computers

MDPI

Article

LLaMA 3 vs. State-of-the-Art Large Language Models: Performance in Detecting Nuanced Fake News

Stefan Emil Repede * and Remus Brad

Department of Computer Science, Electrical and Electronics Engineering, University of Sibiu, 4 Emil Cioran Street, 550025 Sibiu, Romania; remus.brad@ulbsibiu.ro

* Correspondence: stefan.repede@ulbsibiu.ro

Abstract: This study investigates the effectiveness of a proposed version of Meta's LLaMA 3 model in detecting fake claims across bilingual (English and Romanian) datasets, focusing on a multi-class approach beyond traditional binary classifications in order to better mimic real-world scenarios. The research employs a proposed version of the LLaMA 3 model, optimized for identifying nuanced categories such as "Mostly True" and "Mostly False", and compares its performance against leading large language models (LLMs) including Open AI's ChatGPT versions, Google's Gemini, and similar LLaMA models. The analysis reveals that the proposed LLaMA 3 model consistently outperforms its base version and older LLaMA models, particularly in the Romanian dataset, achieving the highest accuracy of 39% and demonstrating superior capabilities in identifying nuanced claims, over all the compared large language models. However, the model's performance across both languages highlights some challenges, with generally low accuracy and difficulties in handling ambiguous categories by all the LLMs. The study also underscores the impact of language and cultural context on model reliability, noting that even state-of-the-art models like ChatGPT 4.o and Gemini exhibit inconsistencies when applied to Romanian text and more than a binary true/false approach.

Keywords: fake news detection; large language models; natural language processing; disinformation management; transformer architecture; bilingual NLP

Citation: Repede, S.E.; Brad, R. LLaMA 3 vs. State-of-the-Art Large Language Models: Performance in Detecting Nuanced Fake News. *Computers* **2024**, *13*, 292. https://doi.org/10.3390/computers13110292

Academic Editor: Ming Liu

Received: 8 October 2024
Revised: 4 November 2024
Accepted: 6 November 2024
Published: 11 November 2024

1. Introduction

The proliferation of digital media and the widespread use of social networks have drastically transformed the way information is disseminated and consumed. However, this rapid expansion has also led to the unintended consequence of an increase in the spread of misinformation and fake news, which pose significant risks, including the lack of societal trust, low trust in democratic processes, and public safety. For instance, platforms like Facebook and Twitter enable information to reach millions of users within minutes; a study from 2021 [1] found that 59% of U.S. adults receive news from social media at least occasionally, and misinformation on platforms like Facebook can reach over 100 million users in a single day. During the COVID-19 pandemic, false claims about the virus spread rapidly online, with one study estimating that misinformation reached nearly 25% of all Twitter users within the first few weeks of the outbreak. More drastically, since 2023 until the present, the Israeli–Hamas conflict further demonstrates the role of social media in rapidly spreading both information and misinformation. *The Washington Post* found that viral videos and unverified claims regarding the conflict spread widely across platforms like X and TikTok within hours, reaching millions of users globally [2]. For instance, a miscaptioned video that falsely depicted an event as part of the conflict was viewed over 1.2 million times before fact-checkers corrected it, illustrating the challenge of controlling misinformation during crises.

During these troubling times, the current project proposes a large language model approach using a LLaMA 3 architecture for a bilingual (Romanian and English) multiple-class automatic claim detection as opposed to a binary class detection of fake versus

true claims. Binary class (true vs. false) detection strategies have been attempted by more research teams using natural language processing, transformer models, and similar machine learning models [3], obtaining quite varying accuracy scores over controlled datasets.

Our research aims to explore the capability of LLMs, with a centered approach on Meta's LLaMA 3 LLM [4], when confronted with claims that belong to four classes (true (Adevărat), false (Fals), mostly true (Parțial Adevărat), mostly false (Parțial fals/Trunchiat)) and are made in two different languages (English and Romanian). The motivation for the added parameters is to more accurately mimic some of the real-world problems posed by disinformation campaigns [5]. Multi-class and multilingual fake news classifications are increasingly recognized as essential for effectively identifying the nuanced nature of misinformation in an intercultural ecosystem [6]. Traditional binary classification models, which categorize claims simply as "True" or "False", often fall short in capturing the full spectrum of misinformation types that exist in real-world scenarios as fake news frequently involves partial truths, exaggerations, or subtle misinformation that does not fit neatly into a true-or-false binary [7]. For instance, claims that are nuanced, like "Mostly True" or "Mostly False", represent a significant proportion of misinformation [8] as they include factual elements but are distorted or presented in misleading ways. In these cases, binary models may either misclassify or overlook such claims, potentially allowing misleading information to spread unchecked [9]. Multi-class classification, in contrast, allows models to distinguish between these varying levels of truthfulness, making it possible to address a broader range of misinformation types [10]. Even the term "fake news" has become overly broad and is considered technically imprecise, as it encompasses a wide range of media types. A more targeted approach, focusing on misinformation and disinformation, currently includes at least five distinct categories [11]. This nuanced approach aligns more closely with the real-world complexity of fake news, where claims may be exaggerated, partially accurate, or lacking critical context. By categorizing misinformation into different classes, such as "True", "Mostly True", "Mostly False", and "False", multi-class models provide a more granular understanding, which can significantly enhance efforts to combat fake news. Such classification contributes to more effective decision making in fact checking, content moderation, and public communication and thus adds to the overall reliability and transparency of information in digital media.

Related Works: The automatic detection of fake news has gained significant attention in recent years [12], benefitting by the development of advanced AI models that could be adapted to identify false information [13]. Initially, different research teams adapted models used in identifying deception or similar supervised learning models for fake news analyses [14]. The deep learning model architecture proved to have a greater success in this field than the other adapted methods [15] as transformer-based architectures emerged and evolved [16]. Such models are designed to identify deception trough natural language processing methods [17] and have shown success in scenarios like the COVID-19 crisis [18] by using different methods for detecting fake media that include text classification [19], authorship labeling, propagation, or a sentiment analysis [20]. On a more recent note, the introduction of large language models (LLMs) [21] showed promise in the automatization of different fact-checking sub-tasks like the classification of news [22] and may become the next step in refining the automation of fake news detection. LLMs are a family of advanced machine learning models designed to understand and generate human language and represent a significant advancement in NLP [23]. Their ability to process and generate human language with accuracy and flexibility makes them indispensable tools in the field of AI. These models are built on deep learning architectures involving billions of parameters, enabling them to perform a wide range of natural language processing (NLP) tasks such as text generation, translation, summarization, and question answering [24]. The development of LLMs can be traced back to the introduction of neural networks and the advancement of deep learning techniques in the early 2010s [25]. The evolution of these models has been marked by significant milestones like the introduction of Word2Vec (2013) [26] and GloVe (2014) [27] or Recurrent Neural Networks (RNNs) [28] and their

variants, such as Long Short-Term Memory (LSTM) networks [29]. The introduction of the transformer architecture revolutionized NLP by eliminating the need for sequential data processing, allowing for parallelization and greater scalability, and became the foundation for most modern LLMs [30]. Following the transformer, models like BERT (Bidirectional Encoder Representations from Transformers, 2018) [31] and GPT (Generative Pre-trained Transformer) [32] emerged. The architecture of LLMs is primarily based on the transformer model, which includes a self-attention mechanism, Multi-Head Attention, Feed-Forward Neural Networks, Layer Normalization and Residual Connections, and positional encoding [33].

Following these considerations, the research questions posed in this project involve how well does the LLM model family perform in identifying partial fakes and misplaced facts.

