LLMisinfo: Large Language Models for fact-checking online misinformation

Multidisciplinary project A.A. 2023/2024

1st Leonardo Brusini
Computer Science and Engineering
Politecnico di Milano
Mantova, Italy
leonardo.brusini@mail.polimi.it

2nd Javier García
Computer Science and Engineering
Universidad Politécnica de Valencia
Valencia, España
jgarcer@upv.edu.es

3rd Alessandro Masini
Computer Science and Engineering
Politecnico di Milano
Parma, Italy
alessandro.masini@mail.polimi.it

4th Miguel Nogales
Signal Theory and Communications
Universidad de Sevilla
Sevilla, España
mnogales@us.es

I. Introduction

In an era where information travels at incredible speed, the spread of **misinformation** has become a critical challenge. False claims, whether intentional or accidental, can have significant consequences, from influencing public opinion to jeopardizing public health. The rise of social media platforms has only accelerated the dissemination of unverified information, making manual **fact-checking** efforts increasingly difficult to scale. As a result, the need for automated systems capable of verifying the accuracy of claims in real-time has never been more urgent. **Claim verification**, once the domain of human experts, now presents an opportunity for technological innovation to step in and address this growing issue.

With this work, we aimed to build a pipeline that uses Large Language Models (**LLMs**) with the help of **Google search results** to fact-check a claim. Several models are tested, by classifying sampled datasets of claims into one of the following labels: *false*, *mostly false*, *mixture*, *mostly true*, *true*. The models used different prompt styles to test their effectiveness, in three different languages: English, Italian, and Spanish.

II. Materials

The original dataset used in this study is the **ClaimReview** dataset provided by **Fact-Check Insights**, a collection of claims and their corresponding assessments from various fact-checking organizations worldwide. The dataset includes information on the **claim**, **claimant**, **assessment** (sometimes presented as a label, and other times as a detailed explanation), the **date** of publication of the assessment, and the URL (**domain**) of the fact-checking organization's article. We mapped the most common labels to a fixed set of five labels (as shown in Table I), which were used for the final evaluation. From this dataset, we focused on **English**, **Italian**, and **Spanish** claims, from which we have sampled different sub-sets of 150 claims each. Additionally, we sampled 150 English claims published in 2024, after the training cutoff date of LLaMA 3.1, being **December 2023**, to address how information in the training dataset of the model would help it assess the claims. Table II shows the number of claimants and fact-checking domains for each of the sampled datasets. It is worth noting that the claimant is not always reported, and the number of domains in the original dataset is higher. This is due to both the sampling process and the decision to convert only the most common and easily convertible labels/assessments into the fixed set of labels used for classification, to reduce the likelihood of interpretation biases.

Original Labels	**Converted Label**
False, Fake, Labeled Satire, Trolling, Fake News, Fake Tweet	False
Mostly False, Misleading	Mostly False
Mixture, Half True	Mixture
Mostly True	Mostly True
True, Correct Attribution	True

TABLE I
Static conversion of English labels

III. Methodology

A. Models

In our study, we chose to utilize **LLaMA 3.1**, focusing on both the **8B** and **70B** parameter versions due to its recent release and the limited literature specifically evaluating it on

Language	#Claimants	#Domains	#Claims
English - Before 2024	44	21	150
English - After 2024	65	12	150
Italian	43	4	150
Spanish	46	7	150

TABLE II
NUMBER OF CLAIMANTS AND FACT-CHECK WEBSITES FOR EACH LANGUAGE

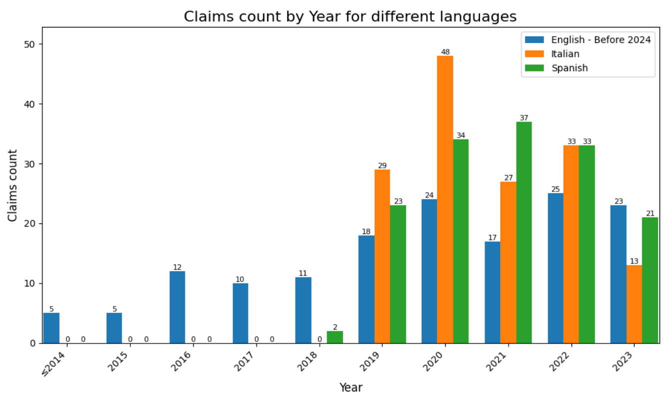

Fig. 1. Claims distribution over the years for the datasets containing claims before the training cutoff date of LLaMA 3.1

fact-checking tasks. While the majority of existing research on this topic tends to focus on models from the GPT family, such as GPT-3.5, we aimed to explore the capabilities of LLaMA 3.1, which remains underexplored in this area. Furthermore, in the context of fact-checking, models with hundreds of billions of parameters are commonly used, so even the 70B version of LLaMA 3.1 can be considered "small" when compared to the SOTA models used in similar studies. By comparing the 8B and 70B versions, we sought to determine whether a smaller, more efficient model could still perform competitively, or if a substantial increase in parameters was necessary to achieve strong fact-checking performance. This comparison allowed us to provide valuable insights into the trade-offs between model size, efficiency, and task-specific performance. For a more deterministic behavior of the models during the testing process, we set the temperature parameter to 0.

B. Prompting styles

In recent literature, a variety of prompting styles have emerged as methods to query LLMs for fact-check claims. These prompting techniques range from simple (like zero-shot) to more complex, multi-step prompts aimed at guiding the model through reasoning. We decided to compare multiple prompting styles in our study. By systematically evaluating the effectiveness of each style, we aim to identify which prompting method yields the best performance for our task.

Standard: this prompting method follows the pipeline suggested by Quelle and Bovet [1], in which LLM is asked to either generate queries for Google searches or to give the final verdict, classifying the claim.

ReAct: short for Reasoning and Acting [2], breaks down the claim verification process into multiple stages. The LLM iterates over a loop of thoughts, actions, and outputs. This allows the model to alternate between reasoning through thoughts and performing actions, such as querying external databases or APIs.

