Explainable Fact-Checking with LLMs

How do different LLMs compare in their rationales?

Matei Bordea[1]

Supervisor(s): Pradeep Murukannaiah, Shubhalaxmi Mukherjee

[1]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: Matei Bordea
Final project course: CSE3000 Research Project
Thesis committee: Pradeep Murukannaiah, Shubhalaxmi Mukherjee, Xucong Zhang

Abstract

Large Language Models (LLMs) are becoming more commonplace in today's society. However their adoption rate, especially in the fact checking field, is being slowed down by the distrust in their thinking process and the rationales leading to the results. In crucial moments the justifications behind a verdict are more important than the verdict itself. However, LLMs often produce explanations that are not grounded in the provided evidence, leading to hallucinations and reduced trust in their outputs. This paper aims to show exactly the level the LLMs have reached in both the faithfulness of their explanations, based on some provided facts, and the correctness of their explanations. To investigate this, multiple LLMs are asked to assign a label to a claim based on some evidence provided from two datasets of varying complexity: HoVer and QuanTemp. The outputs are then evaluated both manually and by another LLM to evaluate how well the LLM relates to the evidence and if the LLM hallucinates in some parts of its responses. The results reveal that while some models demonstrate high correctness in label assignment, faithfulness in explanations varies significantly across models and evidence types. The outcomes of this experiment aim to inform both LLM developers and fact-checking researchers about the current limitations of LLMs in response quality while also showing which areas require further improvements to become mainstream.

1 Introduction

LLMs have recently demonstrated good results in a lot of natural language processing tasks, including summarization, translation, question answering and fact-checking. In most high-stakes contexts, it is not enough for an AI system to simply label a claim as true or false. Users expect a clear and faithful explanation of why a claim is accurate or inaccurate. This explanation should be based on evidence or at least demonstrate a competent thought process or a suitable rationale behind the assigned label. Without this explanation or assurance, fact-checking risks being perceived as untrustworthy making people stay away from it.

Although there have been improvements in accuracy, LLMs often generate explanations that are not grounded in the input evidence or add information that is false and state it as a fact—an issue also known as hallucination. These hallucinations can involve fabricated facts, omissions, or contradictions, destroying the very trust that explanatory fact-checking is meant to build and needs to be adopted. In particular, some LLMs will produce convincing explanations that are unsupported or even contradictory to the provided evidence, creating baseless reasoning that undermines the reliability of fact-checking outputs. As pointed out by different research studies (Feher et al., 2025; Adlakha et al., 2024), even when models are trained to generate natural-language rationales, evaluating the quality and consistency of those explanations remains challenging. Additionally, another paper aims to show that reasoning with quantitative claims across multiple documents introduces further complexity, highlighting how faithfulness is deeply tied to task and evidence type (Venktesh et al., 2024) .

1.1 Related works

Prior work has proposed new models, datasets, and evaluation metrics, but less attention has been paid to the comparison between the rationales generated by different LLMs under shared conditions. For example, Lanham et al. (2023) show that even chain-of-thought rationales can be unfaithful to the model's actual reasoning process. Similarly, other works have shown that while reinforcement learning can improve factual consistency, it may reduce informativeness or lead to more extractive summaries. These trade-offs depend on various factors such as regularization strength and sampling strategy, and require careful balancing to work as needed (Roit et al., 2023). Evaluation methods like G-EVAL (Liu et al., 2023) provide human-aligned assessments, but their application to comparative analysis of rationales across LLMs is still limited. Furthermore, benchmarking efforts such as Benchmarking the Generation of Fact-Checking Explanations (Russo et al., 2023) often focus on single-model settings or assume outside templates, without investigating how different models behave under identical conditions.

Recent research has shown that LLMs can serve as strong evaluators across various tasks. For example, Kocmi and Federmann (Kocmi and Federmann, 2023) showed that LLMs such as GPT-3.5 can get better results, when evaluating translations, than traditional metrics like BLEU. Similarly, Zheng et al. (Zheng et al., 2023) showed that LLMs can compare chatbot responses using the given instructions, which shows that we should be able to use LLMs as capable evaluators as well.