Hypothesis 1. *Our proposed LLaMA 3 model will achieve higher accuracy and precision across various performance metrics in fake news detection compared to its predecessors and similar LLMs.*

Hypothesis 2. *Our proposed LLaMA 3 model will provide a higher capability of identifying more nuanced categories.*

2. Materials and Methods

In this section, we outline the methodologies employed to obtain significant results. We begin with a review of the relevant literature to establish the background and context of our study. Following this, we provide a detailed description of the specific methods utilized in this study.

2.1. Model Description

The LLaMA 3 model is an LLM still based on the transformer architecture, known for its ability to handle complex language tasks through self-attention mechanisms. It includes models of varying sizes, with the most prominent being the 8B and 70B parameter versions [4]. The architecture of the model builds on the transformer backbone, known for its ability to handle complex language tasks through self-attention mechanisms. This foundational design allows the model to process and generate language efficiently by focusing on different parts of the input text based on relevance. A key feature in this model is Grouped-Query Attention (GQA), an optimization technique specifically designed to improve inference efficiency. GQA reduces the computational complexity of self-attention by grouping similar queries together, allowing the model to make more efficient use of resources without compromising accuracy. This is particularly beneficial in scenarios requiring high-speed processing, such as real-time fact checking. In addition to GQA, the model employs Low-Rank Adaptation (LoRA) for fine tuning. LoRA modifies only the most impactful parameters, particularly in linear layers, rather than adjusting the entire model, making the fine-tuning process more efficient. This approach, paired with mixed-precision (FP16) training and Int8 quantization, reduces the model's computational footprint, enabling deployment in resource-constrained environments while preserving performance. To further enhance stability during training, the model uses Layer Normalization and Residual Connections, which help to stabilize gradients and improve convergence. Positional encoding is also incorporated to provide information about the relative position of tokens, which aids the model in understanding sentence structure, especially in complex multilingual tasks [34].

2.2. Methods

2.2.1. Datasets

Data Collection

The models in this project, as represented in Figure 1, were fine-tuned on 2 sets of data. *An English dataset* [35] contained 19,422 records of various data scrapped from Politifact.com

(accessed on 11 April 2021), with 12 columns describing the person associated with the quote, the date of the quote, the platform/setting in which the information was provided, the specific claim/statement, the reviewer of the claim, the analysis date, the classification of the claim (the 6 classes are defined as true, mostly true, half-true, false, mostly false, and pants-on-fire), the link to the fact-checking article, the headline of the article, its complete text, and the associated tags. *A Romanian dataset* was obtained from different sources, based on the https://huggingface.co/datasets/readerbench/ro_fake_news (accessed on 20 August 2024) Dataset, containing 982 columns. The datasets used in this study were curated from reputable fact-checking sources to ensure a high level of reliability. However, as with any dataset, there is the potential for biases that could influence the model's performance, especially in the context of political content or skewed representations. Efforts were made to minimize these biases by sourcing data from independent fact-checking organizations, which adhere to standards aimed at objective reporting. Nevertheless, the presence of implicit biases, such as those related to regional or cultural perspectives, remains a possibility as they may reflect the dominant political narratives or social issues pertinent to the US or Romania.

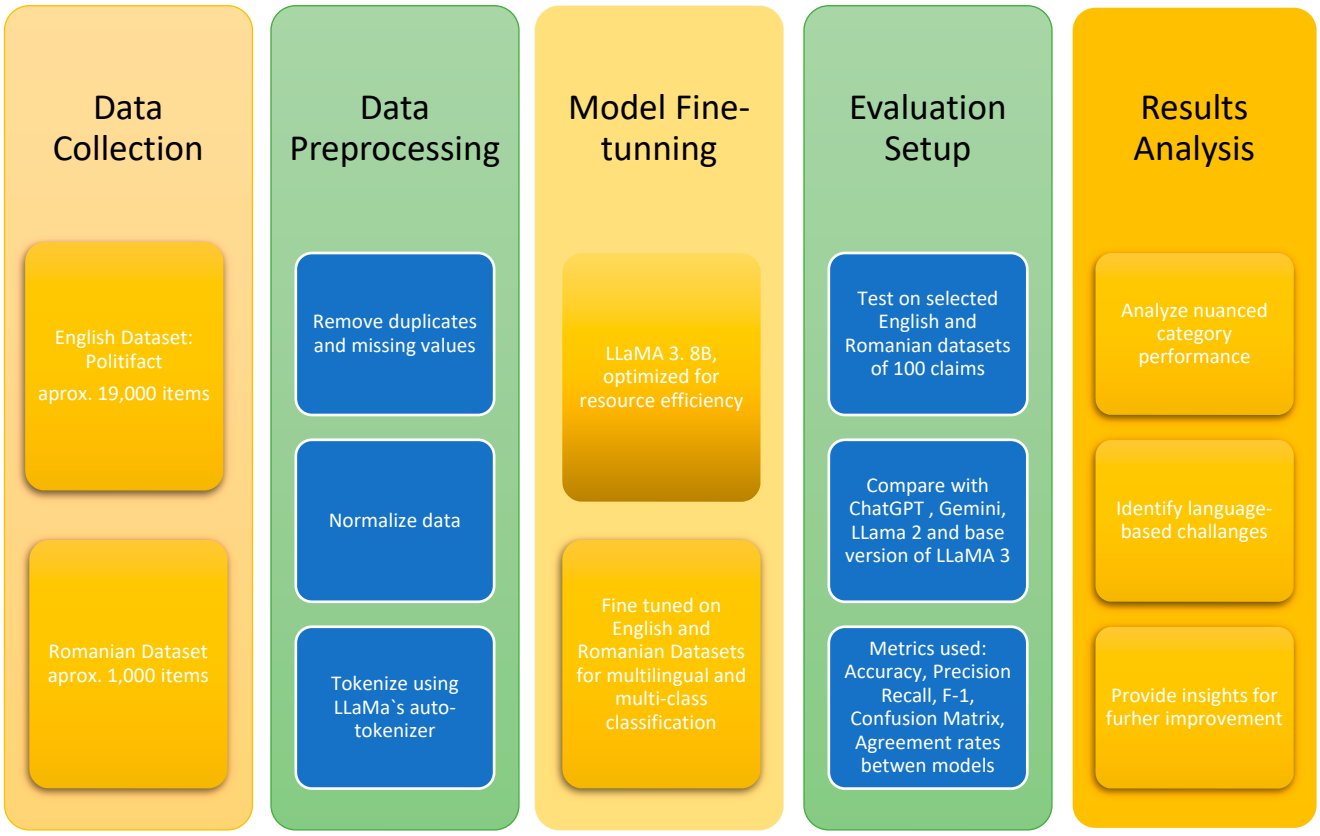

Figure 1. This figure represents the overall methodology employed in this research, detailing the sequential approach taken to compare a specialized version of the LLaMA 3 model with other leading LLMs for detecting nuanced fake news. The figure provides a visual roadmap, covering the principal stages presented in the following sections such as dataset preparation, model fine tuning, testing, and evaluation methods.

Data Processing

Both datasets were optimized for Supervised Fine Tuning of the 3 8B LLaMA variant with the goal of adapting the model to perform better by learning from the labeled examples. The datasets were cleaned by removing duplicates and missing values, normalized, padded, and tokenized using LLaMA's 3 auto-tokenizer. The fact columns were transformed into categorical values.