HiSS: proposed by Zhang and Gao [3], the *Hierarchical Step by Step* prompting method asks the LLM to decompose the claim into a list of atomic sub-claims, each assessed independently by the model and then combining all the information into the final classification.

C. Search

Retrieving the correct information from a web search is crucial to guiding the model in making the correct classification. Our approach consisted in making Google searches with the query generated by the LLM interacting with **Chrome** using **Selenium**, appending *before:YYYY-MM-DD* to the query if the published date of the claim was available, to avoid feeding the model with information coming from pages in which the claim is already being assessed or debunked. The URLs of the result page were then collected and filtered to avoid social media websites and the page of the fact-checking organization assessing the claim. We then collected the text for each result page and split it using a different sentence tokenizer based on the language. Using the sentence-transformer model *all-MiniLM-L6-v2*, we generate the embeddings of each sentence. Finally, we sorted the scraped sentences based on the cosine similarity between the sentence and the query embeddings, keeping the three most similar sentences from each of the first five search results

D. Performance metrics and evaluation

We initially evaluated the model on a **5-label classification** task, with labels *false*, *mostly false*, *mixture*, *mostly true*, and *true*. However, we observed that the models rarely output the *mixture* label (see section V), likely due to the inherent ambiguity of this category. As a result, we also evaluated the model's performance on a **4-label classification** task, excluding the *mixture* label to provide clearer distinctions between the remaining labels. Additionally, we evaluated the model in a **binary classification** setting, where *true* and *mostly true* were combined into a single class, as were *false* and *mostly false*. This binary evaluation was conducted to assess the model's ability to judge the veracity of claims without considering subjective nuances or shades of truth, thereby focusing on a clearer distinction between true and false claims.

The following are the performance metrics adopted for this study:

Accuracy, **Precision**, **Recall**, **F1-Score**: the main metrics used for evaluating the general performances on the classification task.

Relaxed Accuracy: Considering the 5-label classification task, we define *relaxed accuracy* as the accuracy where a classification is considered correct if the predicted label differs by no more than one position from the ground truth label. This allows for minor classification discrepancies (e.g., predicting 'mostly false' instead of 'mixture') to be counted as correct, accounting for the inherent subjectivity or closeness between certain labels.

Support: Since the different prompting methods, like HiSS and ReAct, can often be too complex for the model, especially for the 8B parameter version of LLaMA, strictly following their structure can make the model unable to output the final classification. We define as support the fraction of claims in the datasets actually classified by the LLM. Other metrics are computed with respect to the support.

IV. Prompting Methods

In this section, we describe the structure in detail for the different prompting styles and discuss the reasons for these choices. Some examples for each prompting method are provided in the appendix (section VIII).

A. Simple Few Shots

Our first prompting method was the most basic one, we asked the LLM to simply perform a veracity classification of the claim that we provided, only based on its embedded knowledge and without explicitly manifesting its internal chain of thoughts. To make this prompting style as basic as possible the few shots provided also didn't allow the model to perform any claim decomposition into sub-claims or to perform any search over the internet for missing information, therefore entirely relying on the model's embedded knowledge from the training.

B. Standard

This prompting method asks the LLM to assess a claim given by the user. Once the user inputs the claim, along with the author of the claim and the date, the model has the option to either generate a query replying in the format *Query: <query>* or by giving the final verdict classifying the claim, replying in the format *Reasoning: <reasoning>. Final verdict: <label>*. If the model chooses to generate a query, the search result will be appended to the chat as a user message and the LLM will take the decision again. This method allows Llama to get external information on the topic that can allow more informed decisions but also allows the model to autonomously decide to return the final verdict when confident enough, even without doing any search, just relying on information embedded within the model.

C. ReAct - Reasoning and Acting

ReAct stands for Reason + Act, and it prompts LLMs to generate both reasoning traces (thoughts) and actions (steps to solve a task) at the same time. This combination helps the model build more effective strategies, especially for tasks that require understanding and interacting with external information.

Reasoning Traces: The model generates thoughts that represent logical steps for solving a problem. This is heavily inspired by the Chain of thought (CoT) prompting style [4], [5], which allows the model to have internal conversations to develop its answers and thus add reasoning. Actions: The model interacts with the environment. These *actions* are diverse, one of the most common is retrieval, which is performed over knowledge bases such as databases and all types of data storing spaces. LLMs that use this mechanism are usually denoted as RAGs (Retrieval Augmented Generation) [6] and are widely used these days for large data management. Another action that can be performed is web search, which is used in this project. The model provides a query that is then retrieved by an additional module of our project with some web agent, so the data is then returned to the LLM after being processed. This can be performed via API's for this matter or via web scrapping.

In the project, during a task such as verifying a fact, ReAct may retrieve information from an external source (web search) and then use such information to adjust its reasoning process, ensuring that the decisions it makes are grounded in real data.

1) How it works: ReAct is implemented by integrating reasoning and action generation in an interleaved manner. It is implemented as a loop which takes always three steps to complete. In this subsection, it will be explained as it has been used in the project. The way to make the models follow this ReAct scheme is via few-shot examples, specifically three, which are provided to the model before starting to loop, and then the claim to be verified itself is presented to the LLM.
Each iteration of the loop is made up of these three different parts:

- Thought: It represents the reasoning trace that the model generates at each step. It reflects the model's internal process of considering its current state, goals, or next steps. These thoughts are verbalized and guide the model toward making decisions or performing actions. For instance, a thought might involve breaking down a complex task into smaller steps or identifying missing information needed to complete the task. The thought helps the model structure its approach to solving a task, enabling it to plan its next actions or update its strategy if needed. This allows the model to better understand its environment and to make decisions that are more grounded and contextually aware. Thanks to the few-shot examples, it can perform not only claim decomposition in easier ones but also plan which information is needed to search next.
- Action: The task-specific operations or steps that the model decides to perform based on its current thoughts. An action can be for example querying an external knowledge source (e.g., a Wikipedia search). Actions are the model's way of interacting with its environment, in this case by a short query that is searched in Google and then scraped.
- Result: The result part is referred to the information given by the search query. It shows a specified number of webs interacted and a summary of the most relevant part to the LLM's. In the project, five webs are shown for each search, and up to three sentences are shown, as said,

summaries from the whole web.