Research Questions To address these gaps in knowledge, our research focuses on how different LLMs behave when tasked with generating fact-checking explanations based on the same claims, evidence under different prompting styles. Specifically, we aim to answer the following questions:

- **RQ1**: To what extent do different LLMs maintain factual consistency between the provided evidence and their generated explanations?
- **RQ2**: How do different LLMs treat different types of evidence?
- **RQ3**: Can automatic evaluation correlate with human judgment of faithfulness for LLM explanations?
- **RQ4**: Are there systematic patterns in the hallucinations or inconsistencies produced by different LLMs?

Main Contributions. In this paper, we evaluate how well LLMs generate explanations for fact-checking tasks that are faithful (only use information found in the evidence), correct (factually accurate), and evidence-grounded (supported by source material). Using tools like OLLAMA and LangChain, and testing on datasets of varying complexity such as QuanTemp and HoVer, we analyze how models perform under shared conditions and assess the quality of their rationales through both automatic (using another LLM as an evaluator) and manual evaluation. The findings provide insights into which models are better in generating explanations, how evidence influences hallucinations, and whether LLMs can be used as evaluators in these types of tasks. We make our dataset, code and most results available here https://github.com/MateeiB/ResearchProject .

Structure of the Paper. Section 2 outlines the methodology used. Section 3 introduces our contributions in more depth. Section 4 describes the experimental setup. Section 5 presents the results of our research, including an objective analysis. Section 6 offers a discussion of the implications and the conclusions drawn from our research, and Section 7 concludes with key takeaways and directions for future work, while Section 8 presents the current limitations of our research.

2 Methodology

This research seeks to analyze how different models behave in controlled fact-checking settings and whether their rationales align with evidence across tasks and datasets.

To investigate this, our study is split into two parts. The first part is an experimental approach to compare the explanations generated by different LLMs under different conditions. We used two datasets of varying complexity: HoVer, which requires multi-hop textual reasoning across multiple documents, and QuanTemp, which focuses on challenging numerical and temporal claims. These datasets were chosen to test the models' ability to handle both complex reasoning and precise fact verification across different types of content. Four models were evaluated (LLaMA2, Mistral, Gemma, Phi), which were selected for their diversity and availability.

Model outputs were analyzed using both automatic and manual methods. The automatic method used was asking the LLM to analyze the responses and report back which response was the best and why. Manual evaluation involved assessing whether each explanation contained hallucinated information, was consistent with the evidence, and if the LLM-generated label is correct, in the applicable parts. All the analysis was logged and then conclusions were drawn about the different LLM capabilities. The automatic evaluation also used chain-of-thought prompting to encourage more structured and reasoned assessments from the evaluating LLMs.

3 Faithfulness Analysis of LLM Explanations

This research adds to the ongoing research on evaluating and understanding how LLMs justify their labels through natural-language explanations. Unlike prior studies that focus solely on label accuracy or use simple overlap-based metrics, this paper focuses on explanation faithfulness meaning the extent to which generated answers align with the given evidence.

The core of this research is an experimental analysis of four open-source LLMs: LLaMA2, Mistral, Gemma, and Phi, which were selected to test across different architectures and training strategies while remaining accessible enough for controlled evaluation on a local machine. To ensure that any difference in answers is caused by the models under test and not by any parameter changes, all models were tested using the same tools and under the same conditions which will be further presented in Section

4.

Since there are a lot of different tasks that LLMs are expected to complete with different requirements multiple types of claims were considered in order for the results to show a better picture. The two main types of claims are complex natural language claims and numerical claims such as statistical or temporal claims that require more complex reasoning. This approach shows what the weak points of the LLMs are, which will allow further experiments to not run into unexpected results and behaviors from the LLMs.

The evaluation follows a two stage approach. First, an automatic evaluation is done by each of the LLMs to compare all four justifications following the chain of thought prompting framework. Chain-of-thought is a technique where the model is asked to reason in steps before giving a final answer. This step-by-step reasoning improves accuracy, especially in this case where the task of analyzing justifications is very complex (Liu et al., 2023) For the more challenging QuanTemp dataset, a full 4×4 evaluation is conducted twice — once for each prompting style — where each model evaluates the justifications generated by all four models. For the simpler HoVer dataset, a more lightweight 2×2 setup is used, where two of the LLMs are used for one prompting style, while the other 2 justify and then they swap roles. Then a manual check is done to validate the results of the first step as well as find any patterns in the answers of the LLM, any hallucinations or any places where the LLM might have missed something.