2.2.2. Proposed Model Architecture

The LLaMA 3 8B model was fine-tuned, as illustrated in Figure 1, for the task of fake news fact checking using a comprehensive and methodologically sound approach. The model was adapted through Parameter-Efficient Fine Tuning (PEFT), leveraging Low-Rank Adaptation (LoRA) to efficiently integrate additional parameters, specifically targeting the model's linear layers. This fine-tuning process was optimized using mixed precision (FP16) to balance computational efficiency and accuracy, alongside Int8 quantization to reduce the model's footprint, making it more suitable for deployment in resource-constrained environments [36]. The fine tuning employed the AdamW optimizer with a linear learning rate scheduler, set to an initial learning rate of 0.00003 with a warmup phase covering 10% of the training. To manage the gradient flow, gradient accumulation steps were set to 4, with a max gradient norm of 1 to prevent instability [36]. Training was conducted over 4 epochs with small batch sizes to ensure careful learning from the dataset, allowing the model to adapt effectively to the nuances of fake news detection [36].

This combination ensured that the proposed LLaMA 3 model was not only fine-tuned effectively but also optimized for practical deployment scenarios, ensuring that it can handle the complexities of multilingual, multi-class fake news detection with high efficiency.

2.2.3. Environment

For the model training, the NVIDIA A10G computing environment provided by Hugging Face was used. The testing of all the LLaMA models was achieved using the same NVIDIA L4 GPU-powered environment, featuring 24 GB of memory, in order to have the exact same conditions for all the tested models. The 3.12 Python version was used.

2.3. Evaluation Methods and Compared Models

2.3.1. Models Used in the Research

In the evaluation of the proposed model, as shown in Figure 1, several state-of-the-art large language models were included for comparison to assess their performance in the context of fake news detection. The models evaluated alongside it include

- Meta's (Menlo Park, CA, USA) LLaMA 2 13B: A predecessor in the LLaMA series, LLaMA 2 13B has a larger parameter count at 13 billion, providing enhanced capabilities in natural language understanding and generation [37]. This model serves as a benchmark to assess the improvements made in the LLaMA 3 series, particularly in how the newer architecture handles complex tasks like multilingual and multi-class claim checking.
- Meta's LLaMA 3 8B base model, an intermediate-sized model from Meta's LLaMA 3 family, featuring 8 billion parameters, which was used for fine-tuning our proposed model, was included as a benchmark in this research [4].
- OpenAI's (San Francisco, CA, USA) ChatGPT 4: The fourth iteration of OpenAI's GPT series, ChatGPT 4, is one of the most known models in the LLM family for its ability to handle complex language tasks with improved accuracy and coherence. It features billions of parameters that enable it to generate contextually rich and semantically accurate text [38].
- ChatGPT 4.o: The state-of-the-art version in the ChatGPT series is often distinguished by fine tuning for specific use cases or enhanced features in certain environments. This version may focus on specific optimizations or enhancements for improved inference speed or specific task performance, making it a relevant comparison point in understanding the capabilities of our proposed model [39].
- Google's (Menlo Park, CA, USA) Gemini is a robust language model developed with a focus on multilingual capabilities and enhanced inference efficiency; known for great performance across different languages, including lesser-resourced ones, Gemini is particularly valuable in settings requiring diverse linguistic understanding and rapid adaptation to various text-based tasks [40].

The LLaMA models were tested under the same environmental conditions and the ChatGPT models and Gemini were tested using a black box testing method in their native platforms.

2.3.2. Testing Datasets

For the testing, the 2 datasets were compiled, one in English, and one in Romanian, from different fact-checking organizations like Politifact.com or Factual.ro. Each dataset consisted of 100 claims, organized in 4 categories (25 claims per category)—true (Adevărat), false (Fals), mostly true (Parțial Adevărat), and mostly false (Trunchiat). All news items being provided to the LLMs dated from 2021 to 2024. While the test selection is balanced for both datasets, one must consider that the Romanian dataset used for the creation of the test sample includes a considerably smaller volume than the similar English dataset. This discrepancy in dataset size reflects a wider limitation in available Romanian-language resources and is more likely to introduce variability in model performance across the two languages. While efforts were made to balance the classes within each dataset, certain categories, particularly "Mostly True" and "Mostly False", are generally underrepresented in Romanian. This underrepresentation is partly due to the difference in the number of fact-checking organizations that address English versus Romanian fake news and are able to cope with the complexity and subjectivity involved in evaluating claims that fall within nuanced categories. Romanian specialized fact-checking groups are much fewer than their English counterparts. This is also why, in contrast, the English dataset has a more uniform distribution of classes, benefiting from the extensive fact-checking efforts in English-speaking regions. The imbalances within the Romanian dataset have a higher chance to affect the model's ability to generalize, potentially leading to overfitting or reduced accuracy for underrepresented classes, and thus could impede the model's ability to handle partial truths or subtler inaccuracies, reducing the overall robustness in multi-class scenarios. Also, news items that have not been verified by independent fact-checking agencies, regardless of the media outlet that posted them, are excluded from the pool of news items used in this study. This measure is implemented to minimize framing bias that could be introduced by media organizations, whether they are state-owned, publicly owned, or privately operated for profit. For the purposes of this experiment, we rely solely on independent fact checkers as the definitive source of truth. The datasets can be found in [41].

2.3.3. Testing Procedure

The LLMs were evaluated using a set of prompts derived from the collected news headlines, with each prompt crafted to elicit a response that could be classified as true, false, or partially true/false. These prompts were presented to the LLMs in a randomized order to prevent any potential order effects. The legitimacy of the test items was categorized into four distinct groups. Unlike similar studies that often rely on binary classification, our simulation utilized these four categories—True (Adevărat), False (Fals), Mostly true (Parțial Adevărat), and Mostly False (Trunchiat)—reflecting the fact-checked content typically produced by third-party agencies. Although the "Mostly True" and "Mostly False" categories can be ambiguous and potentially confusing, which might lead to lower accuracy scores, our objective was to specifically assess the LLMs' ability to navigate these ambiguous classes. To address the potential for classification ambiguity in the LLMs' responses, we explicitly instructed each LLM to choose from these four categories (true, false, mostly true, and mostly false) within the prompts, ensuring that the responses were as clear and specific as possible [42].

2.3.4. Evaluation Metrics

For this project, we used the following metrics to compare the six models:

The accuracy score was used to determine the performance of the 6 models included in the benchmark as it measures the proportion of correctly identified news by each model in

relation to the Correct Category (refers to the accurate classification of a claim according to its intended label, such as "True", "False", "Mostly True", or "Mostly False"; for each claim in the dataset, a predefined label represents the "ground truth", or the correct classification, based on the fact checking and analysis). It is the most straightforward statistic to compare model performance [43].

$$\text{Accuracy} = (\text{Number of Correct Predictions})/(\text{Total Number of Predictions}) \quad (1)$$

The precision, recall, and F1 score allow us to consider both false positives and false negatives and are calculated as shown in Equations (2)–(4) [29].

$$\text{Precision} = (\text{True Positives})/(\text{True Positives} + \text{False Positives}) \quad (2)$$

$$\text{Recall} = (\text{True Positives})/(\text{True Positives} + \text{False Negatives}) \quad (3)$$

$$\text{F1-Score} = 2 \times (\text{Precision x Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

The confusion matrix is calculated as shown in Equation (5). The scores for each model underline and improve the understanding of the specific types of errors (false positives, false negatives) that each model is making [31].

$$\text{Formula: [[True Positives, False Positives], [False Negatives, True Negatives]]} \quad (5)$$

Agreement rates between models show how often different models agree with each other and with the correct classification. Their Calculation (6) would provide insight into the consistency and reliability of all these models [44].

$$\text{Agreement Rate} = (\text{Number of Times Models Agree (on correct label)})/(\text{Total Number of Predictions}) \quad (6)$$

3. Results

3.1. English Dataset

Table 1 shows that ChatGPT 4 achieved the highest overall accuracy at 54%, followed by Gemini at 51%, while the LLaMA 3.1 8B base had the lowest accuracy at 26%, indicating that it struggled more than the other models. Our proposed model outscored the other versions in the overall score. The range of accuracy suggests that while all models performed similarly, their abilities to classify correctly have a large improvement range.

