Too Good to Be True? Generalization Challenges in LLM-Based Fake News Detection

Gustavo Gambarra
gustavo.gambarra@ccc.ufcg.edu.br
Federal University of Campina Grande - UFCG
Campina Grande, Paraíba, Brazil

Caio L. M. Jerônimo
caio.jeronimo@servidor.uepb.edu.br
State University of Paraíba - UEPB
Patos, Paraíba, Brazil

Claudio E. C. Campelo
campelo@computacao.ufcg.edu.br
Federal University of Campina Grande - UFCG
Campina Grande, Paraíba, Brazil

ABSTRACT

Although recent studies report high accuracy in fake news classification using supervised models, such models often rely heavily on spurious correlations in training datasets and fail to generalize to unseen contexts. A smaller body of work has explored more realistic approaches based on linguistic cues, such as subjectivity, aiming to reduce the reliance on dataset-specific content learned during training. In this study, we assess the capabilities of four state-of-the-art large language models (LLMs) — Llama3.1-70B, Claude 3.5 Sonnet, GPT-4o, and Mistral Large 2 — in identifying fake news based solely on linguistic features, without access to external verification. Two experiments were conducted: one with generic prompts and another with fine-grained instructions and examples, including classification of false excerpts into categories like untrue facts, exaggerations, and incorrect named entities. Despite the well-known linguistic capabilities of modern LLMs, our results show consistently low performance across both experiments. This outcome supports the hypothesis that detecting disinformation based solely on content or linguistic cues, without external factual grounding, is far more challenging than commonly reported, and that current evaluation protocols may overestimate the generalization capabilities of fake news classifiers.

KEYWORDS

fake news, large language models, fake news classification

1 INTRODUCTION

The proliferation of fake news poses a significant threat to public perception, democratic processes, and collective decision-making [1, 20]. The ease with which misleading content can be produced and disseminated online has intensified concerns about the reliability of information ecosystems and the urgent need for effective detection mechanisms [18].

In recent years, the task of fake news detection has attracted considerable attention, particularly through the use of supervised machine learning models trained on labeled datasets [17, 25]. Many of these studies report impressive classification results, suggesting that distinguishing between true and false information is a largely solved problem. However, growing evidence suggests that these models often achieve high performance by exploiting superficial and spurious correlations in the training data — such as frequent mentions of particular named entities or recurrent topics — rather than learning to identify intrinsic signals of deception [2, 4, 15]. Consequently, such models tend to generalize poorly to out-of-domain data or real-world scenarios, where these shortcuts are no longer available.

A crucial limitation of most existing approaches is the lack of interpretability and the absence of fine-grained linguistic analysis. Fake news classification is frequently reduced to a binary decision task (true vs. false), overlooking the fact that disinformation is a complex, multifaceted phenomenon. Misleading content may involve partially false claims, exaggerations, unverifiable statements, or incorrect references to named entities — each with distinct rhetorical and linguistic characteristics. Moreover, very few studies aim to pinpoint which parts of the news are false, or why they are false, limiting the explanatory power of the models and their applicability in practical fact-checking settings [6].

To address these challenges, some researchers have proposed alternative approaches that reduce reliance on factual content and instead focus on linguistic features such as subjectivity, coherence, hedging, and sensationalism [7, 13, 14, 21]. Such strategies aim to capture structural and rhetorical cues of misinformation, which are more stable across domains and time. In this context, Large Language Models (LLMs) — such as GPT-4, Claude, Mistral, and LLaMA — have emerged as powerful tools capable of capturing deep linguistic and pragmatic structures, and have shown strong performance across a variety of natural language understanding tasks [12, 19].

Despite these capabilities, our study shows that even state-of-the-art LLMs struggle to detect fake news reliably when restricted to purely linguistic information, without access to external factual knowledge. We conduct two experiments using a carefully curated dataset of 184 news excerpts — balanced between true and fake content — and prompt four cutting-edge LLMs (Llama3.1-70B, Claude 3.5 Sonnet, GPT-4o, and Mistral Large 2) to analyze the news using only intrinsic textual properties. In the first experiment, generic prompts are used to assess the models' ability to identify the veracity and location of falsehoods. In the second, more detailed instructions and examples are provided to encourage few-shot learning and fine-grained categorization of the types of falsehood (e.g., untrue facts, exaggerations, and misattributed entities).

Our findings show consistently low performance across all models and both experiments, highlighting the limitations of current content-based classifiers and reinforcing the hypothesis that success in previous studies may be largely driven by dataset artifacts and overfitting. These results provide empirical support for a more

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

cautious interpretation of current benchmarks in fake news classification and call for more rigorous and interpretable methodologies.

Finally, the dataset and artifacts developed in this study are publicly available[1], with the goal of fostering more robust evaluations and contributing to a deeper understanding of the linguistic structures underlying misinformation.

2 RELATED WORK

With the rise of LLMs, many studies are being conducted on the use of these tools for fake news detection [5, 12]. A comparative study [16] compares the performance of different LLM based approaches, like BERT-like models and autoregressive decoder-only models (e.g. Llama, Mistral and GPT series). In the study, the authors demonstrate that the BERT-like models outperform the autoregressive models in classification tasks. However, the same study finds that the autoregressive models were more robust against text perturbations. This kind of results demonstrate the complexity involved in the fake news classification task. Wu et al.[23] propose the SheepDog, a style-robust fake news detector, that focus in the documents content, instead of the writing style of the documents to determine the veracity of the news.

