not all reasoning strategies are beneficial for claim verification. Strategies such as Chain-of-Thought or Cross-Lingual Prompting, which aim to alter the model's reasoning process, often had minimal or negative effects on success rates. In contrast, decoding strategies such as self-consistency show potential as a future research direction.

We also discovered a surprising correlation between higher model accuracy and lower language internet content, indicating that models performed better on low-resource languages. Further investigation is needed to understand the causes and

[3] Note that the function of $\rho \in [-1, 1]$.

extent of this relationship.

For future work, we plan to delve deeper into the relationship between model performance and the extent of a language's internet corpora. We will also develop and test new, custom-designed prompting techniques and decoding strategies specifically tailored to enhance claim-verification performance. Additionally, we aim to experiment with other leading models such as Claude 3 Opus, Gemini-1.5 Pro, and the Llama 3.1 model series. We will expand our study to include more high and low-resource languages from the X-Fact dataset, such as French, Russian, Indonesian, and Romanian.

Limitations

Although our study represents progress in understanding LLM fact-checking capabilities and reveals interesting results, it is affected by several potential limitations. The dataset we used, X-Fact, was published in 2021 and may be present in the pre-training data of some of the models we tested. Additionally, because the dataset is from 2021, some temporally evolving claims might contribute to noise in our final datasets, as the factual status of certain statements may have changed since the dataset's creation. We also tested a relatively limited set of languages and models. To make more definitive statements about model performance concerning language resources, we would need to test on a much larger range of languages.

Additionally, we began testing on GPT-4-Turbo and Gemini 1.0 Pro, but due to budget constraints and runtime issues, we were unable to complete all of the experiments. However, the results of the experiments we were able to run on both of these models are presented in Appendix C.

Ethics Statement

All data used in this research were obtained from publicly available sources, ensuring no privacy violations or ethical breaches. This study aims to enhance the capabilities of fact-checking in multiple languages using large language models (LLMs) and combat misinformation. We acknowledge several potential risks associated with our work. First, we acknowledge the possibility of the LLMs tested being misused to generate disinformation or fake profiles, which could exacerbate the spread of false information, particularly in low-resource languages with limited fact-checking resources. Second, inherent biases in the models might lead to unfair outcomes, disadvantaging speakers of less-represented languages and further exacerbating existing inequalities. Third, our work involves publicly available datasets, but we ensure that no sensitive or private information is inadvertently included in the testing process. Finally, we acknowledge that the models could be vulnerable to adversarial attacks, where manipulated input data could deceive the model into making incorrect fact-checking decisions.

To mitigate these risks, we propose several strategies. We emphasize the importance of clear usage guidelines to prevent the misuse of LLMs (Dong et al., 2024). Continuous monitoring for bias and the implementation of fairness-aware pre-training techniques can help mitigate bias and ensure more equitable performance across different languages (Gallegos et al., 2024). Strict data handling protocols should be implemented to protect privacy, including anonymization and data minimization techniques (Mozes et al., 2023). Developing and integrating robust defenses against adversarial attacks is crucial to safeguarding the integrity of fact-checking systems.

We advocate for ongoing research to improve the accuracy and fairness of LLMs, especially in multilingual contexts. Our research aligns with promoting social good and advancing natural language processing to benefit diverse linguistic communities.

Acknowledgments

We would like to thank Rishab Kumar Jain for his grant, which helped cover the costs associated with this project.

References

Anthropic. 2023. Claude 3 model card. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study.

Recep Firat Cekinel, Pinar Karagoz, and Cagri Coltekin. 2024. Cross-lingual learning vs. low-resource finetuning: A case study with fact-checking in turkish.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking.

Eun Cheol Choi and Emilio Ferrara. 2024. Fact-gpt: Fact-checking augmentation via claim matching with llms. *arXiv preprint arXiv:2402.05904*.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*, 2 edition. Lawrence Erlbaum Associates.

Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Building guardrails for large language models.

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

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

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging chatgpt for efficient fact-checking. *PsyArXiv. April*, 3.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts?

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact-checkers? *arXiv preprint arXiv:2006.04102*.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D. Griffin. 2023. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-06-11.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality.

W3Techs. 2024. Usage statistics and market share of content languages for websites, june 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models — arxiv.org. https://arxiv.org/abs/2201.11903.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024. Do we need language-specific fact-checking models? the case of chinese.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

A Testing Datasets

The X-fact dataset (Gupta and Srikumar, 2021) was utilized as our primary data source for the claims. Gupta and Srikumar provide the dataset for public use under the MIT License[4]. All personal and offensive information was anonymized and/or removed by Gupta and Srikumar. We double-checked and ensured that all personal and offensive information was anonymized and/or removed from our final datasets.

[4] https://opensource.org/license/mit

A.1 Data Preprocessing

1. Filtering: We filtered the dataset first to include claims labeled as either "true" or "false" and then "mostly true", "mostly false", or "partly true/misleading" if the number of fully true or false claims fell short. Claims with other labels or those lacking verification were excluded from the finalized dataset.

2. Combining Splits: After filtering, the claims from the Dev, Train, In-domain Test (α_1), Out-of-domain (α_2), and Zero-Shot (α_3) splits in the X-Fact dataset were randomly shuffled and combined to form a final dataset of 500 (250 true and 250 false) claims for our experiments.

A.2 Spanish Dataset

The claims in the final dataset for Spanish were sourced from `chequeado.com`, an Argentinian fact-checking website.

A.2.1 Dataset Composition

Table A1 shows a breakdown of the total number of Spanish claims in the X-Fact dataset and the number of Spanish claims filtered for the finalized dataset.

A.2.2 Label Distribution Percentage

True Claims: 34.0%
False Claims: 19.6%
Mostly True Claims: 16.0%
Mostly False Claims: 0.0%
Partly True/Misleading Claims: 30.4%

A.3 Portuguese Dataset

The claims in the final dataset for Portuguese were sourced from `piaui.folha.uol.com.br`, a Brazilian monthly magazine, and `poligrafo.sapo.pt`, a Portuguese newspaper dedicated to fact-checking.

A.3.1 Dataset Composition

Table A2 shows a breakdown of the total number of Portuguese claims in the X-Fact dataset and the number of Portuguese claims filtered for the finalized dataset.

A.3.2 Label Distribution Percentage

True Claims: 35.2%
False Claims: 36.2%
Mostly True Claims: 14.8%
Mostly False Claims: 0.0%
Partly True/Misleading Claims: 13.8%

A.4 Italian Dataset

The claims in the final dataset for Italian were sourced from `pagellapolitica.it`, an Italian fact-checking organization that verifies the accuracy of statements made by politicians, and `agi.it`, an Italian news agency that provides news coverage of national and international events.

A.4.1 Dataset Composition

Table A3 shows a breakdown of the total number of Italian claims in the X-Fact dataset and the number of Italian claims filtered for the finalized dataset.

A.4.2 Label Distribution Percentage

True Claims: 28.0%
False Claims: 26.2%
Mostly True Claims: 22.0%
Mostly False Claims: 0.0%
Partly True/Misleading Claims: 23.8%

A.5 Turkish Dataset

The claims in the final dataset for Turkish were sourced from `dogrulukpayi.com`, a Turkish fact-checking platform that evaluates the accuracy of statements made by Turkish politicians and public figures, and `teyit.org`, an independent fact-checking organization based in Turkey.

A.5.1 Dataset Composition

Table A4 shows a breakdown of the total number of Turkish claims in the X-Fact dataset and the number of Turkish claims filtered for the finalized dataset.