Table 1. English Dataset. Overall Accuracy Scores by Model.

Model	Overall Accuracy
ChatGPT 4	**54.00%**
ChatGPT 4.o	45.00%
Gemini	51.00%
LLaMA 2–13b	40.00%
LLaMA 3–8B base	26.00%
LLaMA 3–8B fine-tuned	43.00%

According to Table 2, while ChatGPT 4 performed exceptionally well in the "True" and "False" categories, achieving 100% and 96% accuracy, respectively, the model struggled significantly in the nuanced categories. Surprisingly, four versions outclassed the 4.o version in every category. Gemini demonstrated a more balanced performance, with decent accuracy across the first three classes, but also failed to correctly classify any "Mostly false" cases.

Table 2. English Dataset. Accuracy Scores by Model for each Class.

Model/Accuracy of Category	True	False	Mostly True	Mostly False
ChatGPT 4	**100.00%**	**96.00%**	20.00%	0.00%
ChatGPT 4.o	76.00%	84.00%	20.00%	0.00%
Gemini	64.00%	**96.00%**	44.00%	0.00%
LLaMA 2–13b	48.00%	76.00%	24.00%	12.00%
LLaMA 3–8B base	28.00%	68.00%	8.00%	0.00%
LLaMA 3–8B fine-tuned	40.00%	68.00%	**48.00%**	**16.00%**

Our proposed model showed the best accuracies when dealing with the nuanced classes (48% for mostly true and 16% for mostly false) and outclassed the other LLaMA models compared. The "Mostly false/Trunchiat" category appears to be particularly challenging for all models, with no model scoring above 16% accuracy, indicating a significant area for improvement.

The metrics shown in Table 3 provide further insight into the performance of the models, with F1 scores reflecting the balance between precision and recall scores showing the ability of the model to identify all relevant instances of a class. The LLaMA 3–8B fine-tuned model has a balanced F1 score and recall, indicating that it is relatively good at identifying correct categories but with some trade-offs in precision when compared to the three commercial models but is superior to its base version and the older version, even if the latter has more training parameters.

Table 3. English Dataset. Overall F1, Precision, and Recall Scores by model.

Model	F1 Score	Precision Score	Recall Score
ChatGPT 4	0.423	0.473	0.54
ChatGPT 4.o	0.363	0.326	0.45
Gemini	0.437	0.430	0.51
LLaMA 2–13b	0.368	0.379	0.40
LLaMA 3–8B base	0.196	0.173	0.26
LLaMA 3–8B fine-tuned	0.409	0.422	0.43

According to Figure 2, ChatGPT 4 has the most accurate classifications for "True" and "False", with minimal misclassifications across the board, showing strong performance in binary classifications, but struggles with the nuanced classes where it scored 0 across the board. ChatGPT 4.o and Gemini have similar patterns but show more misclassifications, particularly between "True" and "Mostly True" categories, as well as 0 across the board with "Mostly False". LLaMA 2–13b and the LLaMA 3–8B base show more spread in their misclassifications, indicating difficulties in distinguishing between all categories. Notably, these models often confuse "Mostly False" with "False" and "True".

Our proposed model has obtained a slightly more balanced confusion matrix than the base version, but it still struggles, particularly in distinguishing "True" from "False", mostly by replacing them with the nuanced categories, handling "Mostly False" the best between the compared models but still struggling with this category. On the other hand, its ability to correctly identify "Mostly True" statements is a standout, particularly against models like ChatGPT 4.o and the LLaMA base models, which struggled more in this area.

As shown in Table 4, the highest agreement is between ChatGPT 4 and Gemini (0.553), indicating that these two models tend to classify in similar ways more often than other pairs. The agreement rates between ChatGPT models and the LLaMA models are generally lower, with the lowest being between ChatGPT 4.o and the LLaMA 3–8B base (0.145), suggesting that these models have quite different classification behaviors. Among the LLaMA models, the agreement is moderate but still relatively low, with the highest being Gemini vs. LLaMA 3–8B fine-tuned (0.312). These agreement rates suggest that while there is some consistency between the models, particularly within the same family (e.g., ChatGPT models), there are significant differences in how they classify the same data.

