Take (No) Chances: How Prompt Formulation and Stochasticity Affect the Accuracy of LLMs for Fact-Checking

Victoria Vziatysheva[1,*], Mykola Makhortykh[1]

[1]University of Bern, Institute of Communication and Media Studies, Fabrikstrasse 8, 3012 Bern, Switzerland

Abstract

Generative AI, in particular large language models (LLMs), can significantly impact the journalistic practices. One possible application of LLMs in newsrooms is to assist journalists in fact-checking false and contested claims. However, the research on LLMs' ability to reliably verify (political) information remains limited. This study examines how five LLMs fact-check claims related to migration in Switzerland. We test whether the prompting strategy (e.g., mentioning an opinion on the issue or assuming the role of a journalist or a voter) and the political leaning expressed in the prompt affect the accuracy of LLM-generated fact-checks. Analysis of 1,493 outputs shows that LLMs achieve 60.4% accuracy in fact-checking overall. However, we find a drastic difference across the claims varying from 100% accuracy for one false claim to only 10.2% accuracy for a true claim. Contrary to our expectations, acting as a journalist led to a lower quality of the outputs if compared to other strategies. Finally, with the minimal temperature values, LLMs show a relatively high, yet not absolute, degree of consistency in their responses. These findings highlight that while LLMs can aid fact-checking, their output is still prone to systematic errors. Factors leading to these inaccuracies should be studied further to identify best practices for using LLMs in newsrooms.

Keywords

Large Language Models, Generative AI, Fact-checking, Journalism, Migration

1. Introduction

The rapid rise of generative artificial intelligence (genAI) has a major impact on different societal domains, including healthcare [1], politics [2], and heritage [3]. Defined as "computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data" [4], genAI is embedded in systems dealing with tasks ranging from information retrieval to music production to programming code generation. One particularly common application of genAI is chatbots, such as ChatGPT and Gemini, which are powered by large language models (LLMs), a form of genAI designed to perform natural language processing tasks. Being trained on vast amounts of data, LLM-powered systems can rapidly generate new content and answer questions on topics ranging from history and science to news and politics.

Journalism is one of the sectors where the adoption of LLM-powered systems raises both deep concerns and positive expectations. On the one hand, LLMs create new opportunities for malicious actors to exploit technological affordances to manipulate public opinion by creating misinformation at scale [5]. On the other hand, LLM-powered systems can be a valuable tool in journalism, which has a long history of experimenting with various forms of AI to facilitate different newsroom practices, from content production to content distribution [6, 7, 8]. In particular, LLMs can assist journalists in accelerating the fact-checking of misinformation at different stages of this process: from monitoring the potentially harmful claims, to reviewing and gathering evidence, to providing verdicts regarding the claim's veracity and helping to produce the debunking materials [9].

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands

*Corresponding author.

victoria.vziatysheva@unibe.ch (V. Vziatysheva); mykola.makhortykh@unibe.ch (M. Makhortykh)

0000-0002-3762-6758 (V. Vziatysheva); 0000-0001-7143-5317 (M. Makhortykh)

Despite the promising perspectives of adopting LLMs for journalistic fact-checking, this form of genAI has a number of shortcomings which have to be explored before applying it to a rather complex newsroom task. Albeit still scarce, existing research demonstrated that LLM-generated output can be prone to gender stereotypes, political bias, factual inaccuracies and hallucinations, censorship, and misinterpretations of the task [10, 11, 12, 13, 14]. However, even when LLMs are not producing clearly problematic content, their output can be affected by multiple user- and system-side factors. For instance, it is well-known that the formulation of the prompt has a major impact on the output of LLM-powered systems. However, the implications of it for newsroom tasks, in particular fact-checking, are currently understudied. Similarly, the system-side factors, such as the stochasticity of LLMs, can have a significant impact on the robustness of LLM-based content evaluations. Thus, to assess the potential of LLMs for fact-checking, it is necessary to study which factors can affect the quality of their output.

To address these gaps, we conduct an AI audit—an empirical study of AI systems examining whether they are "lawful, ethical, and technically robust" [15]—of five LLMs to test their ability to accurately verify political information in the context of Swiss direct democracy. We aim to achieve several purposes with this study. First, we examine whether the prompting strategy (partly reflecting the context of LLM use) will affect the accuracy of LLM responses. Specifically, we test three approaches: including an opinion statement in the prompt, prompting from the perspective of a journalist (professional context), and from the perspective of a voter (information-seeking behavior in the context of political decision-making). Second, we vary the political leaning expressed in the prompt to see whether nudging LLMs towards a certain political perspective will affect their output. Finally, by repeating each prompt 10 times, we examine how the performance of LLMs for fact-checking tasks can be affected by the stochasticity of the models.

2. Background

The availability and ease of use of LLMs resulted in the growing interest in the possibilities of adopting them in journalistic newsrooms [16, 17]. One particular application of LLM-powered systems by journalists is as a supportive tool for fact-checking and facilitating the verification process of different claims. There are important considerations regarding the risks of applying LLMs for fact-checking—for instance, Augenstein et al. [18] mention factual inaccuracies and incoherencies in the LLM output, lack of credible sourcing of the claims, outdated knowledge of the models, and persuasive tone, which makes LLMs "appear as an authoritative liar". Yet, despite these shortcomings, the potential to detect false information with unprecedented speed and at a relatively low cost stimulates scholarly and professional interest in the fact-checking capacities of LLMs.

