frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 07 February 2024
DOI 10.3389/frai.2024.1341697

OPEN ACCESS

EDITED BY
Rafael Berlanga,
University of Jaume I, Spain

REVIEWED BY
Ismael Sanz,
University of Jaume I, Spain
Kevin Matthe Caramancion,
University of Wisconsin-Stout, United States

*CORRESPONDENCE
Dorian Quelle
dorian.quelle@uzh.ch

RECEIVED 20 November 2023
ACCEPTED 22 January 2024
PUBLISHED 07 February 2024

CITATION
Quelle D and Bovet A (2024) The perils and promises of fact-checking with large language models. *Front. Artif. Intell.* 7:1341697. doi: 10.3389/frai.2024.1341697

The perils and promises of fact-checking with large language models

Dorian Quelle [1,2]* and Alexandre Bovet [1,2]

[1]Department of Mathematical Modeling and Machine Learning, University of Zurich, Zurich, Switzerland, [2]Digital Society Initiative, University of Zurich, Zurich, Switzerland

Automated fact-checking, using machine learning to verify claims, has grown vital as misinformation spreads beyond human fact-checking capacity. Large language models (LLMs) like GPT-4 are increasingly trusted to write academic papers, lawsuits, and news articles and to verify information, emphasizing their role in discerning truth from falsehood and the importance of being able to verify their outputs. Understanding the capacities and limitations of LLMs in fact-checking tasks is therefore essential for ensuring the health of our information ecosystem. Here, we evaluate the use of LLM agents in fact-checking by having them phrase queries, retrieve contextual data, and make decisions. Importantly, in our framework, agents explain their reasoning and cite the relevant sources from the retrieved context. Our results show the enhanced prowess of LLMs when equipped with contextual information. GPT-4 outperforms GPT-3, but accuracy varies based on query language and claim veracity. While LLMs show promise in fact-checking, caution is essential due to inconsistent accuracy. Our investigation calls for further research, fostering a deeper comprehension of when agents succeed and when they fail.

KEYWORDS

fact-checking, misinformation, large language models, human computer interaction, natural language processing, low-resource languages

1 Introduction

Fact-checking has become a vital tool to reduce the spread of misinformation online, shown to potentially reduce an individual's belief in false news and rumors (Morris et al., 2020; Porter and Wood, 2021) and to improve political knowledge (Nyhan and Reifler, 2015). While verifying or refuting a claim is a core task of any journalist, a variety of dedicated fact-checking organizations have formed to correct misconceptions, rumors, and fake news online. A pivotal moment in the rise of fact-checking happened in 2009 when the prestigious Pulitzer Prize in the national reporting category was awarded to Politifact. Politifact's innovation was to propose the now standard model of an ordinal rating, which added a layer of structure and clarity to the fact check and inspired dozens of projects around the world (Mantzarlis, 2018). The second wave of fact-checking organizations and innovation in the fact-checking industry was catalyzed by the proliferation of viral hoaxes and fake news during the 2016 US presidential election (Bovet and Makse, 2019; Grinberg et al., 2019) and Brexit referendum (Mantzarlis, 2018). Increased polarization (Flamino et al., 2023), political populism, and awareness of the potentially detrimental effects of misinformation have ushered in the "rise of fact-checking" (Graves and Cherubini, 2016).

Although fact-checking organizations play a crucial role in the fight against misinformation, notably during the COVID-19 pandemic (Siwakoti et al., 2021), the process of fact-checking a claim is an extremely time-consuming task. A professional

fact-checker might take several hours or days on any given claim (Hassan et al., 2015; Adair et al., 2017). Due to an ever-increasing amount of information online and the speed at which it spreads, relying solely on manual fact-checking is insufficient and makes automated solutions and tools that increase the efficiency of fact-checkers necessary.

Recent research has explored the potential of using large artificial intelligence language models as a tool for fact-checking (He et al., 2021; Caramancion, 2023; Choi and Ferrara, 2023; Hoes et al., 2023; Sawiński et al., 2023). However, significant challenges remain when employing large language models (LLMs) to assess the veracity of a statement. One primary issue is that fact-checks are potentially included in some of the training data for LLMs. Therefore, successful fact-checking without additional context may not necessarily be attributed to the model's comprehension of facts or argumentation. Instead, it may simply reflect the LLM's retention of training examples. While this might suffice for fact-checking past claims, it may not generalize well beyond the training data.

Large language models (LLMs) like GPT-4 are increasingly trusted to write academic papers, lawsuits, news articles,[1] or to gather information (Choudhury and Shamszare, 2023). Therefore, an investigation into the models' ability to determine whether a statement is true or false is necessary to understand whether LLMs can be relied upon in situations where accuracy and credibility are paramount. The widespread adoption and reliance on LLMs pose both opportunities and challenges. As they take on more significant roles in decision-making processes, research, journalism, and legal domains, it becomes crucial to understand their strengths and limitations. The increasing use of advanced language models in disseminating misinformation online highlights the importance of developing efficient automated systems. The 2024 WEF Global Risk Report ranks misinformation and disinformation as the most dangerous short-term global risk as LLMs have enabled an "explosion in falsified information" removing the necessity of niche skills to create "synthetic content" (World Economic Forum, 2024). On the other hand, artificial intelligence models can help identify and mitigate false information, thereby helping to maintain a more reliable and accurate information environment. The ability of LLMs to discern truth from falsehood is not just a measure of their technical competence but also has broader implications for our information ecosystem.

A significant challenge in automated fact-checking systems relying on machine learning models has been the lack of explainability of the models' prediction. This is a particularly desirable goal in the area of fact-checking as explanations of verdicts are an integral part of the journalistic process when performing manual fact-checking (Kotonya and Toni, 2020). While there has been some progress in highlighting features that justify a verdict, a relatively small number of automated fact-checking systems have an explainability component (Kotonya and Toni, 2020).

Since the early 2010's, a diverse group of researchers have tackled automated fact-checking with various approaches. This section introduces the concept of automated fact-checking and the different existing approaches. Different shared tasks, where research groups tackle the same problem or dataset with a defined outcome metric, have been announced with the aim of automatically fact-checking claims. For example, the shared task RumourEval provided a dataset of "dubious posts and ensuing conversations in social media, annotated both for stance and veracity" (Gorrell et al., 2019). CLEF CheckThat! prepared three different tasks, aiming to solve different problems in the fact-checking pipeline (Nakov et al., 2022b). First, "Task 1 asked to predict which posts in a Twitter stream are worth fact-checking, focusing on COVID-19 and politics in six languages" (Nakov et al., 2022a). Task 2 "asks to detect previously fact-checked claims (in two languages)" (Nakov et al., 2022c). Lastly, "Task 3 is designed as a multi-class classification problem and focuses on the veracity of German and English news articles" (Köhler et al., 2022). The Fact Extraction and VERification shared task (FEVER) "challenged participants to classify whether human-written factoid claims could be SUPPORTED or REFUTED using evidence retrieved from Wikipedia" (Thorne et al., 2018b). In general, most of these challenges and proposed solutions disaggregate the fact-checking pipeline into a multi-step problem, as detection, contextualization, and verification all require specific approaches and methods (Das et al., 2023). For example, (Hassan et al., 2017) proposed four components to verify a web document in their ClaimBuster pipeline. First, a claim monitor that performs document retrieval (1), a claim spotter that performs claim detection (2), a claim matcher that matches a detected claim to fact-checked claims (3), and a claim checker that performs evidence extraction and claim validation (4) (Zeng et al., 2021).