A.5.2 Label Distribution Percentage

True Claims: 35.2%
False Claims: 25.4%
Mostly True Claims: 14.8%
Mostly False Claims: 7.2%
Partly True/Misleading Claims: 17.4%

A.6 Tamil Dataset

The claims in the final dataset for Tamil were sourced from `youturn.in`, an Indian fact-checking website that debunks misinformation on social media.

A.6.1 Dataset Composition

Table A5 shows a breakdown of the total number of Tamil claims in the X-Fact dataset and the number of Tamil claims filtered for the finalized dataset.

X-Fact Dataset Split	Total Number of Spanish Claims	Filtered Number of Spanish Claims				
		True Claims	False Claims	Mostly True Claims	Mostly False Claims	Partly True/Misleading Claims
Dev	126	17	11	8	0	19
Train	1011	127	78	60	0	107
In-domain Test (α_1)	195	26	9	12	0	26
Out-of-domain Test (α_2)	0	0	0	0	0	0
Zero-Shot Test (α_3)	0	0	0	0	0	0
Total	**1332**	**170**	**98**	**80**	**0**	**152**

Table A1: Summary of the dataset splits before and after filtering the claims for Spanish

X-Fact Dataset Split	Total Number of Portuguese Claims	Filtered Number of Portuguese Claims				
		True Claims	False Claims	Mostly True Claims	Mostly False Claims	Partly True/Misleading Claims
Dev	718	17	17	6	0	9
Train	5418	137	135	57	0	47
In-domain Test (α_1)	1073	20	24	11	0	7
Out-of-domain Test (α_2)	471	2	5	0	0	6
Zero-Shot Test (α_3)	0	0	0	0	0	0
Total	**7680**	**176**	**181**	**74**	**0**	**69**

Table A2: Summary of the dataset splits before and after filtering the claims for Portuguese

X-Fact Dataset Split	Total Number of Italian Claims	Filtered Number of Italian Claims				
		True Claims	False Claims	Mostly True Claims	Mostly False Claims	Partly True/Misleading Claims
Dev	120	4	15	12	0	10
Train	909	84	83	80	0	94
In-domain Test (α_1)	185	12	15	18	0	14
Out-of-domain Test (α_2)	250	40	18	0	0	1
Zero-Shot Test (α_3)	0	0	0	0	0	0
Total	**1464**	**140**	**131**	**110**	**0**	**119**

Table A3: Summary of the dataset splits before and after filtering the claims for Italian

X-Fact Dataset Split	Total Number of Turkish Claims	Filtered Number of Turkish Claims				
		True Claims	False Claims	Mostly True Claims	Mostly False Claims	Partly True/Misleading Claims
Dev	105	19	9	10	4	3
Train	827	80	44	57	26	44
In-domain Test (α_1)	162	19	7	7	6	10
Out-of-domain Test (α_2)	610	58	67	0	0	30
Zero-Shot Test (α_3)	0	0	0	0	0	0
Total	**1704**	**176**	**127**	**74**	**36**	**87**

Table A4: Summary of the dataset splits before and after filtering the claims for Turkish

X-Fact Dataset Split	Total Number of Tamil Claims	Filtered Number of Tamil Claims				
		True Claims	False Claims	Mostly True Claims	Mostly False Claims	Partly True/Misleading Claims
Dev	140	27	23	0	0	2
Train	1054	178	164	0	0	30
In-domain Test (α_1)	209	45	26	0	0	5
Out-of-domain Test (α_2)	0	0	0	0	0	0
Zero-Shot Test (α_3)	0	0	0	0	0	0
Total	**1403**	**250**	**213**	**0**	**0**	**37**

Table A5: Summary of the dataset splits before and after filtering the claims for Tamil

A.6.2 Label Distribution Percentage

True Claims: 50.0%
False Claims: 42.6%
Mostly True Claims: 0.0%
Mostly False Claims: 0.0%
Partly True/Misleading Claims: 7.4%

B Results Breakdown

The tables in this section summarize each LLM's performance with every prompting method and language tested in this study.
Table B1 presents the results for each prompting method and LLM for Spanish.