Another study [10] propose the use of RAG (Retrieval-Augmented Generation) framework with the support of real-time articles acquisitions to perform fake news classifications. The potential of online LLMs (e.g. Gemini and GPT-4) in performing fake news classifications is also explored [24], comparing these models with offline ones.

However, most of these studies treat the task in a binary way, without explaining the inferences made by the models, which compromises the reliability of the results. In addition, a challenge in using such models to fake news detection is the dynamic nature of misinformation, that changes both in format and content, pushing any classification model to the edges of performance.

An alternative to mitigate these problems is to couple real-time data retrieval tools to the models, as proposed in [8, 9], or to use a content-based approach [3], as is done in this study.

Some authors [7, 21] found that fake news tends to present a higher level of subjectivity, which differentiates it from true news. In addition, [21] propose an annotation protocol to classify fake news into categories such as **untrue fact**, **unverifiable**, **incorrectly named entity**, and **exaggeration**. This protocol was adopted in this work to structure the analysis of fake news.

Thus, this study starts from the premise that fake news has a higher level of subjectivity [22] and, based on this starting point, evaluates the news without using previously known facts, focusing only on the intrinsic characteristics of the text. In addition, a comparative analysis is carried out between state-of-the-art models to evaluate the reliability of these tools in the task of detecting fake news.

3 METHODOLOGY

3.1 Dataset

In this study, a dataset of 184 English-language news documents was manually collected, divided equally between true and fake news. Regarding the fake news subset, we have 45 documents labeled as partially fake and 47 labeled as entirely fake. The true news documents were collected from reputable media platforms, such as *Reuters* and *BBC News*, ensuring the quality and reliability of the information. The fake news, in turn, were obtained from fact-checking websites, such as *FactCheck.org* and *Snopes*.

The collection of fake news documents from these fact-checking websites is important, as it indicates that these documents had a large impact, possibly reaching a wide audience. Furthermore, these fake news documents present many different characteristics in terms of style, structure and content. The news topics included in the dataset cover areas such as culture, sports, health and politics, providing a broad view of the different forms and contexts in which misinformation can occur.

During the data collection, it was observed that some of the fake news articles are no longer available on the web, which highlights the volatility of this type of online content.

3.2 News Annotation Protocol

To analyze the fake news, the same annotation protocol used in [21] was adopted. This protocol allows for the identification and categorization of false excerpts within the news, facilitating a detailed analysis of how misinformation is constructed. The annotation was done using tags similar to HTML tags to mark the false excerpts, including additional information relevant to the research.

The annotation protocol was defined as follows:

- **url_fake_news**: Located in the news header, this variable contains the URL of the original source from which the fake news was collected.
- **url_true_news_[1-9]+**: One or more variables located in the news header, containing URLs of true news that show which fragments of the news can be defined as false.
- **<fake></fake>**: Tag placed around a false excerpt identified in the news. This tag is used to isolate and precisely identify the parts of the text that contain false information.
- **type**: Attribute within the `<fake>` tag that indicates the category of falsehood of the annotated excerpt. The four categories used are:
  - **false**: Fragments that present completely false information (117 annotations);
  - **unverifiable_fact**: Excerpts of dubious origin, where there is no true news to prove the falsity of the content (13 annotations);
  - **exaggeration**: When a fact is distorted through exaggeration (19 annotations);
  - **incorrectly_named_entity**: When names of people, places, or institutions are used misleadingly in fake news (2 annotations).

Figure 1 shows an example of a fake news story annotated according to this annotation protocol.

3.3 Fact-Checking Control

Since this study uses a content-based approach, it is necessary to instruct the models not to perform any type of fact-checking. In other words, the models were instructed to ignore any prior knowledge about the facts presented in the news and to consider only

[1] https://github.com/gustavogambarra/llm-fake-news-artifacts

url_fake_news="https://www.disneydining.com/magic-kingdom-for-adults-with-children-only-bb1/#:~:text=Beginning%20August%207%2C%202023%2C%20Guests,1%20through%20August%2031%20annually."

url_true_news_1="https://www.snopes.com/fact-check/magic-kingdom-disney-adults/"

<fake type="false" url_true_news_1>Disney's Magic Kingdom to Welcome ONLY Adults With Children, Beginning Late Summer 2023**</fake>**

For all the progress that's been made in the campaign for the advancement of Disney Adults and Childless Disney Guests, the efforts were simply not enough to stave off a barrage of snide comments, tacky social media posts, and degrading online essays against these Disney Parks enthusiasts that has finally **<fake type="false" url_true_news_1>**resulted in Disney World's decision to bar these Guests from visiting Magic Kingdom on weekends and over the summer, beginning in August.**</fake>**

Since mid-2019, when a female Guest at Magic Kingdom flew into a terrifying fit of rage after her son was forced to wait in line behind other Guests who had the audacity to arrive at the Disney World theme park without any children in tow, Disney Adults–especially the childless ones–have come under attack by Disney lovers and haters alike.

