Claim Verification Leveraging In-Context Learning and Retrieval Augmented Generation

Giuseppe Fenza, Domenico Furno, Mariacristina Gallo, Vincenzo Loia, and Pio Pasquale Trotta

University of Salerno, Fisciano, SA, 84084, Italy
{gfenza, dfurno, mgallo, loia}@unisa.it, p.trotta11@studenti.unisa.it

Abstract. The proliferation of digital information has escalated the necessity for efficient and scalable fact-checking methods to combat misinformation. Existing solutions mainly rely on customized learning models that leverage ad-hoc training data and do not fit well in different domains. Indeed, claim verification stresses the availability of an updated and open knowledge base for the designed model. This paper proposes a novel approach that integrates Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to enhance claim verification, leveraging in-context learning to assess the veracity of input claims. The methodology involves a two-step process, the evidence retrieval — including web document summarization and claim-focused relation extraction — and claim validation mainly consisting of triple relation extraction and comparison. Evidence retrieval, by filtering information sources, guarantees the reliability of the verdict, enabling the claim verification feasibility in multiple domains. Experimental activities are conducted using the FEVER dataset, and the results demonstrate that the proposed framework significantly improves claim verification accuracy at the state-of-art.

Keywords: Retrieval Augmented Generation (RAG) · Claim Verification · Fact-Checking · Large Language Model · Information Disorder.

1 Introduction

The proliferation of digital information has made claim verification a critical task in ensuring the accuracy of disseminated knowledge. The rapid spread of misinformation across various platforms underscores the necessity for robust and efficient methods to evaluate the veracity of claims. Traditional fact-checking processes, while effective, are labor-intensive and cannot scale to meet the demands of the current information landscape. Advanced computational approaches have been developed to address these challenges, among which Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) stand out due to their innovative architectures and promising results [10].

Retrieval-Augmented Generation (RAG) combines the strengths of information retrieval and generative modeling to enhance the accuracy and contextual relevance of claim verification. RAG employs a two-step process: first, it uses retrieval mechanisms to fetch relevant documents from extensive datasets, and second, it utilizes these documents as context to generate informed responses. This approach allows RAG to ground its outputs in up-to-date and specific information, addressing a significant limitation of purely generative models, which often rely solely on their training data and may lack current knowledge [13],[18].

Large Language Models (LLMs), such as GPT-4 [1] and BERT [14], have demonstrated exceptional natural language understanding and generation capabilities. These models are trained on diverse and massive datasets, enabling them to comprehend and generate human-like text across a wide range of topics. However, in the context of claim verification, LLMs face a significant challenge. Their reliance on static training data poses challenges in dynamic and rapidly evolving information environments, where the veracity of claims can change quickly [11], [29], [9]. This limitation underscores the need for a more adaptive and context-aware approach to claim verification [17]. Therefore, to further enhance the accuracy and reliability of claim verification, the proposed approach also leverages LLMs and few-shot learning to incorporate the extraction of triple relations, which is a task where Large Language Models showed strong performance [27] from open documents and the claims themselves [12]. Triple extraction involves identifying subject-predicate-object relationships within the text, which provides a structured representation of information. This structured data helps improve the match between claims and evidence, facilitating more precise verification [23]. Moreover, due to the unreliable nature of many web sources, documents are also filtered considering Newsguard's reliability ratings, which allow to exclude those sources known for spreading online low-quality information or are affected by political bias, conflict of interests, etc. [7].

In the current work, experiments using the Fact Extraction and VERification (FEVER) dataset have been conducted to explore the feasibility and validity of the proposed framework for claim verification. The FEVER dataset provides a comprehensive collection of claims and corresponding evidence from Wikipedia, enabling rigorous testing of claim verification models [15].

Promising experimentation results demonstrate that the proposed RAG-based framework, strengthened by triple relation extraction, enhances the accuracy, reliability, and efficiency of automated claim verification systems, providing a scalable solution to combat misinformation in the digital age.

The remaining parts of the paper are organized as follows. Section 2 explores the existing literature in the field of claim verification. The proposed framework is described in Section 3, detailing the Evidence Retrieval and Claim Verification phases and their sub-phases. The experimental results are compared with baselines on the same benchmark in Section 4. Finally, the limitations (Section 5) and conclusions (Section 6) of the work are presented.

