The proliferation of misinformation across digital platforms necessitates robust and swift verification methods to counteract its rapid spread. This study evaluates the efficacy of large language models (LLMs) such as OpenAI's GPT-3.5 and GPT-4, alongside smaller open-source models like Llama-2 and Mistral 7B, in multilingual fact-checking tasks. We reveal significant variations in model performance, where GPT-4 showed high accuracy but conservative verdict delivery. Conversely, GPT-3.5, previously reported to achieve 72% accuracy on English claims, significantly underperformed on multilingual data, highlighting the challenges of applying LLMs across diverse linguistic contexts. Similarly, Llama-2, along with Mistral 7B, despite its decisiveness, demonstrated limited accuracy. The feature-specific analysis identifies key predictors of accuracy across models, noting annual performance improvements and the consistent positive impact of languages like Bengali and German. These findings underscore the critical need for model-specific tuning and more nuanced labeling strategies to enhance the practical application of LLMs in addressing the complexity of real-world claims, thus advancing their utility in the fight against misinformation.

The urgency for automating the fact-checking process stems from the alarming speed at which falsehoods can propagate. Studies reveal that false claims travel significantly faster than truthful ones, and note that the spread of viral assertions often peaks within minutes of their online posting. This velocity outstrips the pace at which traditional, manual fact-checking can operate, pointing to a clear need for more instantaneous methods. Selecting which claims to verify is an intricate endeavor. The manual process of fact-checking is not only slow but requires meticulous effort to determine the checkability and veracity of a claim. Journalists and professional fact-checkers spend hours, sometimes days, validating a single claim by consulting multiple sources and gauging their reliability. Moreover, the multilingual nature of misinformation adds another layer of complexity, necessitating tools that can recognize and translate claims across different languages.

Automated fact-checking tools could significantly expedite this process. They have the potential to swiftly access and sift through vast repositories of data from trusted sources, saving fact-checkers valuable time and enabling them to counter misinformation effectively. This is particularly beneficial when the evidence is buried in extensive documents, multimedia content, or is in an unfamiliar language. Thus, automating the fact-checking process is not only a measure to enhance efficiency but a strategic move to bolster the integrity of information in the digital age.

My research builds upon the work of Hoes et al., who tested political claims in English using ChatGPT and reported 72% accuracy. I extend this research by evaluating additional models and expanding the scope to multilingual datasets covering a broader range of topics. The study assesses various evaluative dimensions of LLM performance, including temporal performance evaluation, performance across languages, and methodological variations such as simple base prompts, few-shot prompting, and setting the temperature to 0.

One of the most significant findings from this study was that GPT-3.5 failed to replicate the 72% accuracy previously reported for English US-based political claims, managing only a maximum accuracy of 61.52% under the best-performing setting. This result underscores the challenges associated with applying large language models to multilingual datasets, where linguistic

diversity and contextual nuances can dramatically impact performance. In contrast, GPT-4 demonstrated superior capabilities, achieving an accuracy of 79.11% in its best setting, although it notably refrained from providing a verdict on approximately 68% of the claims, indicating a possibly conservative approach in its processing or a higher standard of certainty for response generation. Additionally, our exploration into smaller open-source models revealed a less encouraging picture: Llama-2's highest accuracy was only 61.75% in Mode 4a, and Mistral 7B's performance was near baseline levels at 50.25%, which is marginally above the threshold of random chance success. Mistral 7B, however, performed better on English-only claims, suggesting specific limitations in its current configurations when dealing with a broader linguistic spectrum.

These results collectively highlight the importance of model choice, configuration, and the inherent trade-offs between decisiveness and accuracy in the application of language models for fact-checking across diverse linguistic contexts. Although few-shot prompting emerged as the best-performing setting for GPT-3.5, attempts to replicate this success with Llama-2 and Mistral 7B were unsuccessful. Both models exhibited a significant bias influenced by the order of examples in the prompt. Notably, Llama-2 consistently adhered to the verdicts of the example claims, regardless of changes in the order of the presented claims. This behavior underscores a critical challenge in using few-shot techniques: some models may be inherently more susceptible to biases introduced by the structure and content of their input prompts. This variation in susceptibility highlights the necessity for careful prompt design and perhaps model-specific adaptations to mitigate such biases effectively.

All models struggled with "partly true/misleading" claims, which points to a significant challenge in the field: the ability of models to handle nuance and ambiguity. This difficulty suggests that current models are still far from understanding the subtleties of human language and the complex ways in which information can be presented to mislead. Our study uncovered significant disparities in how different methodological settings - namely the five modes - impacted model performance on the same set of claims. Notably, certain configurations yielded better results for true claims, while others were more effective at identifying false claims. This variance was also observed across different models; for example, GPT-3.5 and Llama-2 demonstrated superior accuracy with true claims, while GPT-4 was notably more adept at discerning false claims. These findings highlight the importance of understanding the strengths and weaknesses of each model configuration, which can guide more informed decisions when deploying models in diverse fact-checking scenarios. This approach emphasizes the need for a flexible deployment strategy that can adapt to the nuanced nature of model performance, which depends heavily on the chosen methodological approach.

The analysis over time revealed improvements in accuracy for more recent years, possibly reflecting the models' better handling of contemporary language use and current events. However, performance varied significantly across languages, with languages like Russian, German, and Indonesian showing consistently strong performance, while others such as Azerbaijani and Romanian lagged. Notably, Bengali stood out as an exception, achieving over 80% accuracy across three models - GPT-3.5, GPT-4, and Llama-2 - yet only attaining accuracy close to that of random chance on Mistral 7B. These disparities highlight the importance of diverse and extensive training datasets that are representative of the global linguistic landscape,

underscoring the variability in model effectiveness with different language characteristics and the need for model-specific tuning to address these variations.

The research focused on evaluating how models like GPT-3.5, GPT-4, Llama-2, and Mistral 7B handle the complex task of discerning truth across various languages and factual contexts. The findings reveal a stark reality: despite their advanced capabilities, these models often fell short of the accuracy levels necessary for reliable fact-checking. Notably, GPT-3.5 did not meet the accuracy benchmarks previously documented for simpler datasets, and even the more advanced GPT-4, while achieving higher accuracy, left a significant portion of claims unchecked due to its conservative processing approach. These findings highlight a critical trade-off between accuracy and utility, underscoring the need for diverse training datasets to enhance model reliability across varied linguistic contexts.

The compelling fluency of LLM outputs, while persuasive, frequently does not equate to factual accuracy, illustrating the risks associated with their use in critical decision-making contexts such as fact-checking. As these models become more integrated into daily applications, it is crucial that developers and users approach their deployment with caution, recognizing the potential for misinformation. The potential of LLMs to improve information processing is clear, but their application must be managed carefully to avoid spreading inaccuracies. This research emphasizes the importance of ongoing enhancements to model training and configuration to ensure they serve the public interest, promoting accurate information dissemination in an increasingly automated world.