With this, the loop repeats itself three times. This value in this case is fixed to three, even though it could be arbitrarily decided or left to the LLM to choose, however, leaving this choice to the model sometimes makes it lose its chain of thought and become errant. This is why all the prompts and the model follow a fixed three-iteration ReAct loop.

2) Why it works: This mechanism expands the model's action space to include both task-specific actions (like searching for an entity on Wikipedia) and language-based reasoning (like generating logical steps for answering a question). As the model generates thoughts or takes actions, it updates its internal context, enabling it to keep track of what information has been retrieved, what reasoning has been done, and what steps remain. The LLM can generate thoughts like breaking down complex goals into sub-goals, applying commonsense reasoning, and updating the plan based on new observations.

ReAct uses prompt-based learning, where models are provided with examples that show how to alternate between reasoning and acting. This few-shot learning approach does not require task-specific fine-tuning but instead relies on carefully crafted examples. The prompt format contains sequences of thought-action-observation tuples that demonstrate how to approach tasks, such as verifying facts or solving interactive decision-making problems.

D. HiSS - Hierarchical Step by Step prompting

This, more complex, prompting method aims to make a deeper analysis of the claim by splitting it into simpler and atomic sub-claims that can be assessed independently. This serves the purpose of addressing whether the claim is either completely or only partially true or false. Once the user inputs the claim, the LLM generates the sub-claims, takes the first one, and asks one or more questions to assess its veracity. For each of the generated questions, the following user message is appended: *Tell me if you are confident to answer the question or not. Answer with 'yes' or 'no'*. This question aims to encourage the use of embedded information within the model given by training data. If the answer is yes, then the model would follow with its answer and go on with the following question or sub-claim. Otherwise, the question is used as a query and the answer is given to the model in the form of search results. This process is repeated for every sub-claim. Finally, the LLM will return the final classification.

Each of these prompting methods is taught to the LLM with the help of the system prompt and few-shot examples, ideally one for each label, which has been carefully crafted by hand following the specific prompting style and starting from randomly extracted claims from the original dataset in all three languages.

V. Results

To correctly evaluate the performances of the model, we decided to use a Test Set composed of 150 claims randomly extracted from the ClaimReview dataset, with 30 claims for each of our veracity labels, to have a balanced representation of the performances of our model across many types of possible claims.

Firstly, we wanted to assess what were the performances of the model without performing any complex prompting technique and without allowing the LLM to perform searches over the internet, therefore we started by measuring the performances using **"Simple Few Shots"** prompts, as we can see in table III. We proceeded by checking the performances of our models when using the **standard** prompting technique, so by allowing the LLM to also perform searches over the internet to gather the information it needed.

Model	Metric	Value
Llama 3.1 70B	Precision	0.36
	Recall	0.35
	F1	0.24
	Accuracy	0.35
	Accuracy Relaxed	0.64

TABLE III
Performances of Llama 3.1 70B - Simple Few-Shots
English claims before training cutoff date

To do such an operation, we thought it was important to differentiate whether the claim was made before the knowledge cut-off date of the LLM (this way allowing the model to already have some knowledge regarding the events), or after such date (table IV). Highlighting an increase in performance with respect to the Simple Few Shots due to the information that the LLM is able to retrieve from the internet, allowing in the specific case of claims dated before the knowledge cutoff to have also the much simpler Llama 3.1 8B to perform better than its 70B counterpart.

Model	Metric	Value	
		Before cutoff	After cutoff
Llama 3.1 8B	Precision	0.31	0.30
	Recall	0.33	0.29
	F1	0.3	0.29
	Accuracy	0.33	0.29
	Accuracy Relaxed	0.67	0.63
Llama 3.1 70B	Precision	0.32	0.29
	Recall	0.37	0.32
	F1	0.33	0.29
	Accuracy	0.37	0.32
	Accuracy Relaxed	0.72	0.68

TABLE IV
Performances with standard search prompting
English claims before and after the knowledge cutoff date

Lastly, we could see that even if all the required knowledge should be available over the internet, the model still performs slightly better if it has fully assimilated such knowledge during its training, and it's not relying on a web search to find them instead.

Then we started applying more complex approaches, like the **HiSS** prompting technique.

Both before and after the knowledge cutoff date (table V). This data illustrates a very slight increase in performance due to the more complex prompting technique, which in our opinion is due to the fact that most of the claims that are present in our dataset aren't so complex to require their decomposition into subclaims, since the LLM is already able to implicitly do this by itself (as proven by the very similar results obtained with the standard prompting technique). The very slight increase in performance may be due to that small subset of claims that actually require such decomposition into subclaims to be fully understood by the LLM, or might just be noise.

Model	Metric	Value	
		Before cutoff	After cutoff
Llama 3.1 70B	Precision	0.53	0.27
	Recall	0.44	0.35
	F1	0.36	0.29
	Accuracy	0.43	0.34
	Accuracy Relaxed	0.73	0.69

TABLE V
PERFORMANCES WITH HISS BEFORE AND AFTER THE KNOWLEDGE CUTOFF DATE

Subsequently, we started to explore what were the possible weaknesses of our model while using the HiSS prompting technique. Our first idea was to investigate whether our search module over the internet was not up to the task and was degrading the model's performance, therefore we started to collect metrics regarding the predictions of our model both when it deemed necessary to use our search function and when it instead just relied on its knowledge (table VI).