Figure 1: Example of a fake news annotated using the protocol we propose.

the intrinsic characteristics of the text, such as internal coherence, quality of argumentation, and use of sensationalist language. If the model uses fact-checking, the results of the study could be compromised, since models that have prior knowledge about the facts would have an advantage when classifying the news.

To ensure that the model did not perform fact-checking, tests were performed with different instruction formats. In each test, three true news and three fake news documents were sent to the model, asking it to classify these documents as true or fake and to explain its answer. The model's responses were analyzed to observe how it reached its conclusion, evaluating whether it used previously known facts or focused exclusively on the textual content.

The instruction set with the best performance was as follows:

> *For this analysis, you cannot perform any type of fact-checking, ignore any prior knowledge or known facts about the content and base your evaluation exclusively on the textual characteristics present in the document, such as:*
> (1) *Internal Coherence and Consistency: Whether the text has a logical narrative without contradictions or inconsistent statements.*
> (2) *Quality of Argumentation: If the arguments are well-supported and if there is adequate backing for the claims made in the text.*
> (3) *Use of Sensationalist Language: Whether the text uses exaggerated or alarming terms that might indicate emotional manipulation or an attempt to deceive.*
> (4) *References and Citations: Whether the text cites sources and if these are presented clearly and transparently.*
> (5) *Structure and Writing Style: Whether the text follows an organized structure and if the writing style is professional and objective.*

These instructions were incorporated into the prompts used in the experiments, with the aim of directing the models to focus on the textual content, without resorting to external factual information.

3.4 First Experiment: General Detection

The first experiment was designed to evaluate the models' ability to detect the veracity of news, identify the level of falsehood and correctly mark false excerpts present in the news. To do this, the following prompt was used:

> *You are a fake news detector tasked with analyzing the news enclosed within triple backticks and classifying it into one of two main categories: "True" or "False." If the news is classified as "False," further determine if it is "Partially False" (if the news contains both true and false segments) or "Entirely False" (if the news is composed entirely of false segments). For this analysis, you cannot perform any type of fact-checking, ignore any prior knowledge or known facts about the content and base your evaluation exclusively on the textual characteristics present in the document, such as:*
> (1) *Internal Coherence and Consistency: Whether the text has a logical narrative without contradictions or inconsistent statements.*
> (2) *Quality of Argumentation: If the arguments are well-supported and if there is adequate backing for the claims made in the text.*
> (3) *Use of Sensationalist Language: Whether the text uses exaggerated or alarming terms that might indicate emotional manipulation or an attempt to deceive.*
> (4) *References and Citations: Whether the text cites sources and if these are presented clearly and transparently.*
> (5) *Structure and Writing Style: Whether the text follows an organized structure and if the writing style is professional and objective.*
>
> *Your response should be a JSON object with the following properties:*
> - ***news_with_tags**: A string that contains the news itself with the following annotation protocol: the false segments should be enclosed in* `<fake></fake>` *tags, similar to HTML tags, as shown in the following example: "NASA Confirms Evidence of Liquid Water on Mars. NASA has announced that its Mars Reconnaissance Orbiter (MRO) has detected evidence of liquid water on the surface of Mars.* `<fake>` *The agency also noted that this water is suitable for human consumption without any need for treatment.* `</fake>` *"*
> - ***news_veracity**: A boolean indicating whether the news is "True" (true) or "False" (false).*
> - ***falsity_level**: A string with possible values "Partially False" or "Entirely False". This property should only be included if news_veracity is "False". If the entire content of the news in the news_with_tags property is enclosed in* `<fake></fake>` *tags, the value should be "Entirely False." If there are segments that are not enclosed in* `<fake></fake>` *tags, the value should be "Partially False."*

The prompt instructs the model to return a JSON object containing three main attributes: *news_veracity*, a boolean indicating whether the news document is true or fake; *falsity_level*, a string specifying whether the fake news document is "Partially False" or

"Entirely False"; and *news_with_tags*, which contains the text with the identified false excerpts. These three elements are used to measure the model's performance in the tasks of classifying veracity, identifying the level of falsehood, and flagging incorrect passages.

3.5 Second Experiment: Categorized Detection

The second experiment was designed to deepen the analysis performed in the first experiment, adding an extra layer of detail in the identification and classification of false excerpts. In addition to evaluating the veracity and level of falsehood of the news, the model was instructed to classify the false excerpts into one of the four categories mentioned. The prompt we use consists of all the instructions present in the first experiment, with the addition of the following instructions, which describe the categories of falsehood, accompanied by illustrative examples for each type of excerpt:

> *Each opening* `<fake>` *tag should have an attribute named "type", which can be one of the following values:*
> - ***false**: Category for news fragments that have no truth at all (completely false). Example:* `<fake type="false">` *NASA Confirms the Presence of Advanced Alien Civilizations on Mars* `</fake>`.
> - ***unverifiable_fact**: Category for dubious fragments where there is no true news to prove falsehood. Example:* `<fake type="unverifiable_fact">` *NASA Planning Secret Manned Mission to Mars to Investigate Water Sources* `</fake>` *(It is not confirmed whether NASA is planning such a mission, and there is no concrete evidence to verify or disprove this claim.).*
> - ***exaggeration**: Category for when a fact is distorted by the presence of some exaggeration. Example:* `<fake type="exaggeration">` *NASA Claims Liquid Water Flows Freely Across the Entire Surface of Mars* `</fake>` *(The actual findings indicate that liquid water is present in localized streaks, not flowing freely across the entire surface, making the claim an exaggeration.).*
> - ***incorrectly_named_entity**: Category for when an entity is incorrectly named. Example:* `<fake type="incorrectly_named_entity">` *The European Space Agency (ESA) has reportedly discovered evidence of liquid water on Mars.* `</fake>` *(The Mars Reconnaissance Orbiter (MRO) from NASA was actually responsible for the discovery of liquid water on Mars, not the European Space Agency (ESA), making the entity mentioned incorrect.)*

The goal of this experiment is to evaluate whether adding more detailed examples and instructions about falsehood categories such as **false**, **unverifiable_fact**, **exaggeration**, and **incorrectly_named_entity** would improve the model's performance compared to the first experiment. Furthermore, this experiment not only tests the model's capabilities already evaluated in the first experiment, but also introduces a more granular analysis of the types of falsehood present in the texts.

3.6 Metrics

To evaluate the model's performance, we use two metrics: accuracy and recall. Accuracy was used to measure the model's overall performance, considering the total number of correct answers in relation to the complete data set. Recall, in turn, is applied to analyze the model's performance in relation to specific attribute values.

For example, for the attribute "news veracity", we will have an overall accuracy measure that represents the model's performance in inferring results for both fake and true news. In addition, we will have a specific recall for fake news and another for true news, which allows us to measure the model's performance in inferring these specific cases. This ensures a more detailed evaluation of the ability to distinguish between the two types of news, since the recall of both classes provides a comprehensive view of how the model behaves in each scenario.

This methodology was applied consistently across all analyses, adapting to the various attributes and their values present in the study, providing a comprehensive and detailed understanding of the model's performance across different aspects and levels of granularity.

Regarding the *news_with_tags* attribute, which assesses the model's ability to correctly identify false excerpts in the documents, we concatenate all the false excerpts that we annotate from the dataset, and also the excerpts identified by the model. Then, the **Jaccard coefficient** [11] was used to measure the similarity between the concatenated excerpts from the manual annotations and those tagged by the model.

3.7 Jaccard coefficient

To evaluate the model's ability to correctly identify the false excerpts from the fake news documents, the **Jaccard coefficient** [11] was used as a similarity metric. The Jaccard coefficient is a measure that quantifies the similarity between two sets, and is calculated by the ratio between the size of the intersection of the sets and the size of their union. The formula for the Jaccard coefficient is:

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}$$

where A and B represent the sets of manually annotated false excerpts and the excerpts returned by the model, respectively.

A **word-based tokenization** approach was employed to compare the snippets, ignoring punctuation marks. Thus, the texts were divided into words, and the similarity was measured based on the matching words between the identified excerpts. For the purposes of this study, the model was considered to be correct when the Jaccard coefficient resulted in a value greater than 0.8, indicating a high similarity between the excerpts.

In the second experiment, in addition to comparing the general similarity of the excerpts, a more detailed analysis was performed **by category of falsehood**. The excerpts were grouped by category — such as **untrue fact**, **unverifiable**, **exaggeration**, and **incorrectly named entity** — and the Jaccard coefficient was again used to measure the similarity between the annotated excerpts and those returned by the model, this time considering each category separately.

4 RESULTS AND DISCUSSION

4.1 News Veracity

The recall for fake news measures the proportion of correctly identified documents among all fake documents. Observing Figure 2, it is possible to see that for the first experiment, all models have promising results, fluctuating between 94.6% and 100%.

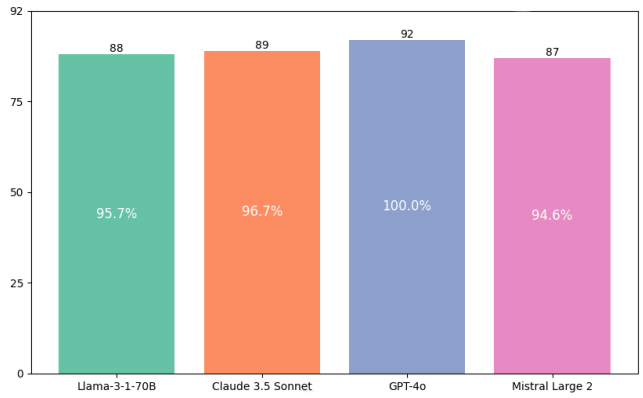

Figure 2: Chart representing Recall for fake news in the first experiment

In the second experiment, there is no great variation in the results, as can be seen in Figure 3, the models maintain good results, with only subtle variations, reaching a maximum drop of 4.3% in GPT-4o, and a maximum growth of the same value in Mistral Large 2.

