Explainable Fact-Checking with Large Language Models

How Prompt Style Variation affects Accuracy and Faithfulness in Claim Justifications

Marina Serafeimidi

Supervisor(s): Pradeep Murukannaiah, Shubhalaxmi Mukherjee

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Marina Serafeimidi
Final project course: CSE3000 Research Project
Thesis committee: Pradeep Murukannaiah, Shubhalaxmi Mukherjee, Xucong Zhang

Abstract

Large Language Models (LLMs) such as GPT-4 and LLaMA have demonstrated promising performance in fact-checking tasks, particularly in labeling the veracity of claims. However, the real-world utility of such fact-checking systems depends not only on label accuracy but also on the faithfulness of the justifications they provide. Prior work has explored various prompting strategies to elicit reasoning from LLMs, but most studies evaluate these styles in isolation or focus solely on veracity classification, neglecting the impact on explanation quality. This study addresses that gap by investigating how different prompt styles affect both the accuracy and the faithfulness of LLM-generated claim labelling and justifications. Seven established prompting strategies such as Chain-of-Thought, Role-Based, or Decompose-and-Verify, were tested across two datasets (QuanTemp and HoVer) using two efficient models: LLaMA 3.1:8B and GPT-4o-mini. Additionally, two novel prompt variants were introduced and all styles were tested under three label conditions to assess bias and explanation drift.

1 Introduction

The rise and accessibility of Large Language Models (LLMs) in everyday use has introduced an inevitable spike in the spread of misinformation (Hanley and Durumeric, 2024). Although the accessibility of these systems can have positive applications, it has certainly heightened the need for robust and transparent fact-checking systems. LLMs such as GPT-4 and LLaMA models have at times shown strong performance in claim verification - meaning, in deciding whether a claim is true or false (Quelle and Bovet, 2024; Cheung and Lam, 2023). However, in practice, the true value of real-world fact-checking lies not just in labeling a claim's veracity, but in providing a faithful justification for that label. Currently, LLMs have been shown to at times generate explanations that can be factually inconsistent, hallucinated, or biased, even when provided with the context and/or direct evidence for the claim(s) in question (Huang et al., 2025). This can have severe societal implications given that people who frequently use LLMs for facts and information tend to over-rely on and, often blindly, highly trust these systems (Si et al., 2024; YouGov, 2023).

The current research landscape has explored the task of fact checking with LLMs, however most tend to only evaluate the veracity of claims rather than also taking into account the quality of justifications (Kuznetsova et al., 2025; Setty, 2024). A 'Faithful' justification is one in which the model's generated reasoning is factually supported by the evidence, coherent, non-hallucinated, and relevant to the claim. Recent work by Russo et al. (2023) highlights the challenge of generating faithful fact-checking explanations, showing that state-of-the-art LLMs often hallucinate or misrepresent the reasoning behind fact labels, even on benchmark datasets. However, a knowledge gap exists in understanding how *prompt phrasing* - the way a user would attempt to formulate reasoning questions - affects the models' faithfulness to provided evidence. Prior work has introduced a variety of different approaches to prompting, with varying degrees of success, but the bulk of these studies typically evaluate each style in isolation or on narrow metrics (e.g., solely label accuracy) rather than conducting head-to-head comparisons of both accuracy and explanation faithfulness (Dmonte et al., 2025). This project attempts to build a better understanding of how different styles of prompting influence LLM fact-checking quality as a whole.

To address this gap, I pose the following Research Question:

"How does variation in prompt style affect the accuracy and faithfulness of LLM-generated justifications in claim verification?"

To further structure the research, I decompose the Research Question into the following sub-questions:

- SQ1: How does changing the structure, order, and phrasing of prompts impact the predicted labels and the accompanying justifications?
- SQ2: How does the presence and correctness of a supplied label influence the model's reasoning, justification quality, and susceptibility to label bias?
- SQ3: Does the accuracy and faithfulness of LLM justifications change depending on the type or complexity of the claim?
- SQ4: What prompt styles or LLM usage practices consistently maximize accuracy and

faithfulness, leading to higher factual alignment with evidence?

2 Prompt Formulation

A targeted review was conducted of recent studies on LLM-based claim verification (2023–2025) to identify the range of prompting strategies, styles, and templates that have been explored in practice in the context of LLM-based claim verification. Prompting styles were selected based on appearing repeatedly across multiple published papers and being evidently different in terms of either reasoning structure, role framing, evidence ordering, model instruction, etc.

This resulted in the following seven prompting strategies, capturing the most widely studied and structurally distinct approaches to LLM-based claim verification found in current research. While other variations exist, in terms of instruction tuning, output format, or evidence retrieval, they are typically based on one or more of these paradigms, not distinct, stand-alone approaches.

1. **Minimal:** A minimal instruction to label and justify a claim with no specific rationale. (baseline)

2. **Few-Shot In-Context Learning (ICL):** Includes a set of worked examples before the to-be-evaluated prompt to guide the model's prediction format (Singhal et al., 2024)

3. **Zero-Shot Role-Based Prompting:** Instructs the model to assume a specific role, often one of a professional fact checker or journalist, to guide its rationale and encourage critical reasoning (Li and Zhai, 2023)

4. **Chain-of-Thought:** Instructs the model to "think step-by-step" before answering, encouraging explicit reasoning (Wei et al., 2023), (Kojima et al., 2023), (Wang and Shu, 2023)

5. **Decompose and Verify:** Attempts to break down the claim into smaller 'reasoning units', such as sub-claims, predicates, sub-questions, etc. and expects the model to evaluate each of them before giving its overall judgement. (Zhang and Gao, 2023, HiSS), (Wang and Shu, 2023, FOLK)

6. **Evidence-First:** Requires the model to process the evidence first and draw conclusions about the relationship between the claim and evidence, such as the extent of correlation, or relevant facts between the claim and evidence (Tan et al., 2025, CorXFact), (Jafari and Allan, 2024, FactDetect)

7. **Multi Agent Debate:** Assigns different roles to multiple model agents, allowing them to critique each other's reasoning while working towards a common verdict (Kim et al., 2024, MADR), (Du et al., 2023)

Out of these strategies, I chose to evaluate structured approaches that are feasible to implement in such a comparative experiment, given the timeframe and available computing power. As a result, the Multi Agent Debate technique was taken out of consideration, given that managing multiple distinct LLM processes and synchronizing inter-agent communication introduces design and runtime overheads that place it outside the practical scope of this experiment.

