# Automated Fact-checking based on Large Language Models: An application for the press

Bogdan Andrei Baltes[1], Yudith Cardinale[1] and Benjamín Arroquia-Cuadros[1]

[1]*Centro de Estudios en Ciencia de Datos e Inteligencia Artificial (ESenCIA), Valencian International University, C/Pintor Sorolla 21, 46002 València, Spain*

**Abstract**

The current proliferation of digital media for the dispersion of news represents advantages given the ease of access but also challenges as the different sources might not necessarily be reliable or fully consistent with each other. Existing solutions for contrasting information include knowledge bases with previously verified information that are often lacking updated information or insightful details. In this context, we propose a framework for enhancing information retrieval from the press to make information more digestible and with the ultimate goal of reducing misinformation. The proposed framework, at the interconnection of automated fact-checking, AI-based reasoning, and ethics, consists of a tool that combines information from several sources and allows users to verify a claim given the information as a knowledge base. The work explores the reasoning capabilities of Large Language Models (LLM) as a new way of automating fact-checking, creating a flexible and dynamic solution. The framework returns a verdict about the claim, as well as a justification and references, building trust for the users. The performance is rigorously evaluated achieving a score of 70% accuracy of classification and justification production for the top-performing models. Equally important, the work studies the ethical challenges of building a framework that changes the way that information from the press is consumed by society. The underlying ethics of the project are discussed both from a perspective for final users and publishing companies, offering guidance for large-scale implementation of the framework. This research poses challenges as well, mainly regarding the capabilities of current and future LLM and the commercial partnership dynamics with publishing companies.

**Keywords**

Automated Fact-Checking, Large Language Models, Artificial Intelligence, Ethics

## 1. Introduction

Disinformation and fake news have been combatted through the manual work of journalists at traditional media and fact-checking outlets [1]. Tasks related to fact-checking procedures include contacting the original source via phone or e-mail, consulting alternative sources, and writing and rating the claim and publishing it [2]. While this workflow is complete and consistent, the workforce is often insufficient to monitor every piece of published information, so it is often users' mission to verify whether something they read or heard is true or false [3].

With or without deliberation to spread false information, there are often a variety of sources that are not fully consistent with each other. It is generally not possible for a person to read the same news in many different media to find complementary or contradictory information to get the full picture. Because of this, fact-checking is needed. In addition to the manual efforts of journalists, automated fact-checking (AFC) techniques are being developed mostly by nonprofit fact-checking entities [4]. The limitations of AFC have traditionally been the sensitivity to context that impedes the full automation of fact-checking systems, requiring human supervision. Another direction that AFC has been taking is that of identifying claims and constructing a database of verified claims [5], which is useful for assisting the fact-checker although with a static context.

To make information more digestible, in this work we present a framework that processes information from different sources, solves the user's original question, and indicates where the information comes from, being able to consider the context that is given to the system. An important aspect is that the user of the system is fully aware of the contents of its context and can check it if necessary, adding

*Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 40–53, October 20, 2024, Santiago de Compostela, Spain*

trustworthiness. The aim is to help users get a broader perspective on the news from different sources, in order to fight misinformation. This is pursued through the creation of a system that compiles information from the press to be able to verify claims based on the knowledge base, providing a reasoned answer, and having the ability to reference the employed sources, supported by AI-based reasoning and ethics.

In this work, we explore a new approach for automating fact-checking: through the reasoning capabilities of Large Language Models (LLM). We carry out a complete implementation of the framework, starting from a knowledge base crafted from news from Spanish media until the interface where final users can make use of the framework. The functioning of the proposed system is tested, both from a technical and functional perspective, rigorously carrying out an evaluation achieving a score of 70% accuracy of classification and justification production, but also from an ethical standpoint, studying the underlying ethics of the change that people would undergo if the framework were implemented at a large scale and the way media is consumed were modified.

This research revolves around solving problems derived from misinformation and disinformation. The former is defined as "false or inaccurate information", while the latter is adding the notion of the "false or misleading information peddled deliberately to deceive, often in pursuit of an objective" [6]. In particular, this system is intended for journalists as primary end-users. Journalists at fact-checking agencies continuously track claims made by politicians and evaluate the veracity of them, publishing the results for the general public.

The rest of the paper is organized as follows. Section 2 describes recent studies on automated fact-checking and reasoning capabilities of LLM. The proposed Assisted Fact-Checking Framework is presented in Section 3 and the obtained results are discussed in Section 4. The ethics of the proposed work are studied in Section 5 and we present the conclusions in Section 6.

## 2. Related Work

In this section, some studies related to fact-checking and reasoning capabilities of LLM are described.

Fact-checking, in its simplest form, is a practice that verifies whether a claim is true or not. It has, of course, more complex definitions, given that a claim can be technically true but written in a misleading way, or only partially true. The most common workflow when doing the task of fact-checking is searching through multiple sources that can be used to verify the veracity of the claim, assess their reliability, and make a decision on the original claim based on the evidence found in the sources [7].

Traditionally, it consists of manual work carried out by journalists, whether to fact-check published claims in other agencies' work or to assess the correctness of works before being published [2].

The need to automatise fact-checking processes arises from the inability of journalists to verify everything they publish, since this manual work is oftentimes a task that can take up to several days [2]. Sources are not always accessible in literature or on the Internet. While there are official databases or reports such as the National Institute of Statistics (INE) in Spain, there is also manual work to be done when information needs to be verified directly calling an institution, like the Government or the Police.