In their summary of automated fact-checking (Zeng et al., 2021) define entailment as "cases where the truth of hypothesis h is highly plausible given text t." More stringent definitions that demand that a hypothesis is true in "every possible circumstance where t is true" fail to handle the uncertainty of Natural Language. Claim verification today mostly relies on fine-tuning a large pre-trained language model on the target dataset (Zeng et al., 2021). State-of-the-art entailment models have generally relied on transformer architecture such as BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019). Hoes et al. (2023) tested GPT-3.5's claim verification performance on a dataset of PolitiFact statements without adding any context. They found that GPT-3.5 performs well on the dataset and argue that it shows the potential of leveraging GPT-3.5 and other LLMs for enhancing the efficiency and expediency of the fact-checking process. Novel large language models have been used by Sawiński et al. (2023) in assessing the check-worthiness. The authors test various models ability to predict the check-worthiness of English language content. The authors compared GPT-3.5 with various other language models. They find that a fine-tuned version of GPT3.5 slightly ourperforms DeBerta-v3 (He et al., 2021), an improvement over the original DeBERTa architecture (He et al., 2020). Choi and Ferrara (2023) use fact-checks to construct a synthetic dataset of contradicting, entailing or neutral claims. They create the synthetic data using GPT-4 and predict the entailment using a smaller fine-tuned LLM. Similarly, Caramancion (2023) test the ability of various LLMs to discern fake news by providing Bard, BingAI, GPT-3.5, and GPT-4 on a list of 100 fact-checked news items. The authors find that all LLMs achieve

1 https://cybernews.com/news/academic-cheating-chatgpt-openai/; https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

performances of around 64–71% accuracy, with GPT 4 receiving the highest score among all LLMs. Cuartielles Saura et al. (2023) interview fact-checking platforms about their expectations of ChatGPT as a tool for both misinformation fabrication, detection, and verification. They find that while professional fact-checkers highlight the potential perils such as the reliability of sources, the lack of insights into the training process, and the enhanced ability of malevolent actors to fabricate false content, they nevertheless view it as a useful resource for both information gathering and the detection and debunking of false news (Cuartielles Saura et al., 2023).

While earlier efforts in claim verification did not retrieve any evidence beyond the claim itself (for example, see Rashkin et al., 2017), augmenting claim verification models with evidence retrieval has become standard for state-of-the-art models (Guo et al., 2021). In general, evidence retrieval aims to incorporate relevant information beyond the claim. For example, from encyclopedias (e.g., Wikipedia Thorne et al., 2018a), scientific papers (Wadden et al., 2020), or search engines such as Google (Augenstein et al., 2019). Augenstein et al. (2019) submit a claim verbatim as a query to the Google Search API and use the first ten search results as evidence. A crucial issue for evidence retrieval lies in the fact that it implicitly assumes that all available information is trustworthy and that veracity can be gleaned from simply testing the coherence of the claim with the information retrieved. An alternative approach that circumvents the issue of the inclusion of false information has been to leverage knowledge databases (also knowledge graphs) that aim to "equip machines with comprehensive knowledge of the world's entities and their relationships" (Weikum et al., 2020). However, this approach assumes that all facts pertinent to the checked claim are present in a graph. An assumption that (Guo et al., 2021) called unrealistic.

Our primary contributions in this study are 2-fold. First, we conduct a novel evaluation of two of the most used LLMs, GPT-3.5, and GPT-4, on their ability to perform fact-checking using a specialized dataset. An original part of our examination distinguishes the models' performance with and without access to external context, highlighting the importance of contextual data in the verification process. Second, by allowing the LLM agent to perform web searches, we propose an original methodology integrating information retrieval and claim verification for automated fact-checking. By leveraging the ReAct framework, we design an iterative agent that decides whether to conclude a web search or continue with more queries, striking a balance between accuracy and efficiency. This enables the model to justify its reasoning and cite the relevant retrieved data, therefore addressing the verifiability and explainability of the model's verdict. Lastly, we perform the first assessment of GPT-3.5's capability to fact-check across multiple languages, which is crucial in today's globalized information ecosystem.

We find that incorporating contextual information significantly improves accuracy. This highlights the importance of gathering external evidence during automated verification. We find that the models show good average accuracy, but they struggle with ambiguous verdicts. Our evaluation shows that GPT-4 significantly outperforms GPT-3.5 at fact-checking claims. However, performance varies substantially across languages. Non-English claims see a large boost when translated to English before being fed to the models. We find no sudden decrease in accuracy after the official training cutoff dates for GPT-3.5 and GPT-4. This suggests that the continued learning from human feedback may expand these models' knowledge.

2 Materials and methods

2.1 Approach

This paper contributes to both the evidence retrieval and claim verification steps of the automated fact-checking pipeline. Our approach focuses on verifying claims—assessing if a statement is true or false, while simultaneously retrieving contextual information to augment the ability of the LLM to reason about the given claims. We use large artificial intelligence language models GPT-3.5 and GPT-4. We combine state-of-the-art language models, iterative searching, and agent-based reasoning to advance automated claim verification.

GPT-3.5 and GPT-4 are neural networks trained on vast amounts of textual data to generate coherent continuations of text prompts (Vaswani et al., 2017). They comprise multiple layers of transformer blocks, which contain self-attention mechanisms that allow the model to learn contextual representations of words and sentences (Vaswani et al., 2017). LLMs are trained using self-supervision, where the model's objective is to predict the next token in a sequence of text. The GPT models are trained on vast amounts of unstructured textual data like the common crawl dataset, which is the largest dataset to be included in the training. Common Crawl is a web archive that consists of terabytes of data collected since 2008 (Buck et al., 2014; Brown et al., 2020). GPT-3.5 and GPT-4 are additionally trained with reinforcement learning from human feedback (Rlhf), where human feedback is incorporated to enhance the models' usability. OpenAI states that they regularly update their models based on human feedback, potentially leading to knowledge of current events that expands upon the initial training regime, which was stopped in September of 2021.[2]

We are evaluating the performance of GPT-3.5 and GPT-4 based on two conditions. First, we query the models with the statement, the author, and the date of the statement. The model does not possess any means of retrieving further information that might enable it to make an informed decision. Rather this approach relies on the model having knowledge of the events described in the claim. In the second condition, we enable the LLM to query a Google Search engine to retrieve relevant information surrounding the claim. To prevent the model from uncovering the fact-check itself, we filter the returned Google search results for all domains present in the dataset. We present the LLM with information returned from the Google Search Engine API, comprising previews of search results. These previews included the title of the website, the link, and an extract of relevant context, mirroring the typical user experience on Google. To refine our approach, we experimented with integrating additional information from the full HTML content of websites. We quickly realized that this voluminous data was overwhelming the LLM's context window,

2 https://openai.com/blog/chatgpt

leading to suboptimal performance. To address this, we employed BM25, an information retrieval function (Robertson and Zaragoza, 2009), to distill the most critical parts of each website. We found no improvement to the performance of the LLMs, as Google already uses machine learning methods to identify the most important information to include in the preview. We equip the agent with the ability to query Google by leveraging the Reasoning and Acting (*ReAct*) framework, proposed by Yao et al. (2023), which allows an LLM to interact with tools. The goal of the ReAct framework is to combine reasoning and take actions. Reasoning refers to the model planning and executing actions based on observations from the environment. Actions are function calls or API calls by the model to retrieve additional information from external sources. The model is initially prompted with the claim and then decides if it needs to take actions. As we equip both models with the ability to query Google, the model then is able to retrieve information and receives its next observation. Based on this observation, the model can decide to either return a final answer or retrieve information with a different query. If the model fails to answer after three iterations of retrieving information, we terminate the search. The model retrieves 10 Google search results per iteration. To test the claim verification in a realistic scenario, any result from a fact-checking website present in the dataset is removed from the results. If the model was terminated because it reached the maximum number of iterations, it is prompted to provide a final answer based on all of its previous Google searches. In this paper, we employ the LangChain library (Chase, 2022) to create an agent within the ReAct framework. We show in Figure 1 the workflow we implemented and an example of the treatment of a claim.

We employed the 16k context window which is standard for the GPT API. Since our experiments larger context windows have been introduced (OpenAI, 2023)[3], which could enable future LLM powered fact-checking applications to incorporate more search engine results or to include more content from each website.

To illustrate the capabilities of the proposed system, Figures 2, 3 showcase two correctly classified and two incorrectly assessed verdicts. The full transcripts from these examples, including all retrieved Google links, are available in the Supplementary material.

2.2 Experiment

We conduct two experiments to evaluate how well our agents can fact-check different claims. First, we evaluate whether GPT-4 is significantly better at fact-checking than GPT-3.5. GPT-4 has been shown to outperform GPT-3.5 on a variety of benchmarks, particularly in zero-shot reasoning tasks (Espejel et al., 2023). We therefore compare the performance of GPT-3.5 and GPT-4 on two datasets, a dataset of US political fact-checked claims provided by PolitiFact (Misra, 2022) and a dataset of multilingual fact-checked claims provided by Data Common. As GPT-3.5 is openly available to the public for free, it is used substantially more used than GPT-4, thereby increasing the importance of understanding its performance.