The main point of the research is a systemic comparison of how models vary in reasoning capabilities across different model sizes and types of data. This includes whether hallucinations are consistent, what kinds of evidence are ignored or overused, and analyzing how models balance factual accuracy with alignment to user expectations. The results showed that some models will add information when the evidence is incomplete or will commit to a label because of the phrasing even if that is wrong. By analyzing all the outputs some strengths or some blind spots can be seen. These will be used in future work as a starting point to show if LLMs can be trusted with the justification behind the label or if more training or development needs to be done before reaching that point.

4 Experimental Setup

An overview of the experiment pipeline is shown in Figure 1. We start by filtering the claims and evidence in a human readable format. Then we choose a prompting style and the LLM that is tested and ask it to generate an explanation. Then we aggregate all results and perform both a manual evaluation and an automatic one to formally obtain the results.

4.1 Datasets

Two publicly available datasets were used to evaluate model performance under varying levels of complexity and different types. The HoVer dataset (Jiang et al., 2020) consists of claims supported or refuted by multiple Wikipedia sentences depending on the number of hops, requiring basic multi-hop reasoning. In contrast, the QuanTemp dataset (Venktesh et al., 2024) has numerical claims grounded in real-world articles and demands more complex reasoning. These datasets were chosen to evaluate how LLMs handle different types of evidence and reasoning styles. The exact distribution of the combined datasets can be seen in Table 1, with the Statistical, Comparison, Interval and Temporal claim types coming from the QuanTemp dataset and the Supported and Not Supported coming from the HoVer dataset.

Claim Type	Count
Statistical	220
Comparison	73
Interval	44
Temporal	13
Supported	82
Not Supported	68

Table 1: Distribution of original labels across all used data.

4.2 Models

The following four open-source language models were evaluated: LLaMA2, Gemma, Mistral, and Phi. These models vary in architecture, size, training objectives, and intended use cases, providing a diverse sample of LLM behavior under a common experimental setup.

LLaMA2 is designed to have a good performance across a wide range of tasks. It is trained on publicly available datasets and optimized for general-purpose reasoning. It is known for its

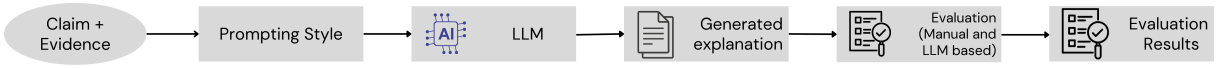

Figure 1: Experimental setup

strong few-shot capabilities and balanced performance on reasoning and language understanding tasks. In this study the 7B parameters version was used with 4-bit quantization to make it possible to run locally.

Gemma is designed to run efficiently on everyday computers. It is designed with alignment and safety in mind, focusing on producing helpful and most importantly harmless outputs. In this study the 8.5B parameters version was used with the same 4-bit quantization.

Mistral is developed with a focus on factual accuracy and robustness. It is trained on selected data highlighting truthfulness and has shown very good performance in tasks like fact-checking or reducing hallucinations, making it particularly relevant for this study and expected to have the best performance in this experiment. In this study the 7B parameters version was used with the 4-bit quantization.

Phi is the smallest model in this comparison, more compact than the others having only 3B parameters. Despite its small size, it is optimized for reasoning and educational use cases. Phi is trained using a "textbook-quality" data strategy, focusing on clean and structured examples to extract maximum reasoning ability from fewer parameters as mentioned by (Li et al., 2023).

This diversity in architecture size and training style makes them a diverse comparative set for evaluating correctness and faithfulness in justification generation.

4.3 Tools and Environment

For all the experiments LangChain was used as an interface compatible with all four models. All the experiments were done on a local machine with 32GB RAM and 6GB VRAM. The coding was done in PyCharm using Python 3.12.3 and the LangChain version used was 0.3.25. Prompts were constructed in three different styles: normal zero-shot prompting, providing the LLM with the label and asking only for the justification, and role-based prompting, as inspired by prior work on structured and context-aware prompting strategies (Sathe et al., 2023; Zhou et al., 2022). In all cases, models were used in the mode they came with no fine-tuning or retraining applied.