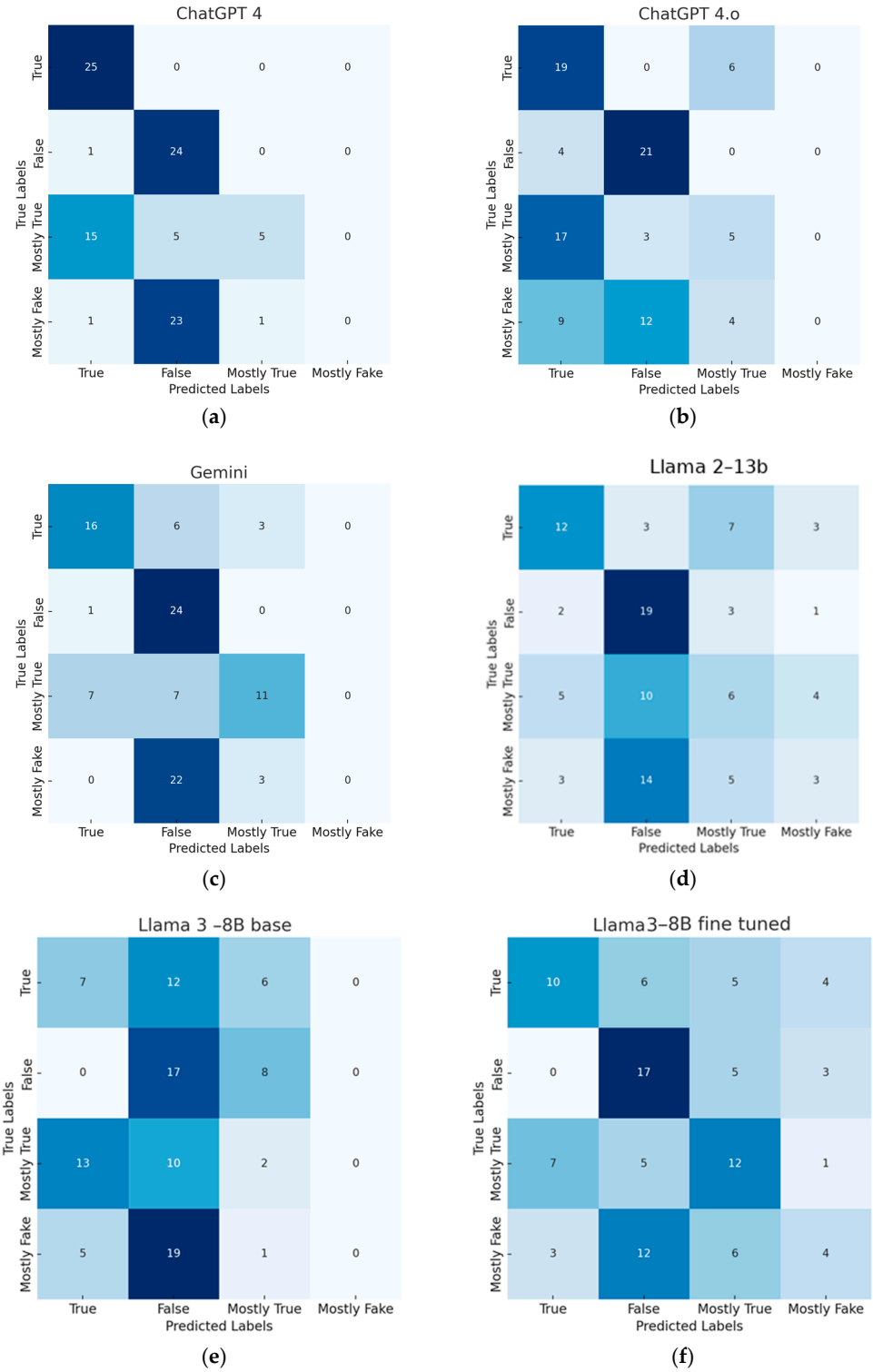

Figure 2. The English dataset. Confusion matrices by model. Panels: (**a**) ChatGPT 4; (**b**) ChatGPT 4.o; (**c**) Gemini; (**d**) LLaMA 2–13b; (**e**) the LLaMA 3–8B base; (**f**) LLaMA 3–8B fine-tuned. The confusion matrix is a valuable tool for interpreting model performance, showing the model's strengths and misclassification patterns across the four categories. For instance, ChatGPT 4 performs accurately in the "True" and "False" categories but shows frequent misclassifications in the nuanced "Mostly True" and "Mostly False" categories (often defaulting these to binary categories) while the proposed fine-tuned LLaMA 3 model is more accurate with "Mostly True" claims, but shows difficulty in reliably distinguishing "Mostly False" statements, though it achieves a more balanced spread in its misclassifications compared to other models.

Table 4. English Dataset. Agreement Rates between Models.

Model/Agreement Rates	ChatGPT 4.o	Gemini	LLaMA 2–13b	LLaMA 3–8B Base	LLaMA 3–8B Fine-Tuned
ChatGPT 4	0.515	0.553	0.268	0.194	0.262
ChatGPT 4.o	x	0.364	0.209	0.145	0.260
Gemini	x	x	0.344	0.209	0.312
LLaMA 2–13b	x	x	x	0.237	0.200
LLaMA 3–8B base	x	x	x	x	0.224

3.2. Romanian Dataset

Table 5 shows that for the Romanian dataset, at 39%, our proposed model outperforms all other models, including its base version and other models like Gemini and ChatGPT. The increase in accuracy from 27% (base version) to 39% (fine-tuned) indicates that the proposed model has adapted better to the dataset's characteristics after fine tuning. Both LLaMA 2–13B and the LLaMA 3–8B base have an accuracy of 27%. These scores are in the middle range and indicate that these models are moderately effective in this specific task, though they also leave room for improvement, while ChatGPT 4.o has the lowest accuracy. The notable decrease in accuracy for the Romanian dataset can be attributed to several factors that impact the model's performance. The training data available for Romanian fake news are substantially smaller and less diverse than its English counterpart and the unique linguistic challenges presented by the Romanian language, such as complex inflections, flexible syntax, and culturally specific idioms, which are less prevalent in English, introduce greater variability and reduce classification accuracy. The presence of nuanced categories like "Mostly True" and "Mostly False" further complicates this task, as these distinctions require a more sophisticated understanding of partial truths—something that the model, trained primarily on English, may not fully capture in Romanian [34]. Addressing factors like expanding the Romanian dataset, diversifying class representations, and incorporating more language-specific pre-processing steps would likely enhance the model's ability to handle Romanian claims with more accuracy and improve performance in future work.

Table 5. Romanian Dataset. Overall Accuracy Scores by Model.

Model	Overall Accuracy
ChatGPT 4	25.00%
ChatGPT 4.o	21.00%
Gemini	33.00%
LLaMA 2–13b	27.00%
LLaMA 3–8B base	27.00%
LLaMA 3–8B fine-tuned	**39.00%**

According to Table 6, the ChatGPT 4 model shows a bias on true versus false statements for this dataset (60% versus 0%) and performs poorly on the nuanced classes with 20% accuracy each, showing challenges in identifying partially true or truncated information while the 4.o model shows a balanced, though generally low, performance across all classes. Similarly to ChatGPT 4, it struggles with "Mostly True" (8%) but has the highest score in "Mostly False" (32%). Gemini exhibits the highest accuracy on "False" with 64% but fails entirely on "Mostly False" with 0% accuracy, indicating difficulty in this category. The LLaMA 2–13b shows a balanced but low performance across all categories. The base version of LLaMA 3 shows a balanced performance with the highest accuracy for "True" (56%) but fails entirely on "Mostly false", similar to Gemini. The proposed model shows the most balanced and generally better performance across all categories. It performs particularly well on "Mostly True" with 68% accuracy and shows the second-best score for but struggles somewhat with "Mostly False" (20%). Overall, this model seems to offer the most balanced performance, showing improvement over its base model, even if it still has issues, similar to the English dataset, in identifying clear truths and clear falsehoods.

Table 6. Romanian Dataset. Accuracy Scores by Model for each Class.

Model/Accuracy of Category	True	False	Mostly True	Mostly False
ChatGPT 4	**60.00%**	0.00%	20.00%	20.00%
ChatGPT 4.o	20.00%	24.00%	8.00%	**32.00%**
Gemini	36.00%	**64.00%**	32.00%	0.00%
LLaMA 2–13b	24.00%	44.00%	24.00%	16.00%
LLaMA 3–8B base	**56.00%**	40.00%	12.00%	0.00%
LLaMA 3–8B fine-tuned	32.00%	36.00%	**68.00%**	20.00%

The metrics shown in Table 7 show that the proposed model is the standout performer, with the highest precision, recall, and F1 scores. This indicates that fine tuning significantly enhanced the model performance, making it a more robust choice for accurate classification tasks. Gemini and LLaMA 2–13b perform well, particularly in recall, but they fall short of the fine-tuned model in precision. ChatGPT 4 and ChatGPT 4.o lag behind, indicating possible biases.

Table 7. Romanian Dataset. Overall F1, Precision, and Recall Score by model.

Model	F1 Score	Precision Score	Recall Score
ChatGPT 4	0.199	0.188	0.25
ChatGPT 4.o	0.206	0.208	0.21
Gemini	0.287	0.257	0.33
LLaMA 2–13b	0.264	0.269	0.27
LLaMA 3–8B base	0.216	0.202	0.27
LLaMA 3–8B fine-tuned	**0.377**	**0.424**	**0.39**

According to Figure 3, both ChatGPT 4 and ChatGPT 4.o tend to misclassify a large number of instances into other categories, especially "False" and "Mostly True". This is particularly evident in ChatGPT 4, where there are no correct predictions for "False". Google's Gemini model performs better, particularly in identifying "False" and "Mostly True" categories. However, it still shows confusion between "Mostly True" and "Mostly False". The two base LLaMA models, LLaMA 2–13b and the LLaMA 3.0–8B base, show a moderate level of confusion across categories, particularly between "True" and "False". The LLaMA 3 base version struggles more with "Mostly False", often misclassifying these as "False". Our proposed model, the LLaMA 3–8B fine-tuned, shows the most balanced performance, correctly identifying a higher number of instances across all categories. However, it still shows some confusion between the nuanced "Mostly False" and "Mostly True" categories, but proves great improvement over its base version.

As shown in Table 8, the agreement rates between the models mostly fall in a relatively narrow range, from 31% to 34%. This indicates that while there is some variation, the models are not drastically different in their predictions. The proposed model has a moderate level of agreement with the other models; its highest agreement rate is with the "Gemini" model at 33%, which may suggest some similarities between their prediction patterns. However, the agreement rate between "LLaMA 3–8B fine-tuned" and the "LLaMA 3–8B base" is notably lower than what might be expected between a base model and its fine-tuned version (around 32%), indicating that the fine-tuning process introduced significant changes. The "ChatGPT 4" and "ChatGPT 4.o" models show the highest agreement with each other, which makes sense as they are likely very similar versions. This suggests once again that the updates from 4 to 4.o might not have drastically changed the model's predictions. Overall, the models share some commonalities but there are meaningful differences, especially with the fine-tuned model, which appears to have a unique approach compared to both its base version and the other models. Compared to the English dataset, the agreement rates are inferior across the table.