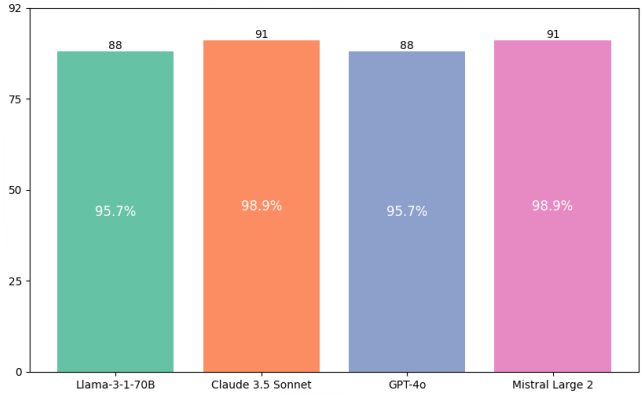

Figure 3: Chart representing Recall for fake news in the second experiment

In terms of the true news, the scenario changes, as can be seen in Figure 4, only Llama-3-1-70B maintained a recall above 90%. Claude 3.5 Sonnet is in second place, with a recall of 57.6%, followed by GPT-4o and Mistral Large 2, which presented results below 40%.

Consequently, it can be seen that the models face greater difficulty in identifying true news, tending to incorrectly classify them as fake, a problem that was aggravated in the second experiment, where the models' results showed significant drops compared to

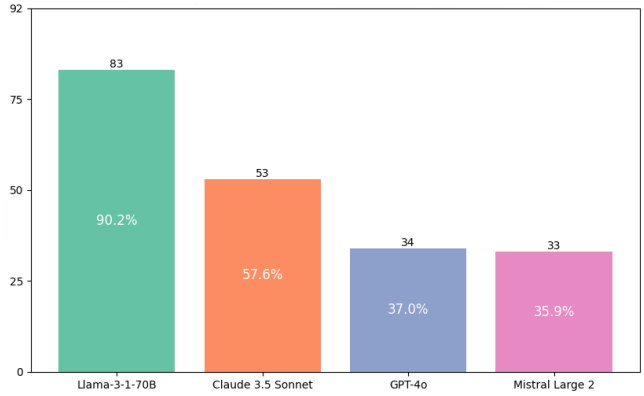

Figure 4: Chart representing Recall for true news in the first experiment

the first one. Except for GPT, which showed an increase of 11.9% in its recall. The other models faced reductions that reached around 28% in Llama and Claude, as shown in Figure 5.

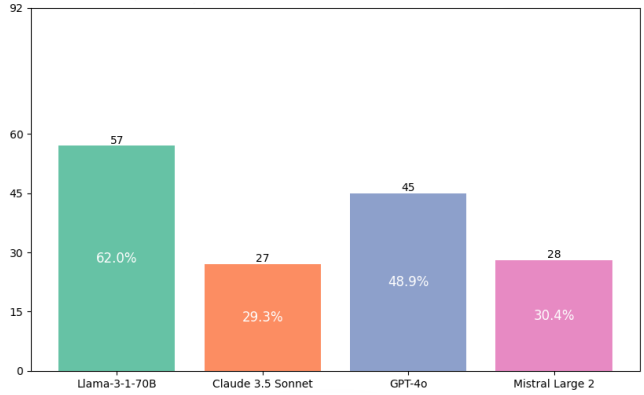

Figure 5: Chart representing Recall for true news in the first experiment

In terms of accuracy, Figure 6 shows that, in the first experiment, LLama obtained the best results, with an accuracy of 92.9%, followed by Claude, with 77.2%.

In the second experiment, it can be seen in Figure 7 that most models presented results inferior to those of the first experiment, with the exception of GPT, which obtained an improvement of 3.5%, reaching 72%. However, even with a drop of 13.9%, Llama continues to lead with the best results, obtaining an accuracy of 79%, which makes it the most reliable model, considering the news veracity attribute.

4.2 Falsity Level

Regarding the level of falsehood of the news labeled as fake in the first experiment, when examining Table 1 we can conclude that for the "partially false" news, Llama and GPT obtained the best results, with recall of 0.84. However, for the "entirely false" news, GPT had the worst performance, with recall of 0.38. In this scenario, Claude

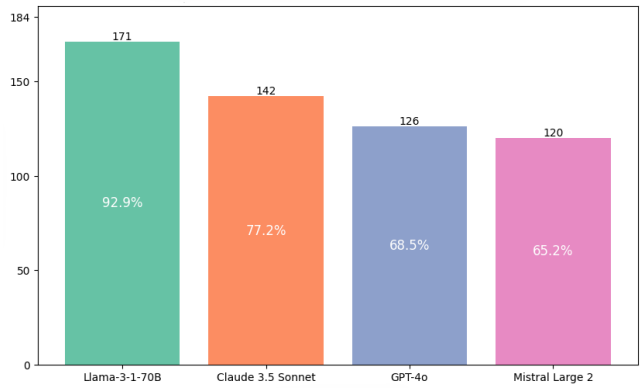

Figure 6: Chart representing the accuracy of the news veracity in the first experiment