2.2.1 Experiment on the PolitiFact dataset

The dataset used in this experiment consists of a database of fact-checked claims by PolitiFact (Misra, 2022). Each observation has a statement originator (the person who made the statement being fact-checked), a statement (the statement being fact-checked), a statement date, and a verdict. The verdict of a fact-check is one of six ordinal categories, indicating to which degree a statement is true or false: True, Mostly-True, Half-True, Mostly-False, False, and Pants-Fire. Pants-Fire indicates that a statement is utterly false. In total, the dataset has 21,152 fact-checks, spanning a time-frame of 2007–2022. The number of fact-checks per month is shown in Figure 4. We sampled 500 claims for each response category, leading to a total of 3,000 unique fact-checks to be parsed. We thereby ensured an equal distribution of outcome labels in the final dataset.

While comparing the two models, we also compare the accuracy of the model with and without additional context. We therefore run four conditions, GPT-3.5 with context and without context, and GPT-4 with and without context. The *No-Context* condition refers to the model being prompted to categorize a claim into the veracity categories, without being given any context, or the ability to retrieve context. In addition to the veracity labels, the model is able to return a verdict of "uncertain" to indicate, that it isn't capable of returning an assessment. In the *Context* condition, the agent is capable of formulating Google queries to retrieve information pertinent to the claim. When the model is able to retrieve contextual information, any results from PolitiFact are excluded from the results.

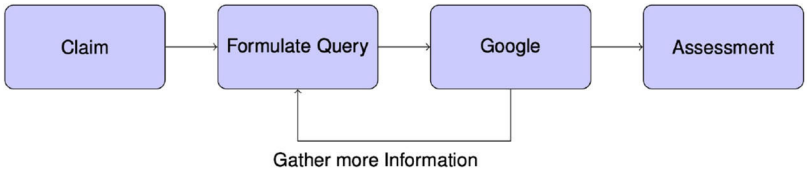
2.2.2 Experiment on the multilingual dataset

Secondly, we investigate whether the ability of LLMs to fact-check claims depends significantly on the language of the initial claim. While both GPT-3.5 and GPT-4 exhibit impressive multilingual understanding, a variety of empirical results have shown that large language models struggle to adequately understand and generate non-English language text (Bang et al., 2023; Jiao et al., 2023; Zhu et al., 2023). However, misinformation is a problem that is not restricted to high-resource languages. Conversely, fact-checking has become a global endeavor over the last decade, with a growing number of dedicated organizations in non-English language countries.

The dataset used is a fact-check dump by Data Commons. Each fact-check has an associated author and date. Additionally, it contains the *Claim* that is being fact-checked, a *Review*, and a *verdict*.

The dataset contains a breadth of fact-checking organizations and languages. In total, we detect 78 unique languages in the dataset. By extracting the domain of each fact-check link we find that it contains fact-checks from 454 unique fact-checking organizations. The largest domains, factly (7,277), factcrescendo (5,664), youturn (3,582), boatos (2,829), dpa-fact-checking (2,394), verafiles (1,972), uol (1,729), and tempo (1,133), have more than one thousand fact-checks,

3 OpenAI (2023). *Models—openAI api*.



Prompt: You're a fact-checking expert. Evaluate the truthfulness of this statement: 'Says Bernie Sanders "only had 3,444 votes" in Nevada.' made by Viral image on 3/6/2020. Use Google for research. Consider your sources, context, and date while assessing. To answer return 'Final Answer: {verdict}, {reason}. You must respond with a valid verdict ('pants-fire','false', 'mostly-false', 'half-true', 'mostly-true', or 'true') or 'uncertain', providing reasoning and citing sources by providing the domain of pertinent search results.
Google: Bernie Sanders Nevada 2020 votes count

1. nytimes.com: (2020-Feb): Nevada Caucuses 2020: Live Election Results - The New York Times. Feb 22, 2020 ...
2. nbcnews.com: Sanders wins Nevada Democratic caucuses with wave of young and ....
3. nytimes.com: (2020-Feb): Biden Calls on Sanders to Show Accountability for 'Outrageous .. Feb 16, 2020
4. vox.com: (2020-Feb): Nevada caucus results: 3 winners and 2 losers - Vox. Feb 22, 2020
5. washingtonpost.com: 2020 Nevada Caucuses Live Election Results | The Washington Post. Feb 24, 2020 ... 2020 Nevada Democratic presidential caucuses ; Bernie Sanders, 6,788, 46.8 ; Joe Biden, 2,927, 20.2 ; Pete Buttigieg, 2,073 ....
6. theintercept.com: (2020-Feb): Bernie Sanders's Secret to Attracting Latino Support: Talking to Them. Feb 20, 2020
7. pbs.org: (2020-Feb): Bloomberg qualifies for debate, Sanders leads ahead of Nevada
8. wikipedia.org: 2020 Nevada Democratic presidential caucuses - Wikipedia.
9. politico.com: Iowa Election Results 2020 | Live Map Updates | Voting by County
10. tufts.edu: (2020-Feb): Exclusive Analysis: In Nevada, young people once again force

Final Answer: False, Bernie Sanders received more than 3,444 votes in the Nevada Democratic presidential caucuses according to The Washington Post.

FIGURE 1
Workflow showing how we enable LLM agents to interact with a context to assess the veracity of a claim **(top)**. Example of the treatment of a specific claim **(bottom)**.

and make up 73% of the dataset. In contrast, there are 266 different domains which only have one associated fact-check. Figure 4 shows the total amount of fact-checks in the dataset per month. While the oldest fact-check in the dataset was published in 2011, we see that very few fact-checks were uploaded until early 2019. Since then the dataset contains a relatively stable amount of around eight hundred fact-checks per month.

To use the Data Commons dataset we need to standardize the dataset. We reduced the number of unique verdicts, as many different fact-checking organizations use different ordinal scales that are hard to unify. By mapping all scales onto a coarse 4-level scale ("False," "Mostly False," "Mostly True," and "True"), we can then compare how the ability of our model to fact-check depends on the correct verdict assigned by the fact-checker. We translated all present verdicts from their original language to English. We then manually mapped all verdicts which appeared at least twice in the dataset (n = 468) to the four categories. We removed all observations that we were unable to map to one of these four categories. This included observations with a verdict such as "Sarcasm," "Satire," or "unconfirmed." Subsequently, we discarded all languages, that did not have at least 50 observations and languages that did not have at least 10 "true" or "mostly true" observations. From the remaining languages, we sampled up to 500 observations. To compare the ability of our approach across languages, we then used Googletrans,[4] a free python library that utilizes the Google Translate API, to translate all claims from their original language to English.

In this experiment, we compare the performance of GPT-3.5 both with context and without context in English and in the original language. In the context condition, we removed

4 https://pypi.org/project/googletrans/

Statement: 'Ted Cruz is mathematically out of winning the race.' made by Donald Trump on 4/20/2016.

Google Search Query: Republican primary delegate count April 20 2016

- 2016 Delegate Count and Primary Results - The New York Times Republican Delegates ; Donald J. Trump · 95 ; Ted Cruz · 9 ; Marco Rubio. 167 ; John Kasich. 161 ... https://www.nytimes.com/interactive/2016/us/elections/primary-calendar-and-results.html
- 2016 Republican Party presidential primaries - Wikipedia Fifty-six primary contests were conducted to choose 2,472 delegates. In 50 states and territories the delegates were allocated to candidates by popular vote ... https://en.wikipedia.org/wiki/2016_Republican_Party_presidential_primaries
- [...]

Google Search Query: Ted Cruz Donald Trump delegate count April 20 2016

- 2016 Delegate Count and Primary Results - The New York Times According to the Associated Press, Donald J. Trump and Hillary Clinton have each won enough ... Donald J. Trump. Earned. Unpledged. 1,447. 95. Ted Cruz. https://www.nytimes.com/interactive/2016/us/elections/primary-calendar-and-results.html
- The Real Magic Number That Donald Trump Needs to Hit Is Likely ... Experts weigh in on the delegate count needed for the GOP nomination. ByMEGHAN KENEALLY. April 20, 2016, 8:26 AM. 4:02 ... Ted Cruz or Gov. https://abcnews.go.com/Politics/real-magic-number-donald-trump-hit-lower-1237/story?id=38539256
- [...]