2 Related works

In the literature, the claim validation problem is mainly approached through automated processes that typically involve three primary steps treated collectively or individually [10]: (i) *claim detection*, (ii) *evidence retrieval*, and (iii) *claim validation* or *verification*. Often, evidence retrieval and claim validation are treated as a unified task. Moreover, claim validation may involve providing an *explanation* or *justification* for the verdict on a claim.

Claim detection is the process of identifying which assertions require fact-checking, thereby reducing the volume of content to be analyzed. Many contemporary approaches involve fine-tuning existing language models such as BERT and T5. For instance, [8] highlights the role of tweet-meta-features in detecting check-worthy tweets, integrating these features into the classification process. Another method for claim detection focuses on identifying claims that have been previously fact-checked. This involves verifying if the claim exists in a database and can be addressed by referencing an earlier fact-check. Ranking methods are utilized to evaluate the similarity between the input claim and those in the database. In [23], a learning-to-rank method is suggested. Additionally, Alhindi et al. [2] created an annotated corpus of checkworthiness in the context of climate change.

Evidence retrieval is essential for locating pertinent information confirming or refuting a claim. Including this step in the fact-checking process enhances comprehension of the claim assertions, facilitating effective comparisons with more reliable sources. This can be achieved by utilizing traditional information retrieval techniques [24], employing search engines such as Google [28] or relying on existing knowledge graphs (e.g., Wikidata) [31], [6]. However, an interesting approach is presented in [4] where evidence search focuses on documents available before the claim making, modeling the realistic scenario of emerging claims.

Claim validation focuses on assessing the veracity of an input claim. The simplest methods used binary classification with labeled datasets [20]. More advanced techniques involve referencing textual sources or knowledge bases and utilizing the outcomes of evidence retrieval to make informed decisions, as proposed by Pankovska et al. [21]. Additionally, some methods use graphs to facilitate claim validation, as shown in [19]. Recent works propose different approaches employing large language models. In [26], [5], [6], LLMs are fine-tuned on specific data in order to achieve better results; however, this process may become really expensive in terms of computational and resource expenditure. Alternatively, [30] leverages the in-context learning, demonstrating the strong capabilities of LLMs through few-shot prompting for the described task. Lastly, Lee et al. [16] propose an approach based on perplexity scores calculated by pretrained language models using claim-evidence pairs to determine if the claim supported or not. Also, they fine-tune other models (e.g., BERT-B, BERT-L, RoBERTa, XLNet) to achieve other results and use them as baselines and make additional comparisons.

This manuscript proposes a claim verification solution that implements a Retrieval-Augmented Generation (RAG)-based approach in which an evidence retrieval phase collects relevant information and structures it. Then, a claim verification phase inquiries a Large Language Model to classify the claim by leveraging the structured pieces of evidence.

3 Methodology

As aforementioned, the claim validation problem is usually dealt with three-step processes devoted to identifying the claim to validate (i.e., *Claim detection*), locating pieces of evidence confirming or refuting the claim (i.e., *Evidence Retrieval*), and assessing the verdict for the claim (i.e., *Claim Validation*).

The proposed methodology covers two of these steps: it starts from a given claim (so without passing through the *Claim Detection* phase) and exploits web documents to feed a Retrieval-Augmented Generation framework which gives the final classification for the claim. In particular, as expressed in Fig. 1, the **Evidence Retrieval** is achieved through two additional sub-phases. In *Web Document Summarization*, relevant web pages intercepted by a search engine are scraped in order to extract the content and summarize it through the support of an LLM. In *Claim-focused Relation Extraction*, an LLM constructs triple relations from summaries previously extracted.

The **Claim Verification** phase deals with the extraction of triple relations from the input claim through the LLM (i.e., *Claim Relation Extraction* sub-phase) and comparing such relations against ones from summaries of the previous phase to produce a final Verdict (i.e., *Relation Comparison* sub-phase).

The following subsections detail each of the aforementioned phases.