2.1 Prompt Creation

To begin, a 'base prompt' was created to serve as a bare-boned template which could then be manipulated while keeping the underlying structure as similar as possible, in order to properly isolate the prompting style variations:

> "You are given a Claim and corresponding Evidence. (This Claim is considered [LABEL]). Please label the Claim as True, False, or Conflicting based solely on the Evidence, without using other internal or external knowledge or information, and provide an explanation for your decision.
>
> Claim: [CLAIM]
>
> Evidence: [EVIDENCE]
>
> Please format your answer as:
>
> Label: (True, False, or Conflicting)
>
> Explanation: (your explanation)"

Keeping these instructions identical allowed to observe the effects of the actual strategy change on the output, rather than potentially allowing superficial details to affect the reasoning.

The selected prompting strategies were then used to adapt this base template into six distinct

prompt templates, into which each claim and evidence could be "plugged into" to obtain results. Beyond the aforementioned strategies, two novel techniques were explored, which can be considered variants of the "Evidence-First" template. Both techniques specifically instruct the model to extract supportive and refuting parts of the evidence in relation to the claim, before making its veracity decision. Specifically, this was explored in two ways:

- a **"Support-Refute"** technique in which the model is instructed to identify supporting and refuting parts of the evidence, if such parts exist, before labeling the claim.
- a **"Arguments"** technique in which the model is instructed to identify and present 3 arguments for and 3 against the claim, from the evidence, before labeling it.

The exact prompt templates used for each strategy can be found in Appendix A.

Additionally, in order to further understand how the presence of label information affects model output, each prompt strategy was tested in three distinct conditions. These include one case where *no label* was given to the model, one where the *correct label* was given, and a case where a *fixed label* was always given, regardless of the original label of the claim. Specifically, this resulted in three fixed-label cases for the QuanTemp dataset (either **True**, **False**, or **Conflicting**) and two for the HoVer dataset (either **SUPPORTED** or **NOT_SUPPORTED**). Overall, this experimental setup allowed for observing the effect of two factors:

1. **Prompt Style:** how differences in wording or structure influence the model's response.
2. **Label Injection:** how the presence or absence of a supplied label for the given claim affects response and bias, using either
   a. no label
   b. the correct ("true") label
   c. a deliberately incorrect ("fake") label

3 Datasets and Models

Models I evaluate two near-state-of-the-art LLMs for this task: the 8B-parameter LLaMA 3.1 and OpenAI's GPT-4o-mini. These models were chosen for their performance-efficiency trade-off, making them ideal for this application. The LLaMA model showcases relatively strong reasoning capabilities while remaining small enough to be run locally (Meta, 2025). Its open-source availability is also a favorable aspect. The GPT model, a cost-efficient variant of GPT-4, achieves robust reasoning performance, surpassing GPT-3.5 Turbo and other equivalent small models on multiple academic benchmarks, while maintaining a significantly reduced per-token cost (OpenAI, 2024).

Datasets Experiments are run on filtered versions of two datasets, namely QuanTemp (Venktesh et al., 2024), and HoVer (Jiang et al., 2020). These datasets were chosen as they complement each other in covering a wide range of fact-checking challenges: QuanTemp provides real-world, journalist-verified claims with grounded evidence, while HoVer introduces multi-hop reasoning complexity, requiring models to combine information across multiple sources. Specifically:

► **QuanTemp** The QuanTemp dataset is comprised of numerical real-world claims, each associated with a label and evidence. Its test split was filtered down to 350 numerical real-world claims, all sourced from PolitiFact, an award-winning fact-checking website (The Pulitzer Prizes, 2009). Duplicates and non-English entries were removed. The split was restricted to PolitiFact to ensure consistency and reliability, given its verdicts having been produced by expert journalists. Finally, to avoid bias and allow the model to draw its own conclusions, the part of the evidence in which PolitiFact states its verdict was removed, leaving only the supporting evidence for each claim.

► **HoVer** HoVer is a multi-hop evidence extraction and fact verification dataset. A "hop", in this context, refers to the number of distinct evidence sentences (or Wikipedia pages) that must be combined in order to verify a claim. The dataset divides the claims by number of hops, which range from 1 to 5. For our purposes, 50 claims were taken per 2, 3 and 4 hops respectively, resulting in a reduced set of 150 claims. This filtering was done to reduce the computational complexity of the experiments.

Figure 1: Experimental Setup Workflow

4 Experimental Setup

A comparative experiment was conducted to evaluate the eight total prompting strategies outlined above across two datasets (QuanTemp and HoVer), two models (GPT-4o-mini and LLaMA 3.1:8B), and three label conditions.

4.1 Processing

An overview of the experiment pipeline can be found in Figure 1. Each claim-evidence pair was inserted into all eight prompt templates, for both datasets. Each template was run on five label injection conditions on QuanTemp (No Label, Real Label, and three fixed fakes) and four on HoVer (No Label, Real Label, two fakes) by appending "This claim is considered [LABEL]" to the prompt (omitted in the No-Label case). Each model therefore answered every claim once per template-condition combination. Scripts then parsed the raw response to obtain the generated Label and Justification, and evaluated the results to compare strategies.

Tools and Technologies: Langchain was used to facilitate the model communication, both with the LLaMA and GPT model. Experiments using the LLaMA model were run on a local machine with 16GB RAM. The GPT API was used for experiments using the GPT model. All code is made available on https://github.com/yuanzexiong/llm-fact-check.

4.2 Evaluation

Accuracy in this context was measured in terms of the amount of veracity labels generated by the model that matched the ground truth label assigned in each data set.

Faithfulness, defined as the degree to which each explanation matches the evidence, was measured using G-Eval (Liu et al., 2023), an LLM-based framework that prompts models to act as expert raters. It first generates evaluation steps to assess explanation faithfulness. Then, it rates the faithfulness of each explanation on a scale from 1 to 5 based on the provided claim and evidence, using the generated evaluation steps. G-Eval was chosen because it shows better performance and aligns more strongly with human judgments than traditional n-gram metrics such as BLUE/ROUGE, as shown by Liu et al.

5 Results

5.1 Prompt Strategy Comparison

The "No Label" condition provides the clearest view of how each prompt strategy alone influences model output. Figures 2 and 3 plot accuracy for QuanTemp and HoVer, while Figures 4 and 5 report the corresponding G-Eval faithfulness scores. Across both datasets, the GPT-4o-mini model is consistently more accurate and faithful than the LLaMA-3.1:8B model, regardless of strategy. The GPT model shows relative stability, as its accuracy ranges between 55-69% across both datasets. The LLaMA model is more volatile, with accuracy ranging from 38-63%.