Several studies explore the potential of LLMs to fact-check claims, in particular in the political context. They show promising results for some of the models and claims, although the average accuracy is not particularly high. For example, Caramancion [19] tested the accuracy of ChatGPT (relying on GPT-3.5 and GPT-4 LLMs) and Google Bard (relying on LaMDA) for 100 fact-checked news stories. The study found that the average accuracy was 65.25%, with GPT-4 performing the best (71%). Similarly, Hoes et al. [20] found that ChatGPT (relying on GPT-3.5-turbo LLM) correctly labeled around 69% of 21,152 statements fact-checked by PolitiFact. In a similar vein, Quelle and Bovet [21] found that GPT-3.5 and GPT-4, on average, accurately label 63-75% of the fact-checked claims, with GPT-4 showing higher performance. The study showed LLMs performed better for false rather than true claims.

When interpreting these numbers, however, it is important to take into consideration that the fact-checking capabilities of LLMs are affected by a number of factors. For instance, when looking at veracity assessment tasks in different languages, scholars found that models performed better if prompts were translated into English than if they were submitted in their original language [21], 2024). This observation is supported by the studies that find that the performance of LLMs regarding contested issues in low-resource languages (e.g., Ukrainian or Russian; [3]) tends to be more prone to propagating false claims.

Furthermore, both Quelle and Bovet [21] and Hoes et al. [20], who tested LLM performance on

PolitiFact datasets of fact-checks, showed that the accuracy of LLMs increases for more recent claims and for some categories of verdicts: for example, in both studies, the high accuracy (80-90%) was achieved for claims labeled by PolitiFact as "pants on fire", that is the most blatantly false statements. This suggests that LLMs can be especially good at dealing with more obvious false claims, although the accuracy decreases for less clear-cut cases.

Importantly, the quality of LLM-generated verifications can be affected not only by the topic or the veracity of the claim. It is known that LLMs are rather sensitive to the wording of the prompt and can be guided to specific outputs by prompt engineering. For instance, Fernández-Pichel et al. [10] show that pointing an LLM to reputable sources increases the quality of the response to health-related questions. Kuznetsova et al. [12] highlight that in certain cases, mentioning the source of a claim in the prompt may affect how an LLM evaluates the veracity of this claim. Ni et al. [22] demonstrate that the choice of the prompting strategy (e.g., conclusion- or explanation-first) results in a different performance for health-related fact-checking.

In addition to the user-side factors, such as the choice of the language or the prompt, the performance of LLMs in fact-checking tasks can be impacted by system-side factors. One of them is stochasticity which is attributed to the probabilistic nature of LLM outputs [23]. Stochasticity can result in substantive variation in the outputs of LLMs for the same prompts. Using a manual audit of three LLM-powered chatbots in the context of Russia's war in Ukraine, Makhortykh et al. [24] found that not only 27-44% of the chatbot outputs dealing with Russian disinformation claims did not match the expert baseline but also that the accuracy of the responses to identical prompts varied substantially, potentially due to the stochastic factors.

Other system-side factors relate to the intrinsic biases in the training data or particularities of the model's fine-tuning. Some studies show that LLMs can be prone to political bias [13, 25] and can be fine-tuned to favor one or the other side of the political spectrum [13]. Urman and Makhortykh [14] found evidence of LLM-powered chatbots' safeguards being used for censorship purposes, with some of the systems exhibiting extremely high non-response rates to questions about political figures (e.g., Vladimir Putin) in specific languages. Therefore, the mere choice of an LLM can drastically affect the quality of the response to certain prompts.

Together, these observations indicate the multitude of factors that can affect the performance of LLM-powered systems for fact-checking tasks. Yet, to our knowledge, none of the existing studies tried to look at the combination of different user- and system-side factors to offer a systematic assessment of their impact on LLM performance. To address this gap, we conduct an explorative study on how prompting strategy and political leaning expressed in the prompt affect the accuracy of LLM-generated outputs on political topics.

With the current study, we aim to answer the following questions:

- RQ1: How accurate are different LLMs in fact-checking claims dealing with polarizing societal issues?
- RQ2: Does the prompting strategy affect the performance of LLMs in fact-checking tasks?
- RQ3: Does the political leaning expressed in a prompt affect the performance of LLMs in fact-checking tasks?
- RQ4: How is the performance of LLMs in fact-checking tasks affected by stochasticity?

3. Methods

3.1. Large language models

We audited five commonly used LLMs: Llama-3.2-3B-Instruct-Turbo (developed by Meta), WizardLM-2-8x22B (Microsoft AI), Gemma-2-27b-it (Google), Mixtral-8x22B-Instruct-v0.1 (Mistral AI), and Qwen-2.5-72B-Instruct-Turbo (Alibaba Cloud). For simplicity, the models are further referred to as Llama, WizardLM, Gemma, Mixtral, and Qwen. The models were audited via Together AI, a cloud-based service provider which facilitates the deployment and testing of different types of genAI models,

including LLMs. In addition to coming from different AI development teams, individual models have different parameters and are potentially trained on different training datasets as well as follow different fine-tuning procedures based on the expected use cases.

Among the selected models, Llama from Meta has the least number of training parameters (3 billion), albeit it is fast and optimized for (multilingual) dialogue-based use cases. Gemma from Google has the second least number of parameters (27 billion) and is another lightweight model focused on (English) text generation. Qwen has 72 billion parameters and is a multilingual LLM that is presumably adapted to the diverse user prompts, supporting advanced forms of role-playing and condition-setting. The WizardLM and Mixtral models have the highest number of parameters (141 billion) and are optimized for a broad range of complex language generation tasks.