In the last years, the AI community has dedicated efforts to discuss AFC. The most widely accepted structure for this automation was proposed by Vlachos and Riedel [5] [8], in a sequential process that starts by identifying the claims that need to be checked, looking through sources for the evidence needed to support or refute the claims, and taking a decision considering the given evidence.

There are, however, two issues with the knowledge commonly available to most approaches found in the literature [4]: not all available information is trustworthy, and not all needed information is available.

To overcome these problems, researchers have taken the assumptions that the information included in the employed sources is correct and that the evidence is the information that can be retrieved from there. As the evidence is assumed to be correct, veracity will be defined as the coherence of the claim and the evidence.

This common structure for automated fact-checking can - and should - be adapted to the needs of its

end users (mostly journalists). Regarding the research of these systems, there has sometimes been a lack of collaboration between researchers and journalists [9]. A better collaboration could lead to the solution of some of the issues that AFC systems present, although not all of them can be technically solved.

Furthermore, advances in the field of Generative Artificial Intelligence and specifically LLM are contributing to the transition from simpler natural language processing (NLP) techniques to the usage of the reasoning abilities of more complex models. There have been attempts to integrate LLM in the whole framework for automated fact-checking, using it to detect claims, retrieve evidence and finally, predict a verdict and build a conclusion [10]. However, the results obtained are inferior to those of the state of the art models on datasets like FEVER [11] and WiCE [12] and further research is encouraged.

To better understand the purpose of the research of automated fact-checking, a study has shown that there are eight main intended use cases of automated fact-checking [13]. The study has analysed 100 highly-cited papers, with publication dates ranging from 1998 to 2023, with most studies being from the 2010s. These use cases are listed below, specifying the percentage of the 100 papers where the respective use case is pursued: Automated external fact-checking (22%), Assisted external fact-checking (18%), Assisted media consumption (8%), Scientific curiosity (8%), Assisted knowledge curation (7%), Assisted internal fact-checking (4%), Automated content moderation (4%), Truth-telling for law enforcement (1%).

On the other hand, LLM are AI systems that can process and generate text, to solve a variety of tasks, such as summarisation, translation, question answering [14].

These systems have significantly gained popularity over the recent years. One of the main reasons of this rise was the introduction of the Transformer architecture [15]. This technical breakthrough, along with the ever-growing data collection and generation for training, as well as larger computational abilities, triggered a large wave of more capable language models. The paradigm of their creation shifted from task-specific to task-agnostic training, allowing models to perform a wider range of tasks [16].

One of the desired capabilities of LLM is reasoning. It is a cognitive process designed as the process of thinking about something in order to make a decision. At the intersection of psychology, philosophy and computer science, it is a process that benefits individuals to solve problems and take decisions [17].

Language models have a good performance on specific reasoning tasks, although there is no general agreement on whether or not they have the ability to reason [17]. It has been demonstrated, however, that these models' ability to reason improves considerably with their parameter count. Given this, recently released LLM with over 100 billion parameters are better at reasoning [18].

Performance on reasoning tasks, however, is not only a matter of parameter count. It can be heavily improved through multiple methods, which are commonly classified as [19]:

- Strategy Enhanced Reasoning. As LLM usually contain implicit knowledge for reasoning from their pretraining [20], the focus in this method is how to take advantage of this knowledge. The main research area is prompt engineering, which defines how to construct the questions that are fed to the models. It can be single-stage or multi-stage, the latter emulating human reasoning, decomposing a complex problem and reasoning stage by stage. Both cases are also improved by the Chain-of-Thought (CoT) method [21], which generates a series of intermediate reasoning steps by providing demonstrations on the thought process inside the prompt. Other efforts towards Strategy Enhanced Reasoning include Process Optimization [22] and External Engines [23, 24].
- Knowledge Enhanced Reasoning. These methods focus on how to use both implicit and explicit knowledge to assist the model in reasoning. Regarding the implicit knowledge, there has been work to take advantage of the implicit knowledge contained in LLM to generate more knowledge and refine results [25]. As for explicit knowledge, efforts have been directed towards reducing hallucinations (the invention of incorrect facts) [26] and improving information retrieval from external files [27].

It is noteworthy to mention that the answer to better reasoning is not necessarily found with more training parameters. Recent research is also focused on smaller models, easier to use in production

environments, using explanations from bigger LLM to become better reasoners [28] [29].

## 3. Assisted Fact-Checking Framework

The framework proposed aims at assessing the improvement of retrieval and consumption of information from the press, in an attempt to improve fact-checking processes and reduce misinformation through machine learning techniques. Hence, the selected narrative of this work is that of assisted media consumption or assisted fact-checking. As seen in the literature review, these use cases contribute to around 30% of the intended uses of automated fact-checking tools [13]. It is important to have a human interaction in the models without it being fully automated, since some steps in fact-checking sometimes need to be done manually (e.g., calling an official source at a Ministry to verify a fact, which cannot be done online).

The proposed framework, illustrated in Figure 1, starts with building a knowledge base of digital media according to the interests of users: the sources and categories of articles of choice. This flexibility allows the framework to be versatile, as it can be used for any type of fact-checking, with information from the press, official documents or any private document base. For the implementation of this work, several pieces on unemployment from Spanish digital media were used as a knowledge base. We recommend that whenever this framework is used for political fact-checking, the knowledge base should ideally consist of a choice of media agencies having different types of audiences [30], to make sure the contents are diverse and can complement or contrast each other.

