m.serafeimidi@student.tudelft.nl
TU Delft, EEMCS, June 2025

Author: Marina Serafeimidi
Supervisor:
Shubhalaxmi Mukherjee
Responsible Professor:
Pradeep Murukannaiah

EXPLAINABLE FACT CHECKING WITH LLMs:

How **prompt style variation** affects accuracy and faithfulness in claim justifications.

1. BACKGROUND

The recent rise of LLMs introduces both an inevitable increase in misinformation, as well as an equivalent potential in automated **fact-checking**.

Exploring **prompt variation** in this context aims to discover the most effective prompting technique(s) that yield the most accurate labeling and faithful justification of the veracity (truthfulness) of a claim.

This research attempts to explore how **different prompt styles and techniques** affect LLMs' ability to assess the **veracity** of a claim based solely on provided evidence, and **explain** the reasoning behind that assessment.

2. RESEARCH QUESTIONS

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

1. How does changing the **structure**, **order**, and **phrasing** of prompts impact the predicted labels and the accompanying justifications?
2. How does the presence and correctness of a **supplied label** influence the model's reasoning, justification quality, and susceptibility to label bias?
3. Does the accuracy and faithfulness of LLM justifications change depending on the **type** or **complexity** of the claim?
4. What prompt styles or LLM usage practices **consistently maximize accuracy and faithfulness**, leading to higher factual alignment with evidence?

3. METHODOLOGY

Datasets: HoVer, QuanTemp

Filtered and reduced for our needs.
QuanTemp: real-world, PolitiFact-verified claims with article as evidence.
HoVer: multi-hop claims requiring models to combine information across multiple sources.

Prompt Design

Multiple prompt strategies are evaluated, collected from a review of recent literature. Six techniques are chosen to test, including **Role-Based**, **Few-Shot**, **Chain Of Thought**, two different **Decompositional** techniques (**FOLK** and **Correlation**), and a **Minimal**, unguided technique. Two novel techniques are introduced, namely **Support-Refute** and **Arguments**. Both instruct the extraction of relevant to the claim parts of the evidence before assessment.

Label Injection

For each prompt strategy tested, the effects of the **label** (e.g., True, Not Supported) being provided to the model or not are explored.

LLM Experimentation

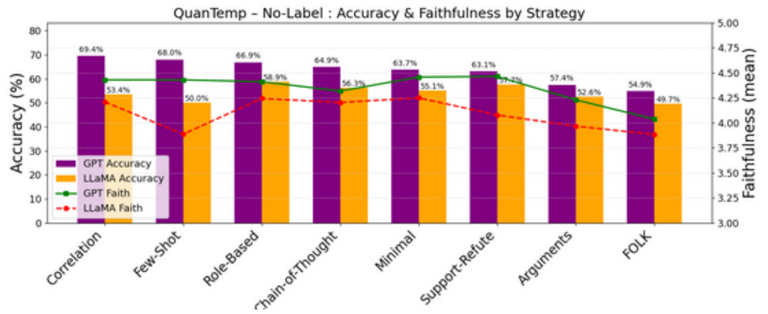
Systematic automated testing on the two datasets, using model **Llama3.1-8B** through the **Ollama** framework, and **GPT-4o-mini** through its API.

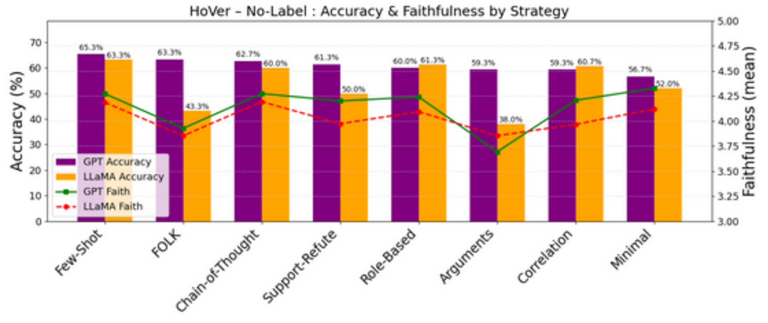
Results Analysis & Evaluation

Measured **Accuracy %** of generated labels, **Label Bias %** between label injection cases, and **Faithfulness** with **G-Eval** on a 1-5 scale.