Model	Metric	Value	
		Before cutoff	After cutoff
Llama 3.1 70B (perform search)	Precision	0.56	0.30
	Recall	0.42	0.35
	F1	0.4	0.31
	Accuracy	0.47	0.38
	Accuracy Relaxed	0.75	0.71
Llama 3.1 70B (doesn't perform search)	Precision	0.31	0.24
	Recall	0.42	0.32
	F1	0.31	0.26
	Accuracy	0.39	0.30
	Accuracy Relaxed	0.72	0.67

TABLE VI
PERFORMANCES WHEN INTERNET SEARCHES HAVE/HAVEN'T BEEN PERFORMED
HISS - LLAMA 70B

These results confirmed our former claim that the model still performs better when the event described in the claim has taken place before its knowledge cutoff date, but also added the new observation that the search over the internet isn't the bottleneck, as the model performs even slightly better when such search is performed.

Still, in an effort to improve the model's performance, we started to look into the accuracy of the LLM over the different veracity labels (table VII). These results showed something interesting, they revealed how the accuracy relaxed of the model wasn't balanced between the different veracity labels, but was very high with the extremes (True and False) and had a huge dip with the Mixture label.

Label	Accuracy	Accuracy Relaxed
True	0.83	0.87
Mostly True	0.25	0.79
Mixture	0.03	0.38
Mostly False	0.19	0.69
False	0.88	0.96

TABLE VII
PERFORMANCES BY VERACITY LABEL
HISS - LLAMA 70B - BEFORE KNOWLEDGE CUTOFF

Therefore we started to investigate the possibility of removing the claims in the dataset whose true label was Mixture and calculating our metrics again, comparing them to the old ones to see if there was an improvement (table VIII). There was a clear improvement across the board, which can be pushed to its extremes if we merge the two couples of remaining labels defining a binary classification task (table VIII). In this scenario, it's worth noting that the "Accuracy Relaxed" metric has become equivalent to the Accuracy itself, as well as the "Accuracy Relaxed" of the 4 labels case, which anyway reaches the highest value we were able to obtain, along with a sharp increase of the other metrics as well.

Metric	With Mixture	Without Mixture	Binary Labels
Precision	0.53	0.57	0.83
Recall	0.44	0.54	0.83
F1	0.36	0.49	0.83
Accuracy	0.43	0.54	0.83
Accuracy Relaxed	0.73		

TABLE VIII
PERFORMANCES ON DIFFERENT SETS OF LABELS
HISS - LLAMA 70B - BEFORE KNOWLEDGE CUTOFF

We also explored how the LLM could deal with other languages besides English, by performing all the former steps and observations with few shots and datasets in Italian or Spanish and measuring the performances of our final version of the model both with and without the Mixture label, as well as with a binary classification. Lastly, we performed the same measurements for the Spanish language, where we used the same technique of trying to remove the Mixture label beforehand and then checking the performance of a binary classification. The performance metrics for these two languages are shown in Table IX.

Our results with the Italian and Spanish languages confirmed the trends we noticed beforehand with English. Specifically, they still performed better before the knowledge cutoff date of the model, when they were actually getting information

Language	Metric	With Mixture	Without Mixture	Binary Labels
Italian	Precision	0.27	0.44	0.77
	Recall	0.38	0.48	0.77
	F1	0.30	0.43	0.77
	Accuracy	0.37	0.47	0.77
	Accuracy Relaxed	0.73		
Spanish	Precision	0.26	0.45	0.64
	Recall	0.34	0.43	0.63
	F1	0.27	0.39	0.62
	Accuracy	0.34	0.46	0.63
	Accuracy Relaxed	0.55		

TABLE IX
PERFORMANCES FOR THE ITALIAN AND SPANISH DATASETS
HISS - LLAMA 70B - BEFORE KNOWLEDGE CUTOFF

from the internet. Like before, they reached their best performances when structured as a binary classification between just True and False.

Now regarding **ReAct**, the same series of experiments was carried out with this prompting style. The reason behind that was due to two main reasons, the first was plain fairness, since the main techniques presented in this work were HiSS and ReAct, so an even ground was needed to fairly test and compare them. The second reason was to have a second validation (or denial) of the conclusions reached using the HiSS technique.

In this way, the first experiment on ReAct was to verify the metrics of the English dataset, before and after the knowledge cutoff (Table X). The results are very similar to the HiSS case, just a bit more balanced between the classes but not in a too significant way. There is, a slight decay in the performance after the cut-off, but very small, even smaller than with HiSS. It is worth remembering that the relaxed accuracy is maintained in both cases.

Model	Metric	Value	
		Before cutoff	After cutoff
Llama 3.1 70B	Precision	0.42	0.37
	Recall	0.37	0.31
	F1	0.35	0.31
	Accuracy	0.37	0.31
	Accuracy Relaxed	0.75	0.75

TABLE X
PERFORMANCES WITH REACT BEFORE AND AFTER THE KNOWLEDGE CUTOFF DATE

Now, with respect to the balance of the veracity labels in the dataset, there is the next experiment, the results of which are shown in Table XI. Please keep in mind that the experiment which takes into account the performance when performing or not an internet search does not proceed with ReAct, as disabling its ability to act would make it a plain Chain of thought, which it is not being taken into account. Anyway regarding the balance of veracity labels, it can be seen a clear bias of the model in the English case towards the False class, which can be due to several reasons, such as the prompts given as few shots to the model to learn how to use ReAct. For example, in such prompts, the final evaluations for the claims are, in this order, "Mostly False", "Mostly True", and "Mostly False", which could induce the model to output "False" or "Mostly False" more frequently. But the same prompts could also be seen the other way around, considering the fact that the order in which the answers are could induce the model to predict now "Mostly True" to follow the existing pattern. We can see that with HiSS the predictions used to be more extreme, e.g. the model used to output more "True" and "False" labels compared to "Mostly True" or "Mostly False", while ReAct instead acts the other way around, which is an advantage when calculating its Accuracy Relaxed.