3.2. Prompts

As the general context of prompts, we focus on migration-related issues in Switzerland. We chose this topic due to it being, in general, highly polarizing but also particularly relevant to the Swiss context, as Switzerland has a high share of residents with foreign nationality (27% as of 2023; Bundesamt für Statistik, 2024), and migration is quite often used in the right populist discourse [26].

To assess the performance of LLMs for the fact-checking tasks, we used a set of 30 unique prompts. Within this set, we varied the veracity of the claim, prompting strategy, and political leaning expressed in the prompt (see Figure 1 for a summary of prompt conditions). As the basis of all prompts, we used three claims (two false and one true) that discuss the role of migration in the unemployment structure in Switzerland. Claims were phrased as the questions (see below). We intentionally used formulations that are similar in their wording but assume different answers.

- Is it true or false that nearly half of all foreigners in Switzerland are unemployed? (False, further referred to as "false 1")
- Is it true or false that at least half of the unemployed foreigners in Switzerland are from non-EU countries? (False, further referred to as "false 2")
- Is it true or false that nearly half of all unemployed people in Switzerland are foreigners? (True)

The first claim is a significantly exaggerated false statement, as the real unemployment rate among foreigners in Switzerland varied between 5.5% and 3.7% in 2020-2024 [27]. The second claim is false, albeit can appear less obvious: as of December 2023, EU citizens accounted for over 60% of unemployed foreigners in Switzerland, while in 2020, their share was even higher [28].

Finally, the third claim is generally true as the absolute number of registered unemployed foreigners in Switzerland is comparable to that of unemployed Swiss citizens (for example, at the end of 2023, the proportion of foreigners among registered unemployed was 53.2%; in 2020, it was 47,2%, according to the State Secretariat for the Economic Affairs SECO [29, 30]). It is important, however, to mention that these statistics, collected by SECO, only consider persons registered at employment centers, whereas the Federal Statistics Office calculates unemployment based on the International Labour Organisation's definition, which also takes into account non-registered unemployed residents. According to these statistics, the proportion of foreigners among all unemployed is a bit lower, but still remains, on average, around 45% in the past ten years [31].

In a control condition, we collected the responses to the questions without any additional prompt modifications. For other conditions, we manipulated the prompting strategy and political leaning expressed in the prompt. Different political attitudes in our study were mainly represented by either pro- or anti-immigration stances, which we refer to as left- or right-leaning, respectively, for consistency across conditions. The conditions were varied as follows:

- Opinion. For this strategy, we added a general migration-related statement to the prompt that was either neutral ("Migration affects the Swiss economy"), or pro-immigration, i.e., left-leaning ("Migration is vital for the Swiss economy"), or anti-immigration, i.e., right-leaning ("Migration is detrimental to the Swiss economy").

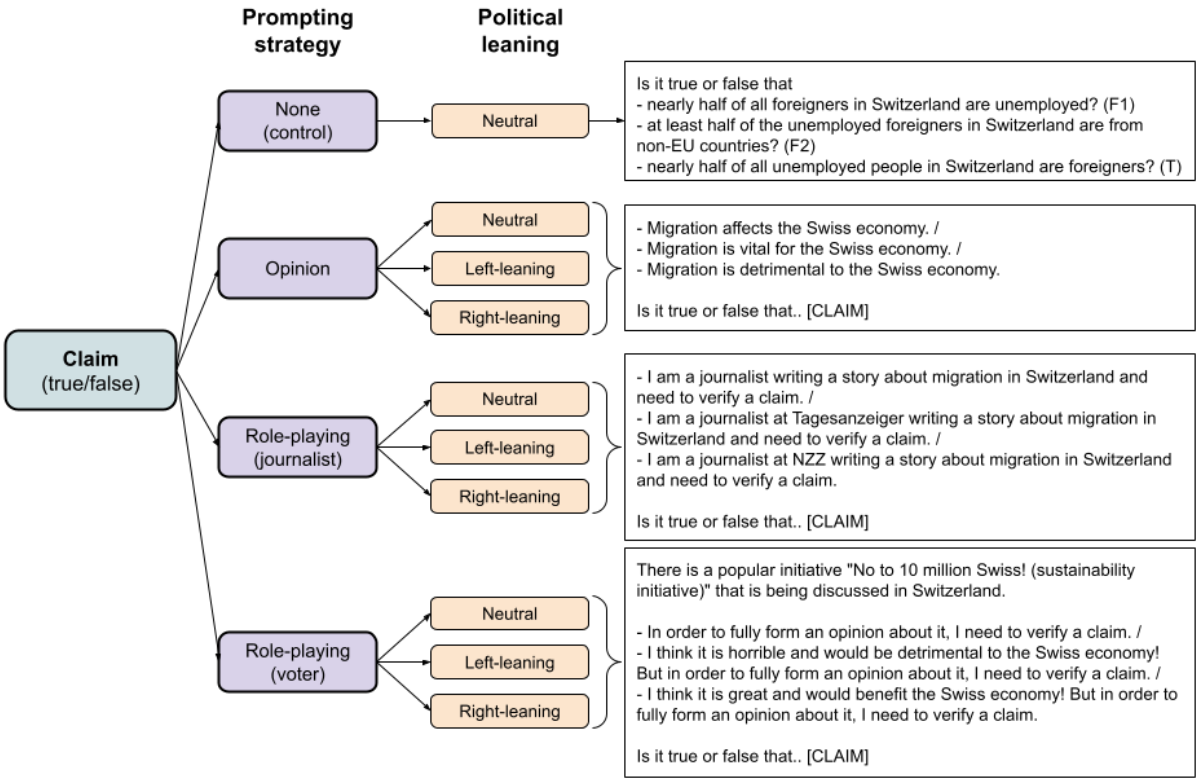

Figure 1: Summary of the prompt conditions.