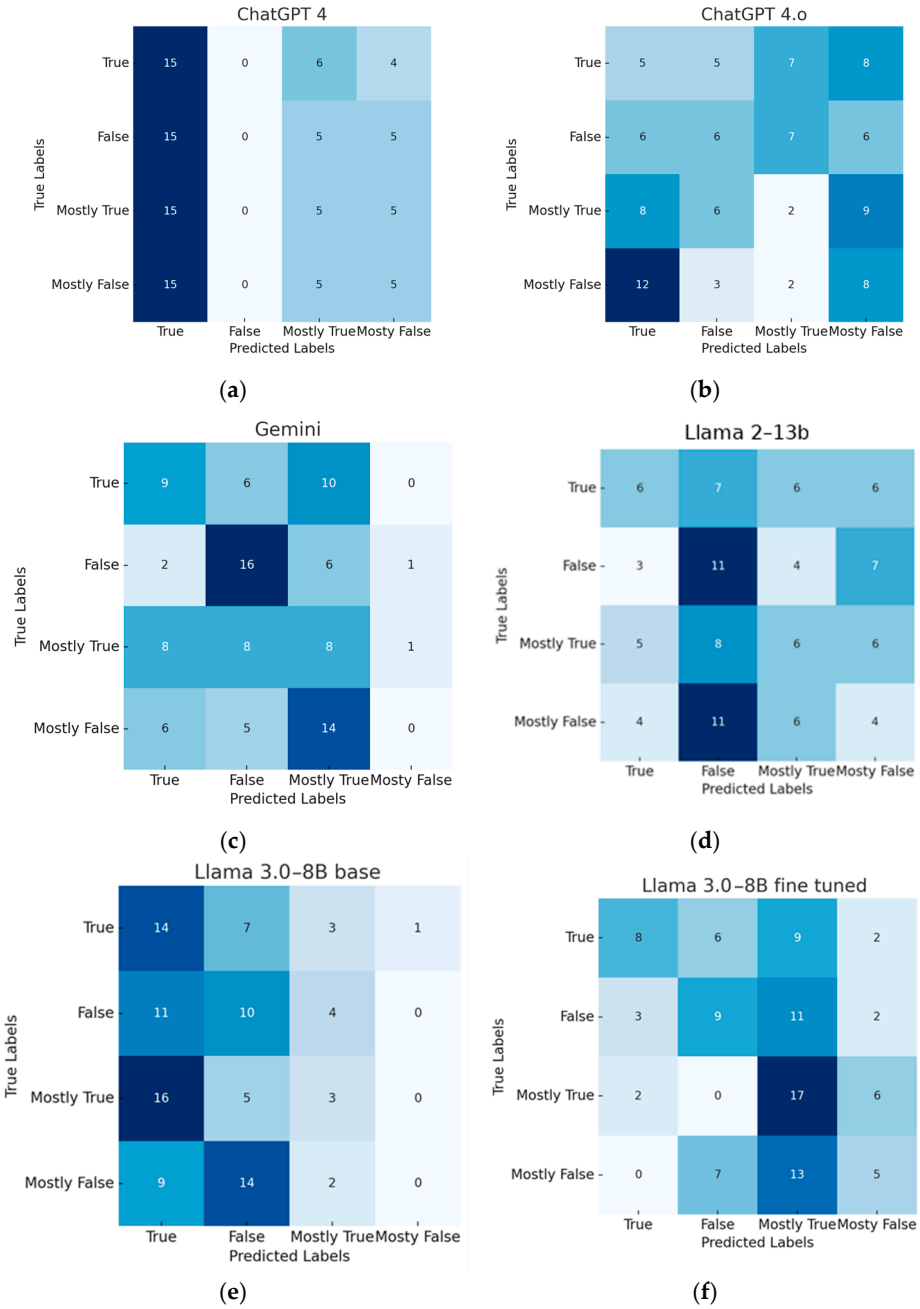

Figure 3. The Romanian dataset. Confusion matrices by model. Panels: (**a**) ChatGPT 4; (**b**) ChatGPT 4.o; (**c**) Gemini; (**d**) LLaMA 2–13b; (**e**) the LLaMA 3–8B base; (**f**) LLaMA 3–8B fine-tuned.

Table 8. Romanian Dataset. Agreement Rates between Models.

Model/Agreement Rates	ChatGPT 4.o	Gemini	LLaMA 2–13b	LLaMA 3–8B Base	LLaMA 3–8B Fine-Tuned
ChatGPT 4	0.34	0.22	0.17	0.34	0.21
ChatGPT 4.o	x	0.24	0.31	0.20	0.25
Gemini	x	x	0.27	0.30	0.33
LLaMA 2–13b	x	x	x	0.32	0.30
LLaMA 3–8B base	x	x	x	x	0.29

4. Discussion

The analysis of the English and Romanian datasets reveals several insights that directly relate to the broader challenges and objectives of this study and offer a better understanding of the performance and limitations of the LLaMA 3 proposed model (not exclusively),

particularly in a bilingual, multi-class claim detection task. The most important fact is that the accuracy scores are very low for both datasets and the models have problems adapting to a real scenario involving more nuanced approaches and to properly evaluate scenarios in a multilingual setup. One such derived observation is the impact of language and cultural differences on model performance. The Romanian dataset, due to its linguistic and cultural nuances, poses a more significant challenge for the models compared to the English dataset as proven by the lower scores obtained by all the models compared. Romanian's complex inflectional morphology and frequent use of compound and inflected words pose unique challenges for language models, especially those primarily trained on English data. Even with fine tuning, the LLaMA 3 model struggled with nuanced categories in Romanian, highlighting a broader challenge in multilingual NLP: models must adapt not only to vocabulary changes but also to the structural and semantic subtleties that define language-specific claims. Romanian, for example, has flexible word order that can shift to emphasize different parts of a sentence. A phrase like "Este adevărat că el a spus asta" (It is true that he said this) can be rephrased as "El a spus asta, este adevărat" to change emphasis, which can affect how the statement's truthfulness or relevance is perceived. Models may struggle to interpret these subtle shifts, as they often alter the intended meaning. Another challenge arises from Romanian's reliance on inflections to convey meaning and context, especially in truth statements. For instance, "adevărat" (true) conveys an objective claim, while "adevăr" (truth) suggests a subjective or generalized truth. Similarly, "Acesta este adevărul" (This is the truth) and "Acesta este adevărat" (This is true) can imply different claim types, which models may overlook. Romanian also commonly uses double negations, adding complexity. The phrase "Nu este neadevărat" (It is not untrue) introduces ambiguity compared to the straightforward "Este adevărat" (It is true), which may lead models to misinterpret or oversimplify the statement. Further complicating matters, cultural and contextual nuances like idioms or the gradation of truthfulness often lacks direct English equivalents, adding interpretive challenges. For instance, "mostly true" may translate as "parțial adevărat" in Romanian, but it implies a higher level of truthfulness. Likewise, "trunchiat" (truncated) is used for "mostly false" but carries connotations of deliberately omitting parts of the truth, suggesting a degree of intent absent in the English version. Another limitation posed by the Romanian dataset may be concerning the limitations in data resources. This further exacerbates model performance issues. While the English language benefits from extensive, better-quality datasets that capture a wide variety of expressions and claim types, Romanian datasets are relatively limited in both volume and diversity. For instance, English datasets used in this study included more than 19,000 records, while Romanian datasets had fewer than 1000. This disparity limits the model's capacity to generalize nuanced claim detection in Romanian, particularly in real-world applications where fine-grained distinctions are essential for reliability. This disparity in model accuracy between the two languages suggests that while LLMs like LLaMA 3 are powerful, their effectiveness can vary significantly depending on the language, highlighting a potential area for improvement in multilingual model training. The consistency of model performance across languages, as shown in the two analyses, illustrates that the agreement rates between models in the Romanian dataset were 0 across all six models and generally lower than in the English dataset. This suggests that even state-of-the-art models like ChatGPT 4.o and Gemini may not be as reliable when applied to Romanian text. This inconsistency indicates that while these models may have robust general language processing capabilities, their performance in specific languages can vary, affecting their reliability in multilingual contexts. The fine tuning applied on the LLaMA 3 model demonstrated a substantial impact on model performance, particularly in the Romanian dataset. The improvement in accuracy from 27% (base version) to 39% (fine-tuned version) illustrates how fine tuning can enhance a model's ability to handle the specific characteristics of a dataset. This is still undermined by the low scores obtained by all the models overall. If we were to launch any of these models in a real disinformation mitigation system, they would be unreliable, as shown by the distribution of categories across the datasets. The models, including the fine-tuned version,