3.1 Evidence Retrieval

The Evidence Retrieval phase aims to obtain relevant, updated, and reliable information regarding the input claim from multiple sources. It is further divided into two sub-phases, as detailed below.

Web Document Summarization The input claim is searched on the Google search engine by adopting the Serper Google Search API[1]. It returns a set of relevant URLs and related snippets. Results are filtered based on a reliability score of their belonging domain to enhance the objectivity of the outcome. In particular, a reliability threshold is applied to NewsGuard's score of the web domain to identify reliable sources. NewsGuard[2] is a service that provides ratings on the reliability and transparency of news and information websites. Ratings are made by a team of in-house journalists who follow a defined set of criteria. Based on Newsguard's rating ranges[3], the threshold has been set to 0.6.

[1] https://serper.dev/

[2] https://www.newsguardtech.com/ratings/rating-process-criteria/

[3] https://www.newsguardtech.com/ratings/rating-process-criteria/

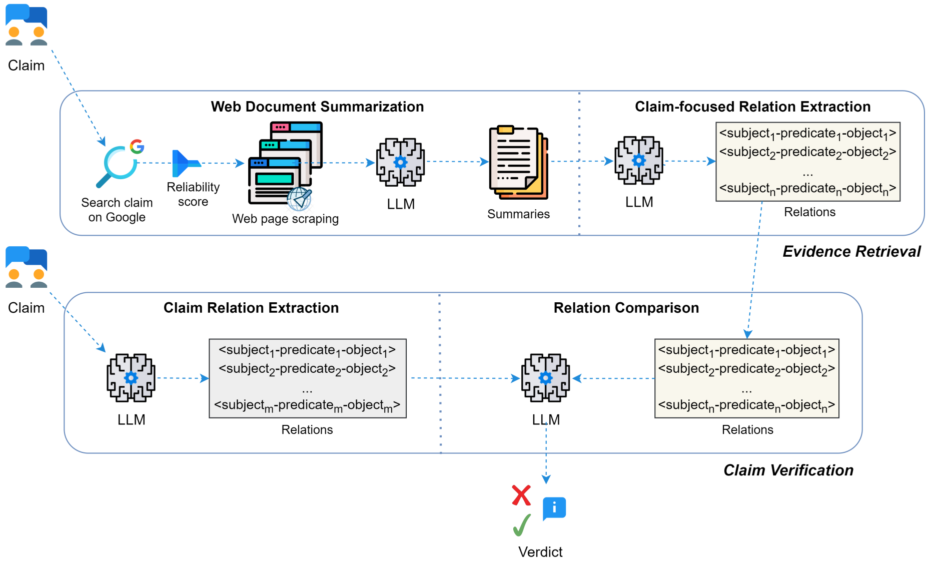

Claim
Web Document Summarization
Search claim on Google
Reliability score
Web page scraping
LLM
Summaries
Claim-focused Relation Extraction
LLM
<subject1-predicate1-object1>
<subject2-predicate2-object2>
...
<subjectn-predicaten-objectn>
Relations
Evidence Retrieval
Claim
Claim Relation Extraction
LLM
<subject1-predicate1-object1>
<subject2-predicate2-object2>
...
<subjectm-predicatem-objectm>
Relations
Relation Comparison
LLM
<subject1-predicate1-object1>
<subject2-predicate2-object2>
...
<subjectn-predicaten-objectn>
Relations
Claim Verification
Verdict

Fig. 1. Methodology overview. The proposed framework presents two main phases: (1) *Evidence Retrieval* for the collection of useful evidence supporting the outcome; (2) *Claim Verification* in which relations extracted from pieces of evidence and from the claim are compared by the LLM to decide the verdict.

Once sources are filtered, a scraping process based on the Trafilatura tool[4] [3] retrieves website contents, creating a collection of heterogeneous claim-related documents (e.g., news articles, Wikipedia pages, blog posts, etc.). When scraping is unsuccessful, the related snippet is considered instead of the website content. Finally, each retrieved document is summarized by an LLM, Gemini 1.5 Pro [22], through the Google Cloud Vertex AI API[5]. The idea is to balance the length of extracted information and its value for the subsequent validation of the input claim.