A machine learning model, specifically an LLM in this case, is used by a data actor (mostly journalists) to verify an input, having the knowledge base as a context. The output of the proposed framework consists of a classification of the given input, as well as a justification with citations to the sources supporting or refuting it. Prompt engineering and retrieval techniques are used to control the behaviour of the language model, in an effort to restrict its context to the given knowledge base without hallucinating information and giving false information to the user [31].

As for the evaluation of the performance of this framework, traditional benchmarks are not useful since the accuracy of the responses are not commonsense reasoning capabilities but depend on specific information from the context. Moreover, human evaluation has shown reproducibility limitations and instability towards the execution of NLP tasks [32]. Hence, the approach shifts from the usage of traditional benchmarks to the evaluation through elements inspired by the LLM-as-a-judge method [33]. In this case, given a fixed knowledge base, an LLM creates a series of potential claims given its context, as well as their classification (supported by the context, refuted by it, or with no information) and their justification. This serves as an evaluation dataset that needs to be manually revised and then utilised to extract performance metrics from the behaviour of the framework, with the same context and several combinations of prompts.

The implementation of the framework was done through LangChain, an open-source framework aimed at developing applications with LLM. Through components of this framework, LLM from OpenAI (gpt-3.5-turbo-1106, gpt-4-0125-preview[1]) and Cohere (Command[2]) were integrated. The embeddings used for this work were also from OpenAI (text-embedding-ada-002) and the LLM were used through API calls to providers offering them at no cost or at a limited one. To perform the evaluation, data consist on several pieces of information on unemployment data from Spanish digital media, in the Spanish language, from the following sites: El Plural, ABC, El Mundo, and Okdiario. These data were stored in an open-source vector database: Faiss. Lastly, the results of the framework were shown through the Gradio interface.

Several techniques were combined to improve the prompt composition [34]:

- Specifying the role: "You are a fact-checker."
- Explicitly asking to only use knowledge from the context provided.

[1] https://platform.openai.com/docs/models/
[2] https://docs.cohere.com/docs/models

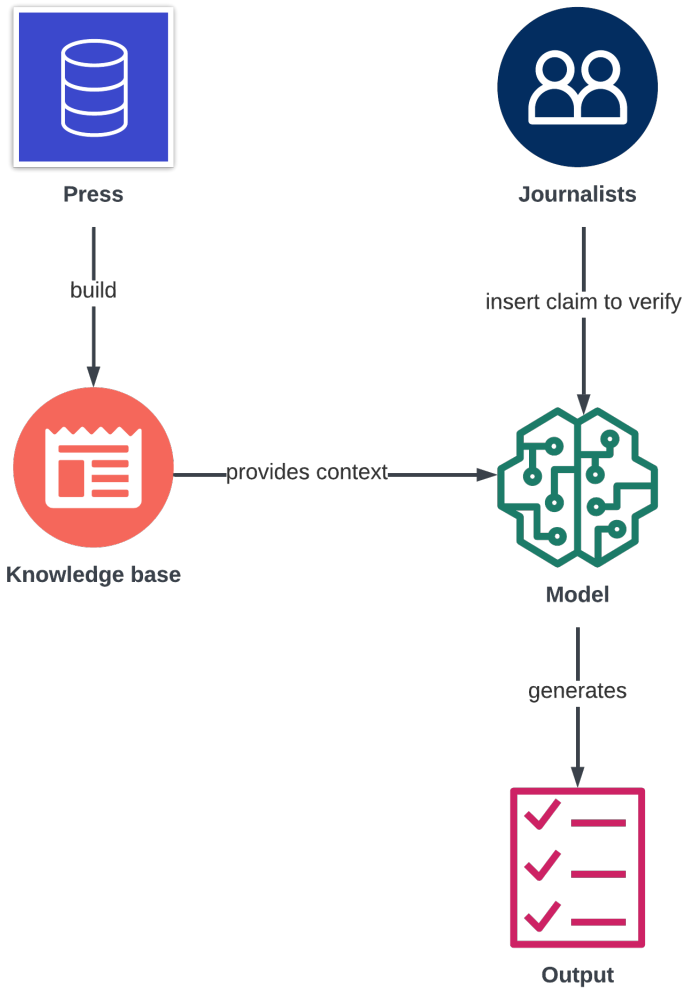


**Figure 1: Framework for Assisted Fact-Checking from the Press.**

- Explaining the format of the desired output: verdict, justification, and passages of each of the sources either supporting or refuting the given claim.
- Chain-of-Thought [21]: providing an example of how a claim can be verified.

The workflow of the implementation is depicted in Figure 2, through the diverse parts of the described process, leading to the different parts of the output being generated.

The model is invoked twice. In the first call to the model, the prompt only contains instructions about the justification. Specifically, it is told that it is a fact-checker that can only base its answers on the provided context. Then, a description of each of the categories that the system needs to classify the claim into is depicted, however, it is only asked to provide the justification. It is given also an example of how it should work. Finally, the Markdown format to return is specified, and the model is once again reminded to not produce anything that is not in the given context.

Next, the model is invoked for a second time. In this call, it is only asked to create a classification based on the justification from the response of the first call. The prompt once again explains the different values that the categories can have, and the output format (in Markdown) is specified.

As there is no single ground truth for the use case of this work, there is no standard traditional machine learning evaluation method. However, besides the qualitative evaluation, which serves only to identify whether some specific examples were functioning correctly, an additional evaluation is needed to asses the general behaviour of the framework.