Google Search Query: Republican primary delegate count by candidate April 20 2016

- 2016 Delegate Count and Primary Results - The New York Times Delegate totals include unpledged delegates, also known as superdelegates, who are free to support any candidate at the party conventions. https://www.nytimes.com/interactive/2016/us/elections/primary-calendar-and-results.html
- 2016 Republican Party presidential primaries - Wikipedia However, Trump scored landslide victories in New York and five northeastern states in April, before taking every delegate in the Indiana primary on May 3. https://en.wikipedia.org/wiki/2016_Republican_Party_presidential_primaries
- [...]

Final Output: Mostly-true, the statement made by Donald Trump on 4/20/2016 that "Ted Cruz is mathematically out of winning the race" is mostly true. According to the delegate count from the New York Times (nytimes.com), Donald Trump had significantly more delegates than Ted Cruz. However, it's important to note that the race isn't officially over until a candidate reaches the required number of delegates to secure the nomination, which in 2016 was 1,237 delegates (abcnews.go.com). At the time of the statement, Trump had not yet reached this number, so technically Cruz was not mathematically eliminated. However, his path to victory was highly unlikely.

PolitiFact Verdict: Trump said, "Ted Cruz is mathematically out of winning the race." Cruz would need at the very least 678 more delegates to win the race. It's impossible for him to reach that number in the remaining primaries, given that there are just 674 delegates left to win. However, he could conceivably find more supporters in the states who've already held primaries, given that a few dozen are thus far uncommitted. Trump's statement is largely accurate but needs additional information. We rate it Mostly True.

Statement: 'President Trump's limousine driver did burnouts to entertain Daytona 500 spectators during a rain delay.' made by Viral image on 2/21/2020'

Google Search Query: President Trump limousine Daytona 500 burnouts

- 'Start your engines!' President Trump revs up Daytona 500 crowd ... - Feb 16, 2020 ... He was the first of any president to take a lap in the armored presidential limo, dubbed "The Beast," FOX News reported. Trump and first lady ... https://www.fox2detroit.com/news/start-your-engines-president-trump-revs-up-daytona-500-crowd-does-historic-lap-in-presidential-limo
- Denny Hamlin wins 3rd Daytona 500; Ryan Newman hospitalized - Feb 17, 2020 ... President Donald Trump and first lady Melania Trump ride in the presidential limousine as they take a pace lap ahead of the start of the ... https://www.news4jax.com/sports/2020/02/17/daytona-500-resumes-without-president-pomp-packed-house/
- Ryan Newman hospitalized in horrific wreck as Hamlin wins ... - Feb 17, 2020 ... Denny Hamlin won a crash-filled Daytona 500 on Monday after a ghastly final-lap ... Trump takes lap in armored presidential limo 'the Beast'. https://www.theguardian.com/sport/2020/feb/17/daytona-500-denny-hamlin-ryan-newman-crash
- NASCAR 2020 Thread | Steeler Nation - Feb 16, 2020 ... President Trump becomes first president named Daytona 500 grand marshal. DAYTONA BEACH, Fla. ... do a burn-out. the-beast-president-limo.jpg ... https://www.steelernation.com/forums/threads/nascar-2020-thread.34154/

Google Search Query: Republican primary delegate count by candidate April 20 2016

- Ryan Newman hospitalized in horrific wreck as Hamlin wins ... - Feb 17, 2020 ... ... 500 at Daytona International Speedway. Photograph: Peter Casey/USA Today Sports. Nascar ... Trump takes lap in armored presidential limo 'the ... https://www.theguardian.com/sport/2020/feb/17/daytona-500-denny-hamlin-ryan-newman-crash
- Denny Hamlin wins 3rd Daytona 500; Ryan Newman hospitalized - Feb 17, 2020 ... ... burnouts that Hamlin learned the accident was bad. "It's a weird balance of ... Trump takes Daytona 500 warmup lap in presidential limousine. https://www.news4jax.com/sports/2020/02/17/daytona-500-resumes-without-president-pomp-packed-house/
- 'Start your engines!' President Trump revs up Daytona 500 crowd ... - ... !' President Trump revs up Daytona 500 crowd, does historic lap in ... Deputy gives Dream Cruise driver a spanking after he did burnout on Woodward .. https://www.fox2detroit.com/news/start-your-engines-president-trump-revs-up-daytona-500-crowd-does-historic-lap-in-presidential-limo
- [...]

Final Output: False, there is no evidence from reliable sources to suggest that President Trump's limousine driver performed burnouts at the Daytona 500. The sources only confirm that Trump took a lap in his limousine.

PolitiFact Verdict: Sorry to disappoint, but no. A Facebook post shares an altered image of such a scene, along with a caption that says, "the presidential limo driver entertains the crowd during the rain delays at the Daytona 500 with smokey burnouts." [...] The image then got passed around as authentic. If such a thing happened in the First Car, we would all know about it. We rate this altered photo an extra-smoky Pants on Fire!

FIGURE 2
Examples of PolitiFact statements for which the model returns correct responses. The LLM is tasked with verifying a statement made by Donald Trump, indicating that Ted Cruz is mathematically out of the race. The LLM uses Google to retrieve information on the delegate count and correctly concludes the statement is mostly true. We show the Google queries performed by the LLM and the first results of each query. In the second example, the LLM is tasked to verify a statement claiming Donald Trump's driver "did burnouts" during a race. The LLM finds information that Donald Trump did a lap around the race but correctly concludes that no information indicates that he did "burnouts." The full examples, including all Google results, are shown in the Supplementary material.