Table B2 presents the results for each prompting method and LLM for Portuguese.

Table B3 presents the results for each prompting method and LLM for Italian.

Table B4 presents the results for each prompting method and LLM for Turkish.

Table B5 presents the results for each prompting method and LLM for Tamil.

C Miscellaneous Results

C.1 GPT-4 Turbo

We ran experiments on GPT-4 Turbo for Tamil, excluding self-consistency for 0-shot. The results are presented in Table C1.

C.2 Gemini-1.0 Pro

We ran experiments on Gemini-1.0 Pro for Spanish and Tamil, excluding self-consistency for 0-shot, and for Turkish where we excluded both self-consistency on English CoT and self-consistency for 0-shot.

The results for Spanish are presented in Table C2.

The results for Turkish are presented in Table C3.

The results for Tamil are presented in C4.

C.3 Two-Way ANOVA

Table C5 details the two-way ANOVA results for the LLMs and prompting techniques tested on model accuracy.

Table C6 details the two-way ANOVA results for the LLMs and non-self-consistency prompting techniques tested on model accuracy.

Table C7 details the two-way ANOVA results for the LLMs and self-consistency prompting techniques tested on model accuracy.

D Computational Details

D.1 Expenditure

Across all of the experiments[5], we spent \$175 worth of OpenAI credits to run GPT-3.5 Turbo, GPT-4o, and GPT-4 Turbo[6], and \$30 worth of Anthropic credits to run Claude 3 Haiku[7]. To run the Llama 3 series of models, we used the Groq API[8], which is free as the models are open source. We conducted our experiments primarily on Intel Core i7 processors and Google Colab TPUs, totaling approximately 80 hours of runtime.

D.2 Software Packages Used

To build our datasets, we used conventional Python 3.12.3 libraries to take a subset of the X-Fact dataset. For our data and result analysis, we used Matplotlib (version 3.9.0) and Seaborn (version 0.13.2). For our statistical analysis, we used SciPy (version 1.13.1).

[5] Most of the computational experiments we ran were on privately owned LLMs. Therefore, we were unable to report the exact number of parameters for some of the LLMs used in our experiments (GPT-3.5 Turbo, GPT-4o, and Claude 3). However, the Llama 3 series of models is open source. Specific details about the models can be found at the following: https://llama.meta.com/llama3/

[6] OpenAI Pricing: https://openai.com/api/pricing/

[7] Anthropic Pricing: https://www.anthropic.com/api

[8] Groq API documentation: https://console.groq.com/docs/quickstart

Model	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
GPT-3.5 Turbo					
0-shot	245	173	82	49.00%	–
SC 0-shot	281	175	44	**56.20%**	7.20%
EN-CoT	163	125	212	32.60%	-16.40%
SC EN-CoT	160	94	246	32.00%	-17.00%
CLP	177	146	177	35.40%	-13.60%
SC CLP	155	125	220	31.00%	-18.00%
GPT-4o					
0-shot	210	191	99	42.00%	–
SC 0-shot	195	152	153	39.00%	-3.00%
EN-CoT	265	160	75	53.00%	11.00%
SC EN-CoT	245	182	73	49.00%	7.00%
CLP	270	182	48	**54.00%**	12.00%
SC CLP	268	168	64	53.60%	11.60%
Llama 3 70B					
0-shot	209	183	108	41.80%	–
SC 0-shot	225	187	88	45.00%	3.20%
EN-CoT	192	151	157	38.40%	-3.40%
SC EN-CoT	181	136	183	36.20%	-5.60%
CLP	251	241	8	**50.20%**	8.40%
SC CLP	217	209	74	43.40%	1.60%
Llama 3 8B					
0-shot	210	167	123	42.00%	–
SC 0-shot	254	220	26	**50.80%**	8.80%
EN-CoT	172	145	183	34.40%	-7.60%
SC EN-CoT	201	150	149	40.20%	-1.80%
CLP	249	244	7	49.80%	7.80%
SC CLP	200	182	118	40.00%	-2.00%
Claude 3 Haiku					
0-shot	184	131	185	36.80%	–
SC 0-shot	197	141	162	39.40%	2.60%
EN-CoT	225	179	96	45.00%	8.20%
SC EN-CoT	228	154	118	**45.60%**	8.80%
CLP	191	126	183	38.20%	1.40%
SC CLP	179	114	207	35.80%	-1.00%