4.4 Input Formatting

For each claim-evidence pair, the ground truth was extracted and then a natural-language prompt was used to query the LLM. Both prompting styles were tested across all four LLMs with very little to none fine-tuning to form a comprehensible image on how each LLM thinks and also how each LLM evaluates other LLMs. The exact prompt templates used for justification generation and comparative evaluation are documented in Appendix A.

4.5 Evaluation Methods.

Model outputs were analyzed using both automatic and manual methods. The automatic method used was asking the LLM assigned as evaluator, to analyze and report back which response was the best and why. Manual evaluation performed by us involved assessing whether each explanation contained hallucinated information, was consistent with the evidence and if the LLM generated label and the original label of the claim matched, where it was applicable. All the analysis was logged and then conclusions were drawn about the different LLM capabilities.

5 Results

We analyze the correctness of the labels assigned to claims, the faithfulness of the explanations generated by each model, and patterns in the types of hallucinations observed. We also analyze how each of the LLMs performed as an evaluator for the others.

5.1 Correctness

Figure 2 shows how accurately each model classified different types of claims from the Quantemp dataset based on the provided evidence. Each bar represents the accuracy of the model. This breakdown helps reveal not just overall performance, but also which types of truth values are more difficult for each LLM. The same type of figure for the Hover dataset can be seen in Figure 3 where claims are split into only two categories supported and not supported.

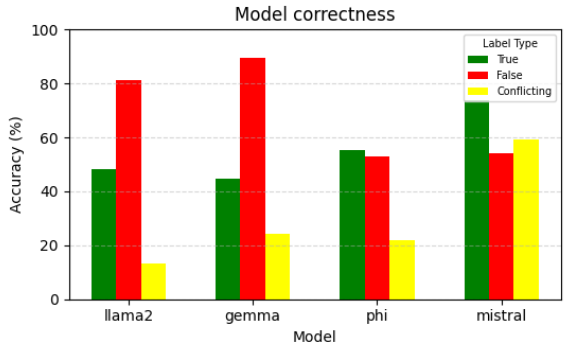

Figure 2: Correctness of labels generated by tested LLMs across True, False and Conflicting types of claims.

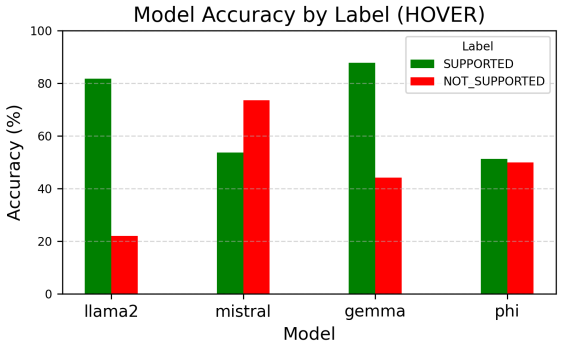

Figure 3: Correctness of labels generated by tested LLMs across both SUPPORTED and NOT SUPPORTED types of claims.

We observe that Mistral demonstrates strong and relatively balanced performance across all claim types, correctly classifying around 60-70% of all claims across both the datasets and showing balanced results across all types, unlike the other LLMs. In contrast, Gemma and LLaMA2 show a steep performance drop when handling false and conflicting or supported and not supported claims: Gemma has high accuracy on false claims but drops below 30% accuracy on conflicting ones, with the same pattern for not supported claims dropping its accuracy to about half of the supported claims. LLaMA2 follows a similar pattern, with weak performance on not supported and conflicting claims. This may be explained by a tendency of the models to default to assertive labels even when evidence or resulted label is ambiguous. Lastly Phi shows a more balanced result for both datasets, even though the performance is lower in total, around 40-50% average correctness. This result is due to the varying complexity of the claims as Phi shows better performance on simpler claims than on more complex ones.

5.2 Faithfulness

The second part of the evaluation process focuses on how well the generated justification aligns with the provided evidence. To get better results, models were tested using two main prompting styles: with and without the label provided in the prompt. Evaluations covered multiple categories of claims, including statistical, comparison, temporal, interval claims and the supported/ not supported claims.