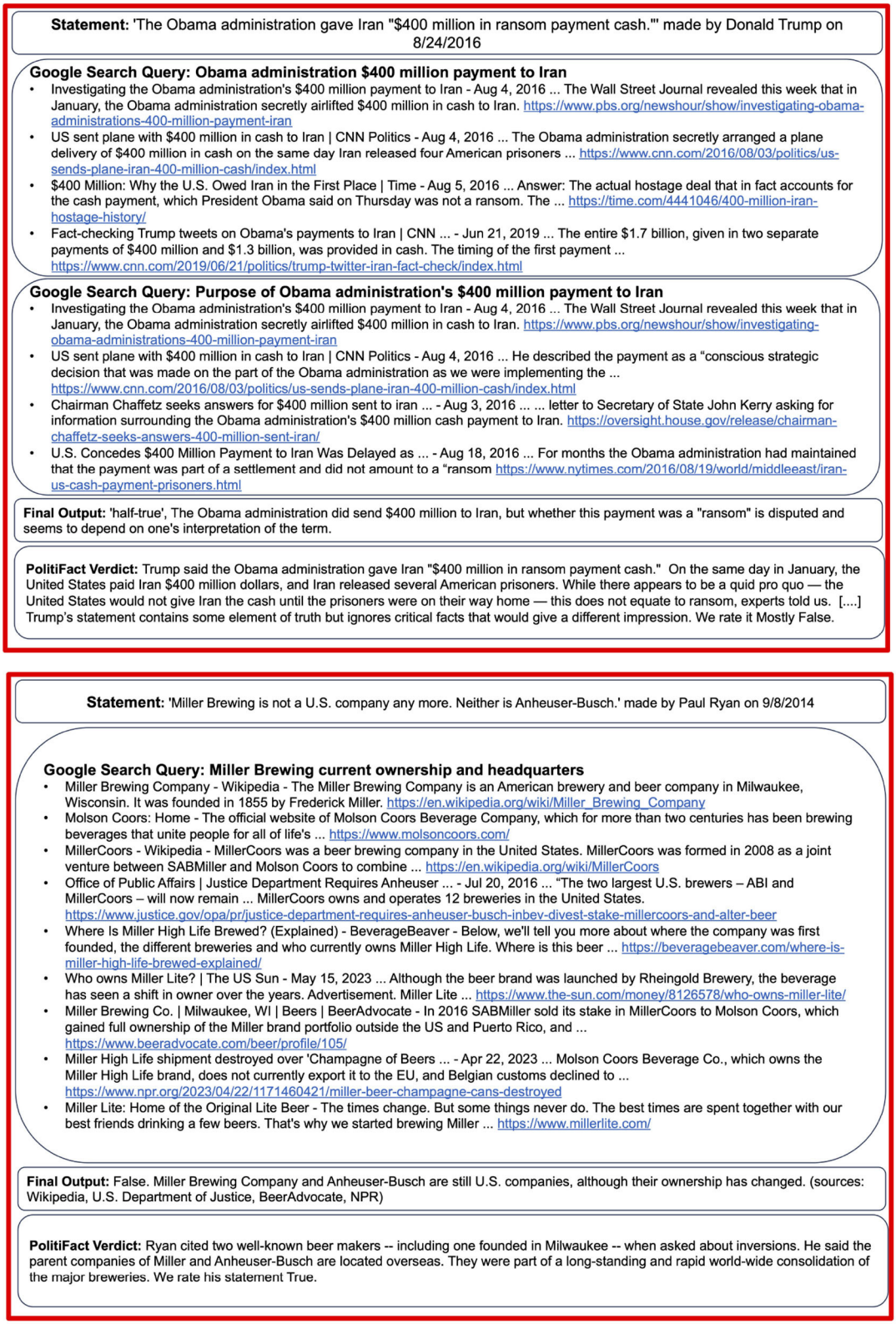


FIGURE 3
Examples of PolitiFact statements for which the model returns incorrect responses. We show the Google queries performed by the LLM and the first results of each query. The LLM is asked to verify whether the Obama administration paid a ransom payment to Iran. The LLM finds information on the payment but can't conclusively confirm the purpose of the payment. It concludes that the statement is half-true. PolitiFact argues that the statement is mostly-false, as the payment is not necessarily a ransom payment. In the second example, the LLM is asked to verify whether a beer brand is American. It finds information indicating that the company is American and returns False. The company has, however, been bought by foreign investors, making the statement true. The full examples, including all Google results, are shown in the Supplementary material.

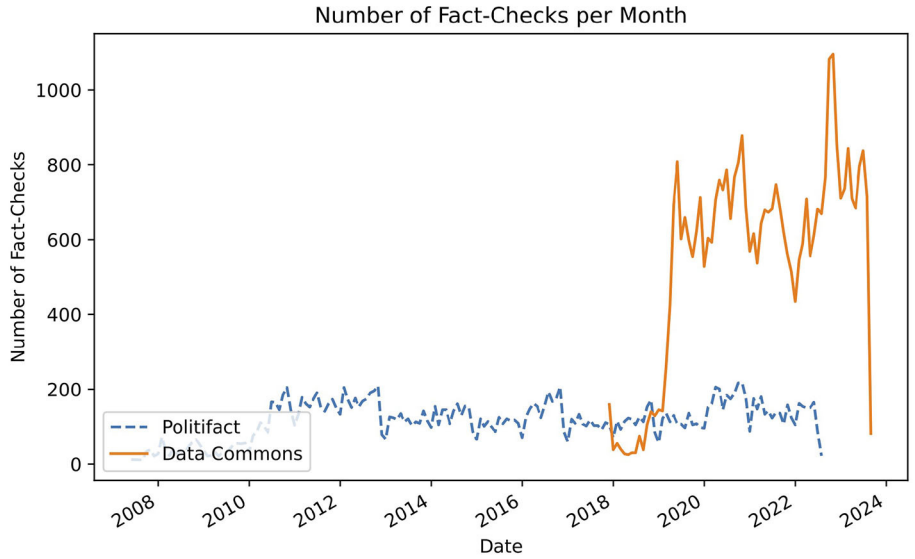


FIGURE 4
Number of fact-checks per month in the Data Commons and PolitiFact datasets. Number of fact-checks Per Month in the Data Commons and PolitiFact datasets. In blue (dashed) the number of fact-checks in the PolitFact Dataset are shown. The orange (solid) line indicates the number of fact-checks in the Data Commons dataset.

TABLE 1 Comparison of accuracy of all conditions on the PolitiFact dataset.

	No context		Context			No context		Context	
Correct verdict	GPT-3.5	GPT-4	GPT-3.5	GPT-4		GPT-3.5	GPT-4	GPT-3.5	GPT-4
Pants fire	92.19	91.80	93.42	93.92		25.81	28.94	0.00	12.17
False	81.63	88.08	88.75	86.13		64.17	47.49	86.32	55.82
Mostly false	79.64	68.70	71.11	55.82		8.82	50.00	10.67	40.75
Half true	36.16	51.90	51.58	67.26		2.75	11.58	0.00	3.57
Mostly true	42.35	79.29	67.27	80.88		9.41	59.91	36.82	61.13
True	49.11	71.30	67.61	84.92		37.05	22.22	33.80	33.00

For the table on the left, accuracy is calculated as the percentage of time predicting the correct overall category. Predicting any of pants-fire, false, or mostly false is correct when the claim is labeled as pants-fire, false, or mostly false. Conversely, predicting any half-true, mostly true, or true is correct when the claim is labeled as one of these categories. For the table on the right, accuracy is calculated as the percentage of time predicting the PolitiFact category. For example, predicting "half-true" when the claim is labeled as "mostly-true" is treated as a false prediction.

all fact-checking websites from the dataset from the Google search results.

3 Results

3.1 Experiment on the PolitiFact dataset

Table 1 presents a comparative analysis of the GPT-3.5 and GPT-4 models in the task of automated fact-checking using the PolitiFact dataset. The table categorizes the performance of both models across different veracity labels ranging from "pants-fire" to "true." In the left table, the accuracy of the model is computed by considering only two categories, i.e., grouping "pants-fire," "false," and "mostly-false" together to false and "half-true," "mostly-true," and "true" to true. In the right table, the accuracy is computed using all the PolitiFact categories.

In the no-context condition, considering the overall accuracy (left table), GPT-4 generally outperforms GPT-3.5. GPT-3.5 predicts the veracity of a claim to be false in 58.2% of the cases (compared to 22.89% for GPT-4) achieving an accuracy of 36–49% in the true-label categories. In the false-label categories, the difference between the two models is significantly smaller and GPT-3.5 outperforms GPT-4 in the mostly false category. Both models achieve an accuracy of over 90% for claims that are labeled with "pants-on-fire." In the context condition, GPT-3.5 is significantly better calibrated, meaning it exhibits a more balanced accuracy between true and false verdicts. While the accuracy in the false categories only differs insignificantly, GPT-3.5 achieves significantly better results in predicting true verdicts in the context condition than in the no-context condition. Similarly, the difference in false categories is smaller for GPT-4 than in true categories, where it achieves an increase of 10.19 percentage points on average. In the context condition, both GPT-3.5 and GPT-4 outperform the no-context conditions on average. Context both increases the ability of the models to discriminate true from false claims, but additionally, calibrates both models better, predicting true for more cases correctly.

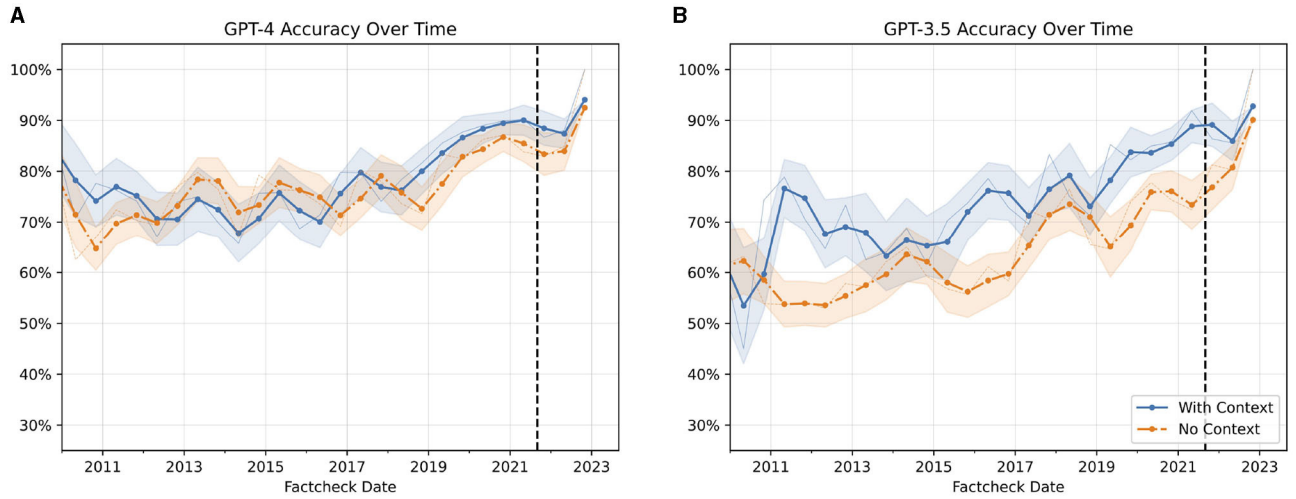


FIGURE 5
Accuracy of GPT-3.5 and GPT-4 overtime on the PolitiFact dataset. Yearly rolling average of the accuracy of LLMs over Time. **(A)** Displays the accuracy of GPT-4. **(B)** Shows the accuracy of GPT-3.5. The blue line indicates the context condition, and the orange line indicates the no-context condition. The vertical line represents the training end date of both models according to OpenAI. A faint line is the 3-month average. The x-axis represents the date of the claim. The bands represent one standard error.