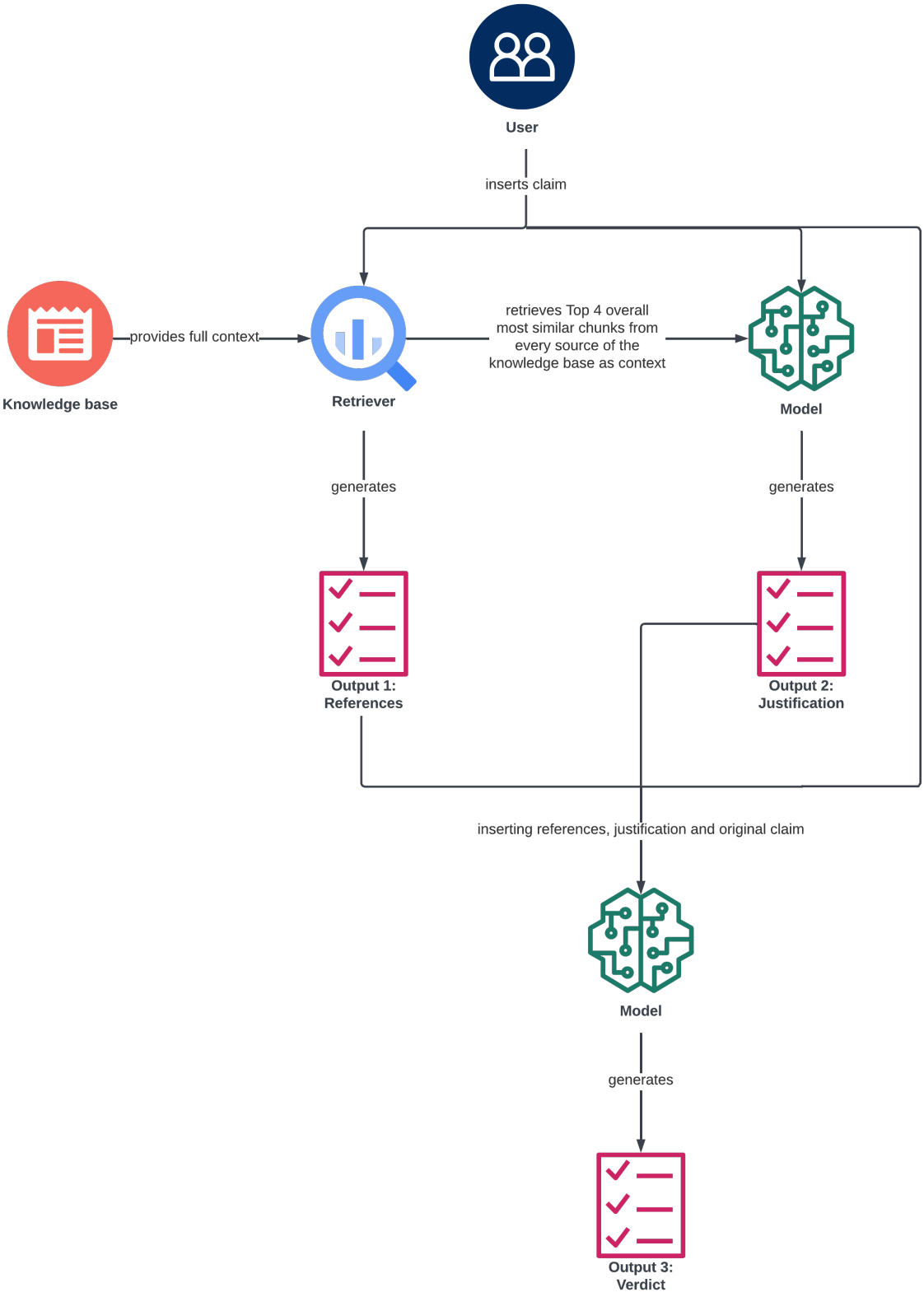


**Figure 2: Workflow for the Final Implementation.**

The evaluation was carried out in a mixed approach: automated and manual. The automated stages were the evaluation dataset creation and initial classification grading. On the other side, the classification adjustment and the justification grading were done manually.

GPT-4 was used to generate an evaluation dataset. Iteratively, each of the media pieces was passed to the model with a prompt asking to generate 40 claims from each of the news pieces. The prompt also specified the the claim can have - Supported, Refuted, Half-supported and No information - and a description of each of them, with the demand to classify the generated claim as well. Additionally, it was specified that there should be an equal number of each of the categories to not have unbalanced classes. The output format was demanded to be a Python dictionary in order to use it to create a dataframe afterwards, with three columns: claim, classification, and source (to keep traceability of where the claim was generated from).

After the generation of the evaluation dataset, each model (gpt-3.5-turbo-1106, gpt-4-0125-preview, Cohere Command) was invoked with the same prompt that is used in the final implementation, feeding it as a claim each generation of the evaluation dataset at a time. Therefore, a loop was created, iterating over the 160 claims that were generated for evaluation in total. The method then saved the results of each invocation, appending a column in each row with the verdict (or classification), justification, and the generated references. The results were afterwards opened in a Google Sheets file, where the automatic classification grading was done. A function was implemented to compare the column of the classification created as part of the evaluation dataset, with the one that was extracted from the output through a Google Sheets function. This would result, for each row, in a score of 0 or 1, with the latter being an exact match.

Moving to the manual stage, each row from the 480 generated in total (160 per model) was manually revised to find incorrect classifications from the evaluation dataset - changing the ground truth to an adjusted, correct version - and to find incorrect classifications of each model that could be accepted upon revision, if the justification was supporting it. The accuracy metric is created by the sum of the column of the classification adjustment, dividing by the number of rows evaluated (initially, 160). This would return a result between 0 and 1, later presented as percentage.