The scores given to each of the models can be seen in Figure 4 where the models are split into two runs: the first one with the label hidden and the second one with the label given. Figure 5 shows the scores for the HoVer datasets for each of the LLMs.

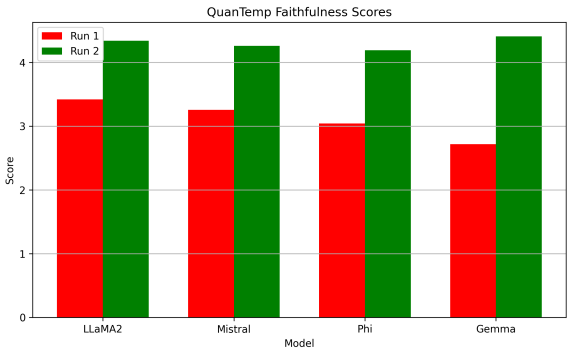

Figure 4: Faithfulness scores assigned to each of the models for the QuanTemp dataset

In the QuanTemp dataset, Mistral and LLaMA2 produced the most grounded and structured justifications for all claim types. Their responses show careful citation of figures, time ranges, and comparative elements, particularly in the statistical and interval categories. When the label was not provided, these two models had strong independent reasoning capabilities, while Gemma and Phi often struggled more with either vagueness or oversimplification, with cases where the claim was just parroted in the answer with little to no explanations added. Phi often generated concise and accurate explanations in simpler contexts but was less reliable in complex claims. We have also seen multiple places where it skipped parts of the evidence or the claim, most often in temporal or interval claims. Gemma had the tendency to add a lot of abstract or general information not grounded in the evidence, leading to a decrease in faithfulness, most often in

interval-based reasoning. When the label was included, all models improved drastically, LLaMA2 and Mistral kept a high performance, particularly in comparison and statistical justifications. Notably, Gemma obtained the highest score in the second run showing the closest similarity to the provided evidence. We believe this score was achieved due to Gemma's tendency to include a lot of the evidence in the answer and generate longer answers.

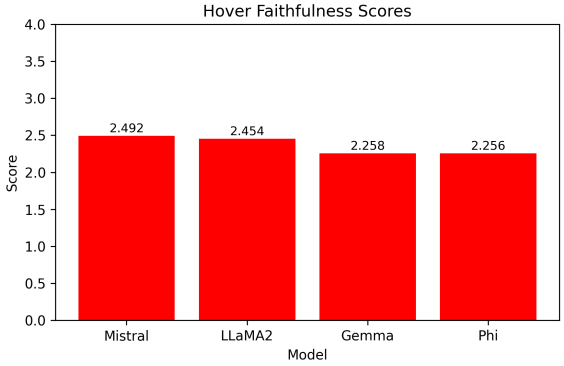

Figure 5: Faithfulness scores assigned to each of the models for the HoVer dataset

In the HoVer dataset, similar results were observed. Phi often had its consistency and brevity in simpler cases, but its justifications became too general as complexity increased. However, in some cases the claim was too complex for Phi. In those cases the LLM apologized for not being able to give a response and justified it by saying that it did not understand the task. We tried across multiple runs and while sometimes Phi would try to give a general response, the most complex claims remained mostly unanswered. LLaMA2 and Mistral again had the best alignment with the evidence, though in this case LLaMA2 sometimes oversimplified and Mistral occasionally introduced extra context not from the evidence which made their scores drop in comparison to what they got in the QuanTemp runs. Gemma continued to produce complex justifications but had a tendency to drift from the provided facts, affecting its faithfulness score. It can be seen that the average scores are smaller in general and we believe that this is because the evidence in the HoVer dataset is shorter. This caused the evaluator LLM to deduct points from the justifications simply because too many words were used.

Across both datasets, Mistral and LLaMA2 consistently produced justifications that demonstrated the highest alignment with the evidence, particularly when claim complexity increased. Phi achieved moderate overall performance, showing reliable behavior on simpler claims where it produced concise and coherent justifications. However, its performance declined on more complex examples, where explanations were either incomplete or absent. Gemma showed a pattern of producing longer, more elaborate justifications, but frequently introduced abstract or unrelated content, which negatively impacted its faithfulness scores. These trends remained consistent across both prompting styles and suggest that larger or more complex models may benefit from improved reasoning guidance when evidence is more compressed, as for the HoVer dataset.