Table B1: Results for each LLM and prompting method in Spanish.'% Increase' denotes the percentage increase in model performance from the baseline (0-shot).

Model	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
GPT-3.5 Turbo					
0-shot	213	149	138	**42.60%**	–
SC 0-shot	208	131	161	41.60%	-1.00%
EN-CoT	206	119	175	41.20%	-1.40%
SC EN-CoT	185	141	174	37.00%	-5.60%
CLP	192	119	189	38.40%	-4.20%
SC CLP	183	128	189	36.60%	-6.00%
GPT-4o					
0-shot	280	199	21	56.00%	–
SC 0-shot	286	194	20	57.20%	1.20%
EN-CoT	287	197	16	57.40%	1.40%
SC EN-CoT	283	192	25	56.60%	0.60%
CLP	315	153	32	**63.00%**	7.00%
SC CLP	296	152	52	59.20%	3.20%
Llama 3 70B					
0-shot	245	190	65	49.00%	–
SC 0-shot	245	176	79	49.00%	0.00%
EN-CoT	205	152	143	41.00%	-8.00%
SC EN-CoT	187	110	203	37.40%	-11.60%
CLP	257	241	2	**51.40%**	2.40%
SC CLP	234	138	128	46.80%	-2.20%
Llama 3 8B					
0-shot	195	198	107	39.00%	–
SC 0-shot	262	213	25	**52.40%**	13.40%
EN-CoT	196	186	118	39.20%	0.20%
SC EN-CoT	214	191	95	42.80%	3.80%
CLP	245	250	5	49.00%	10.00%
SC CLP	205	181	114	41.00%	2.00%
Claude 3 Haiku					
0-shot	201	137	162	42.20%	–
SC 0-shot	247	149	104	**49.40%**	7.20%
EN-CoT	239	180	81	47.80%	5.60%
SC EN-CoT	242	184	74	48.40%	6.20%
CLP	193	157	150	38.60%	-3.60%
SC CLP	207	145	148	41.40%	-0.80%

Table B2: Results for each LLM and prompting method in Portuguese.'% Increase' denotes the percentage increase in model performance from the baseline (0-shot).

Model	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
GPT-3.5 Turbo					
0-shot	247	183	70	**49.40%**	–
SC 0-shot	229	175	96	45.80%	-3.60%
EN-CoT	193	153	154	38.60%	-10.80%
SC EN-CoT	189	165	146	37.80%	-11.60%
CLP	185	134	181	37.00%	-12.40%
SC CLP	170	126	204	34.00%	-15.40%
GPT-4o					
0-shot	243	225	32	48.60%	–
SC 0-shot	258	216	26	51.60%	3.00%
EN-CoT	258	232	10	51.60%	3.00%
SC EN-CoT	258	226	16	51.60%	3.00%
CLP	270	219	11	**54.00%**	5.40%
SC CLP	256	214	30	51.20%	2.60%
Llama 3 70B					
0-shot	262	202	36	**52.40%**	–
SC 0-shot	250	221	29	50.00%	-2.40%
EN-CoT	234	200	66	46.80%	-5.60%
SC EN-CoT	216	194	90	43.20%	-9.20%
CLP	260	236	4	52.00%	-0.40%
SC CLP	239	215	46	47.80%	-4.60%
Llama 3 8B					
0-shot	244	222	34	50.41%	–
SC 0-shot	255	205	40	**51.00%**	0.59%
EN-CoT	195	216	89	39.00%	-11.41%
SC EN-CoT	207	183	110	41.40%	-9.01%
CLP	231	257	12	46.20%	-4.21%
SC CLP	210	212	78	42.00%	-8.41%
Claude 3 Haiku					
0-shot	229	177	94	45.80%	–
SC 0-shot	241	196	63	**48.20%**	2.40%
EN-CoT	228	196	76	45.60%	-0.20%
SC EN-CoT	222	207	71	44.40%	-1.40%
CLP	193	157	150	38.60%	-7.20%
SC CLP	196	163	141	39.20%	-6.60%