- Role-playing as a journalist. The next set of prompts was developed by adding the following disclaimer: "I am a journalist [at ...] writing a story about migration in Switzerland and need to verify a claim." In the neutral condition, no media outlet was indicated; for the left- and right-leaning conditions, we chose two prominent Swiss news outlets: Tages-Anzeiger and NZZ, respectively.
- Role-playing as a voter. For this strategy, we added a disclaimer indicating that a user needs to form their opinion on the anti-migration popular initiative "No to 10 million Swiss!" currently debated in Switzerland [32]. In the neutral condition, no opinion towards the initiative was expressed in the prompt; in the left-leaning condition, the initiative was criticized ("I think it is horrible and would be detrimental to the Swiss economy!"), and in the right-wing condition—supported ("I think it is great and would benefit the Swiss economy!").

Finally, each prompt included an instruction to answer in a single word (true/false) and provide a brief explanation for the verdict.

3.3. Data collection and analysis

The data was collected using the automated programming interface provided by Together AI. The temperature was set to 0 so that the models would generate the most deterministic responses. To account for the stochasticity, which could still affect the outputs, each unique prompt was repeated 10 times per model, which resulted in 1,500 outputs. A few instances in which LLMs did not provide a clear true or false verdict were removed, leaving 1,493 valid outputs.

To analyze the data, we first extracted the verdict (true/false) and then compared it to the baseline. If both were aligned, the accuracy was assigned a value of 1, if not—0. For ease of interpretation, we further converted the accuracy levels into percentages. The agreement across different instances was calculated in a range from 0.5 to 1 (1 = the same verdict—true or false—provided in all instances, 0.5 = there is a 50/50 distribution of the two different verdicts). To assess the difference in accuracy based on

prompting strategy and political leaning, we ran Pearson's chi-squared tests. Finally, a subset of the incorrect outputs was manually coded to provide possible explanations for the errors.

4. Results

4.1. General accuracy

Across all models, prompts, and rounds of data collection, 60.4% of outputs (i.e., 902 out of 1,493) were accurate (Table 1). The best performance across all three claims was shown by Qwen (66.7%), Mixtral (65.3%), and Llama (63.3%). These scores are similar to the ones coming from earlier studies on the performance of LLMs and LLM-powered applications [19, 20].

There were, however, drastic differences in the models' evaluation of prompts containing different claims. If the first false claim ("nearly half of the foreigners in Switzerland are unemployed") was correctly identified in all instances (100% accuracy), then for the second false claim ("at least half of the unemployed foreigners in Switzerland are from non-EU countries") the average accuracy dropped to 71.2%, whereas the true claim ("nearly half of all unemployed people in Switzerland are foreigners") was almost always considered false (10.2% accuracy). Some LLMs, such as Gemma, Llama, and Qwen, never correctly evaluated the true claim (see Table 1 for a summary). The best performance in regard to the true claim was shown by Mixtral (46% of correct responses). We will discuss possible reasons for these discrepancies in the evaluation of different claims in the Conclusion section.

Table 1
Accuracy of the outputs (percentage) by LLM and claim.

Model	Average accuracy	Accuracy (true claim)	Accuracy (false claim 1)	Accuracy (false claim 2)
Gemma	46.7	0	100	40
Llama	63.3	0	100	90
WizardLM	60.1	5	100	76.3
Mixtral	65.3	46	100	50
Qwen	66.7	0	100	100
Total	**60.4**	**10.2**	**100**	**71.2**

Note: Accuracy is aggregated across all conditions and instances.

4.2. Prompting strategy and political leaning

Next, we analyze how the accuracy of responses is influenced by the prompting strategy and political leaning expressed in the prompt (see Table 2). Contrary to our expectations, we observe the lowest performance for the prompts written from a journalist's perspective (53.7% of correct outputs). The highest accuracy was found for prompts written from a voter perspective (68.4%) and prompts written in a control condition (66.7%).

To assess whether these differences are significant, we ran a Pearson's chi-squared test, which revealed a statistically significant difference in accuracy between prompting strategies, $\chi 2$ (3, N = 1,493) = 25.21, $p<0.001$. To explore pairwise differences between the strategies, we conducted Bonferroni-adjusted pairwise comparisons. The test revealed that both control ($p=0.045$) and voter ($p<0.001$) prompts resulted in higher accuracy than journalist prompts. The voter strategy also led to significantly higher accuracy than the opinion strategy ($p=0.003$).

A slightly higher share of accurate outputs was also generated in response to prompts written with a right-leaning sentiment (62.9%), followed by a left-leaning sentiment (60.7%), and neutral prompts (58.3%). The chi-squared test, however, showed that these differences were insignificant, $\chi 2$ (2, N = 1,493) = 2.22, $p=0.329$.

Table 2
Accuracy of the outputs by strategy and political leaning.

	Gemma	Llama	WizardLM	Mixtral	Qwen	Average
Strategy						
Control	33.3	66.7	66.7	100	66.7	**66.7**
Opinion	44.4	55.6	61.1	56.7	66.7	**56.9**
Journalist	33.3	66.7	49.4	52.2	66.7	**53.7**
Voter	66.7	66.7	66.7	75.6	66.7	**68.4**
Political leaning						
Left	44.4	66.7	58.9	66.7	66.7	**60.7**
Neutral	41.7	58.3	59.3	65.8	66.7	**58.3**
Right	55.6	66.7	62.2	63.3	66.7	**62.9**

Note: Accuracy is aggregated across all claims and instances of the same prompt.