Accuracy: QuanTemp The GPT model performs best with the Correlation strategy, at 69.4% accuracy. It also scores above 63% for 6/8 total strategies, with the exception of the Arguments and FOLK techniques. LLaMA scores 5-18 percentage points lower than GPT across all strategies, as is evident in Figure 2. It performs best under the Role-Based prompt with an accuracy of 58.9%, and also does well with Support-Refute and Chain-of-Thought. The FOLK strategy scores lowest for both models, at 54.9% and 49.7% accuracy for GPT and LLaMA respectively.

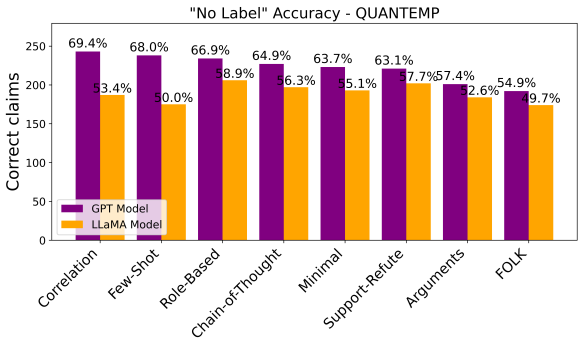

Figure 2: Accuracy per strategy on QuanTemp dataset.

Accuracy: HoVer On the tougher HoVer dataset, GPT's accuracy declines overall, while LLaMA improves on some prompts, as can be seen in Figure 3. However, its results are more uneven, showing both higher peaks and greater dips than on QuanTemp. Both models achieve highest accuracy with the Few-Shot model, at 65.3% for GPT and 63.3 for LLaMA. GPT maintains relatively consistent scores across strategies, but LLaMA shows steep dips in performance in decompositional-style prompt techniques such as FOLK (43.3%), Support-Refute (50%), and Arguments (38%).

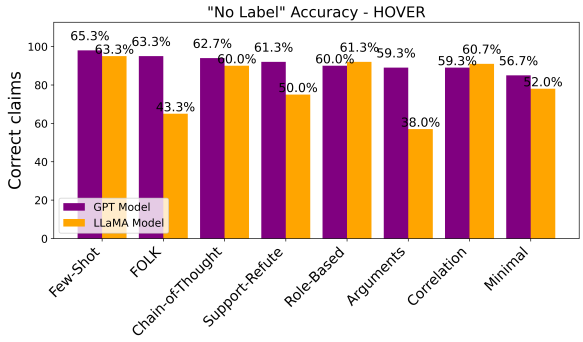

Figure 3: Accuracy per strategy on HoVer dataset.

Faithfulness: QuanTemp The Minimal approach yields highest the faithfulness across both models, with a score of 4.46 out of 5 for GPT and 4.25 for LLaMA. However, as can be seen in Figure 4, it is important to note that the differences are very small, as GPT scores above 4.4 for 5/8 strategies, and LLaMA above 4.2 for 4/8. The FOLK approach yields the worst results, with a faithfulness score of 4.04 using the GPT model, and 3.88 using LLaMA.

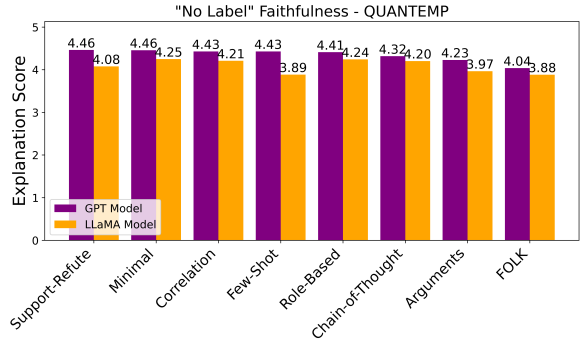

Figure 4: Faithfulness per strategy on QuanTemp dataset.

Faithfulness: HoVer Again, the HoVer dataset introduces a general drop in performance across both models. The Minimal strategy yields the best results for faithfulness with the GPT model, achieving a score of 4.33, while the Chain-of-Thought strategy is best with LLaMA with a score of 4.19. Again, the differences are mostly small. GPT scores above 4.2 for 6/8 strategies, while LLaMA shows similar scores above 4.05 for 4/8 strategies, including the Minimal one, as shown in Figure 5.

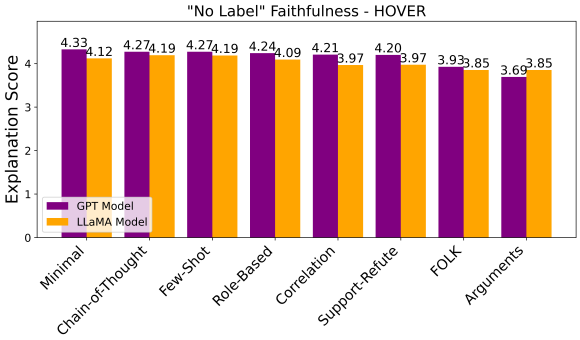

Figure 5: Faithfulness per strategy on HoVer dataset.

5.2 Label Injection Effect

Real Label Given Unsurprisingly, providing either model with the Real Label of the claim resulted in a significant improvement in accuracy compared to the case where no label was provided. This is shown clearly in Figure 6, which plots the improvement in accuracy for each strategy between the two conditions, for both models. Similar results are found for the HoVer dataset, however, when using the LLaMA model, four strategies actually introduced a decline in accuracy. A figure showcasing this spread can be found in Appendix B.

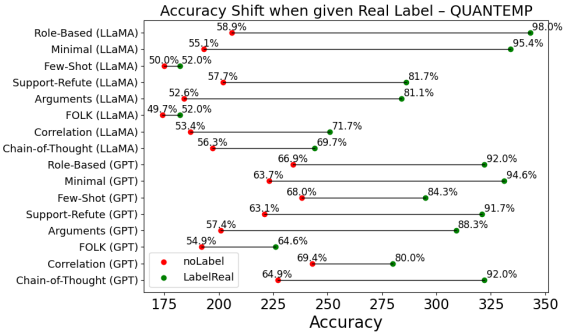

Figure 6: Accuracy improvement per strategy on QuanTemp dataset between the No Label and Real Label conditions.