4. RESULTS & DISCUSSION

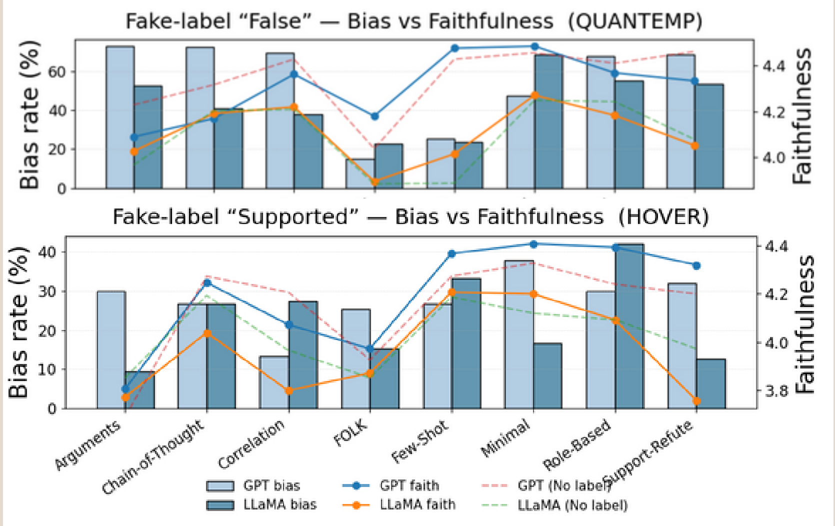
Prompt Strategy Comparison





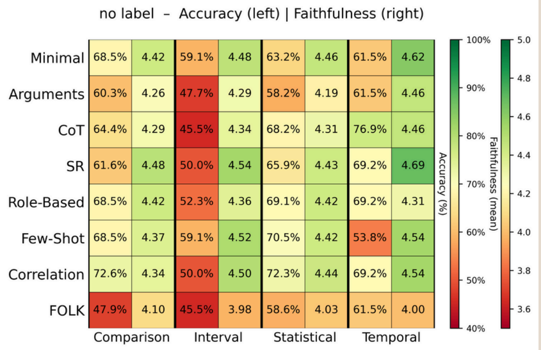
- Generally, GPT showed better, more stable results in accuracy and faithfulness than the smaller LLaMA model.
- LLaMA struggles with more complex instructions and forced decomposition, showing drop in accuracy and faithfulness.
- Faithfulness trends are similar between the two models, accuracy varies more.
- The most accurate strategies are not the most faithful and vice versa: **tradeoffs** required for best results.
- On QuanTemp, GPT is most *accurate* with the **Correlation** technique (69.4%), while LLaMA is most *accurate* with the **Role-Based** technique (58.9%). On HoVer, both models yield best accuracy with the **Few-Shot** technique (65.3% and 63.3%).
- On QuanTemp, both models produce the most *faithful* explanations with the **Minimal** technique (GPT: 4.46, LLaMA: 4.25), while on HoVer, LLaMA is most *faithful* with **Few-Shot** and **Chain-of-Thought** (4.19) and GPT again with **Minimal** (4.33).
- **As such, generally, the Few-Shot technique appears most balanced on both accuracy and faithfulness, across both datasets and models.**

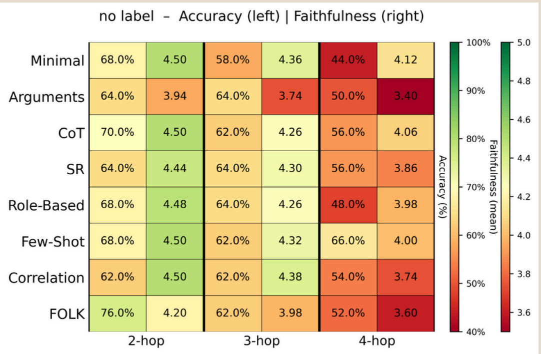
Label Injection Effect



- Injection of the Real Label generally improved accuracy significantly (+1-40%), but often reduced faithfulness.
- The **FOLK** strategy was the most robust against injection of a fake label (3-23% bias), but generally performed poorly on both accuracy and faithfulness when no label was provided .
- **Few-Shot** was similarly robust (3-31%), but generally more accurate and faithful in all cases.
- Faithfulness does not consistently penalise bias: i.e. biased strategies still achieve high faithfulness scores, e.g. **Minimal**, **Role-Based**.
- HoVer shows more inconsistent results, Few-Shot underperforms in the 'Supported' case (27-33% bias). 'Not Supported' shows similar trends with QuanTemp.
- **Again, Few-Shot appears the most balanced, being both unbiased and faithful under Label Injection.**

Claim Type/Complexity Analysis





- On QuanTemp, claim **taxonony type** has little effect on faithfulness. **Interval** type claims show a general decline in Accuracy across all strategies.
- On HoVer, a higher **hop-number**, i.e. a more complex claim, leads to worse performance on both accuracy and faithfulness across all strategies. **Few-Shot handles complexity best.**

5. CONCLUSIONS & FUTURE WORK

- Prompt style **does** have an effect on accuracy and faithfulness in claim justifications, but varying trends appear acoss models and datasets.
- The **Few-Shot** approach yields the most balanced results on all aspects explored in the research. Given that only a bare-bones, rationale-free Few-Shot approach was tested, a recommendaion for future work is to compare few-shot variants of all and more strategies by appending concrete examples.
- **FOLK** showed remarkable resistance to label injection bias despite poor performance on the No-Label case.
- Highest faithfulness was generally achieved using a **Minimal** strategy, indicating that complicated instructions negatively impact explanation quality despite raising accuracy.
- Higher complexity yields worse results, and **Interval** claims are 'harder' across strategies.