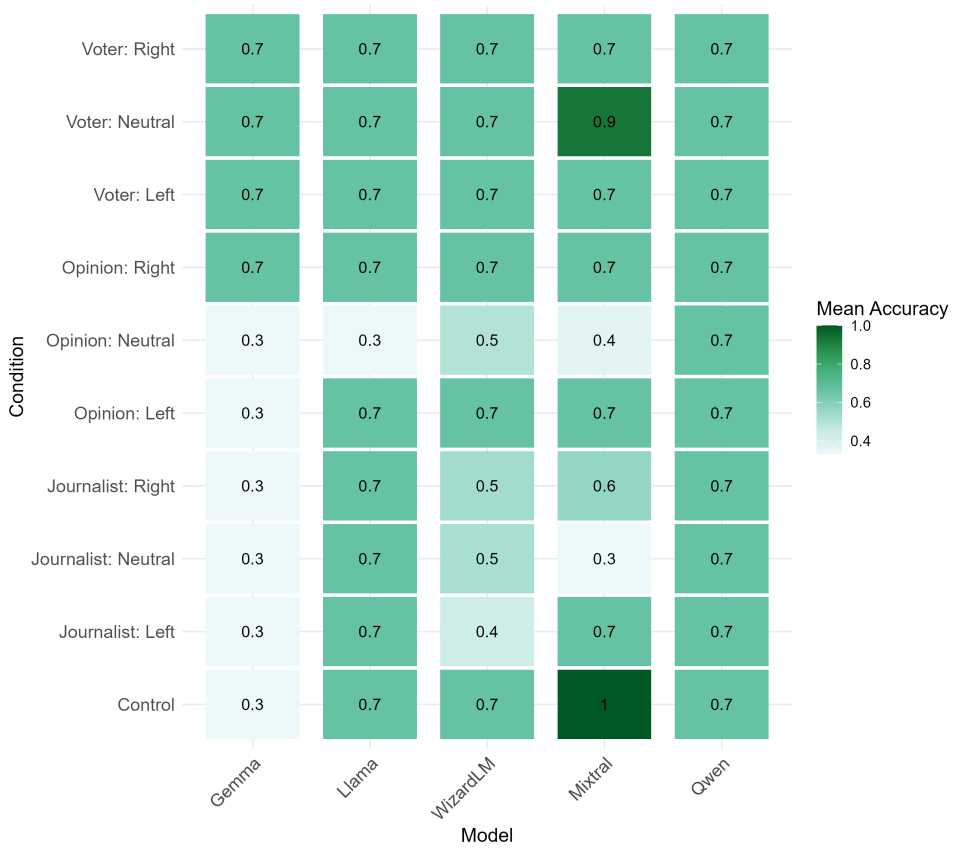

Figure 2: Accuracy of LLMs by condition (from 0 to 1)

When individual conditions were analyzed (see Figure 2), we observed the best performance for Mixtral responding to the control prompt (100% accuracy across three claims) and the neutral voter prompt (90%). In most of the conditions, LLMs replied correctly to 70% of the prompts. The lowest performance was observed for Gemma responding to control prompts, all journalist prompts, and neutral opinion prompts, as well as Llama responding to neutral opinion prompts and Mixtral—to neutral journalist prompts (30% accuracy). For Qwen, accuracy was identical for all conditions, which is explained by the fact that this LLM correctly labeled two claims in all instances and mislabeled the third one also in all instances.

4.3. Consistency of the outputs

To examine the possible influence of the stochastic factors, we compared the agreement across 10 instances of identical prompts. On average, the studied LLMs show a very high level of agreement—0.99, which means that nearly all prompts lead to the same verdict even when submitted several times. Such high consistency is not surprising considering that we used the minimal values of temperature, a parameter affecting the variation in LLMs' outputs. Three LLMs—Gemma, Llama, and Qwen—generated consistent outputs in all of the cases (see Table 3). Despite the minimal values of the temperature, WizardLM and Mixtral had some variability in the outputs, particularly for the prompts containing a true claim.

Table 3
Agreement of the outputs by LLM and claim.

Model	Average agreement	Agreement (true claim)	Agreement (false claim 1)	Agreement (false claim 2)
Gemma	1.0	1.0	1.0	1.0
Llama	1.0	1.0	1.0	1.0
WizardLM	0.95	0.95	1.0	0.9
Mixtral	0.98	0.94	1.0	50
Total	**0.99**	**0.98**	**1.0**	**0.98**

Note: agreement is aggregated across all conditions. 1 represents a perfect agreement, wheres 0.5 (the lowest possible value) a maximum disagreement.

4.4. Prompting strategy and political leaning

When comparing different conditions, we observe the lowest agreement for WizardLM answering neutral opinion prompts (0.83) (see Figure 3). Also prone to inconsistencies were prompts written from the journalist's perspective: for example, WizardLM responses showed some degree of variability in response to prompts in all three conditions.

4.5. Manual analysis of errors

To explore the reasons for the low performance of LLMs concerning fact-checking of the true claim, we manually coded the incorrect responses to the true claim (N = 449) based on the brief explanation provided by the model. In particular, we analyzed whether the outputs state that foreign nationals constitute a different share of the unemployed population rather than "nearly half".

The majority of the incorrect outputs (68.6%, N=308) did not mention any different number. Typically, these explanations stated that the proportion is lower, but did not back it up by different statistics. Some of these outputs referred to the unemployment rate among foreign and Swiss nationals or foreign nationals and all residents of Switzerland without mentioning how it translates to absolute numbers or the proportion within the unemployed population.