The final part of the evaluation is the justification grading. Each justification was graded with an answer correctness metric, assigned a score from 0 to 5. Score 0 corresponds to an output where none of the justification is correct, or it is classified as "No information" even though there exists relevant information in the context to provide a classification, while Score 5 is a justification that is entirely correct.

Justification grading was done manually following the guidance of the criteria above. To avoid inconsistencies, two rounds of grading were conducted, on different days, shuffling the order of the claims in the evaluation dataset. Afterwards, the scores of both rounds were compared, in case any of the claims were graded differently, and a final decision was taken in those cases where the scores varied. This methodology added rigorousness to an otherwise potentially subjective evaluation process.

At the end, after the grading, the score is divided by 5 in order to be a score between 0 and 1. The column of the normalised justification grading is added up and divided by the number of rows that are being evaluated (once again, initially, 160), so the resulting score for the answer correctness metric is also assigned a result between 0 and 1, later presented as percentage. It was deemed important to use both metrics, since both evaluate important functions of the framework that might be used independently.

## 4. Results and Discussion

The final results are shown in Table 1. Results show that the model with the best performance in terms of average of both metrics of classification and justification score is GPT-3.5, scoring over 70% in both metrics. Perhaps counter-intuitively, GPT-4 achieves significantly lower performance in terms of claim classification, although it achieves the top results in terms of justification. Lastly, the justification score of Cohere Command could not be calculated since it had issues with the justification language and format.

The evaluation dataset consisted initially of 160 claims and it underwent a manual revision of quality of generated claims and their classification by the judge model. As described in Section 3, this dataset was

**Table 1**
Quantitative results from evaluation.

| Model | Classification Score (%) | Justification Score (%) |
|---|---|---|
| GPT-3.5 | ***71.70*** | 70.82 |
| GPT-4 | 63.52 | ***73.58*** |
| Cohere Command | 44.65 | - |

generated by GPT-4, creating balanced categories of classifications (Supported, Refuted, Half-supported and No information) from each of the four sources (El Plural, ABC, El Mundo, and Okdiario).

As seen from the evaluation results, GPT-3.5 is the best-performing model among the three supported. Its strong points are that it had the best scores in classification and a close second place for justification production. Moreover, out of the 114 claims that were correctly classified, 84 of them (73.68%) also got the best score for its justification production.

Furthermore, there has been exactly one case of a claim that was incorrectly classified, but got the maximum score for the justification. It is, in fact, similar to one of the cases mentioned in the incorrect classifications from the evaluation dataset and it has to do with double negations and antonyms. More exactly, the claim was about the unemployment rate improving with regards to the one from 2022. It is probable that the justification was correctly created, but there was confusion with the model with the concept of improvement for terms like unemployment rate. The unemployment rate improved *for society* as the rate declined, but it seems that the model might have understood this phenomenon as an improvement if the rate actually increased. This example showcases the importance of considering both the classification and the justification in the usage of this proposed framework.

Additionally, there have been 17 claims rated with a score of 4 in justification production instead of the maximum of 5. Most of these claims' loss of the final point were related with inexact quantities or approximations. The output of the framework is almost correct in terms of justification, but it has been observed that in some cases, the model starts comparing several mentioned quantities as if they were completely different and not just a mere approximation. Prompting the model to behave in a specific manner when it had to do with approximations was tried in previous implementations in the experimental phase, although it did not have the expected result. If the behaviour of the framework when dealing with approximations improved and the results rated with a 4 were given a score of 5 instead, the results produced by GPT-3.5 could improve an additional 2.14% in terms of justification production, achieving a score of 72.96%. However, it is worth noting that there are cases where approximations are correctly handled. For instance, a claim that said there were 20 million employed people by the end of 2023 was correctly classified as "refuted" by the framework, since the correct number is 21.24 million and it is not a valid approximation.

Figure 3 shows an example of the functioning of the framework powered by GPT-3.5. The user input claims that the number of unemployed people increased in Spain in 2023. The answer given by the system classifies the claim as "Refuted", which is correct as 3/4 sources in the knowledge base support the contrary, whereas there is no relevant information in the fourth source. The justification is an accurate summary of the reasons why the system refutes the claim. Moreover, the references created by the system are also correct.

GPT-4 achieved short of 60% in the classification score, and more than 73%, surpassing GPT-3.5 for the justification production. These results are lower than expected for classification, as both models are from OpenAI but GPT-4 is a newer, bigger model with better results than GPT-3.5 on most benchmarks.

As seen from the contrast of its two scores, it is highlighted that it has lower classification capabilities in this current implementation. There are six claims classified as "No information", while the justification received the highest score possible, creating a faithful and complete reasoning on the claim. It is unknown why this behaviour occurs, since in the final implementation the methodology was changed. Instead of allowing the model to decide directly in the first invocation the classification of a claim, it had to do it in a second call, based only on the justification it previously created. Therefore, these cases should be reduced with this implementation.