5.3 Evaluation Capabilities

In addition to generating justifications, the models were also used as evaluators to test if LLMs can be used as automatic evaluators in the future.

To assess each model's capability as an evaluator, we analyzed their outputs based on four consistent dimensions: explanation structure, evidence sensitivity, bias (including self-bias) and evaluation clarity for a human reader. Gemma consistently favored structured and well-cited responses, with a preference for clarity, complexity and step-by-step reasoning. It showed minimal self-bias, occasionally critiquing its own outputs for vagueness or simplification. However, it often under-analyzed small logical changes or nuanced phrasing, limiting its effectiveness on more complex justifications. LLaMA2 prioritized evidence sourcing and faithfulness. Its evaluations were often verbose and thorough, but it displayed a slight bias toward its own outputs. From the experiments we say that Mistral was often reluctant to offer strong praise and often applying detailed searches to all models. It emphasized sourcing and citations. In the same way as Gemma, Mistral did not consistently favor its own justifications. Notably, Mistral's evaluations frequently aligned with LLaMA2's style, and it criticized Phi and Gemma more for going off-topic. Lastly, Phi focused heavily on surface-level fidelity. It penalized justifications that added extra information or made assumptions beyond the provided evidence, even when such reasoning was accurate. Phi's strict criticism and size limited its ability to reward nuanced or inferential responses, but it remained fair in self-evaluation.

5.4 Hallucinations

Across all models and all datasets, hallucinations were significantly more common in the first run where the label was not provided in the prompt.

We have observed some model specific hallucinations that show the strengths and weaknesses of each LLM. Gemma occasionally added unsupported claims, often citing some sources that were not mentioned in the evidence. This was mostly seen in statistical or four hop political contexts. Phi had almost no hallucinations, it only struggled a bit with the interval and temporal claims by making some assertions that were not present in the evidence but were logical. LLaMA2 also hallucinated infrequently, but when it did, the mistakes came from adding extra context or making assumptions about parts of the claim that were not actually mentioned in the evidence. Finally, Mistral was the least prone to hallucination, its errors were mostly things it missed, not added context or sources.

To conclude, certain claim types—such as interval and statistical claims—were more prone to triggering hallucinations across models.

6 Discussion

The results highlight clear differences in how LLMs perform when tasked with generating, justifying, and evaluating claims. While all models demonstrated basic competence, their performance was mostly dependent on the complexity of the claim as well as the dimensions of evaluation: faithfulness or correctness.

When analyzing the performance by claim type, clear patterns emerged across the four categories present in the QuanTemp dataset: Statistical, Comparison, Interval, and Temporal. The most challenging type of claim for the models overall was Interval-based. These required interpreting conditions or specific time spans. From these claims it can be derived that LLMs tend to struggle with interval logic with most of them considering even the smallest overlap between the claim and the evidence to make the claim true. Temporal and Comparison claims were handled well by the models. There were no big problems with locating events in time. However, while the label accuracy was present for the Comparison claims, models often failed to emphasize differences clearly. Statistical claims made up the biggest part of the dataset and proved to be the most accessible for faithfulness as models demonstrated strong reasoning tied to numeric evidence. The HoVer dataset had its own complexity scale-the number of hops. Claims that contained modifiers such as dates or qualifiers such as “in the summer of”, proved more difficult than the rest as models often ignored or glossed over these specifics. Also, claims that required the aggregation of two distinct parts were more error-prone especially when multiple sources had to be aggregated and compared.

In terms of label correctness, most models performed better on false claims than on true or conflicting ones, with conflicting ones having the lowest performance. This was especially surprising as past works have consistently stated that LLMs have a bias towards the "true" value. We believe that this is largely because the LLMs tend to overthink and sometimes in the evidence the whole context is not given so the LLM decides that if a statement is not 100% true it is false. Models that were trained specifically for analytical tasks, like Mistral, showed more balanced outputs, suggesting they are better at handling uncertainty and do not treat claims as strictly true or false, but rather assess them based on parts of the evidence. In the Hover dataset LLMs were a bit more balanced but still showed better overall performance on the supported claims.