Table B3: Results for each LLM and prompting method in Italian.‘% Increase’ denotes the percentage increase in model performance from the baseline (0-shot).

Model	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
GPT-3.5 Turbo					
0-shot	269	167	64	53.80%	–
SC 0-shot	267	164	69	53.40%	-0.40%
EN-CoT	228	157	115	45.60%	-8.20%
SC EN-CoT	223	131	146	44.60%	-9.20%
CLP	274	171	55	**54.80%**	1.00%
SC CLP	262	159	79	52.40%	-1.40%
GPT-4o					
0-shot	291	149	60	58.20%	–
SC 0-shot	301	151	48	60.20%	2.00%
EN-CoT	291	173	36	58.20%	0.00%
SC EN-CoT	299	160	41	59.80%	1.60%
CLP	322	150	28	**64.40%**	6.20%
SC CLP	316	144	30	63.20%	5.00%
Llama 3 70B					
0-shot	294	177	29	**58.80%**	–
SC 0-shot	291	186	23	58.20%	-0.60%
EN-CoT	261	177	62	52.20%	-6.60%
SC EN-CoT	257	152	91	51.40%	-7.40%
CLP	259	235	6	51.80%	-7.00%
SC CLP	272	188	40	54.40%	-4.40%
Llama 3 8B					
0-shot	267	209	24	**53.40%**	–
SC 0-shot	262	222	16	52.40%	-1.00%
EN-CoT	226	185	89	45.20%	-8.20%
SC EN-CoT	225	170	105	45.00%	-8.40%
CLP	262	230	8	52.40%	-1.00%
SC CLP	232	210	58	46.40%	-7.00%
Claude 3 Haiku					
0-shot	255	157	80	51.00%	–
SC 0-shot	277	165	58	**55.40%**	4.40%
EN-CoT	270	177	53	54.00%	3.00%
SC EN-CoT	277	161	62	**55.40%**	4.40%
CLP	239	133	128	47.80%	-3.20%
SC CLP	226	135	139	45.20%	-5.80%

Table B4: Results for each LLM and prompting method in Turkish.‘% Increase’ denotes the percentage increase in model performance from the baseline (0-shot).

Model	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
GPT-3.5 Turbo					
`0-shot`	282	201	17	56.40%	–
`SC 0-shot`	304	179	17	**60.80%**	4.40%
`EN-CoT`	226	185	89	45.20%	-11.20%
`SC EN-CoT`	263	163	74	52.60%	-3.80%
`CLP`	281	181	38	56.20%	-0.20%
`SC CLP`	281	178	41	56.20%	-0.20%
GPT-4o					
`0-shot`	339	137	24	67.80%	–
`SC 0-shot`	375	113	12	**75.00%**	7.20%
`EN-CoT`	324	169	7	64.80%	-3.00%
`SC EN-CoT`	346	143	11	69.20%	1.40%
`CLP`	307	135	58	61.40%	-6.40%
`SC CLP`	310	133	57	62.00%	-5.80%
Llama 3 70B					
`0-shot`	330	168	2	**66.00%**	–
`SC 0-shot`	322	168	10	64.40%	-1.60%
`EN-CoT`	285	189	26	57.00%	-9.00%
`SC EN-CoT`	284	180	36	56.80%	-9.20%
`CLP`	292	200	8	58.40%	-7.60%
`SC CLP`	258	193	49	51.60%	-14.40%
Llama 3 8B					
`0-shot`	299	188	13	**59.80%**	–
`SC 0-shot`	286	207	7	57.20%	-2.60%
`EN-CoT`	252	194	54	50.40%	-9.40%
`SC EN-CoT`	268	206	26	53.60%	-6.20%
`CLP`	269	221	10	53.80%	-6.00%
`SC CLP`	226	206	68	45.20%	-14.60%
Claude 3 Haiku					
`0-shot`	314	146	40	62.80%	–
`SC 0-shot`	319	164	17	**63.80%**	1.00%
`EN-CoT`	291	173	36	58.20%	-4.60%
`SC EN-CoT`	296	177	27	59.20%	-3.60%
`CLP`	291	139	70	58.20%	-4.60%
`SC CLP`	309	125	66	61.80%	-1.00%