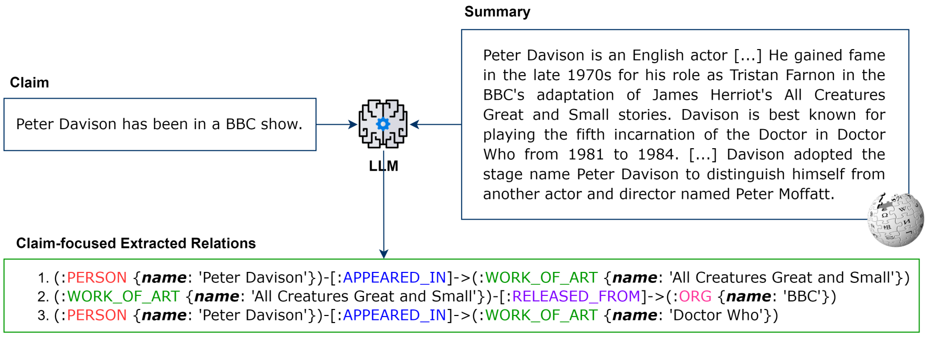
Claim-focused Relation Extraction A few-shot prompt is configured to analyze summaries of the previous step and extract triple relations relevant to the given claim. Specifically, this prompt to Gemini 1.5 Pro contains instructions for extracting relations from produced summaries and includes a list of possible entities categorized into distinct types, such as $PERSON$, GPE, LOC, $EVENT$, accompanied by brief descriptions that facilitate a deeper understanding of each entity. It also provides examples of relations between those entities (e.g., $BORN\_IN$, $OCCUPATION$, $DIRECTED$, etc.). This structured approach enables the extraction of meaningful relationships between entities, allowing for a more accurate and comprehensive analysis of the claims in question.

[4] https://trafilatura.readthedocs.io/en/latest/

[5] https://cloud.google.com/vertex-ai/docs/reference/rest

In particular, in order to cover the wide range of heterogeneous relations from summaries and to exploit LLM in-context learning capabilities, a 10-shot prompt is employed to provide the LLM with examples of relations extracted considering a given claim and a specifically related summary. If relevant relations cannot be extracted from a given summary or the snippet for a specific claim, the LLM is encouraged to output the message *No Relevant relations for the given claim*. In such a case, the claim is not considered during the *Relation Comparison* sub-phase.

The example of relation extraction results in Fig. 2 shows that, given the claim and the summary of the Wikipedia page resulting from the previous phase, the LLM extracts three relations.



Peter Davison has been in a BBC show.

Peter Davison is an English actor [...] He gained fame in the late 1970s for his role as Tristan Farnon in the BBC's adaptation of James Herriot's All Creatures Great and Small stories. Davison is best known for playing the fifth incarnation of the Doctor in Doctor Who from 1981 to 1984. [...] Davison adopted the stage name Peter Davison to distinguish himself from another actor and director named Peter Moffatt.

1. (:PERSON {***name***: 'Peter Davison'})-[:APPEARED_IN]->(:WORK_OF_ART {***name***: 'All Creatures Great and Small'})
2. (:WORK_OF_ART {***name***: 'All Creatures Great and Small'})-[:RELEASED_FROM]->(:ORG {***name***: 'BBC'})
3. (:PERSON {***name***: 'Peter Davison'})-[:APPEARED_IN]->(:WORK_OF_ART {***name***: 'Doctor Who'})

Fig. 2. Example of extraction of claim-focused relations. The summary of web page content (a Wikipedia page, in this case) is given in input to the LLM together with the claim to produce pertinent relations.