Label	Accuracy	Accuracy Relaxed
True	0.40	0.57
Mostly True	0.27	0.40
Mixture	0.07	0.87
Mostly False	0.60	0.90
False	0.53	1.00

TABLE XI
PERFORMANCES BY VERACITY LABEL
REACT - LLAMA 70B - BEFORE THE KNOWLEDGE CUTOFF

Differently from HiSS, dropping the mixture label could be thought to be a bad idea for ReAct as the accuracy relaxed in this case is very good. This is kind of true, in fact in Table XII the results when deleting this class are shown. From there we can see that even though it achieves better metrics (even for accuracy), its accuracy relaxed (which as we said previously is the same as the accuracy for the binary case) decreases. Furthermore, a notable difference in performance is clear between HiSS and ReAct, where the latter is compromised. The reason behind the not-so-good performances, in this case, is due to the fact that differently from HiSS, ReAct benefits a lot from the existence of the Mixture label for what regards its accuracy relaxed, since as we said, the model outputs very often "Mostly True" or "Mostly False", which grants it a very high performance when the true label of the claim is "Mixture" since the accuracy relaxed classify it as a correct prediction. Therefore removing this label damages the model's Accuracy Relaxed

Metric	With Mixture	Without Mixture	Binary Labels
Precision	0.42	0.52	0.78
Recall	0.37	0.45	0.69
F1	0.35	0.45	0.69
Accuracy	0.37	0.45	0.69
Accuracy Relaxed	0.75		

TABLE XII
PERFORMANCES ON DIFFERENT SETS OF LABELS
REACT - LLAMA 70B - BEFORE KNOWLEDGE CUTOFF

Lasty, the performance of the Italian and Spanish datasets can be checked in Table XIII. As in the last case, the drop of the mixture class and conversion to a binary problem does not help as much as with the HiSS prompting (and instead damages the model's accuracy relaxed), both in Spanish and in Italian. Performances in both languages are similar, which is interesting because the Spanish case has much lower performances than the Italian one with HiSS. This could be due to the language of the browsers used for the two experiments being different.

Language	Metric	With Mixture	Without Mixture	Binary Labels
Italian	Precision	0.29	0.36	0.67
	Recall	0.22	0.27	0.54
	F1	0.16	0.20	0.46
	Accuracy	0.22	0.47	0.54
	Accuracy Relaxed	0.63		
Spanish	Precision	0.27	0.43	0.60
	Recall	0.23	0.29	0.53
	F1	0.18	0.26	0.46
	Accuracy	0.23	0.29	0.53
	Accuracy Relaxed	0.61		

TABLE XIII
PERFORMANCES FOR THE ITALIAN AND SPANISH DATASETS
REACT - LLAMA 70B

Overall, results are similar to HiSS but in general a bit worse. This could be due to just HiSS being the better option, or on the other hand, some assumptions in these tests could be biased to favor a bit that method. One reason supporting the latter is that the LLama model used was instruct fine-tuned, which makes it better for chat purposes but worse for arbitrary text completion. This is a subtle difference but could be relevant, anyway both because the HuggingFace API wouldn't allow us to use the regular version of LLama 70B and to use the same model for all the tests, so we were stuck with the instruct one. This assessment was added to highlight that there are potentially several little reasons that can be affecting the performance of a model and that in a real world application they should all be carefully analyzed.

To end this section, a summary of all the performances is presented in Table XIV and XV, to check how HiSS and ReAct achieve different metrics and which one outperforms the other in different cases. Often, HiSS is the preferred option, but if we consider the Mixture label, sometimes things are different.

Dataset	Metric	HiSS	ReAct
English before cutoff	Precision	**0.53**	0.42
	Recall	**0.44**	0.37
	F1	**0.36**	0.35
	Accuracy	**0.43**	0.37
	Accuracy Relaxed	0.73	**0.75**
English after cutoff	Precision	0.27	**0.37**
	Recall	**0.35**	0.31
	F1	0.29	**0.31**
	Accuracy	**0.34**	0.31
	Accuracy Relaxed	0.69	**0.75**
English without mixture	Precision	**0.57**	0.52
	Recall	**0.54**	0.45
	F1	**0.49**	0.45
	Accuracy	**0.54**	0.45
	Accuracy Relaxed	**0.83**	0.69

TABLE XIV
PERFORMANCE SUMMARY USING LLAMA 70B
ENGLISH LANGUAGE

Dataset	Metric	HiSS	ReAct
Italian with mixture	Precision	0.27	**0.29**
	Recall	**0.38**	0.22
	F1	**0.3**	0.16
	Accuracy	**0.37**	0.22
	Accuracy Relaxed	**0.73**	0.63
Italian without mixture	Precision	**0.44**	0.36
	Recall	**0.48**	0.27
	F1	**0.43**	0.20
	Accuracy	0.47	0.47
	Accuracy Relaxed	**0.77**	0.54
Spanish with mixture	Precision	0.26	**0.27**
	Recall	**0.34**	0.23
	F1	**0.27**	0.18
	Accuracy	**0.34**	0.23
	Accuracy Relaxed	0.55	**0.61**
Spanish without mixture	Precision	**0.45**	0.43
	Recall	**0.43**	0.29
	F1	**0.39**	0.26
	Accuracy	**0.46**	0.29
	Accuracy Relaxed	**0.63**	0.53

TABLE XV
PERFORMANCE SUMMARY USING LLAMA 70B
ITALIAN AND SPANISH LANGUAGE

VI. DISCUSSION

Considering the number of different variables in assessing this task, such as language, model size and knowledge, quality of generated queries, and search results, giving a correct and final interpretation of the test result is a complex task. We focused on those variables individually, discussing how they can affect the fact-checking process and how they can be further improved.