For explanation faithfulness, model performance was strongly influenced by prompt design and given information. When the label was provided in the second run, faithfulness increased across the board, meaning that models right now tend to agree with the human or prompter rather than try to tell that they think the label is wrong. Even when prompted with the opposite label the LLM tried to justify it showing that results of LLM queries can be easily manipulated. Another interesting finding when generating explanations was that LLMs simply ignored or did not realize when very small changes were made to the claims. As an example, take the two-hop claim from the HoVer dataset: "The MV Bessel ran aground at the second largest island in the Mediterranean in 1972.", when adding the specific season, in this case summer the LLMs still considered the claim to be supported even though the evidence did not mention anything about the season. We believe this is because of the training strategy used or the number of parameters but the exact reason still needs to be studied.

Evaluation with LLMs had in hindsight good performance. There were however differences in style with some models such as Phi favoring direct-

ness and factual alignment and models like Gemma rating fluency and length basically, way higher than other models. Self-bias, where applicable, is not a big problem as LLMs showed that they did not lose track of the task regardless of the name before the justifications. To end, LLMs can be used as evaluators today, however choosing the right LLM or even the right training method will have a big impact on the results.

The hallucination analysis further supports this difference in style and learning strategy. Models that were trained toward fluency or inferential reasoning such as Gemma and sometimes LLaMA2 included more unsupported information, while more straightforward models hallucinated less but found it more difficult to draw conclusions based on implicit context.

7 Conclusions and Future Work

Our research goal was to investigate how large language models (LLMs) differ in their ability to generate faithful and accurate explanations given different types of claims. We examined four models, Gemma, LLaMA2, Mistral, and Phi, across two datasets of varying complexity, evaluating if they can give the correct label while also checking their thought process to check for patterns or hallucinations.

The findings reveal that while LLMs can predict the correct label around 50-60% of the time on a dataset that they see for the first time. Prompting style affects performance as giving the label in advance improved the faithfulness of justifications. LLMs tend to keep the general facts in their generated answers but struggle with small details, even if those details are important.

Another key conclusion is that LLMs can act as reasonable evaluators in controlled settings, especially if chain-of-thought prompting is used. They showed minimal biases and a strong alignment with the ground truth or the human conclusions. Different AIs have different styles so which training style is best for evaluating in general is something that needs to be researched in the future.

To answer the fourth research question, systematic patterns were observed in the hallucinations produced, for example Gemma's were often inferential with the model drifting off at times, Phi's were rare very small and LLaMA2's hallucinations involved adding correct but unsupported information. This supports the idea that each model has its own consistent template for answering and that each model tries to stick to it.

Future work should explore extending this evaluation to larger models as we expect more parameters will lead to even better results. Future research can also focus more on the training strategy, figuring out which gives the best results either for evaluating or for justifying claims. Lastly, a known problem observed here that needs to be thoroughly researched and fixed is that LLMs are too agreeable and can easily be manipulated.

8 Limitations

Our research also faces several limitations. First, the analysis was limited to four openly available models selected for their accessibility and local deployability. We chose these models using the 8.5B parameters version at most as it was not computationally feasible to choose larger models. This makes the experiment easy to validate but using larger models should give better results.

Second, manual evaluation was conducted on a limited scale, and while it provides important context, it cannot fully substitute for broader human annotation. Lastly, the datasets: QuanTemp and HoVer, focus on specific types of claims, which may not generalize to other domains such as medical or legal fact verification. Also, because of the filtering process the data distribution was uneven, with some claim types such as temporal being fewer. This will not invalidate our overall results but for the specific type it might not give conclusive results either.

Despite these limitations, the findings provide valuable comparative insights and highlight consistent patterns that can inform further researchers.