3.2 Claim Verification

The Claim Verification phase leverages the previously extracted relations to determine whether prompting the LLM (i.e., Gemini 1.5 Pro) with more structured pieces of evidence could improve its reasoning abilities and strengthen the claim verification process. This phase can be divided as follows:

Claim Relation Extraction The first part of the Claim Verification phase involves mapping the input claim in triple relations through Gemini 1.5 Pro. During this phase, a 16-shot prompt is employed to provide illustrative examples of the process. As shown in Listing 1.1, this prompt includes a comprehensive list of potential entities and relations consistent with those delineated in the earlier *Claim-focused Relation Extraction* step. Including the same list of possible entities and relations ensures continuity and coherence throughout the extraction phases. The purpose of utilizing a 16-shot prompt is to enhance the model's ability to understand and identify the complex relationships and entities within the claims and extract relations that can accurately represent the underlying claim structure.

```
Your task is to list the Relations such as [OCCUPATION, LIVE_IN,
WORK_FOR, PLAYED, ACTED_IN, DIRECTED, WROTE, PRODUCED, and other
relations like the already listed] among the following possible
entities [PERSON, NORP (nationalities, religious and political
groups), FAC (buildings, airports etc.), ORG (organizations), GPE
(countries, cities etc.),LOC (mountain ranges, water bodies etc.),
PRODUCT (products), EVENT (event names), WORK_OF_ART (books, song
titles, movies, paintings etc.), LAW (legal document titles),
LANGUAGE (named languages), DATE, TIME, PERCENT, MONEY, QUANTITY,
ORDINAL, TYPE (movies genres, different type of arts, music genres
etc.) JOB and CARDINAL], in the given CLAIM.

CLAIM: Nikolaj Coster-Waldau worked with the Fox Broadcasting
Company.
Relations: [(:PERSON {name: ’Nikolaj Coster-Waldau’})-[:WORKED_FOR
]->(:ORG {name:’Fox Broadcasting Company’})]</s>

...

CLAIM: Adrienne Bailon is an accountant.
Relations:
```

Listing 1.1. Example of few-shot prompt for relation extraction from the input claim.

Relation Comparison The Relation Comparison phase evaluates relations extracted from the input claim and relations from the related summaries to decide the verdict for the claim. In particular, summary-related relations are considered once at a time, and a list of verdicts maintains the output given by the LLM for the relations of each claim-related summary. The framework considers the most frequent verdict to assess whether the claim is supported or refuted and determine the final outcome.

Details about the Relation Comparison process are outlined in Algorithm 1. It is designed to determine the final verdict $\hat{v}_c$ for a given claim c by comparing its relations with those extracted from related summaries. The process begins (Line 1) by initializing the claim c and its associated relations r_c. A set of summaries S and their respective relations R are also established. An empty set V_c is created to store partial verdicts (Line 5), and counters for supporting ns, refuting nr and non-relevant ne verdicts are initialized to zero.

For each summary s_j, from Line 7 to Line 20, the algorithm checks if the extracted relations r_{s_j} are relevant. If r_{s_j} does not contain relevant relations, the partial verdict $NOT\ ENOUGH\ INFO$ is added to V_c, and the non-relevant counter ne is incremented. If relevant relations are found, the algorithm uses the LLM (i.e., Gemini 1.5 Pro) to compare c, r_c, and r_{s_j} to generate a verdict v (Line 12), which is then added to V_c. Depending on whether the verdict v is $SUPPORTS$ or $REFUTES$, the corresponding counter (ns or nr) is incremented.

After processing all summaries (Line 21), the algorithm determines the final verdict $\hat{v}_c$. If all partial verdicts indicate $NOT\ ENOUGH\ INFO$ or the counts of supporting and refuting verdicts are equal, the claim c is excluded from further consideration. Otherwise, the final verdict $\hat{v}_c$ is set to $SUPPORTS$ if ns exceeds nr, or $REFUTES$ if nr exceeds ns. The final verdict $\hat{v}_c$ is then returned (Line 28), concluding the Relation Comparison phase.