Faithfulness was, in part, negatively affected across both datasets by the injection of the Real

label. In QuanTemp, 6/16 cases across both models resulted in less faithful explanations regardless of the presence of the real label. A figure showing the specific strategies affected can be found in Appendix B. In the HoVer set, all but 4 cases showed a relative decline in faithfulness, and 2 of the 4 cases remained identical, as can be seen in Figure 7.

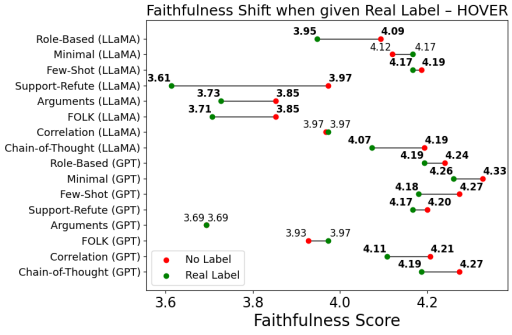

Figure 7: Effect of providing the Real Label on Faithfulness per strategy on HoVer dataset.

Fake Label Given All strategies show relative vulnerability to being provided a fake label, with some notable outliers. Figures 8 and 9 show the 'bias-rate' of each strategy per label condition, i.e. the percentage of matches to the label provided, when that is not the same as the correct label. This aims to investigate the extent to which the model skews its output to match and justify the label provided. The light version of each color corresponds to the GPT output, while the darker version to the LLaMA output. Several observations are made.

Figure 8 shows the effect of the injection of any label on QuanTemp. A 'True' label generally adds little bias, with 8-24% fake label matches in all but one strategy. In contrast, in the 'False' label injection case, the impact becomes striking. For LLaMA, six of the eight prompt styles incorrectly match the injected 'False' label on 38–68% of all claims, while GPT-4o-mini is even more prone, with a 47–73% match-rate. Similar bias is found in the 'Conflicting' injection case, where most strategies sway to match the fake label for 39-58% of claims. Notably, the **FOLK** and **Few-Shot** strategies stand out. They perform remarkably well across all conditions and models. In the 'False' case, they are the only strategies with below 30% bias across both models. In the 'Conflicting' case, FOLK yields at most a 7% bias rate in the GPT model, and 3% in the LLaMA model. The **Few-Shot** technique also remains strong with under 3% bias rate, but only in the LLaMA model, and the **Correlation** strategy performs relatively well across both as well.

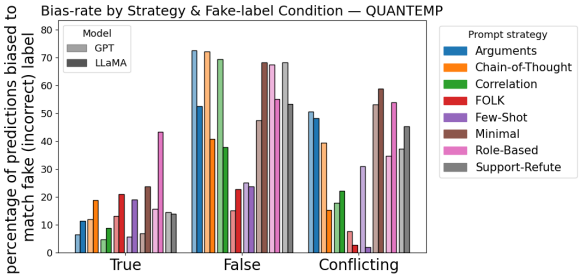

Figure 8: Percentage of matches to fake, incorrect label per strategy and label injection condition, QuanTemp dataset.

In the HoVer dataset, more mixed results are generally observed, as seen in Figure 9. Less susceptibility to label bias is found with the maximum being the Role-Based approach on LLaMA, barely over 40%. Most notably, the **FOLK** strategy appears robust across both models by a large margin once again, when being supplied with a "NOT SUPPORTED" label.

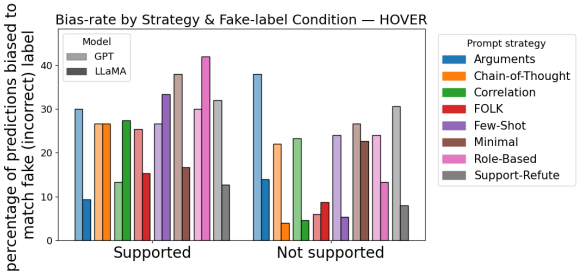

Figure 9: Matches to fake, incorrect label per strategy and label injection condition, HoVer dataset.

Figure 25, overlays faithfulness on the bias-rate of each strategy, for the 'False' label injection condition. In general, even when a strategy shows a very high injection bias of the label, that is, matching the incorrect label in 50 to 70% of the claims (e.g., Minimal, Arguments, Role-Based), its G-Eval score remains high (4.3 to 4.5). Across both models, faithfulness varies by < 0.4 points on the 1–5 scale, underscoring that numerical differences in faithfulness are noticeable, but small. Additionally, relative to the dashed "No-label" baseline, faithfulness shifts by ≤ 0.1 point across most strategies. FOLK and Few-Shot show small improvements in the injection case. The 'True' and 'Conflicting' conditions show generally similar pat-

terns in Faithfulness, and can be found in Appendix C.

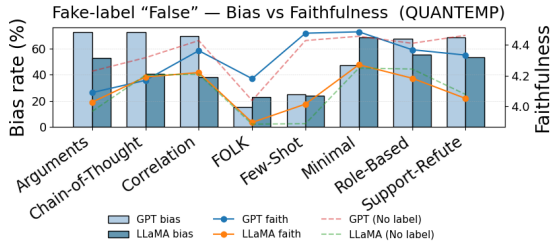

Figure 10: Bias rate and faithfulness in the ’False’ label injection condition, QuanTemp.

In HoVer, the ’SUPPORTED’ label injection introduced more significant bias, as can be also seen in Figure 9. The faithfulness trend is similar to that of QuanTemp, and does not diverge significantly from the No Label baseline, shifting by ≤ 0.1 point as well, as shown in Figure 11. The ’NOT SUPPORTED’ condition can be found in Appendix C.

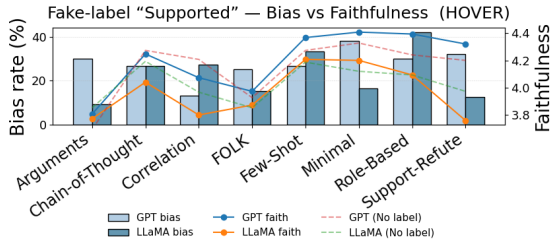

Figure 11: Bias rate and faithfulness in the ’SUPPORTED’ label injection condition, HoVer.

5.3 Claim Type Influence

The QuanTemp dataset labels each claim as statistical, temporal, comparison, or interval (220, 73, 44, 13 items respectively in our 350-claim split). The subset of the HoVer dataset used for this experiment contains 50 each of 2, 3, and 4-hop claims. The *"No Label"* runs using the GPT model are used as baseline to analyze accuracy and faithfulness per claim taxonomy type, or hop number, for both datasets. For those interested in more details, heatmaps of accuracy and faithfulness per taxonomy/hop-count and strategy can be found in Appendix D, for all label injection cases, across both models and datasets .