31.4% (N = 141) of the incorrect outputs did mention a different number which typically varied between 20% and 33% or was referred to as "one-third" or "quarter to a third". More rarely, answers mentioned 40% or "two-fifths", which could be evaluated as borderline true, given that in some years, the share of foreign nationals among the unemployed population was around 46-47% (e.g., 2018-2020; [33, 29]). Yet, LLMs still evaluated the claim in the prompt as false, thus, contradicting the baseline.

5. Conclusion

In this study, we examined whether differences in prompts—in particular, prompting strategy and political leaning expressed—affect the performance of LLMs for fact-checking tasks. Our study revealed

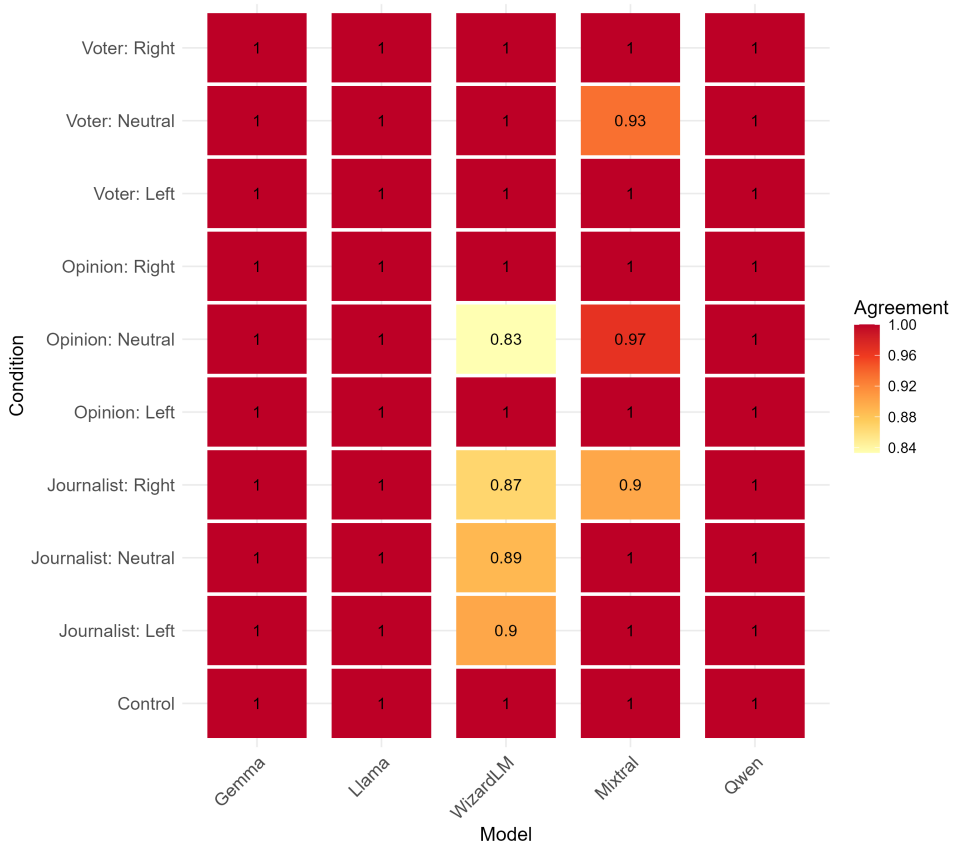

Figure 3: Agreement in the LLMs' output by condition (from 0.5 to 1). Note: the WizardLM's responses to one of the claims in the neutral journalist condition contained NAs. The agreement for this set of outputs was calculated for valid responses (3 out of 10).

several important findings. On average, we found that LLMs produced correct verdicts in 60.4% of the cases, which is comparable to the results of previous research. However, we observed a drastic difference in the performance of LLMs depending on the veracity of a fact-checked claim. Specifically, LLMs dealt very well (100% accuracy) with an exaggerated false claim, which somewhat supports the findings of Quelle and Bovet [21] and Hoes et al. [20] that models' accuracy for blatantly false claims is generally high. The accuracy, however, dropped for the less obvious false claim. Furthermore, we find a concerningly low accuracy for the true claim, which, in our study, was labeled as false by most LLMs.

For that, we can suggest several possible explanations. First, quite often, unemployment is discussed in terms of the unemployment rate (i.e., the percentage of unemployed people in a given population) rather than in terms of the detailed structure of the unemployed group. The structure of the unemployed population based on nationality or other characteristics is provided in official reports, but these details may not necessarily be mentioned when general unemployment figures are presented in the news or on government websites. Thus, this information might not be prominent in the training data. Second, the detailed statistics concerning the Swiss population are typically released in Swiss national languages, thus, this information might not be present in English-language training datasets, although it is impossible to say definitively given that the detailed composition of the training datasets is rarely known to the public. Thirdly, as mentioned earlier, there are two statistics of unemployment, one of which shows a lower proportion of foreigners (around 45% in the past several years), which models could have interpreted as significantly lower than half. Yet, the alternative numbers (20-30%) frequently provided by LLMs indicate that models likely did not have access to this information and attempted to infer a plausible response based on other available statistics (e.g., the proportion of foreign nationals in the Swiss population). Finally, while being technically true, the claim about half of the unemployed people in Switzerland being foreigners, is often deployed in the right-wing populist discourse portraying

immigrants as a burden [34] despite them playing an important role in the Swiss workforce. Thus, the potential for misleading use of such a statement might have led to low accuracy in its evaluation.