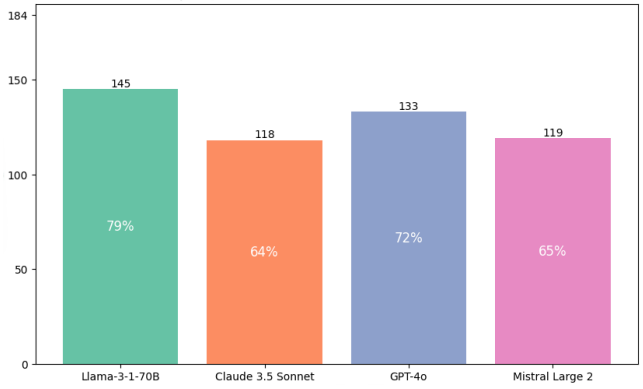

Figure 7: Chart representing the accuracy of the news veracity in the second experiment

Table 1: Table of results related to the "level of falsehood" of the first experiment

Model	Recall of "Partially False"	Recall of "Entirely False"	Accuracy
Llama 3.1-70B	0.84	0.55	0.7
Claude 3.5 Sonnet	0.6	0.87	0.75
GPT-4o	0.84	0.38	0.61
Mistral Large 2	0.56	0.51	0.53

obtained the best result, with a recall of 0.87, making it the best in identifying this type of falsehood. To have a more general view of the performance of the models in this task, we analyze the accuracy of the level of falsehood. In this metric, Claude also obtained the best performance, with 0.75 accuracy, followed by Llama with 0.7. GPT was in third place with 0.61, followed by Mistral with 0.53.

In the second experiment, we can note in Table 2 that the recall value of the "partially false" type increased in most models. On the other hand, the recall of the "entirely false" type showed a significant drop in all results. This means that in this experiment, the models were biased towards responding "partially false", negatively affecting accuracy, which was lower than in the first experiment in all cases, with greater drops for the LLama and GPT models. It is worth mentioning that in addition to having presented the best performance again, Claude was also the model least affected by the drop in performance in this experiment, losing only 0.01 in accuracy compared to the previous experiment. In last place was Mistral, with 0.51 accuracy.

Table 2: Table of results related to the "level of falsehood" of the second experiment

Model	Recall of "Partially False"	Recall of "Entirely False"	Accuracy
Llama 3.1-70B	0.82	0.23	0.52
Claude 3.5 Sonnet	0.8	0.68	0.74
GPT-4o	0.91	0.17	0.53
Mistral Large 2	0.93	0.11	0.51

Regarding the "falsity level" attribute, Claude was the most reliable model, with the best accuracy in both experiments. In contrast, Mistral occupied the last position, with the lowest accuracy in the same experiments.

4.3 Identification of False Excerpts

Considering that a correct answer occurs when the Jaccard coefficient exceeds 0.8, we were able to analyze the recall and accuracy for this attribute. As shown in Table 3, whose results refer to the first experiment, for the "partially false" news, Claude and Llama were tied for the best positions, with 0.38 Recall. Also, for the news with this level of falsehood, the average Jaccard coefficient was higher for Claude, who obtained 0.58. For the "completely false" news, Claude, with a significant advantage, continued to be the best placed, with a recall 0.87. It is important to note that the same metric for the other models did not reach 0.5. Therefore, it is common for the average Jaccard coefficient of Claude to also be the highest among the models, with 0.96, followed by Llama and GPT, respectively, with 0.69 and 0.61.

When analyzing the similarity between the false excerpts for all news documents, we noted that Claude once again occupied the first position with 0.63 accuracy and 0.77 average Jaccard coefficient. The last positions were occupied by GPT and Mistral.

In the second experiment, as shown in Table 4, all models showed drops in all the evaluated metrics, with Claude being the least affected. It shows the best performance, with an accuracy of 0.54 and an average Jaccard coefficient of 0.74. On the other hand, the other models faced significant reductions, obtaining accuracy below 0.2 and an average Jaccard coefficient below 0.5.

Table 3: Table of results related to the identification of false excerpts in the first experiment

Model	Recall of similarity between false excerpts of "Partially False" news	Average Jaccard Similarity Coefficient for "Partially False" News	Recall of similarity between false excerpts for "Completely False" news	Average Jaccard Similarity Coefficient for "Completely False" News	Accuracy of the similarity between false excerpts for all fake news	Average Jaccard Similarity Coefficient for all fake news
Llama3.1-70B	0.38	0.39	0.47	0.69	0.42	0.54
Claude 3.5 Sonnet	0.38	0.58	0.87	0.96	0.63	0.77
GPT-4o	0.18	0.42	0.32	0.61	0.25	0.52
Mistral Large	0.16	0.43	0.36	0.55	0.26	0.49

Table 4: Table of results related to the "identification of false excerpts" in the second experiment

Model	Recall of similarity between false excerpts of "Partially False" news	Average Jaccard Similarity Coefficient for "Partially False" News	Recall of similarity between false excerpts for "Completely False" news	Average Jaccard Similarity Coefficient for "Completely False" News	Accuracy of the similarity between false excerpts for all fake news	Average Jaccard Similarity Coefficient for all fake news
Llama3.1-70B	0.02	0.30	0.21	0.49	0.12	0.40
Claude 3.5 Sonnet	0.29	0.57	0.79	0.91	0.54	0.74
GPT-4o	0.11	0.39	0.19	0.57	0.15	0.48
Mistral Large 2	0.02	0.27	0.06	0.30	0.04	0.29

4.4 Simultaneous hits on the three attributes evaluated

To obtain an overall perspective of which model performed best in the attributes "news veracity", "level of falsehood" and "identification of false excerpts", the number of simultaneous hits in the three attributes was evaluated.