struggled with the nuanced "Mostly True" and "Mostly False" categories, particularly in the Romanian dataset. This difficulty highlights the challenges that models face in dealing with ambiguous or partially true claims, a common issue in misinformation detection. The biases observed, predominantly in the ChatGPT family, show a tendency toward a binary classification of true/false statements. Even if, in some cases, the models were given the exact source of the statement, they would refuse to change their classification of the claim, thus potentially misleading a fact-checker or the user. The performance differences between the models in the English and Romanian datasets may also point to challenges related to translation and interpretation. More times than once, the models would assign a wrong connotation if the prompt pointed exclusively to Romanian classes like "Trunchiat". Models trained primarily on English data may misinterpret or oversimplify claims in Romanian, which leads to lower accuracy and agreement rates. Overall, while the proposed model was proven to have the overall best performance, particularly for the Romanian dataset, all the models gave surprisingly different outputs during the experiments and have many inconsistencies if the Romanian language is used; the insights gained from this comparative analysis of the English and Romanian datasets suggest that while LLMs like LLaMA 3 show promise, particularly after fine tuning, there is still significant room for improvement. Future work should focus on enhancing multilingual capabilities, refining models to better handle nuanced categories, and ensuring that cultural and contextual factors are more effectively integrated into the model training process. One such approach could be the integration of context-aware models that utilize transformers or attention mechanisms tailored to cultural and linguistic nuances. Future models could benefit from leveraging existing cultural knowledge datasets to enhance accuracy. Resources such as multilingual corpora with annotated cultural references [45], regional news datasets, and linguistically diverse fact-checking sources would help capture the unique attributes of each language. Also, a future enhancement could involve transfer learning from high-resource languages to lower-resource languages. By pre-training models on languages with robust datasets, then fine tuning on smaller, culturally relevant datasets (e.g., Romanian fake news), models can retain foundational language, understanding while adapting to the subtleties of less-resourced languages [46].

5. Conclusions

Considering the first hypothesis, that our proposed LLaMA 3 model would achieve higher accuracy and precision across various performance metrics in fake news detection compared to its predecessors and similar LLMs, the results provide partial support. While the fine-tuned LLaMA 3 model demonstrated significant improvement over its base version and older models, particularly within the Romanian dataset, where it achieved the highest accuracy among the models compared, it did not consistently outperform all other models across both datasets and performance metrics. The LLaMA 3 fine-tuned model achieved the highest accuracy in the Romanian dataset (39%), indicating its potential to outperform previous versions in certain contexts. However, in the English dataset, its accuracy was lower compared to models like ChatGPT 4 and Gemini. This suggests that while the LLaMA 3 model shows improvements, particularly after fine tuning, it does not consistently achieve higher accuracy across all datasets and contexts. On the other hand, the proposed model demonstrated strong precision and recall, particularly in the Romanian dataset, where it achieved the highest F1 score (0.377). This supports the idea that fine tuning significantly enhances the model's performance, especially in handling less common languages like Romanian. In conclusion, while the LLaMA 3 model shows notable improvements compared to all its predecessors, it does not achieve higher metrics across all the compared models. Therefore, Hypothesis 1 is partially supported by the evidence, with strong performance in some contexts but not universally across all scenarios.

For Hypothesis 2, that our proposed LLaMA 3 model would demonstrate a higher capability in identifying nuanced categories (such as "Mostly True" and "Mostly False") compared to its predecessors, the findings are fully supported. The fine-tuned LLaMA

3 model outperformed both its base version and earlier LLaMA models in distinguishing between nuanced categories, especially in the Romanian dataset. In the Romanian dataset, it achieved a notably high accuracy of 68% in the "Mostly True" category, which is significantly better than all other tested models. It also outperformed other models in the "Mostly False" category in both datasets, although the absolute accuracy levels in this category remain low across all models. The model's superior performance in these nuanced categories, particularly when compared to the base version and other LLaMA predecessors, indicates that it has a better grasp of the complexities involved in these types of classifications. This suggests that fine tuning has effectively enhanced the model's sensitivity to subtle differences between claims that are partially true or false. Therefore, Hypothesis 2 is supported by the evidence, as the LLaMA 3 model does show a higher capability of identifying nuanced categories compared to both its predecessors and other similar LLMs. This enhanced performance in handling complexity may be useful particularly in tasks that require more nuanced judgment like content moderation in social media (automatic labeling), legal claim or compliance reviews (assessment of the truthfulness and intent of statements), criminology studies (allows criminologists to identify information that might be technically correct but misleading in intent), and psychology (mainly on cognitive biases).

Author Contributions: Conceptualization, S.E.R.; methodology, S.E.R.; software, S.E.R.; validation, S.E.R.; formal analysis, S.E.R.; investigation, S.E.R.; resources, S.E.R.; data curation, S.E.R.; writing—original draft preparation, S.E.R.; writing—review and editing, S.E.R.; visualization, S.E.R.; supervision, R.B.; project administration, S.E.R.; funding acquisition, S.E.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Ro and En Datasets: Available online in Kaggle Datasets; https://www.kaggle.com/datasets/restem/en-ro-datasets-for-llm-testing (accessed on 28 August 2024). Fine-tuned version of Meta-LLaMA-3-8B, trained on two datasets: Available online in Hugging Face Hub; https://doi.org/10.57967/hf/2954 (accessed on 27 August 2024). Dataset compiled for article: Available online; https://doi.org/10.57967/hf/2953 (accessed on 20 August 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shearer, E.; Mitchell, A. News Use Across Social Media Platforms in 2021. Pew Research Center. 2021. Available online: https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/ (accessed on 28 August 2024).
2. Lorenz, T. Why TikTok Videos on the Israel-Hamas War Have Drawn Billions of Views. *The Washington Post*, 10 October 2023. Available online: https://www.washingtonpost.com/technology/2023/10/10/tiktok-hamas-israel-war-videos/ (accessed on 28 August 2024).
3. Repede, Ș.E.; Brad, R. A comparison of artificial intelligence models used for fake news detection. *Bull. "Carol I" Natl. Def. Univ.* **2023**, *12*, 114–131. [CrossRef]
4. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Ganapathy, R. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
5. Broda, E.; Strömbäck, J. Misinformation, disinformation, and fake news: Lessons from an interdisciplinary, systematic literature review. *Ann. Int. Commun. Assoc.* **2024**, *48*, 139–166. [CrossRef]
6. Kaliyar, R.K.; Goswami, A.; Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimed. Tools Appl.* **2021**, *80*, 11765–11788. [CrossRef]
7. Repede, Ș.E. Researching disinformation using artificial intelligence techniques: Challenges. *Bull. "Carol I" Natl. Def. Univ.* **2023**, *12*, 69–85. [CrossRef]
8. Aslam, N.; Ullah Khan, I.; Alotaibi, F.S.; Aldaej, L.A.; Aldubaikil, A.K. Fake detect: A deep learning ensemble model for fake news detection. *Complexity* **2021**, *2021*, 5557784. [CrossRef]
9. Ainslie, J.; Lee-Thorp, J.; de Jong, M.; Zemlyanskiy, Y.; Lebrón, F.; Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv* **2023**, arXiv:2305.13245.
10. Wang, Y.A.; Chen, Y.N. What do position embeddings learn? An empirical study of pre-trained language model positional encoding. *arXiv* **2020**, arXiv:2010.04903.