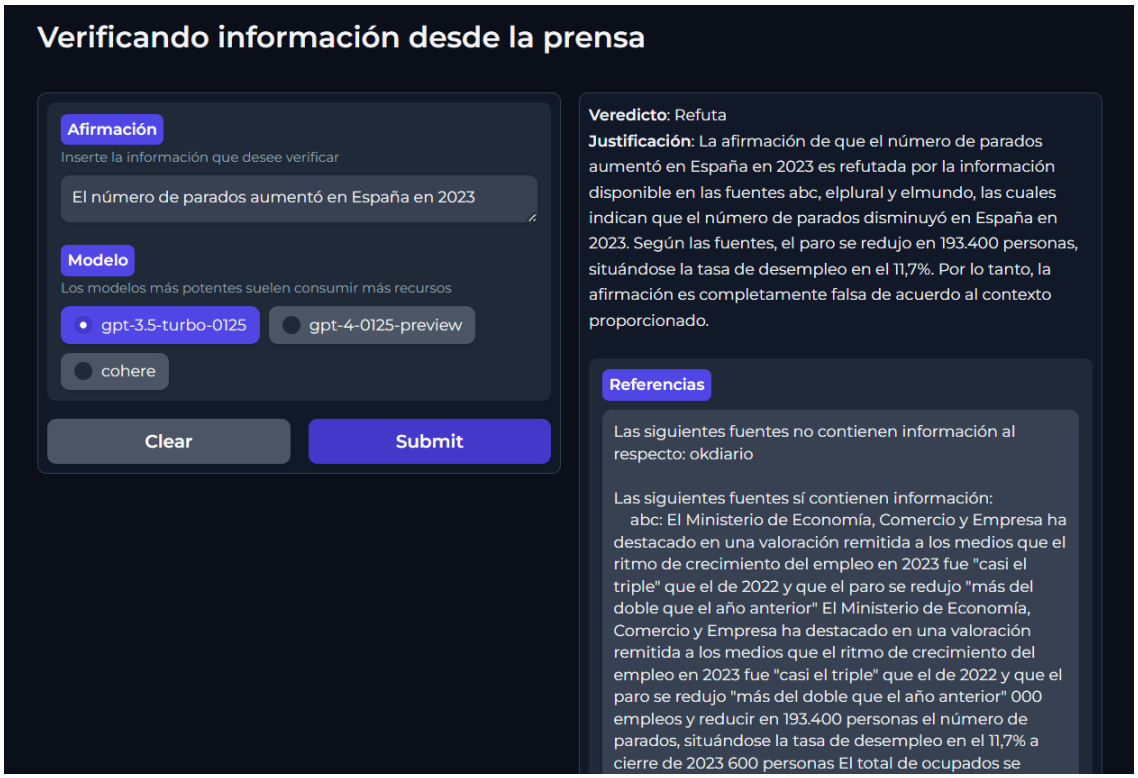


**Figure 3: Output from implementation.**

Additionally, there has been a recurring result that arose for 45 claims that impeded getting both a correct classification and a justification. After invoking the prompts, the output was "Understood, I am ready to start. Please, provide a claim and a justification" in 31 occasions, with another 14 cases returning "Verdict: No information. Justification: [Input given as justification]". After the first run of the evaluation, it had looked like an execution error, so the claims that led to this result were executed again. However, the same result was returned and no explanation was found to explain this behaviour, which leads to indicate an instability in the responses of GPT-4 with these prompts.

It is observed from the evaluation that the different models have varied performance levels. The best-performing model, GPT-3.5, achieves a score higher than 70% in both of the metrics that are evaluated. In this configuration, it is safe to say that the framework can be considered reliable when used as a tool for assisted fact-checking or media consumption, in a setting where a human checks the outputs instead of having a fully-automated environment.

One of the requirements needed for the framework was explainability, which is achieved mainly through the creation of references for each output. This is considered to be assured through the similarity search of the pieces of news given as input to the knowledge base. As the context is created through a more manual procedure, rather than given to the LLM to reason about, it is considered to be more reliable.

The creation of references is one of the strong points towards the trustworthiness of the proposed framework. However, it is worth noting that although there are procedures to avoid hallucinations or the invention of unrelated information in the output of LLM, they are not always completely avoided. This is why it is important to disclose to final users that the framework, if used for fact-checking or assisted media consumption, can be prone to occasionally produce such outputs.

Studies seem to suggest that English LLM trained at a very large scale can have almost as good results in other languages although there is still room for improvement [35]. This might also be the case for this research work: the performance could have been lower as the prompts and input data were designed to be used in Spanish. For instance, the confusion at reasoning might have been avoided in English. However, capabilities of language models are increasing at a considerable speed, therefore

the language shall not be an impediment for the adoption of the proposed framework as a solution for assisted fact-checking or media consumption.

Furthermore, given that the different models produce at times contradictory verdicts and reasoning, an improvement point to increase reliability could be a majority voting, or a weighted majority voting based on the evaluated performance of each method. This technique is used in well-known techniques in traditional machine learning such as Random Forests, where each of the trees have a vote and the final decision depends on the answer with most votes.

All in all, the proposed framework for fact-checking with information from the press provides a reliable solution for automating work that was traditionally manual for journalists, as well as opens new possibilities for non-professionals to consume more contrasted information from the news.

## 5. Ethical Considerations

The proposed framework is aimed to enhance the way information is consumed from the press, either with the intention of assisted fact-checking or mere media consumption in a new form. A new form of consuming information can have a considerable impact on society in case it is established. Therefore, it is important to assess ethical considerations of this proposed framework besides its implementation.