As illustrated in Figure 8, in the first experiment, Claude obtained the best result, hitting all attributes combined in 55.4% of the fake news, followed by Llama with 23.9%. The last performing models were Mistral and GPT, with 21.7% and 20.7%, respectively.

In the second experiment, the Figure 9 shows a decrease in performance for all models, when compared to the previous experiment. In this one, Claude continues ahead of the others, but this time with 46.7% accuracy, followed by Llama and GPT, which tied with 10.%. In last place was Mistral, with 2.2% accuracy.

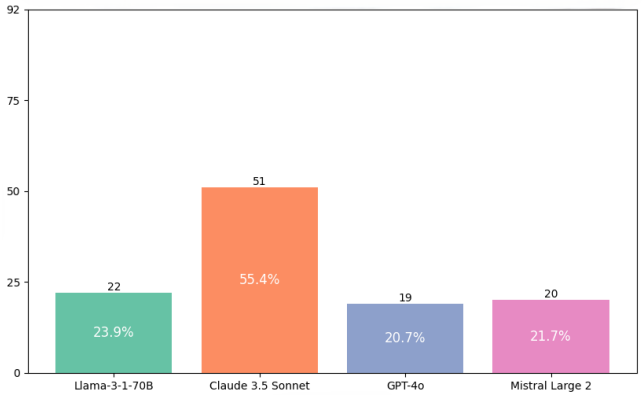

Figure 8: Chart representing the number of simultaneous hits between the attributes "news veracity", "level of falsehood" and "identification of false excerpts" in the first experiment

4.5 Identification of the Falsehood Category in the False Excerpts

With the additional information regarding the falsehood category present in each false excerpt, the second experiment allowed for a more granular analysis of these excerpts.

Regarding the news that contains excerpts of the **untrue fact** type (78), it is possible to observe, in Table 5, that Claude continues to lead in the results, with a recall of 0.35. The worst models were GPT and Mistral, with 0.08 and 0.02, respectively. In this context, precision indicates the number of correct answers of the model over the number of inferences made, with results similar to recall; Claude continues to have the best performance, followed by Llama, GPT and Mistral, respectively.

For the other categories, the following results were obtained:

- **Exaggeration**: Only Claude correctly classified 1 of 13 documents;
- **Unverifiable fact**: Only Claude correctly classified 1 of 7 documents.

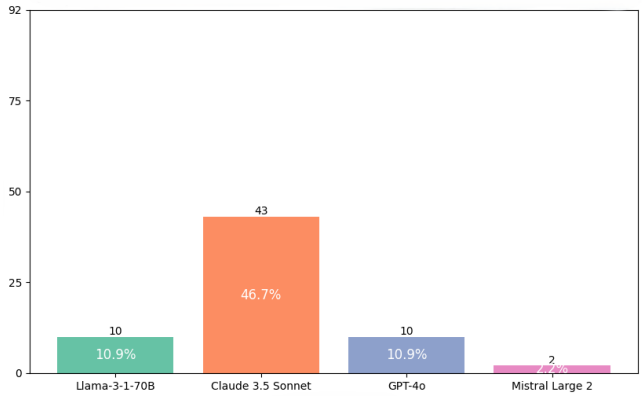

Figure 9: Chart representing the number of simultaneous hits between the attributes "news veracity", "level of falsehood" and "identification of false excerpts" in the second experiment

Table 5: Recall and Precision of the similarity between false excerpts of news that contains the type untrue fact

Model	Recall of the similarity between false excerpts for news that contains the false fact type.	Precision of similarity between false excerpts for news containing the false fact type.
Llama3.1-70B	8/78 = 0.10	0.10
Claude 3.5 Sonnet	27/78 = 0.35	0.33
GPT-4o	6/78 = 0.08	0.08
Mistral Large 2	2/78 = 0.03	0.02

- **Incorrectly named entity**: None of the models correctly classified this category, from a total of 2 news documents.

Due to the small size of the dataset containing false excerpts in the categories of **exaggeration**, **unverifiable fact**, and **incorrectly named entity**, it is not feasible to perform a specific analysis for these categories. However, based on the overall results of this analysis, we conclude that the models did not present promising results, achieving at most 0.35 of recall and 0.33 of precision for Claude, while the other models were even closer to zero. This highlights the challenges of detecting fake news when analyzing attributes with a higher level of specificity.

5 CONCLUSIONS AND FUTURE WORK

In this study, a comparative analysis of LLMs was performed in the detection of fake news, based on different attributes, such as news veracity, level of falsehood, identification of false excerpts, and falsehood categories. Through the two experiments conducted, we reach some important conclusions regarding the limitations and capabilities of the models we analyze.