A Appendix

A.1 Justification Prompts

1. **Label and Justification Prompt (No Label Given to LLM):**

```
You are given a factual claim and an
    evidence passage. Based solely on the
    evidence, determine whether the claim
    is SUPPORTED or NOT_SUPPORTED by the
    evidence.

Your output must include:
1. Justification: Explain your reasoning
    based only on the evidence.
2. Label: One of [SUPPORTED, NOT_SUPPORTED]

Claim: "{claim}"

Evidence:
"""{evidence}"""

Answer:
```

2. **Justification-Only Prompt (Label Provided):**

```
Given the following claim, its correct
    label, and the supporting article text
    (evidence), generate a justification
    that explains why the label is
    appropriate.

Claim: "{claim}"

Label: {correct_label}

Evidence:
"""{doc}"""

Your task is to write a justification for
    this label, based only on the evidence
    provided.
```

A.2 Evaluation Prompts

1. **Comparative Justification Evaluation Prompt:**

```
You are a fact-checking assistant tasked
    with comparing explanations from
    multiple language models for the same
    claim and evidence. You are required to
     think in steps.

Claim: "{claim}"
Claim Type: {taxonomy_label}

Justification from LLaMA2:
{llama2}

Justification from Gemma:
{gemma}

Justification from Mistral:
{mistral}

Justification from Phi:
{phi}

Write a short comparative analysis of the
    justifications above, explaining which
    model(s) provided the most convincing
    and faithful explanation, and why.
```

B Responsible Research

In this section all the ethical and responsible research concerns of this project are being addressed, as guided by the ACL checklist for all the submissions in this format (Review, 2023) and the integrity principles outlined in the Responsible Research lecture from TU Delft.

Transparency and Data Integrity: This project uses public datasets namely HoVer and QuanTemp which are properly cited and documented. All these datasets are impersonal and designed to be used for future academic research. The source of this data as well as the structure are acknowledged in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) data act. In this research there is no misrepresentation of the results to avoid subjective results.

Reproducibility and Replicability: Reproducibility is a major concern in current LLM research. While it is hard for LLM pipelines to be fully deterministic, all experiments in this paper are designed to be reproducible. Models are prompted using the same setup via LangChain and OLLAMA, with shared prompts, fixed model configurations, and consistent hardware. The same datasets, software libraries, and parameters are used across all models, enabling straightforward

reproduction of results. Although identical results cannot be guaranteed, the experimental setup allows the study to be mostly redone with similar outcomes. The different prompting strategies used for evaluation are also explained in-depth in the next part of this section. Future work will be needed to assess whether the same findings hold as new data becomes available.

Bias and Ethical Data Use: The project critically engages with the potential biases present in both datasets and LLM outputs. No data has been collected outside what is already available. Ethical reuse of datasets includes proper citation, as expected by good scientific practice and emphasized by 4TU FAIR Data Management Act.

Plagiarism and Attribution: This paper ensures that all datasets, methods and sources used are clearly cited and proper credit is given to the authors. Generative AI tools, in this case ChatGPT, were only used to polish the sentences and writing style, but not for coding or producing any results. Plagiarism has been consistently avoided in order to match the academic integrity rules.

Human Involvement and Consent: This research does not involve human subjects. Therefore, no consent or ethical review from an institutional board was needed.

Dual Use and Misuse Risks: This paper is not creating new models or tools, however when writing about limitations of our work we acknowledge that some of the data on explanation power of LLMs can be used to craft misleading rationales, however we believe this paper aims to promote accountability rather than encouraging falsification of the results.

Responsible Reporting: All results are reported, this includes the inconclusive or negative results. This was done in order to be in compliance with responsible data practices and discourage prioritizing positive results, as warned in the responsible research lecture.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Darius Feher, Abdullah Khered, Hao Zhang, Riza Batista-Navarro, and Viktor Schlegel. 2025. Learning to generate and evaluate fact-checking explanations with transformers. *Engineering Applications of Artificial Intelligence*, 139:109492.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463. ArXiv:2309.05463 [cs].

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.

ACL Rolling Review. 2023. Responsible nlp research checklist.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, and 1 others. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*.

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.

Trishala Sathe, Shreya Agarwal, Preksha Nema, Niloy Ganguly, and Pawan Goyal. 2023. An exploration of large language models for verification of news headlines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Xuezhi Wang, Swaroop Mishra, Andy Nguyen, Hyung Won Chung, Yi Tay, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 21674–21687.