Table B5: Results for each LLM and prompting method in Tamil.'% Increase' denotes the percentage increase in model performance from the baseline (0-shot).

Prompting Technique	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
0-shot	353	145	2	**70.60%**	–
EN-CoT	310	178	12	62.00%	-8.60%
SC EN-CoT	309	185	6	61.80%	-8.80%
CLP	316	129	55	63.20%	-7.40%
SC CLP	322	127	51	64.40%	-6.20%

Table C1: Results for GPT-4 Turbo on Tamil.'% Increase' denotes the percentage increase in GPT-4 Turbo's performance from the baseline (0-shot).

Prompting Technique	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
0-shot	277	222	1	**55.40%**	–
EN-CoT	236	179	85	47.20%	-8.20%
SC EN-CoT	230	176	94	46.00%	-9.40%
CLP	246	198	56	49.20%	-6.20%
SC CLP	252	192	56	50.40%	-5.00%

Table C2: Results for Gemini-1.0 Pro on Spanish.'% Increase' denotes the percentage increase in Gemini's performance from the baseline (0-shot).

Prompting Technique	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
0-shot	289	211	0	57.80%	–
EN-CoT	273	167	60	54.60%	-3.20%
CLP	293	190	17	58.60%	0.80%
SC CLP	298	162	40	**59.60%**	1.80%

Table C3: Results for Gemini-1.0 Pro on Turkish.'% Increase' denotes the percentage increase in Gemini's performance from the baseline (0-shot).

Prompting Technique	Correct	Incorrect	Inconclusive	% Accuracy	% Increase
0-shot	307	173	20	**61.40%**	–
EN-CoT	282	140	78	56.40%	-5.00%
SC EN-CoT	302	121	77	60.40%	-1.00%
CLP	306	139	55	61.20%	-0.20%
SC CLP	277	105	118	55.40%	-6.00%

Table C4: Results for Gemini-1.0 Pro on Tamil.'% Increase' denotes the percentage increase in Gemini's performance from the baseline (0-shot).

Source	Sum of Squares	Degrees of Freedom	F-statistic	p-value
Technique	0.072164	5.0	2.552192	3.039257e-02
Model	0.263142	4.0	11.632972	3.487599e-08

Table C5: Two-way ANOVA results for the LLMs and prompting techniques on accuracy

Source	Sum of Squares	Degrees of Freedom	F-statistic	p-value
Technique	0.018772	2.0	1.731207	0.184783
Model	0.133341	4.0	6.148595	0.000277

Table C6: Two-way ANOVA results for the LLMs and non-self-consistency prompting techniques on accuracy

Source	Sum of Squares	Degrees of Freedom	F-statistic	p-value
Technique	0.053283	2.0	4.332635	0.016941
Model	0.134711	4.0	5.476887	0.000698

Table C7: Two-way ANOVA results for the LLMs and self-consistency prompting techniques on accuracy