TABLE 2 **Performance on the multilingual dataset without context.**

Language	Accuracy English	Accuracy multilingual	F1 English	F1 multilingual	Number of samples
Turkish	84.19	81.50	83.59	81.91	500
Indonesian	86.68	84.59	87.52	79.59	500
French	74.67	81.67	75.57	72.30	166
English	-	60.98	-	68.65	500
Thai	48.21	54.34	50.41	66.45	69
Portuguese	77.22	65.56	77.02	64.28	500
Tamil	70.33	59.19	67.68	51.59	500
Spanish	56.92	51.26	57.69	49.51	500
Italian	45.25	50.29	48.25	47.07	427
Chinese	64.86	43.75	68.56	42.88	174
Hebrew	49.61	41.17	56.51	40.65	287
Farsi	68.28	40.54	71.96	36.26	259
Telugu	59.79	26.99	67.03	27.24	500
Azerbaijani	37.23	22.89	42.26	15.64	98

Accuracy and F1 score for the fact-checking without context on the multilingual dataset. Accuracy English and F1 English are the scores for the case where the models are provided with English translation of the claims while the other scores are computed in the case where the models are provided with the original claims.

Across all conditions, both models generally fare better in identifying false statements ("pants-fire," "mostly-false," and "false") than true ones ("half-true," "mostly-true," and "true"), mirroring the findings of Hoes et al. (2023) in the no-context settings. This could be attributed to the inherent complexity of verifying a statement's absolute truth compared to identifying falsehoods. Lastly, both models have reduced performance in less extreme categories like "half-true" and "mostly-false." These categories are inherently ambiguous and represent statements that have elements of both truth and falsehood, making them more challenging to classify accurately. The table on the right only considers a claim to be correctly classified when the model predicts the exact category. The extremely low scores show that the models are unable to predict the exact shade of truth that PolitiFact assigned to the statement.

Both GPT-3.5 and GPT-4 were trained with data up to September of 2021.[5] As the training data of both models is private, data leakage is a significant concern for fact-checks published before that date. To investigate, whether the accuracy of the LLMs

5 https://platform.openai.com/docs/models/gpt-4

TABLE 3 Performance on the multilingual dataset with context.

Language	Accuracy English	Accuracy multilingual	F1 English	F1 multilingual	Number of samples
Portuguese	87.42	89.21	75.97	80.78	500
Indonesian	87.16	89.98	77.26	73.36	500
English	-	80.16	-	71.33	500
Telugu	83.15	83.80	77.21	69.97	500
Thai	77.27	50.00	55.66	66.66	69
French	89.55	87.09	79.58	56.13	166
Chinese	75.81	84.78	71.86	55.35	174
Farsi	77.78	59.00	69.15	48.58	259
Turkish	72.28	71.79	70.79	47.32	500
Spanish	82.68	75.18	63.01	46.30	500
Italian	56.93	56.29	40.15	45.54	427
Tamil	80.56	55.22	62.95	34.86	500
Hebrew	73.61	85.71	63.81	34.42	287
Azerbaijani	62.07	44.00	43.43	33.23	97

Accuracy and F1 score for the fact-checking with context on the multilingual dataset. Accuracy English and F1 English are the scores for the case where the models are provided with English translation of the claims while the other scores are computed in the case where the models are provided with the original claims.

differs over time we plot the accuracy of all four conditions (2 models x 2 context conditions) in Figure 5.

The accuracy of all four conditions exhibits an upward trend over time. For the context condition, this improvement can likely be attributed to an increasing availability of relevant information on Google over recent years. More surprisingly, the no-context condition also displays a similar trajectory of improving performance. One potential explanation is that the ground truth labels of some statements in the dataset may have evolved as more facts emerged. We do not observe any sudden decrease in accuracy after the official training cutoff for GPT-3.5 and GPT-4. This suggests that post-training refinements to the models via reinforcement learning from human feedback (RLHF) may introduce new knowledge, particularly for more recent events closer to the RLHF time period.

3.2 Experiment on the multilingual dataset

Table 2 shows the accuracy of GPT-3.5 in evaluating the veracity of fact-checking claims. The columns Accuracy English and Accuracy Multilingual show the percentage of correctly classified claims by language. English refers to the translations of the claims while multilingual refers to the original language. The F1 columns show the F1 score of the models for the Multilingual and English condition. We add the F1 score to the evaluation to account for the fact that the class distribution differs by language.

In all languages, except for Thai, we see an increase in the F1 score of the model when fact-checking the translated claim as compared to the original. This shows that the models that were tested are significantly better at predicting the verdict of a statement when it is presented in English, mirroring prior research indicating that the models struggle to correctly model and classify non-English language text. Similarly, the accuracy was significantly higher for the English language condition.

Table 3 shows the accuracy of GPT-3.5 in evaluating the veracity of fact-checking claims with additional context provided. Similarly to Table 2 the English language condition is significantly more accurate than the Multilingual condition. In all but three languages (Portugese, Thai, and Italian), translating increased the F1 score of the model. This difference was again significant.

Figure 6 summarizes the findings offered in this section and displays a visualization that compares the F1 scores under the two conditions—"No Context" and "Context Retrieval"—across various languages. Each language is represented once on the y-axis, with two corresponding horizontal lines plotted at different heights. The blue dashed lines indicate the performance under the "No Context" condition, while the red solid lines represent the "Context Retrieval" condition. Each line connects two points, corresponding to the F1 scores achieved when the model is trained on either English-only or Multilingual data. The circle marker displays the F1 score in the English language condition, while the square portrays the efficacy under the multilingual condition.

Similarly, to the PolitiFact experiment, we analyzed the performance of the model over time. Figure 7 shows the F1 score of all conditions over time. The confidence intervals represent the standard errors and are bootstrapped by sampling with replacement from the dataset and repeatedly calculating the F1 score. As in the PolitiFact experiment, we again see an increase in the performance of the models over time. Mirroring the previous findings, we do not see a decrease in the performance of the models following the official cut-off date for the training of GPT-3.5.