The Ethics Guidelines for Trustworthy AI developed by the European Commission in 2019 offer a systematic manner to assess the ethical considerations of the proposed framework through a set of requirements that any trustworthy AI system should meet [36]. The system proposed in this work can be systematically assessed in terms of ethics considering the following requirements from the guidelines:

- Human agency and oversight: In terms of human agency and autonomy, it is vital to stress that the proposed framework is a system based on information from the press, which in no case can fully assure that the information on which it relies is totally factual. Therefore, over-reliance shall be avoided. As for the concept of oversight, it is not considered an autonomous system, as it is needed to be overseen by a *Human*-in-the-Loop.
- Technical Robustness and safety: A low level of accuracy could create undesired results from the system. However, as it would have human supervision and reference checking, the consequences could not be damaging. The final levels of accuracy achieved through the final implementation would need to be properly communicated to end-users for them to acknowledge the behaviour and limitations of the system in order to align their expectations. Finally, in relation to the training data and assumptions that the LLM were trained on, they have not been observed to lead to adversarial effects during the experiments, mostly due to the explicit prompting to follow instructions and only rely on the data given as a context.
- Privacy and data governance: The framework does not use any personal data and only uses publicly available data.
- Transparency: There are three main elements that constitute transparency as a requirement. First, traceability is important to track; for this system, the version of the models used, the prompts with which the models are invoked, and the data that make up the knowledge base. Next, explainability is vital for building trust in the AI system, and this is achieved in this framework through the creation of a justification and the references used for that purpose. The last element of transparency is communication. It is clearly communicated that the framework is an AI system and not a human, as well as its benefits, limitations, potential risks, level of accuracy, and error rates. For the sake of transparency, it is also recommended to use open-source models in an industrial implementation of this framework. This research project has only considered closed-source LLM as there was no budget allocated for hardware or API usage. However, transparency would be improved through the usage of open-source models that are presenting performances comparable GPT-3.5 and GPT-4, like models from Llama 2 from Meta[3] and Mixtral from Mistral

[3] https://llama.meta.com/llama2

AI[4].

- Diversity, non-discrimination, and fairness: The intention when implementing this work is to always avoid unfair bias. This has been done by carefully crafting the prompts in order to leave out expressions that could leave room for subjectivity. However, it is necessary to include a disclaimer about biases that could already exist. These biases can either be in the data from the press - as the system needs to be faithful to that information - or in the training data of the LLM, although this is less common as the instructions are clearly defined to not use data from the training. Moreover, as it is advisable to consult stakeholders affected by the AI system throughout the whole life cycle, experts in fact-checking from a Spanish fact-checking start-up were consulted about functional feedback regarding the feedback. This was done in order to ensure that the system's design and development was taking into account the actual needs of professionals that could benefit from this system.
- Societal and environmental well-being: The implementation of the proposed framework, if used at a large scale, could have impact on human work and society at large. It would have the potential to change some aspects from journalism, specifically fact-checking, as evidence retrieval would be faster and journalists and fact-checkers could benefit from the time saved to invest in other tasks. On the other side, the usage of this framework for assisted fact-checking could impact favourably society at large by the reduction of misinformation and disinformation. However, it would also pose a challenge: learning a new way to digest information, as facts and claims would already reach people with a justification created, and critical thinking could decrease.
- Accountability: The functioning of the framework is documented and it can be externally audited. Moreover, it is also well-communicated to end-users about the limitations and data sources of the system, since the framework ultimately verifies if a claim is supported or not by the context provided, not if it is factually true or false. Therefore, the responsibility of the accuracy of the information falls under the data sources.

Through this assessment of requirements, it can be concluded that the proposed framework can be considered a responsible application of AI.

The other ethical aspect that needs to be studied prior to any deployment of the proposed framework in a production environment is where the data come from: What should the process to collect news pieces look like? How should the digital outlets be picked in this regard?

The New York Times, one of the longest-running newspapers in the United States, sued OpenAI in December 2023 over content created by ChatGPT [37]. The lawsuit informed of several issues with this content that the newspaper claimed were hurting the brand and functioning of the New York Times. The two most pressing issues were the regurgitation of full articles from the New York Times that ChatGPT would perform if prompted correctly, and its hallucinations where false or inaccurate information was created and then attributed them to the New York Times. Both of these problems can affect the image and the finances of the publisher.

The proposed framework pursues a new way of consuming information, joining several sources to provide a complete picture of the context of a claim users wish to consult. This is always done attributing the original authors, displaying each of the sources' positions on the claim to be checked. Although credit is important, it might not be enough. It is important to know where the information comes from and provide a way to consult the source data directly in case it is needed. However, in case the framework is adopted at large scale, it could reduce the traffic on the information sources' websites.

For all of this, it is considered that the framework, rather than a framework with direct web-scrapping, should be proposed as a collaboration with diverse digital newspapers or media outlets.

[4] https://mistral.ai/technology/#models

## 6. Conclusions

In this work, we have studied the interconnection between AI, journalism, and ethics in an attempt to create a framework whose ultimate goal is to reduce misinformation in society. The proposed framework is powered by reasoning capabilities of LLM, as it allows users to contrast a given claim with a previously built knowledge base, based on sources of interest. The claim is classified based on its alignment with the knowledge base, and a justification and references are also returned.

The functioning of the framework is evaluated with a mix of automated and manual techniques, ultimately returning a percentage of classification and justification accuracy. The best-performing model out of the several that have been studied - GPT-3.5 - scored over 70% in both metrics.

It is important to consider all parts of the output - verdict, justification, and references - when using the system, since there are cases when only some of the parts are correct. However, even given this limitation, the tool can serve as a companion for professionals and non-professionals when consuming information. The proposed framework can reliably be used as a tool for assisted fact-checking or assisted media consumption.