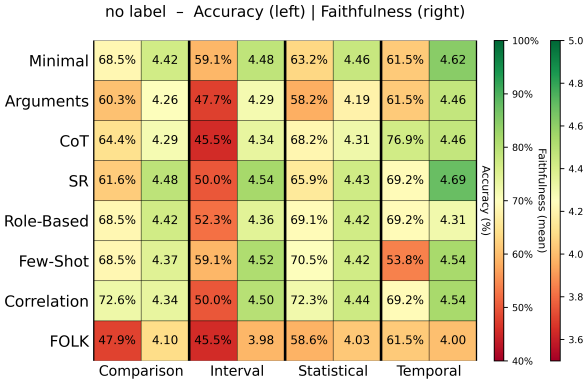

Figure 12: Heatmap of accuracy and faithfulness per taxonomy type, QuanTemp dataset, GPT model.

As shown in the heatmap in Figure 12, faithfulness remains relatively stable across taxonomy types. Temporal claims produce generally more faithful results, but the variation among strategies remains stable: more faithful strategies produce better results across all taxonomy types. In terms of accuracy, Interval claims appear "harder" for all strategies, producing generally less accurate predictions.

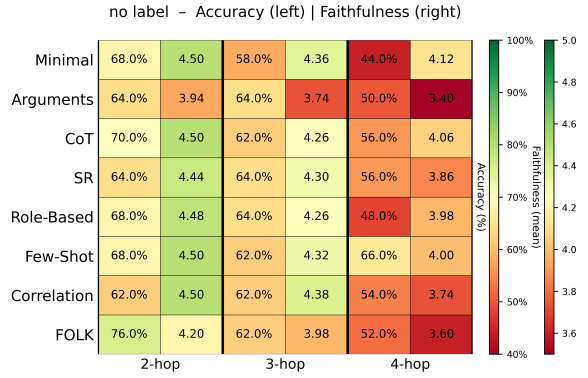

Figure 13: Heatmap of accuracy and faithfulness per taxonomy type, HoVer dataset, GPT model.

Unsurprisingly, higher hop counts give worse results in both accuracy and faithfulness due to the increase in complexity of the claims, regardless of strategy. This can be seen in Figure 13. The **Few-Shot** technique appears to handle 4-hop claims the best out of all strategies, but all appear to struggle more on both fronts as hop-count increases.

6 Discussion

This study aimed to investigate the effects of Prompt Style variation on the accuracy and faithful-

ness of LLM-based fact-checking. Several patterns emerge:

Prompt Strategy effect on Model size: The GPT model showed generally stable outputs on Accuracy across strategies, with variation that didn't exceed 10 percentage points. The smaller LLaMA model appeared more sensitive to prompt style variation. Decompositional-style techniques that force multi-step reasoning such as **FOLK**, **Support–Refute**, or **Arguments** systematically reduced both accuracy and faithfulness on LLaMA, though GPT handled them reasonably well. Adding to this, the LLaMA model often misunderstood instructions, or "got lost" when the strategies were complicated, leading it to misbehave, ignore the desired format, and require manual parsing of its responses.

Accuracy and Faithfulness Trade offs: At first glance, it was found that the strategies that yield the highest accuracy are not the same ones that yield the highest faithfulness and vice versa. However, with small trade-offs, specific strategies can be pinpointed that perform well on both fronts. Specifically, on QuanTemp, the GPT model performs best on accuracy with the **Correlation** and **Few-Shot** strategies, which are only 0.03 percentage points shy of the highest-scoring techniques on faithfulness. Similarly, the LLaMA model performs best on accuracy with the Role-Based approach, which is only 0.01 percentage point shy of the best scoring technique on faithfulness. On HoVer, the **Few-Shot** technique ranks first in accuracy for both models, and faithfulness for the LLaMA model. For the GPT model, it is just 0.06 percentage points off from the best technique for faithfulness. As such, it can be said that **Few-Shot** is generally the most balanced prompt strategy.

Label Injection Bias: The most robust strategy to injection of an incorrect label was **FOLK**, which is interesting given that it scored badly on both Accuracy and Faithfulness in the 'No-Label' case, often worse than every other strategy. Notably, the **Few-Shot** strategy was also nearly as bias-proof on QuanTemp dataset while still delivering high accuracy and faithfulness. It was, however, more inconsistent on the HoVer set.

Faithfulness as measured by G-Eval does not consistently penalise label drift. This means a strategy can be heavily biased toward the injected label yet still earn a high faithfulness score. Furthermore, while the **FOLK** approach was most robust to label injection, its faithfulness dipped across both models, scoring lower than most other strategies. The **Few-Shot** technique, on the other hand, remained both robust and comparatively more faithful, particularly with GPT. For the LLaMA model, the most faithful explanations were generated by the **Minimal** and **Correlation** strategies, but at the cost of larger bias in accuracy.

7 Conclusions & Future work

Overall, it can be concluded that variation in prompt style has an impact on fact-checking accuracy and faithfulness. Across the two datasets, accuracy varies by more than 25 percentage points and faithfulness shifts by roughly 0.4 on the 1-5 scale depending on prompt style. The **Few-Shot** technique appears most balanced across models and datasets in terms of all aspects explored. The **Minimal** technique yielded the highest faithfulness on both models and datasets. Decompositional approaches such as **FOLK**, **Arguments**, and **Support-Refute** negatively impacted the smaller LLaMA model, implying that forcing decomposition on smaller models can over-tax their limited context reasoning. **FOLK** showed remarkable resistance to label injection bias even though it scored low otherwise.

Future Work This study used a very basic Few-Shot approach with no particular rationale. Future work could explore the use of few-shot variants of every strategy, in order to observe the effect on prediction and explanation quality when the model is always given concrete examples. Additionally, replicating the study using larger models, datasets and/or including more strategies would improve generalizability and perhaps unearth more interesting findings.

8 Limitations

Dataset size and taxonomy split: The strategy evaluation was performed on a relatively small number of claims (350 for QuanTemp and 150 for HoVer). In addition, the use of only PolitiFact claims for QuanTemp means the findings may not generalize across different fact-checking platforms, domains, or claim styles. Finally, the taxonomy data in QuanTemp was unevenly split (220, 73, 44, and 13 items for the 4 categories) implying that conclusions on taxonomy may not be reliable.