A. Search results

The potential of a RAG approach for tasks such as fact-checking can be crucial to give the model information about the context of a claim and data that can either help support or debunk it. Despite the ability of the Google search engine to usefully sort the result pages, it is important to filter the most relevant information within each page to avoid filling the context window of the model with noisy data. While our approach, based on ranking sentences based on the embedding's similarity with respect to the query embedding, was able to provide some useful content within the top-scoring sentences, there is much room for improvement. Such improvements should involve a way to search for snippets of text that can reply to the information asked by the query (i.e. with a question-answering deep learning model, or calling another instance of an LLM to summarize the text or search for the relevant sentences within it, at the cost of additional computational burden).

B. Model size and knowledge

The reasoning ability of the model as well as the overfitted knowledge given by the training data can be useful for the fact-checking task, especially for claims related to facts that happened before the training cutoff date. Web search can be useful to compensate for the worst reasoning capability of a smaller-sized model. In Table XVIII, for example, it is shown how LLaMA 8B, with the standard prompting, was able to correctly classify the claim with the help of a search, while the 70B counterpart, although the same prompting style, decided to immediately give the final answer, classifying the claim with the wrong label. At the same time, smaller models are less prone to correctly follow the structure of more complex prompting methods, leading to the inability to return the final verdict, even with the help of the given few-shot examples. This led to our decision to only test the HiSS and ReAct prompting on the 70B version of LLaMA.

C. Subjectivity of assessment process and labels

One of the main goals of fact-checking is to prevent the spread of deceptive or misleading content. However, subjectivity in the fact-checking process can present challenges. While objective facts are essential, some claims may involve interpretations, context, or nuances that lead to disagreements about what is considered "true" or "false." This subjectivity can introduce bias. Therefore, transparency and the use of standardized, rigorous methods are vital to maintaining credibility in fact-checking. Since the original and the sampled dataset contained claims assessed by tens of different fact-checking organizations, different sets of labels and evaluation criteria have been used. The process of statically converting these labels into the fixed set of five labels used in this study may lead to bias in the evaluation. A more comprehensive evaluation, involving a larger number of claims annotated by fact-checkers using a fixed set of labels, would lead to a deeper understanding of an LLM's capabilities in the fact-checking task. This structured approach would allow for a more detailed assessment of how consistently and accurately the model classifies claims. By comparing the model's reasoning in classifying claims with the fact-checkers assessments, researchers could identify patterns in the LLM's decision-making process, uncover potential biases, and better understand the model's interpretative strengths and weaknesses (i.e. no relevant information given by the web search would either be classified as *false*, *mostly false*, or even *mixture*). Additionally, this comparison would highlight discrepancies in reasoning between human annotators and the model, offering insights into areas where the LLM might need improvement, such as grasping nuanced or context-dependent claims.

VII. Conclusion

This study explored the use of Large Language Models (LLMs) for fact-checking online misinformation, focusing on the LLaMA 3.1 model in both 8B and 70B parameter versions. We investigated various prompting techniques, including standard search, ReAct (Reasoning and Acting), and HiSS (Hierarchical Step by Step), across multiple languages. Our findings suggest that:

1) The performance of LLMs in fact-checking tasks improves when they have access to relevant information through web searches, especially for claims dated before their knowledge cutoff.
2) More complex prompting techniques like ReAct and HiSS showed slight improvements over simpler methods, particularly for intricate claims that can benefit from decomposition into sub-claims.
3) The models performed better on binary classification tasks (*true*/*false*) compared to more nuanced multi-label classifications, highlighting challenges in distinguishing between closely related veracity labels.
4) Performance varied across languages, with generally better results for English compared to Italian and Spanish.
5) The subjectivity inherent in fact-checking and the diversity of labeling schemes used by different fact-checking organizations present challenges for creating standardized evaluation metrics.

While LLMs show promise in automating aspects of the fact-checking process, there remains significant room for improvement. Future work should focus on enhancing search result relevance, developing more robust evaluation methodologies, and addressing the challenges posed by subjective or context-dependent claims. As LLMs continue to evolve, their potential role in combating misinformation warrants further exploration and refinement.

References

[1] D. Quelle and A. Bovet, "The perils and promises of fact-checking with large language models," *Frontiers in Artificial Intelligence*, vol. 7, Feb. 2024.

[2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," *arXiv*, Oct. 2022.

[3] X. Zhang and W. Gao, "Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method," *arXiv*, 2023.

[4] X. Wang and D. Zhou, "Chain-of-Thought Reasoning Without Prompting," *arXiv*, Feb. 2024.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *arXiv*, Jan. 2022.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv*, May 2020.

VIII. Appendix

In this section, we will show some examples of fact-checking to highlight what the differences between the different models and prompting styles.

Ground truth: True.
Claim: A video authentically shows UFC fighter Conor McGregor putting his arm around Russian President Vladimir Putin for a photo-op before removing it after someone in Putin's entourage tells him not to.

Few-shot w/o search - **Model**: Both LLaMA 3.1 8B/70B
<assistant> Based on my personal knowledge, it is clear that among False, Mostly-False, Mixture, Mostly-True, and True, the claim is classified as False.

Standard search - **Model**: LLaMA 3.1 8B
<assistant> Query: Conor McGregor Vladimir Putin photo op incident
<user> Search number 1: Google: Conor McGregor Vladimir Putin photo op incident
1. bbc.com: The MMA fighter posted a photo to Instagram of himself with Mr Putin at the World Cup final in Moscow.
Accept and continue According to his Instagram post, Conor was invited as a guest of Mr Putin.
Comments on McGregor's Instagram and Twitter profiles told him to "research the history".
[Other search results]
<assistant> Reasoning: The claim that a video authentically shows Conor McGregor putting his arm around Russian President Vladimir Putin for a photo-op before removing it after someone in Putin's entourage tells him not to is true. The incident occurred at the 2018 World Cup final in Moscow, where McGregor was a guest of Putin. According to multiple sources, including McGregor's Instagram post and reports from reputable news outlets, McGregor did put his arm around Putin, but was asked to remove it by one of Putin's security team members.
Final answer: True.