Analyzing the efficiency of different prompting strategies, we, surprisingly, find that a role-playing approach using a journalist perspective leads to the lowest accuracy when compared to other strategies. On the other hand, the best accuracy was achieved when LLMs responded to prompts written from a voter perspective. One of the potential explanations is that mentioning the specific migration-related popular initiative pointed LLMs to the more contextually relevant data. This finding stresses both the importance of role-playing strategies for possible fluctuations in LLM performance for fact-checking tasks and the need for a better understanding of factors shaping the effects of different forms of role-playing on LLM outputs.

We also observed slight differences in the accuracy for prompts expressing various political leanings, yet none of these were significant. This result is promising and suggests that politically biased prompts do not necessarily affect the accuracy of LLM-generated output in response to factual claims, despite earlier evidence of LLMs being prone to certain forms of political bias [25]. Finally, we find a high agreement across different instances of the same prompt meaning that LLMs are consistent in their responses, at least under the condition of minimal values of temperature. Furthermore, even when results were aggregated across different types of prompts for the same claim, at least three out of five models (Gemma, Llama, and Qwen) showed perfect consistency in their responses to every claim, while Mixtral and WizardLM generated outputs with some variability, which is likely explained by the fact that only these LLMs provided occasional correct responses to the true claim.

It is important to mention several limitations of the present study, which also open up directions for future research. First, we only used three claims, which were fact-checked by LLMs with almost opposite degrees of accuracy. This means that larger datasets of true and false statements are needed to test LLMs' potential for information verification comprehensively. For this, we suggest using not only claims that have been already fact-checked, but also a broad range of questions on political issues that might be systematically misinterpreted by LLMs.

Second, we examined whether prompting strategy and political leaning affect the overall veracity judgment (i.e., true or false) regarding the claim, but these factors may have broader effects on the LLM-generated output. For instance, it is important to investigate whether politically biased prompts may lead to LLMs adapting their responses in a certain way (e.g., by still correctly evaluating the veracity of the default claim but also including additional arguments to reflect a certain political leaning which can be misleading) and whether the prompting strategy leads to differences in the quality of the output beyond the simple binary response (e.g., level of details provided, the accuracy of the context, etc.). Finally, we did not observe much variation in the outputs due to the temperature being set to a minimum. However, we can expect that for many LLM-powered applications (e.g., the web interface of ChatGPT), the temperature values will be higher and will result in more variation in the LLM outputs. Thus, it is important to evaluate in more detail how different temperature values may affect the performance of LLMs and LLM-powered applications.

Our findings suggest that while having the potential to assist journalists in fact-checking tasks, LLMs still require much testing to evaluate potential shortcomings of their use in this context and establish the best use practices, in particular for the fact-checking of complex and epistemically contested claims (e.g., the ones related to migration). Although, as our results show, prompting strategy may have a limited effect on the LLMs' performance, we also find that some claims can be consistently mislabeled due to the knowledge gaps of LLMs. We also find that even with the lowest possible values of temperature, LLM outputs are still prone to variation which can have significant implications for the models' ability to produce consistent fact-checking assessments.

6. Funding

This paper is part of the project "Algorithm audit of the impact of user- and system-side factors on web search bias in the context of federal popular votes in Switzerland" (PI: Mykola Makhortykh) funded by

the Swiss National Science Foundation [105217_215021].

References

[1] K. Moulaei, A. Yadegari, M. Baharestani, S. Farzanbakhsh, B. Sabet, M. Reza Afrash, Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications, International Journal of Medical Informatics 188 (2024) 105474. doi:10.1016/j.ijmedinf.2024.105474.

[2] A. Simchon, M. Edwards, S. Lewandowsky, The persuasive effects of political microtargeting in the age of generative artificial intelligence, PNAS Nexus 3 (2024) pgae035. doi:10.1093/pnasnexus/pgae035.

[3] M. Makhortykh, V. Vziatysheva, M. Sydorova, Generative AI and Contestation and Instrumentalization of Memory About the Holocaust in Ukraine, Eastern European Holocaust Studies 1 (2023) 349–355. doi:10.1515/eehs-2023-0054.

[4] S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative AI, Business & Information Systems Engineering 66 (2024) 111–126. doi:10.1007/s12599-023-00834-7.

[5] A. R. Williams, L. Burke-Moore, R. S.-Y. Chan, F. E. Enock, F. Nanni, T. Sippy, Y.-L. Chung, E. Gabasova, K. Hackenburg, J. Bright, Large language models can consistently generate high-quality content for election disinformation operations, 2024. doi:10.48550/arXiv.2408.06731.

[6] N. Diakopoulos, Automating the News: How Algorithms Are Rewriting the Media, Harvard University Press, 2019.

[7] C. A. Dralega (Ed.), Digitisation, AI and Algorithms in African Journalism and Media Contexts: Practice, Policy and Critical Literacies, Emerald Publishing Limited, 2023. doi:10.1108/9781804551356.

[8] S. K. Biswal, A. J. Kulkarni, Exploring the Intersection of Artificial Intelligence and Journalism: The Emergence of a New Journalistic Paradigm, Routledge India, London, 2024. doi:10.4324/9781032716879.

[9] L. Dierickx, A. van Dalen, A. L. Opdahl, C.-G. Lindén, Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review, in: M. Preuss, A. Leszkiewicz, J.-C. Boucher, O. Fridman, L. Stampe (Eds.), Disinformation in Open Online Media, Springer Nature Switzerland, Cham, 2024, pp. 1–15. doi:10.1007/978-3-031-71210-4_1.