Model size: The two LLMs used are lightweight models with relatively small numbers

of parameters, particularly the LLaMA model. They were chosen on a basis of practical constraints, namely GPU availability and token-cost. While both models are capable, their reasoning capabilities may vary when compared to larger variants, affecting generalizability.

Choice of prompting strategies: Given the limited time-frame of this study, a complete systematic review of all prompting strategies employed in current research was infeasible, meaning there may be strategies that were overlooked or excluded from this experiment.

Metrics: Faithfulness evaluation is done through G-Eval, which is an LLM-based technique on its own. While this method is efficient, and attempts to simulate human judgement, concerns can be raised about potential bias, hallucination, and accuracy.

9 Responsible Research

This research has been conducted in accordance with the principles of responsible research as outlined in the Netherlands Code of Conduct for Research Integrity (Netherlands Code of Conduct for Research Integrity Committee, 2018) and TU Delft's Vision on Integrity (Committee Reassessment Integrity Policy, 2018). Given increasing concerns about the reliability and societal impact of AI-generated information, particularly in fact-checking contexts, this research takes special care to ensure transparency, reproducibility, and scientific integrity.

Throughout the study, I prioritized honesty both in data handling and result interpretation. Models were evaluated using publicly available datasets, and filtering was explained. The codebase is publicly available and has been documented to facilitate future reproducibility. Outputs have been preserved to ensure verification.

Accuracy and faithfulness, the core research subjects, are inherently subjective. To ensure transparency, I carefully defined what 'accuracy' and 'faithfulness' refers to in the context of this research, documented the rationale behind each metric used, and justified the inclusion of the datasets used based on their complexity and relevance.

A crucial aspect of this research project revolves around prompts. I conducted a focused literature review on common prompting strategies used in fact-checking and explanation tasks. Based on this, I crafted a base prompt template, based on which all others were adapted, in order to ensure consistency. Each version was documented and example prompts are included. This process was fully explained to ensure reproducibility and allow for future reanalysis.

In terms of ethics, I adhered to the principle of responsibility by reflecting critically on the societal implications of LLM use in verification contexts. This includes the potential misuse of unverifiable explanations and excessive reliance on automatic tools. In line with TU Delft's integrity policy, I also aimed to be open about the project's limitations, and invite replication and critique through open sharing of results, codebase, and methodology.

Use of Generative AI Generative AI was used briefly during this project, mainly to aid with polishing the writing process. All ideas and concepts are my own. Prompt examples include "how can I make this sentence more clear/concise?" or "is there a better word to fit this context?". Very rarely during the implementation, I may have asked prompts like "Why is this error being thrown?" when stuck.

A Examples of all Prompt Templates

Minimal Prompt Template:

"Claim: [CLAIM]
Evidence: [EVIDENCE]
(This Claim is considered [LABEL].)
Thoughts?
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation)"

Role-Based Prompt Template:

"You are an expert fact-checker, and are provided with a Claim and corresponding Evidence. (This Claim is considered [LABEL])
Please label the Claim as True, False, or Conflicting based solely on the Evidence, without using other internal or external knowledge or information, and provide an explanation for your decision.
Claim: [CLAIM]
Evidence: [EVIDENCE]
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation)"

Evidence-First (Correlation) Prompt Template in Two Steps:

STEP 1:

"You are given a Claim and corresponding Evidence. (This Claim is considered [LABEL]).
Please judge the Correlation between Claim and Evidence based solely on the Evidence.
Pick from the following options:
a) Evidence definitely supports Claim;
b) Evidence definitely contradicts Claim;
c) Evidence indirectly supports Claim;
d) Evidence indirectly contradicts Claim;
e) Evidence partially supports Claim;
f) Evidence partially contradicts Claim;
g) Evidence has no relation with Claim.
Claim: [CLAIM]
Evidence: [EVIDENCE]
Please just state the option picked with no other details. "

STEP 2:

"You are given a Claim and corresponding Evidence. (This Claim is considered [LABEL]).
Please label the Claim as True, False, or Conflicting based solely on the Evidence and the correlation provided, and provide an explanation for your decision.
Claim: [CLAIM]
Evidence & Claim–Evidence Correlation:
[EVIDENCE]
Correlation: [CORRELATION]
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your reasoning)"

Support-Refute Prompt Template:

"You are given a Claim and corresponding Evidence. (This Claim is considered [LABEL].)
Please first identify which (if any) parts of the Evidence support the Claim, and which (if any) parts of the Evidence refute the Claim.
Then, please label the Claim as True, False, or Conflicting and provide an explanation for your decision.
Base your decisions solely on the Evidence, without using other internal or external knowledge or information.
Claim: [CLAIM]
Evidence: [EVIDENCE]
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation)"

Chain-of-Thought Prompt Template:

"You are given a Claim and corresponding Evidence. This Claim is considered label.
Based solely on the Evidence, without using other internal or external knowledge or information, is this Claim True, False or Conflicting?
Claim: claim
Evidence: evidence
Let's think step by step.
Please format your answer as:
Label: (True, False or Conflicting)
Explanation: (your explanation)"

Arguments Prompt Template:

"You are given a Claim and corresponding Evidence. (This Claim is considered [LABEL].)
Please first present 3 arguments for and 3 against the claim from the evidence, if such arguments exist.
Then, please label the Claim as True, False, or Conflicting and provide an explanation for your decision.
Base your decisions solely on the Evidence, without using other internal or external knowledge or information.
Claim: [CLAIM]
Evidence: [EVIDENCE]
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation, including your arguments)"

Decompositional (FOLK) Prompt Template in Two Steps:

STEP 1:

"You are given a Claim and corresponding Evidence. (This Claim is considered label).
The task is to:
1. Define all the predicates in the claim
2. Parse the predicates into followup questions.
3. Answer the followup questions based solely on the Evidence provided, without using other internal or external knowledge or information.
Claim: claim
Evidence: evidence
Please format your answer as:
Predicates: (list of predicates)
Followup Question: (question following from predicate) Answer: (answer to question) for each predicate"

STEP 2:

"Given a Claim and a Context, please label the Claim as True, False, or Conflicting and provide an explanation for your decision.
Claim: claim
Context: info
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation)

Few-Shot ICL Prompt Template:

"You are given a Claim and corresponding Evidence. (This Claim is considered label.)
Please label the Claim as True, False, or Conflicting based solely on the Evidence, without using other internal or external knowledge or information, and provide an explanation for your decision.
Here are some examples:
Example 1:
Claim: "Says Arizona... (This Claim is considered False.)"
Evidence: ...
Response: "Label: False
Explanation: The evidence presented indicates ... the claim is false."
"Example 2:
Claim: "The non-partisan Congressional Budget Office ... (This Claim is considered Conflicting.)"
"Evidence: ...
Response: "Label: Conflicting
Explanation: The claim that ... is conflicting because it misrepresents the nuance ..."
Example 3:
Claim: "More than 50 percent of immigrants... (This Claim is considered True.)"
Evidence: ...
"Response: "Label: True
Explanation: The evidence provided supports ... claim is considered true."
Claim: [CLAIM]
Evidence: [EVIDENCE]
Please format your answer as:
Label: (True, False, or Conflicting)
Explanation: (your explanation)

B Effect of Real Label injection on Accuracy and Faithfulness compared to No Label condition

When using the LLaMA model, certain strategies actually introduced a decline in accuracy on the HoVer dataset. These four cases can be seen in bold in Figure 15.

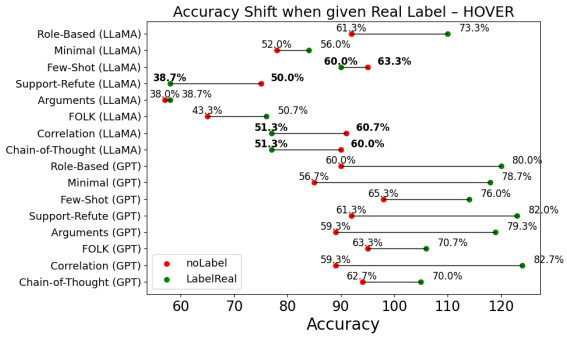

Figure 14: Accuracy improvement per strategy on HoVer dataset between the No Label and Real Label conditions.

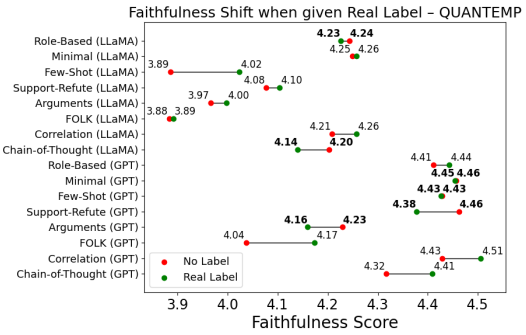

Figure 15: Faithfulness effect per strategy on QuanTemp dataset between the No Label and Real Label conditions.

C Bias rate and Faithfulness in remaining label conditions across both datasets

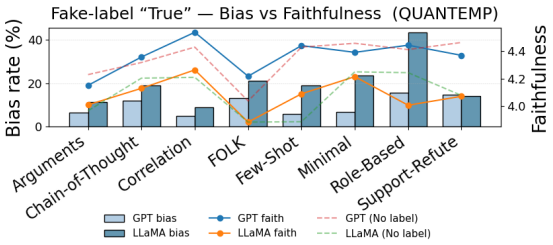

Figure 16: Bias rate and faithfulness in the 'True' label injection condition, QuanTemp.

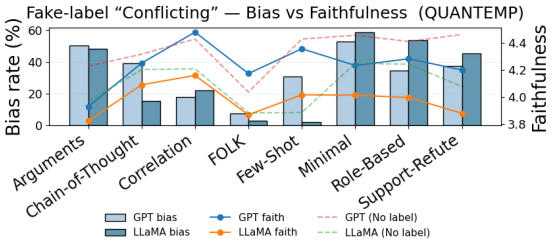

Figure 17: Bias rate and faithfulness in the 'Conflicting' label injection condition, QuanTemp.

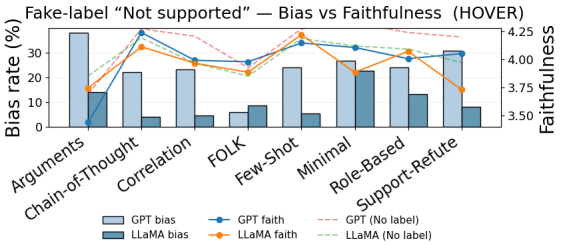

Figure 18: Bias rate and faithfulness in the 'NOT_SUPPORTED' label injection condition, HoVer.

D Heatmaps of Accuracy and Faithfulness by Claim Taxonomy Type/Hop Number

QuanTemp

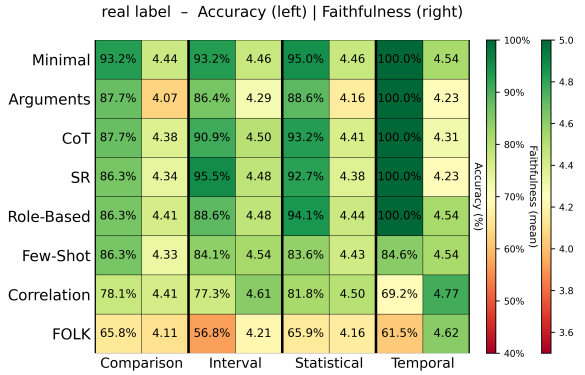

Figure 19: Heatmap of accuracy and faithfulness per taxonomy type, QuanTemp dataset, GPT model in the REAL LABEL case.

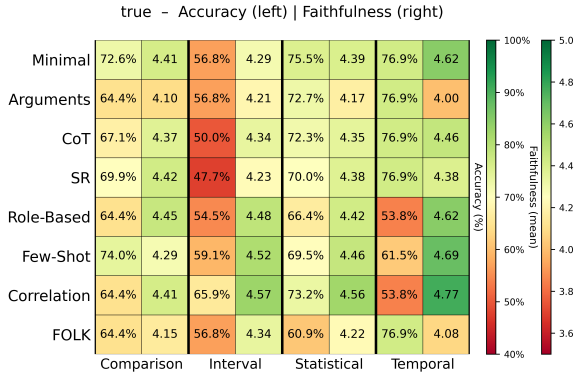

Figure 20: Heatmap of accuracy and faithfulness per taxonomy type, QuanTemp dataset, GPT model in the 'TRUE' LABEL case.