Standard search - **Model**: LLaMA 3.1 70B
<assistant> Reasoning: Without access to the specific video or more information, I couldn't verify the authenticity of the video. However, I can provide some context. Conor McGregor has been known to attend high-profile events, including the 2018 World Cup in Russia, where he was spotted with Vladimir Putin. It's possible that the video is real, but without further evidence or confirmation from reputable sources, it's difficult to say for certain.
Final answer: Mixture

ReAct - **Model**: LLaMA 3.1 70B
<assistant> *Thought 1*: The claim seems suspicious, as there is no credible source or context provided to support the authenticity of the video. It is possible that the video is fabricated or taken out of context to create a sensational story.
Action 1: Search[Conor McGregor Vladimir Putin photo-op]
<user> *Result 1*: Google: Conor McGregor Vladimir Putin photo-op
1. bbc.com: The MMA fighter posted a photo to Instagram of himself with Mr Putin at the World Cup final in Moscow.
[Other search results]
<assistant> *Thought 2*: The search results suggest that the claim is likely false. While Conor McGregor did attend the 2018 World Cup final in Russia and took a photo with Vladimir Putin, there is no evidence to support the specific scenario described in the claim. The photo-op was a public event, and McGregor's spokesperson has confirmed that the fighter often takes photos with event hosts.
Action 2: Search[Conor McGregor Vladimir Putin photo-op debunked]
<user> *Result 2*: Google: Conor McGregor Vladimir Putin photo-op debunked
[*Search results*]
<assistant> *Thought 3*: The search results confirm that the claim is likely false. There is no evidence to support the specific scenario described in the claim, and McGregor's spokesperson has confirmed that the fighter often takes photos with event hosts. The photo-op was a public event, and McGregor has publicly praised Putin as "one of the greatest leaders of our time".
Final answer, choose between the options false, mostly false, mixture, mostly true and true: False.

HiSS - **Model**: LLaMA 3.1 70B
<assistant> A fact checker will decompose the claim into 2 sub-claims that are easier to verify:
1. Conor McGregor and Vladimir Putin were in the same location and took a photo together.
2. Conor McGregor put his arm around Vladimir Putin for a photo-op, but removed it after someone in Putin's entourage told him not to.
To verify subclaim 1, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raises each question and looks for an answer:
Question: Is there evidence that Conor McGregor and Vladimir Putin were in the same location and took a photo together?
<user> Tell me if you are confident to answer the question or not. Answer with 'yes' or 'no':
<assistant> Yes.
Answer: Yes, Conor McGregor and Vladimir Putin were in the same location and took a photo together at the 2018 World Cup in Russia.
To verify subclaim 2, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raises each question and looks for an answer:
Question: Does the video accurately depict Conor McGregor putting his arm around Vladimir Putin for a photo-op, but removing it after someone in Putin's entourage tells him not to?
<user> Tell me if you are confident to answer the question or not. Answer with 'yes' or 'no':
<assistant> Yes.
Answer: Yes, the video accurately depicts Conor McGregor putting his arm around Vladimir Putin for a photo-op, but removing it after someone in Putin's entourage tells him not to.
Based on the answers to these questions, it is clear that among False, Mostly-False, Mixture, Mostly-True and True, the claim is classified as True.

TABLE XVI
Example of claim fact-checking on every different prompting styles.

Ground truth: False.
Claim: purports a World Cup qualifier game is taking place between Israel and Palestine on Nov. 16. Date: 2023-10-29.

Standard search - **Model**: LLaMA 3.1 8B
<assistant> Query: Israel Palestine World Cup qualifier game
<user> Search number 1: Google: Israel Palestine World Cup qualifier game
1. [...]
2. [...]
3. [...]
4. insidethegames.biz: AFC have confirmed that the match between Palestine and Australia on November 21 will now be played at the Jaber Al-Ahmad International Stadium in Kuwait ©Getty Images Palestine had to pull out of the Merdeka Tournament in Malaysia due to the war.
OK Milan Cortina 2026 - 498 days RSS Facebook Twitter LinkedIn Instagram FIFA and the Asian Football Confederation (AFC) has blocked Algerian efforts to host Palestine games, including 2026 World Cup qualifiers, as the deadly war between Hamas and Israel continue.
Palestine are set to begin their 2026 World Cup campaign against Lebanon on November 16.
5. [...]
<assistant> Reasoning: The claim that a World Cup qualifier game is taking place between Israel and Palestine on Nov. 16 is partially true, but the context is missing. According to the search results, Palestine is set to begin their 2026 World Cup campaign against Lebanon on November 16, not against Israel. However, it is mentioned that Palestine had to pull out of the Merdeka Tournament in Malaysia due to the war, and FIFA and the Asian Football Confederation (AFC) has blocked Algerian efforts to host Palestine games, including 2026 World Cup qualifiers, as the deadly war between Hamas and Israel continue.
Final answer: Mostly False.

Standard search - **Model**: LLaMA 3.1 70B
<assistant> Query: Israel Palestine World Cup qualifier game November 16
<user> Search number 1: Google: Israel Palestine World Cup qualifier game November 16
1. [...]
2. insidethegames.biz: AFC have confirmed that the match between Palestine and Australia on November 21 will now be played at the Jaber Al-Ahmad International Stadium in Kuwait ©Getty Images Palestine had to pull out of the Merdeka Tournament in Malaysia due to the war.
Palestine are set to begin their 2026 World Cup campaign against Lebanon on November 16.
OK Milan Cortina 2026 - 499 days RSS Facebook Twitter LinkedIn Instagram FIFA and the Asian Football Confederation (AFC) has blocked Algerian efforts to host Palestine games, including 2026 World Cup qualifiers, as the deadly war between Hamas and Israel continue.
3. [...]
4. [...]
5. [...]
<assistant> Reasoning: According to the search results, Palestine is set to begin their 2026 World Cup campaign against Lebanon on November 16, not against Israel. There is no evidence to suggest that a World Cup qualifier game is taking place between Israel and Palestine on November 16.
Final answer: False

TABLE XVII
EXAMPLE OF CLAIM FACT-CHECKING ON STANDARD PROMPTING, ON DIFFERENT MODELS.