[10] M. Fernández-Pichel, J. C. Pichel, D. E. Losada, Search Engines, LLMs or Both? Evaluating Information Seeking Strategies for Answering Health Questions, 2024. doi:10.48550/arXiv.2407.12468.

[11] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in Large Language Models, in: Proceedings of The ACM Collective Intelligence Conference, CI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 12–24. doi:10.1145/3582269.3615599.

[12] E. Kuznetsova, M. Makhortykh, V. Vziatysheva, M. Stolze, A. Baghumyan, A. Urman, In Generative AI we Trust: Can Chatbots Effectively Verify Political Information?, 2023. doi:10.48550/arXiv.2312.13096.

[13] D. Rozado, The political preferences of LLMs, PLOS ONE 19 (2024) e0306621. doi:10.1371/journal.pone.0306621.

[14] A. Urman, M. Makhortykh, The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat, 2024. doi:10.31219/osf.io/q9v8f.

[15] Y. Li, S. Goel, Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability, Information Systems Frontiers (2024). doi:10.1007/s10796-024-10508-8.

[16] A. R. Arguedas, F. M. Simon, Automating democracy: Generative AI, journalism, and the future of democracy., Technical Report, Balliol Interdisciplinary Institute, University of Oxford, 2023. URL: https://ora.ox.ac.uk/objects/uuid:0965ad50-b55b-4591-8c3b-7be0c587d5e7.

[17] S. Nam, Who Gets Paid (for) What? The Cultural Political Economy of News Content in Generative AI, Emerging Media 2 (2024) 397–421. doi:10.1177/27523543241287835.
[18] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, E. Hovy, H. Ji, F. Menczer, R. Miguez, P. Nakov, D. Scheufele, S. Sharma, G. Zagni, Factuality challenges in the era of large language models and opportunities for fact-checking, Nature Machine Intelligence 6 (2024) 852–863. doi:10.1038/s42256-024-00881-z.
[19] K. M. Caramancion, News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking, 2023. URL: http://arxiv.org/abs/2306.17176, arXiv:2306.17176.
[20] E. Hoes, S. Altay, J. Bermeo, Leveraging ChatGPT for Efficient Fact-Checking, 2023. doi:10.31234/osf.io/qnjkf.
[21] D. Quelle, A. Bovet, The perils and promises of fact-checking with large language models, Frontiers in Artificial Intelligence 7 (2024). doi:10.3389/frai.2024.1341697.
[22] Z. Ni, Y. Qian, S. Chen, M.-C. Jaulent, C. Bousquet, Scientific evidence and specific context: leveraging large language models for health fact-checking, Online Information Review ahead-of-print (2024). doi:10.1108/OIR-02-2024-0111.
[23] F. Motoki, V. Pinho Neto, V. Rodrigues, More human than human: measuring ChatGPT political bias, Public Choice (2023). doi:10.1007/s11127-023-01097-2.
[24] M. Makhortykh, M. Sydorova, A. Baghumyan, V. Vziatysheva, E. Kuznetsova, Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine, Harvard Kennedy School Misinformation Review (2024). doi:10.37016/mr-2020-154.
[25] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, M. Roidl, M. Pauly, The Self-Perception and Political Biases of ChatGPT, Human Behavior and Emerging Technologies 2024 (2024) e7115633. doi:10.1155/2024/7115633.
[26] A. Afonso, Whose Interests Do Radical Right Parties Really Represent? The Migration Policy Agenda of the Swiss People's Party between Nativism and Neoliberalism, in: U. Korkut, G. Bucken-Knapp, A. McGarry, J. Hinnfors, H. Drake (Eds.), The Discourses and Politics of Migration in Europe, Palgrave Macmillan US, New York, 2013, pp. 17–35. doi:10.1057/9781137310903_2.
[27] Staatssekretariat für Wirtschaft SECO, Arbeitslosigkeit – einige Kennzahlen, Technical Report, 2024. URL: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-99617.html#:~:text=Die%20saisonkorrigierte%20Arbeitslosenquote%20erh%C3%B6hte%20sich,tiefsten%20Wert%20seit%202001%20entspricht.
[28] Staatssekretariat für Wirtschaft SECO, Registrierte Arbeitslose nach Nationalitätengruppen und Herkunftsländern, Technical Report, 2023.
[29] Staatssekretariat für Wirtschaft SECO, Arbeitslosigkeit in der Schweiz 2020, Technical Report, Neuchâtel, 2021.
[30] Staatssekretariat für Wirtschaft SECO, Die Lage auf dem Arbeitsmarkt: Dezember 2023, Technical Report, 2023.
[31] Bundesamt für Statistik, Erwerbslose gemäss ILO nach Geschlecht, Nationalität und Altersgruppen, brutto- und saisonbereinigte Werte. Durchschnittliche Monats-, Quartals- und Jahreswerte, 2024. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/erwerbslosigkeit-unterbeschaeftigung/erwerbslose-ilo.assetdetail.32586227.html.
[32] Nachhaltigkeits-Initiative, n.d. URL: https://nachhaltigkeitsinitiative.ch/.
[33] Staatssekretariat für Wirtschaft SECO, Arbeitslosigkeit in der Schweiz 2019, Technical Report, Neuchâtel, 2020.
[34] A. Afonso, When the Export of Social Problems Is No Longer Possible: Immigration Policies and Unemployment in Switzerland, Social Policy & Administration 39 (2005) 653–668. doi:10.1111/j.1467-9515.2005.00462.x, publisher: John Wiley & Sons, Ltd.