The **Llama3.1-70B** model presented the best performance with regard to the **news veracity** attribute, being the most reliable model for this task, with promising results. Its accuracy in correctly identifying true and fake news demonstrates that, for the specific purpose of verifying the veracity of news, Llama is the best option among the models analyzed in this study.

On the other hand, the **Claude 3.5 Sonnet** model showed the best results in most attributes, especially when we evaluated them together. Claude proved to be more balanced, with good performance not only in verifying veracity, but also in identifying false excerpts and analyzing the level of falsehood, being the most robust model in a complete and detailed analysis of fake news.

However, regarding the task of **Identifying the Falsehood Category in the False Excerpts**, although Claude reached the highest results, the overall results were unsatisfactory for all models. This highlights the challenges and limitations faced in detecting fake news when there is greater specificity in inferences, such as identifying whether an excerpt is an exaggeration, an unverifiable fact, or contains an incorrectly named entity. It reveals a situation where the models seems to be able to identify the fake news, but they are not able to precisely identify the reasons that make the document a misleading one. Such results seem to highlight a vulnerability that may affect other fake news classification strategies.

Another relevant point was the poor performance in the **second experiment**, where we use a more detailed set of examples, compared to the first one. In all of the analyses we performed, the results were worse in this scenario. This result can suggest that, even with a set of detailed instructions and examples, the models we use could not clearly distinguish the possible reasons that could be behind a fake news. This result confirms our hypothesis that such models are not yet able to identify the nuances that are behind a misleading document. These results could suggest that more experiments and strategies needs to be explored in order to access the real capabilities of the LLMs to identifying fake news.

5.1 Future Work

To improve investigations on fake news detection with LLMs, we suggest some future work:

- Investigate more deeply the cause of the lower performance in the second experiment compared to the first one, trying to understand which aspects of the examples contributed to the drop in performance.
- Build a set of prompts with more robust examples, improving the capabilities of detecting specific falsehoods, such as exaggerations or incorrectly named entities, to help models make more accurate inferences.
- Fine-tune the models with a more specialized dataset of fake news, to evaluate if it can lead to a significant improvement in overall and granular performance, such as categorizing fake news in excerpts.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. https://doi.org/10.1257/jep.31.2.211

[2] Pepa Atanasova, Isabelle Augenstein, and Christina Lioma. 2020. Diagnostic dataset construction with minimal supervision for interpretable evaluation of fact verification systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1802–1812.

[3] Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. Fakeflow: Fake news detection by modeling the flow of affective information. *arXiv preprint arXiv:2101.09810* (2021).

[4] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1859–1874.

[5] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113.

[6] Ziqing Hu, Yuan Liu, Di Jin, Yucheng Li, Zitao Liu, and Xinyu Lei. 2023. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *arXiv preprint arXiv:2309.12247* (2023).

[7] Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio EC Campelo, Adriano Veloso, and Allan Sales da Costa Melo. 2019. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*. 15–24.

[8] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large Language Model Agent for Fake News Detection. *arXiv preprint arXiv:2405.01593* (2024).

[9] Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection. *arXiv preprint arXiv:2407.08952* (2024).

[10] Mohammad Vatani Nezafat and Saeed Samet. 2024. Fake News Detection with Retrieval Augmented Generative Artificial Intelligence. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. 160–167. https://doi.org/10.1109/FLLM63129.2024.10852474

[11] University of Chicago. 2024. Set-based (Jaccard) similarity. https://ds1.datascience.uchicago.edu/08/1/Set_Based_Similarity.html. Acessado em: 24 de setembro de 2024.

[12] Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet* 16, 8 (2024). https://www.mdpi.com/1999-5903/16/8/298

[13] Veronica Perez-Rosas, Bennett Kleinberg, Akira Lefevre, and Rada Mihalcea. 2021. Disinformation and the language of deception: Linguistic features of fake news stories. *Digital Threats: Research and Practice* 2, 1 (2021), 1–25.

[14] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2921–2927.

[15] Ali Raza, Yash Paul, Abhik De, and Ashutosh Modi. 2024. Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language Models with Generative AI-Annotated Data. *arXiv preprint arXiv:2401.14276* (2024).

[16] Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data. *Knowledge and Information Systems* (2025), 1–26.

[17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM.

[18] Jinjin Su, Xinyi Zhou, and Kai Shu. 2023. Adapting Fake News Detection to the Era of Large Language Models. *arXiv preprint arXiv:2311.04917* (2023).

[19] Jinjin Su, Xinyi Zhou, and Kai Shu. 2023. Fake News Detectors are Biased against Texts Generated by Large Language Models. *arXiv preprint arXiv:2309.08674* (2023).

[20] Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval* 12, 2 (2023), 28.

[21] Lucas Lima Vieira, Caio Libanio Melo Jeronimo, Claudio E. C. Campelo, and Leandro Balby Marinho. 2020. Analysis of the Subjectivity Level in Fake News Fragments. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (São Luís, Brazil) *(WebMedia '20)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/3428658.3430978

[22] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. 34–35.

[23] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 3367–3378.

[24] Ruoyu Xu and Gaoxiang Li. 2024. A comparative study of offline models and online llms in fake news detection. *arXiv preprint arXiv:2409.03067* (2024).

[25] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.