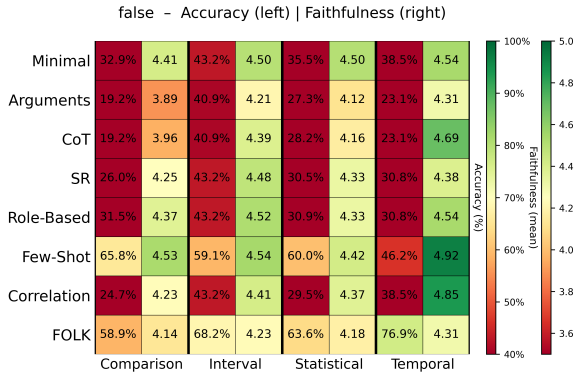

Figure 21: Heatmap of accuracy and faithfulness per taxonomy type, QuanTemp dataset, GPT model in the 'FALSE' LABEL case.

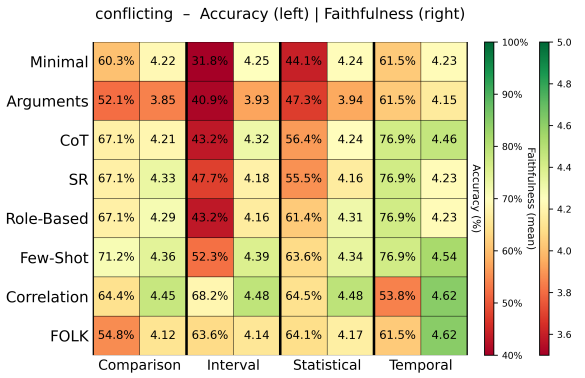

Figure 22: Heatmap of accuracy and faithfulness per taxonomy type, QuanTemp dataset, GPT model in the 'CONFLICTING' LABEL case.

HoVer

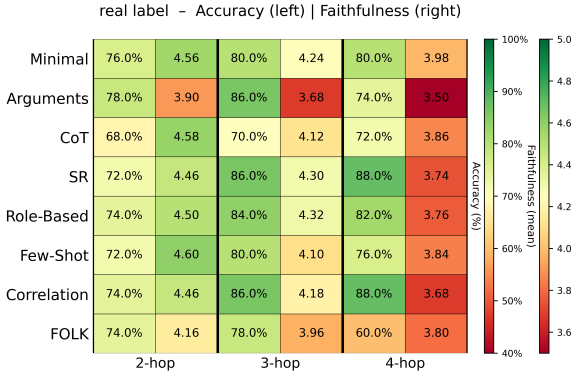

Figure 23: Heatmap of accuracy and faithfulness per taxonomy type, HoVer dataset, GPT model in the REAL LABEL case.

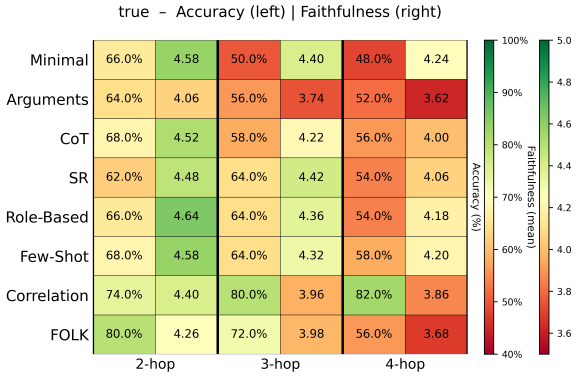

Figure 24: Heatmap of accuracy and faithfulness per taxonomy type, HoVer dataset, GPT model in the 'SUPPORTED' LABEL case.

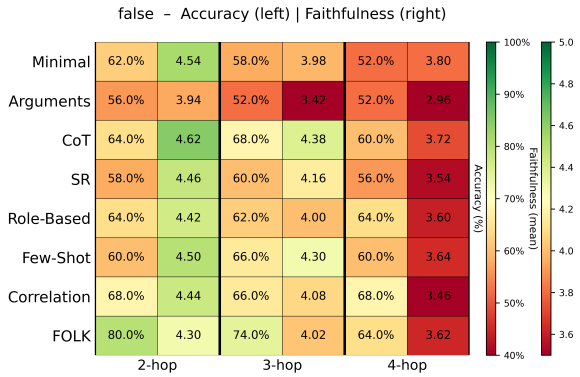

Figure 25: Heatmap of accuracy and faithfulness per taxonomy type, HoVer dataset, GPT model in the 'NOT_SUPPORTED' LABEL case.

References

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. *Preprint*, arXiv:2309.00240.

Committee Reassessment Integrity Policy. 2018. Tu delft vision on integrity 2018–2024. Technical report, Delft University of Technology. September 2018.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. Claim verification in the age of large language models: A survey. *Preprint*, arXiv:2408.14317.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Hans W. A. Hanley and Zakir Durumeric. 2024. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *arXiv preprint*. ArXiv:2305.09820 [cs].

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Nazanin Jafari and James Allan. 2024. Robust claim verification through fact detection. *Preprint*, arXiv:2407.18367.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *Preprint*, arXiv:2402.07401.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Elizaveta Kuznetsova, Ilaria Vitulano, Mykola Makhortykh, Martha Stolze, Tomas Nagy, and Victoria Vziatysheva. 2025. Fact-checking with generative ai: A systematic cross-topic examination of llms capacity to detect veracity of political information. *Preprint*, arXiv:2503.08404.

Yifan Li and ChengXiang Zhai. 2023. An Exploration of Large Language Models for Verification of News Headlines. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 197–206. ISSN: 2375-9259.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Meta. 2025. llama3.1.

Netherlands Code of Conduct for Research Integrity Committee. 2018. Netherlands code of conduct for research integrity. CC-BY 4.0 license; translated version of the "Nederlandse Gedragscode Wetenschappelijke Integriteit".

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

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

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Vinay Setty. 2024. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. *Preprint*, arXiv:2402.12147.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. *Preprint*, arXiv:2310.12558.

Ronit Singhal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using rag and few-shot in-context learning with llms. *Preprint*, arXiv:2408.12060.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving explainable fact-checking with claim-evidence correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics.

The Pulitzer Prizes. 2009. Staff of the *St. Petersburg Times* ("politifact"): 2009 pulitzer prize winner in national reporting. `https://www.pulitzer.org/winners/staff-69`. Accessed 2025-06-02.

V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact checking numerical claims. *Preprint*, arXiv:2403.17169.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. *Preprint*, arXiv:2310.05253.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

YouGov. 2023. Half of americans think artificial intelligence will make the world worse. https://today.yougov.com/technology/articles/46790-most-americans-think-ai-will-make-world-worse. Accessed: 2025-04-22.

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