Overall, the broad implication of the present research work is that it is possible to use an AI-based framework to enhance the retrieval of information from the press in a responsible manner, showing that AI may be considered a promising companion tool to journalists and non-professionals wanting to contrast information.

We are currently working on testing new state-of-the-art Large Language Models, since they have the potential to improve the current performance. New models are getting rolled out at a very fast pace and benchmark scores are improving.

We are putting focus on open-source releases, since there are models already surpassing GPT-3.5 as they are getting support of the open-source community, so they would be worth testing in the scope of the framework. Minimal code modifications are being performed, since the framework is designed to support any LLM.

We also aim at modifying current methods to increase accuracy and faithfulness. There are several methods that might increase the performance of the current framework without more powerful LLM. One of them could be the combination of the answers of several methods, either by a weighted majority-voting based on the evaluated performance, or through a third call giving the outputs of the models as a new context and letting a model decide based on that information.

## References

[1] M. F. Çömlekçi, Why do fact-checking organizations go beyond fact-checking? a leap toward media and information literacy education, International Journal of Communication 16 (2022) 21.

[2] V. Moreno-Gil, X. Ramon-Vegas, M. Mauri-Ríos, Bringing journalism back to its roots: examining fact-checking practices, methods, and challenges in the Mediterranean context, Profesional de la información 31 (2022).

[3] M. Himma-Kadakas, I. Ojamets, Debunking false information: investigating journalists' fact-checking skills, Digital journalism 10 (2022) 866–887.

[4] L. Graves, Understanding the Promise and Limits of Automated Fact-Checking, Reuters Institute for the Study of Journalism, University of Oxford (2018). URL: https://api.semanticscholar.org/CorpusID:13750196.

[5] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[6] Misinformation versus disinformation, explained | The Foundation for Individual Rights and Expression, https://www.thefire.org/research-learn/misinformation-versus-disinformation-explained, ???? Foundation for Individual Rights and Expression.

[7] B. Borel, The Chicago Guide to Fact-Checking, University of Chicago Press, 2016.

[8] A. Vlachos, S. Riedel, Fact Checking: Task definition and dataset construction, in: Workshop on Language Technologies and Computational Social Science, 2014, pp. 18–22.
[9] L. Dierickx, C.-G. Lindén, A. L. Opdahl, Automated Fact-Checking to Support Professional Practices: Systematic Literature Review and Meta-Analysis, International Journal of Communication 17 (2023) 21.
[10] M. Li, B. Peng, Z. Zhang, Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models, 2023. ArXiv:2305.14623 [cs].
[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, 2018, pp. 809–819.
[12] R. Kamoi, T. Goyal, J. Diego Rodriguez, G. Durrett, WiCE: Real-world entailment for claims in Wikipedia, in: H. Bouamor, J. Pino, K. Bali (Eds.), Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7561–7583.
[13] M. Schlichtkrull, N. Ousidhoum, A. Vlachos, The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who, in: Findings of the Association for Computational Linguistics: EMNLP, 2023, pp. 8618–8642.
[14] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, Authorea Preprints (2023).
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, 2017.
[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.
[17] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 1049–1065.
[18] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).
[19] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, H. Chen, Reasoning with Language Model Prompting: A Survey, 2023. ArXiv:2212.09597 [cs].
[20] B. Paranjape, J. Michael, M. Ghazvininejad, H. Hajishirzi, L. Zettlemoyer, Prompting contrastive explanations for commonsense reasoning tasks, in: Findings of the Association for Computational Linguistics, 2021, pp. 4179–4192.
[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).
[23] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, A. M. Dai, Mind's Eye: Grounded Language Model Reasoning through Simulation, in: The Eleventh International Conference on Learning Representations, 2023.
[24] A. Madaan, S. Zhou, U. Alon, Y. Yang, G. Neubig, Language Models of Code are Few-Shot Commonsense Learners, in: Conference on Empirical Methods in Natural Language Processing, 2022, pp. 1384–1403.
[25] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, H. Hajishirzi, Generated knowledge prompting for commonsense reasoning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 3154–3169.

[26] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, Object Hallucination in Image Captioning, in: Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4035–4045.

[27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: 34th International Conference on Neural Information Processing Systems, 2020.

[28] S. Li, J. Chen, Y. Shen, Z. Chen, X. Zhang, Z. Li, H. Wang, J. Qian, B. Peng, Y. Mao, et al., Explanations from large language models make small reasoners better, arXiv preprint arXiv:2210.06726 (2022).

[29] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of gpt-4, arXiv preprint arXiv:2306.02707 (2023).

[30] F. Guerrero-Solé, The ideology of media: Measuring the political leaning of Spanish news media through Twitter users' interactions, Comunicación y sociedad = Communication & Society 35 (2022) 29–43.

[31] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, arXiv preprint arXiv:2104.07567 (2021).

[32] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, arXiv preprint arXiv:2305.01937 (2023).

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Advances in Neural Information Processing Systems 36 (2024).

[34] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[35] J. Armengol-Estapé, O. d. G. Bonet, M. Melero, On the multilingual capabilities of very large-scale english language models, arXiv preprint arXiv:2108.13349 (2021).

[36] Ethics guidelines for trustworthy AI, Publications Office of the European Union, 2019. URL: https://data.europa.eu/doi/10.2759/346720, directorate-General for Communications Networks, Content and Technology (European Commission) and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji.

[37] E. Helmore, K. Paul, New York Times sues OpenAI and Microsoft for copyright infringement, The Guardian (2023). URL: https://www.theguardian.com/media/2023/dec/27/new-york-times-openai-microsoft-lawsuit.