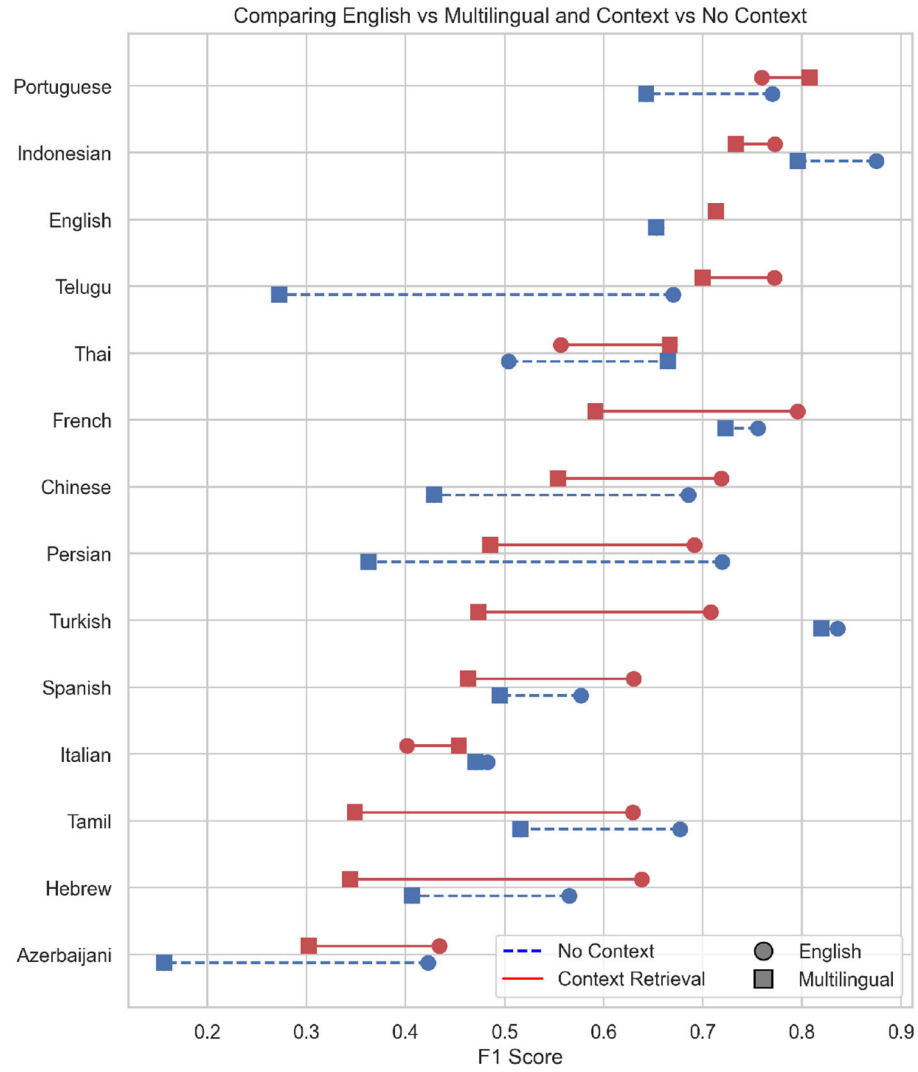


FIGURE 6
Comparison of all conditions on the Data Commons fact-checking dataset. The x-axis displays the F1 score of each language and condition. The y-axis displays each language for which the model was prompted. The blue, dashed line indicates the difference in the performance for the no context condition. The red, solid line shows the difference in performance for the context condition. The circles show the F1 score by language for the translations, the squares show the F1 score for the original scores.

4 Conclusion

In this study, we investigated OpenAIs GPT-3.5 and GPT-4s ability to fact-check claims. We found that GPT-4 outperformed GPT-3.5. Notably, even without contextual information, GPT-3.5 and GPT-4 demonstrate good performance of 63–75% accuracy on average, which further improves to above 80 and 89% for non-ambiguous verdicts when context is incorporated. Another crucial observation is the dependency on the language of the prompt. For non-English fact-checking tasks, translated prompts frequently outperformed the originals, even if the related claims pertained to non-English speaking countries and retrieved predominantly non-English information. The accuracy of these models varied significantly across languages, underscoring the importance of the original language of the claim.

As the training data of the GPT models includes vast amounts of web data, including a filtered version of the Common Crawl Corpus (Brown et al., 2020), there is a significant risk of data leakage for the task of fact-checking. Previous research has shown that misinformation does not exist in isolation, but is repeated (Shaar et al., 2020) across platforms (Micallef et al., 2022) and languages (Kazemi et al., 2022; Quelle et al., 2023). As misinformation is repeated and re-occurs, the ability of models to retain previously fact-checked claims can potentially be seen as a benefit rather than a drawback. Nevertheless, this presents a significant risk for the task of fact-checking novel misinformation. We address this in the paper by testing the ability of the models to detect misinformation after the training end date of the models. We found no reduction in performance for fact-checks after the training end date. It seems that incorporating real-time context for novel misinformation enables the models to reason about

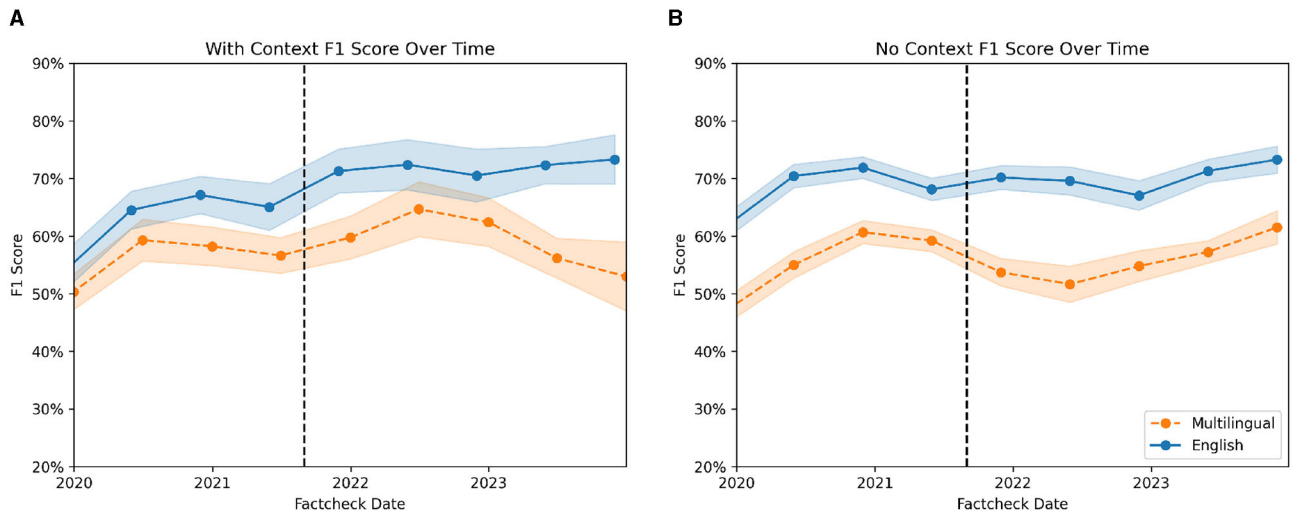


FIGURE 7
F1 score of context **(A)** and no-context-condition **(B)** over time. The orange (dashed) line shows the multilingual condition. The blue (solid) line showcases the F1 score for the English condition. The vertical line indicates the official training end date for GPT 3.5.

novel information. In summary, while data-leakage is a significant concern for any application that tests the abilities of LLMs, we argue that for the task of fact-checking it does not seem to degrade performance and might even be beneficial in view of recurring misinformation.

Comparing the performance of the models across languages is inherently difficult, as the fact-checks are not standardized across languages. Fact-checking organizations differ significantly in their choice of which fact-checks to dedicate time to, with some focussing exclusively on local issues and others researching any claims that are viral on social media. Similarly, some fact-checking organizations focus on current claims, while others debunk long-standing misinformation claims. All of these factors influence the ability of a large language model to discern the veracity of a claim. It is therefore possible that fact-checks in low-resource languages perform better than higher resource languages. The most salient point in the analysis of the multilingual misinformation is that the LLMs outperform the multilingual baseline when prompted with the English translation of the fact-checks. The variability in model performance across languages and the improvement of the accuracy when prompted with English language fact-checks indicates that the training regimen, in which the distribution of languages is highly skewed and English is dominant (Common Crawl Foundation, 2023)[6], significantly impacts accuracy. This suggests that the effectiveness of LLMs in fact-checking is not uniformly distributed across languages, likely due to the uneven representation of languages in training data.

Our results suggest that these models cannot completely replace human fact-checkers as being wrong, even if infrequently, may have devastating implications in today's information ecosystem. Therefore, integrating mechanisms allowing for the verification of their verdict and reasoning is paramount. In particular, they hold potential as tools for content moderation and accelerating human fact-checkers' work.

Looking ahead, it is important to delve deeper into the conditions under which large language models excel or falter. As these models gain responsibilities in various high-stakes domains, it is crucial that their factual reliability is well-understood and that they are deployed judiciously under human supervision.

While our study concentrated on OpenAI's GPT-3.5 and GPT-4, the rapid evolution of the field means newer, possibly fine-tuned, LLMs are emerging. One salient advantage of our methodology, distinguishing it from others, is the LLM Agents' capability to justify their conclusions. Future research should explore if, by critically examining the reasons and references provided by the LLMs, users can enhance the models' ability to fact-check claims effectively.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

DQ: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft, Writing—review & editing. AB: Conceptualization, Supervision, Writing—review & editing, Project administration.