Ground truth: Mostly False.
Claim: Local women demanding ₹8500, as promised by Rahul Gandhi, broke into Congress MP Selja Kumari's office. Made by X users. Date: 2024-06-15.

Few-shot w/o search - **Model**: LLaMA 3.1 8B
<assistant> Based on my personal knowledge, it is clear that among False, Mostly-False, Mixture, Mostly-True and True, the claim is classified as False.

Few-shot w/o search - **Model**: Both LLaMA 3.1 70B
<assistant> Based on my personal knowledge, it is clear that among False, Mostly-False, Mixture, Mostly-True, and True, the claim is classified as True.

ReAct - **Model**: LLaMA 3.1 70B
<assistant> *Thought 1*: The claim seems suspicious, as it lacks specific details and context. Rahul Gandhi is a prominent political figure, and such a promise would likely have been widely reported. Moreover, the exact amount of ₹8500 seems arbitrary, and it's unclear why local women would break into an MP's office over this issue.
Action 1: Search[News articles about Rahul Gandhi promising ₹8500 to local women]
<user> *Result 1*: Google: News articles about Rahul Gandhi promising ₹8500 to local women
[*Other search results*]
5. hindustantimes.com: News / India News / **Rahul Gandhi stresses on big poll promise, says 'will transfer ₹1 lakh to women in poor households'**
PM IST Share Via Copy Link Congress leader Rahul Gandhi made a significant poll promise on Thursday while addressing a public rally in Rajasthan, saying that the party will transfer ₹1 lakh in the account of one woman from every poor household in the country if they win the elections.
Gandhi also attacked BJP during his poll rally, saying farmers are asking for a minimum support price (MSP) for their crops, youngsters are seeking employment, women want relief from inflation, but no one is listening.
<assistant> *Thought 2*: The search results indicate that Rahul Gandhi did make a promise to provide financial assistance to women, but the context and details are different from the claim. The promise was to provide ₹1 lakh per year to poor women, not ₹8500, and it was part of the Congress party's election manifesto. There is no evidence to suggest that local women broke into an MP's office demanding ₹8500.
Action 2: Search[News articles about local women breaking into Congress MP Selja Kumari's office]
<user> *Result 2*: Google: News articles about local women breaking into Congress MP Selja Kumari's office
[*Search results*]
<assistant> *Thought 3*: Based on the search results, it is clear that the claim is false. There is no evidence to suggest that local women broke into Congress MP Selja Kumari's office demanding ₹8500. The promise made by Rahul Gandhi was to provide ₹1 lakh per year to poor women, not ₹8500, and it was part of the Congress party's election manifesto.
Final answer False.

HiSS - **Model**: LLaMA 3.1 70B
<assistant> A fact checker will decompose the claim into 2 subclaims that are easier to verify:
1. Rahul Gandhi promised ₹8500 to local women.
2. Local women demanding ₹8500 broke into Congress MP Selja Kumari's office.
To verify subclaim 1, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raise each question and look for an answer:
Question: Did Rahul Gandhi promise ₹8500 to local women?
<user> Tell me if you are confident to answer the question or not. Answer with 'yes' or 'no':
<assistant> No.
<user> Answer: Google: Did Rahul Gandhi promise ₹8500 to local women?
1. livemint.com: **Ramesh Babu, Congress party's Karnataka media in-charge, said that the Congress has promised ₹1 lakh to women in poor families under the Mahalakshmi scheme.** [...]
2. [...]
3. boomlive.in: By - Hazel Gandhi | 11 Jun 2024 10:14 AM GMT Follow Us A 2020 video showing Muslim women waiting outside a bank in Uttar Pradesh is viral online as recent and being shared with the misleading claim that the women were waiting to avail their promised sum of Rs 8,500 under the Congress' Mahalaxmi scheme.
While the viral video is not recent, several reports published by ABP News and India Today on June 5, 2024, stated how female voters in Lucknow formed queues outside the Congress office to avail the Rs 8,500 promised under the Mahalaxmi scheme.
Taking a cue from this, we ran a search in Hindi and found a report published by News18 on April 20, 2020 titled 'Video of crowd outside Muzaffarnagar bank viral, rumours of money being removed from account viral'
4. thehindu.com: **He said women, who worked more than the men because their responsibilities included domestic chores and childcare responsibilities, would receive ₹8,500 every month if the Congress was voted in**.
Earlier in his speech, Mr. Gandhi had alleged that the BJP wanted the Constitution – that empowered and protected the poor – to be torn up and thrown away and let "20-25 billionaires rule and the rest of the public keeps watching".
The name of one woman from every poor family of India will be selected, and the Congress party will put ₹8,500 into their accounts on the first day of each month.
<assistant> Answer: Yes, Rahul Gandhi promised ₹8500 to women in poor families under the Mahalakshmi scheme, but not to all local women.
To verify subclaim 2, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raise each question and look for an answer:
Question: Did local women demanding ₹8500 break into Congress MP Selja Kumari's office?
<user> Tell me if you are confident to answer the question or not. Answer with 'yes' or 'no':
<assistant> No.
<user> Answer: Google: Did local women demanding ₹8500 break into Congress MP Selja Kumari's office?
[*Search results*]
<assistant> Answer: No, there is no evidence to support this claim.
Based on the answers to these questions, it is clear that among False, Mostly-False, Mixture, Mostly-True and True, the claim is classified as Mostly False.

TABLE XVIII
EXAMPLE OF CLAIM FACT-CHECKING COMPARING REACT AND HISS.