11. Silva, E.C.D.M.; Vaz, J.C. What characteristics define disinformation and fake news?: Review of taxonomies and definitions. *arXiv* **2024**, arXiv:2405.18339.
12. Alghamdi, J.; Luo, S.; Lin, Y. A comprehensive survey on machine learning approaches for fake news detection. *Multimed. Tools Appl.* **2024**, *83*, 51009–51067. [CrossRef]
13. Farhoudinia, B.; Ozturkcan, S.; Kasap, N. Fake news in business and management literature: A systematic review of definitions, theories, methods and implications. *Aslib J. Inf. Manag.* **2023**. *ahead-of-print*. [CrossRef]
14. Vishnupriya, G.; Jeriel K, A.; RNS, A.; Ajay, G.; Giftson J, A. Combating Fake News in the Digital Age: A Review of AI-Based Approaches. In Proceedings of the IEEE 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 5–7 April 2024. [CrossRef]
15. Ayetiran, E.F.; Özgöbek, Ö. A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection. *IEEE Access* **2024**, *12*, 76133–76153. [CrossRef]
16. Kumar, S.; Malhotra, N.; Garg, N.; Shakil, M.A. An In-depth Analysis of Transformer Models for Enhanced Performance in Fake News Detection. In Proceedings of the 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN), Dhulikhel, Nepal, 3–4 July 2024; pp. 158–165. [CrossRef]
17. Saleh, A.O.; Karaoğlan, K.M.; Çakmak, M. A Comprehensive Survey on Automatic Detection of Fake News Using Natural Language Processing: Challenges and Limitations. In Proceedings of the 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkiye, 21–22 September 2024; pp. 1–7. [CrossRef]
18. Alghamdi, J.; Lin, Y.; Luo, S. Towards COVID-19 fake news detection using transformer-based models. *Knowl.-Based Syst.* **2023**, *274*, 110642. [CrossRef]
19. Rohera, D.; Shethna, H.; Patel, K.; Thakker, U.; Tanwar, S.; Gupta, R.; Sharma, R. A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access* **2022**, *10*, 30367–30394. [CrossRef]
20. Cui, L.; Wang, S.; Lee, D. Same: Sentiment-aware multi-modal embedding for detecting fake news. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, BC, Canada, 27–30 August 2019; pp. 41–48. [CrossRef]
21. Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; Qi, P. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 22105–22113. [CrossRef]
22. Węcel, K.; Sawiński, M.; Stróżyna, M.; Lewoniewski, W.; Księżniak, E.; Stolarski, P.; Abramowicz, W. Artificial intelligence—Friend or foe in fake news campaigns. Economics and Business Review. *Sciendo* **2023**, *9*, 41–70. [CrossRef]
23. Yi, Z.; Ouyang, J.; Liu, Y.; Liao, T.; Xu, Z.; Shen, Y. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. *arXiv* **2024**, arXiv:2402.18013.
24. Nazir, A.; Chakravarthy, T.K.; Cecchini, D.A.; Khajuria, R.; Sharma, P.; Mirik, A.T.; Kocaman, V.; Talby, D. LangTest: A comprehensive evaluation library for custom LLM and NLP models. *Softw. Impacts* **2024**, *19*, 100619. [CrossRef]
25. Tanvir, A.A.; Mahir, E.M.; Akhter, S.; Huq, M.R. Detecting Fake News using Machine Learning and Deep Learning Algorithms. In Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019; pp. 1–5. [CrossRef]
26. Di Gennaro, G.; Buonanno, A.; Palmieri, F.A.N. Considerations about learning Word2Vec. *J. Supercomput.* **2021**, *77*, 12320–12335. [CrossRef]
27. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]
28. Liu, Y.; Wu, Y.-F. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
29. Taherdoost, H.; Madanchian, M. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. *Computers* **2023**, *12*, 37. [CrossRef]
30. Raza, S.; Ding, C. Fake news detection based on news content and social contexts: A transformer-based approach. *Int. J. Data Sci. Anal.* **2022**, *13*, 335–362. [CrossRef] [PubMed]
31. Alotaibi, A.; Nadeem, F. Leveraging Social Media and Deep Learning for Sentiment Analysis for Smart Governance: A Case Study of Public Reactions to Educational Reforms in Saudi Arabia. *Computers* **2024**, *13*, 280. [CrossRef]
32. Zong, M.; Krishnamachari, B. A survey on GPT-3. *arXiv* **2022**, arXiv:2212.00857.
33. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access* **2024**, *12*, 26839–26874. [CrossRef]
34. Lin, D.; Wen, Y.; Wang, W.; Su, Y. Enhanced Sentiment Intensity Regression Through LoRA Fine-Tuning on Llama 3. *IEEE Access* **2024**, *12*, 108072–108087. [CrossRef]
35. Repede, Ș.E. Dataset Compiled for the Article. Available online: https://huggingface.co/datasets/Phoenyx83/Politifact-fake-news-6-categories-for-llama3-1 (accessed on 20 August 2024).
36. Repede, Ș.E. Fine Tuned Version of Meta-Llama-3-8B, Trained on 2 Datasets. Available Online on Hugging Face Hub. Available online: https://huggingface.co/Phoenyx83/Meta-Llama-3-8B-Politifact-fake-news (accessed on 20 August 2024).

37. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
38. Li, X.; Zhang, Y.; Malthouse, E.C. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv* **2023**, arXiv:2306.10702.
39. Pang, S.; Nol, E.; Heng, K. ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects. *SSRN Electron. J.* **2023**. [CrossRef]
40. Islam, R.; Ahmed, I. Gemini-the most powerful LLM: Myth or Trut. In Proceedings of the 5th Information Communication Technologies Conference (ICTC), Nanjing, China, 10–12 May 2024; pp. 303–308. [CrossRef]
41. Repede, Ș.E. Ro and En Datasets. Available Online on Kaggle Datasets. Available online: https://www.kaggle.com/datasets/restem/en-ro-datasets-for-llm-testing (accessed on 28 August 2024).
42. Caramancion, K.M. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. *arXiv* **2023**, arXiv:2306.17176.
43. Hu, T.; Zhou, X.H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv* **2024**, arXiv:2404.09135.
44. Ranganathan, P.; Pramesh, C.S.; Aggarwal, R. Common pitfalls in statistical analysis: Measures of agreement. *Perspect. Clin. Res.* **2017**, *8*, 187–191. [CrossRef] [PubMed]
45. Wang, X.; Zhang, W.; Rajtmajer, S. Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey. *arXiv* **2024**, arXiv:2410.18390.
46. Kuntur, S.; Wróblewska, A.; Paprzycki, M.; Ganzha, M. Fake News Detection: It's All in the Data! *arXiv* **2024**, arXiv:2407.02122.