6 Common Crawl Foundation (2023). *Statistics of Common Crawl Monthly Archives—Distribution of Languages*. Available online at: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors thank Fabrizio Gilardi, Emma Hoes, and Meysam Alizadeh for their valuable input that enriched the quality of this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1341697/full#supplementary-material

References

Adair, B., Li, C., Yang, J., and Yu, C. (2017). "Progress toward "the Holy Grail": the continued quest to automate fact-checking," in *Proceedings of the 2017 Computation+ Journalism Symposium* (Evanston, IL).

Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., et al. (2019). Multifc: a real-world multi-domain dataset for evidence-based fact checking of claims. *ArXiv*. doi: 10.18653/v1/D19-1475

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*. doi: 10.48550/arXiv.2302.04023

Bovet, A., and Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* 10:7. doi: 10.1038/s41467-018-07761-2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* 33, 1877–1901. Available online at: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. *LREC* 2:4. Available online at: https://aclanthology.org/L14-1074/

Caramancion, K. M. (2023). News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *ArXiv*. doi: 10.48550/arXiv.2306.17176

Chase, H. (2022). *Langchain*. Available online at: https://github.com/hwchase17/langchain; https://github.com/langchain-ai/langchain/blob/master/CITATION.cff

Choi, E. C., and Ferrara, E. (2023). Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. *ArXiv*. doi: 10.2139/ssrn.4614239

Choudhury, A., and Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of chatgpt: survey analysis. *J. Med. Internet Res.* 25:e47184. doi: 10.2196/47184

Cuartielles Saura, R., Ramon Vegas, X., and Pont Sorribes, C. (2023). Retraining fact-checkers: the emergence of chatgpt in information verification. *UPF Digit. Reposit.* 2023:15. doi: 10.3145/epi.2023.sep.15

Das, A., Liu, H., Kovatchev, V., and Lease, M. (2023). The state of human-centered NLP technology for fact-checking. *Inform. Process. Manag.* 60:103219. doi: 10.1016/j.ipm.2022.103219

Espejel, J. L., Ettifouri, E. H., Alassan, M. S. Y., Chouham, E. M., and Dahhane, W. (2023). Gpt-3.5 vs. gpt-4: evaluating chatgpt's reasoning performance in zero-shot learning. *ArXiv*. doi: 10.48550/arXiv.2305.12477

Flamino, J., Galeazzi, A., Feldman, S., Macy, M. W., Cross, B., Zhou, Z., et al. (2023). Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nat. Hum. Behav.* 7, 904–916. doi: 10.1038/s41562-023-01550-8

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., et al. (2019). "SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, MN: Association for Computational Linguistics), 845–854.

Graves, L., and Cherubini, F. (2016). "The rise of fact-checking sites in Europe," in *Digital News Project Report, Reuters Institute for the Study of Journalism* (Oxford).

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 374–378. doi: 10.1126/science.aau2706

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2021). A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* 10, 178–206. doi: 10.1162/tacl_a_00454

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., et al. (2015). "The quest to automate fact-checking," in *Proceedings of the 2015 Computation+ Journalism Symposium*. Citeseer.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., et al. (2017). Claimbuster: the first-ever end-to-end fact-checking system. *Proc. VLDB Endowment* 10, 1945–1948. doi: 10.14778/3137765.3137815

He, P., Gao, J., and Chen, W. (2021). Debertav3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*. doi: 10.48550/arXiv.2111.09543

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*. doi: 10.48550/arXiv.2006.03654

Hoes, E., Altay, S., and Bermeo, J. (2023). Leveraging chatgpt for efficient fact-checking. *PsyArXiv*. doi: 10.31234/osf.io/qnjkf

Jiao, W., Wang, W., Huang, J., Wang, X., and Tu, Z. (2023). Is chatgpt a good translator? Yes with gpt-4 as the engine. *ArXiv*. doi: 10.48550/arXiv.2301.08745

Kazemi, A., Li, Z., Pérez-Rosas, V., Hale, S., and Mihalcea, R. (2022). "Matching tweets with applicable fact-checks across languages," in *CEUR Workshop Proceedings* (Aachen).

Kenton, J. D. M.-W. C., and Toutanova, L. K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT* (Minneapolis, MN), 2.

Köhler, J., Shahi, G. K., Struß, J. M., Wiegand, M., Siegel, M., Mandl, T., et al. (2022). "Overview of the clef-2022 checkthat! lab task 3 on fake news detection," in *CEUR Workshop Proceedings* (Aachen).

Kotonya, N., and Toni, F. (2020). "Explainable automated fact-checking: a survey," in *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona: International Committee on Computational Linguistics), 5430–5443.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *ArXiv*. doi: 10.48550/arXiv.1907.11692

Mantzarlis, A. (2018). *Fact-Checking 101*. Paris: UNESCO Publishing Paris.

Micallef, N., Sandoval-Castañeda, M., Cohen, A., Ahamad, M., Kumar, S., and Memon, N. (2022). "Cross-platform multimodal misinformation: taxonomy, characteristics and detection for textual posts and videos," in *Proceedings of the International AAAI Conference on Web and Social Media* (Washington, DC), 651–662.

Misra, R. (2022). *Politifact Fact Check Dataset*. Available online at: https://rishabhmisra.github.io/publications/

Morris, D. S., Morris, J. S., and Francia, P. L. (2020). A fake news inoculation? fact checkers, partisan identification, and the power of misinformation. *Polit. Gr. Ident.* 8, 986–1005. doi: 10.1080/21565503.2020.1803935

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Kutlu, M., Zaghouani, W., et al. (2022a). "Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets," in *CEUR Workshop Proceedings* (Aachen).

Nakov, P., Barrón-Cedeño, A., da San Martino, G., Alam, F., Struß, J. M., Mandl, T., et al. (2022b). "Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection," in *International Conference of the Cross-Language Evaluation Forum for European Languages* (Berlin: Springer), 495–520.

Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., and Babulkov, N. (2022c). "Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims," in *CEUR Workshop Proceedings* (Aachen).

Nyhan, B., and Reifler, J. (2015). *Estimating Fact-Checking's Effects. Arlington, VA: American Press Institute*.

Porter, E., and Wood, T. J. (2021). The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2104235118. doi: 10.1073/pnas.2104235118

Quelle, D., Cheng, C., Bovet, A., and Hale, S. A. (2023). Lost in translation–multilingual misinformation and its evolution. *arXiv preprint arXiv:2310.18089*. doi: 10.48550/arXiv.2310.18089

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). "Truth of varying shades: analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA), 2931–2937.

Robertson, S., and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inform. Retriev.* 3, 333–389. doi: 10.1561/1500000019

Sawiński, M., Węcel, K., Księżniak, E. P., Stróżyna, M., Lewoniewski, W., Stolarski, P., et al. (2023). "Openfact at checkthat! 2023: head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy claims," in *CEUR Workshop Proceedings* (Aachen), 3,497.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020). "That is a known lie: detecting previously fact-checked claims," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky, J. Chai, N. Schluter and J. Tetreault (Association for Computational Linguistics) (Stroudsburg, PA), 3607–3618.

Siwakoti, S., Yadav, K., Bariletto, N., Zanotti, L., Erdogdu, U., and Shapiro, J. N. (2021). *How COVID Drove the Evolution of Fact-Checking*. Harvard Kennedy School Misinformation Review, Cambridge, Massachusetts, United States.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). "FEVER: a large-scale dataset for fact extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, LA: Association for Computational Linguistics), 809–819.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018b). "The fact extraction and VERification (FEVER) shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels: Association for Computational Linguistics), 1–9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30, 5998–6008.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., et al. (2020). "Fact or fiction: verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics), 7534–7550.

Weikum, G., Dong, L., Razniewski, S., and Suchanek, F. M. (2020). Machine knowledge: creation and curation of comprehensive knowledge bases. *Found. Trends Databases* 10, 108–490. doi: 10.1561/1900000064

World Economic Forum (2024). *The Global Risk Report*. Cologny.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., et al. (2023). React: synergizing reasoning and acting in language models. *ArXiv*. doi: 10.48550/arXiv.2210.03629

Zeng, X., Abumansour, A. S., and Zubiaga, A. (2021). Automated fact-checking: a survey. *Lang. Linguist. Compass* 15:e12438. doi: 10.1111/lnc3.12438

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., et al. (2023). Multilingual machine translation with large language models: empirical results and analysis. *ArXiv*. doi: 10.48550/arXiv.2304.04675