Algorithm 1 Relation Comparison to obtain the set of partial verdicts for a single claim c and the final verdict $\hat{v}_c$ for the claim.

```
 1: $c \leftarrow$ claim
 2: $r_c \leftarrow$ claim relations
 3: $S \leftarrow \{s_1, \ldots, s_i\}$                    ▷ Set of summaries for $c$
 4: $R \leftarrow \{r_{s_1}, \ldots, r_{s_i}\}$          ▷ Set of relations for each summary
 5: $V_c \leftarrow \emptyset$                             ▷ Set of partial verdicts for $c$
 6: $ns, nr, ne \leftarrow 0, 0, 0$                        ▷ Initialize counts
 7: for $j \leftarrow 1$ to $i$ do
 8:     if $r_{s_j} =$ No Relevant relations then
 9:         $V_c \leftarrow V_c \cup \{$Not Enough Info$\}$
10:         $ne \leftarrow ne + 1$
11:     else
12:         $v \leftarrow \mathrm{LLM}(c, r_c, r_{s_j})$
13:         $V_c \leftarrow V_c \cup \{v\}$
14:         if $v =$ SUPPORTS then
15:             $ns \leftarrow ns + 1$
16:         else if $v =$ REFUTES then
17:             $nr \leftarrow nr + 1$
18:         end if
19:     end if
20: end for
21: if $ne = \mathrm{length}(V_c)$ or $ns = nr$ then
22:     exclude $c$
23: else if $ns > nr$ then
24:     $\hat{v}_c \leftarrow$ SUPPORTS
25: else if $nr > ns$ then
26:     $\hat{v}_c \leftarrow$ REFUTES
27: end if
28: return $\hat{v}_c$
```

4 Experiments & Analysis

This section details experimentation activities carried out to measure the feasibility of the proposed approach. It gives details about the adopted dataset and achieved results. In particular, the resulting performance is compared with different baselines and an ablation study, in which the process excludes operations of relation extraction, is presented.

4.1 Dataset

The framework evaluation passed through the adoption of the Fact Extraction and VERification (FEVER) dataset [25], one of the most used datasets for the claim verification task where claims are labeled as $\{SUPPORTS, REFUTES, NOT\ ENOUGH\ INFO\}$ by human annotators and created by reshaping sentences from Wikipedia. It also contains (for each claim) sentences extracted from Wikipedia pages used by the annotators to carry out the labeling process. The work relies on the FEVER Development dataset composed of $19,998$ claims.

The system is evaluated by considering claims labeled as $SUPPORTS$ or $REFUTES$ from the FEVER Development dataset for a 2-way classification for a total of $13,332$ claims.

4.2 Binary Classification Baselines

The method is compared with other state-of-the-art approaches for FEVER binary classification. In particular, as detailed below, comparisons involved methods based on Large Language Models:

- A perplexity-based method [16], called PPL, leverages conditional perplexity scores to establish a threshold for classification. Specifically, those scores are produced by pre-trained language models and are useful to assess whether a claim-evidence pair is supported. If the perplexity score of a given claim-evidence pair exceeds the threshold, the pair is assigned a *Supports* label; conversely, if the score falls below the threshold, it is assigned a $Refutes$ label.
- Fine-tuned models, such as $BERT - B_{ft}$ and $XLNET_{ft}$ [16] for binary classification.

4.3 Evaluation Metrics

The performance of the proposed method is evaluated in terms of two key metrics: Accuracy and F1-macro. Focusing on two distinct labels, *Supports* and $Refutes$, Accuracy is the ratio of correctly predicted instances to the total instances. It is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The F1-macro score is a type of F1 score that is calculated for each class and then averaged to provide a single performance metric. It treats all classes equally, regardless of their frequency, making it particularly useful when dealing with imbalanced datasets. The F1 score for a single class is the harmonic mean of Precision and Recall, defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1-macro score is then computed by averaging the F1 scores of all classes:

$$\text{F1-macro} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i \tag{5}$$

4.4 Results

This section presents the results of the proposed method compared with the best scores chosen from the baselines described above.

Table 1. Accuracy and F1-Macro of the proposed method compared with the baselines.

Model	Accuracy (%)	F1-macro (%)
$BERT - B_{ft}$	52.18	38.82
$XLNET_{ft}$	49.18	48.42
$PPL_{GPT2\text{-}XL}$	73.67	71.71
Ours	**84.23**	**84.23**

As shown in Table 1, the proposed method outperforms the other approaches in Accuracy and F1-macro score, achieving a value of 84.23% in both metrics. The $BERT - B_{ft}$ model achieves an accuracy of 52.18% and an F1-macro score of 38.82%, indicating difficulties with class imbalance despite adopting an approach based on fine-tuning. $XLNET_{ft}$ records an Accuracy of 49.18% and an F1-macro score of 48.42%, showing better handling of class balance despite lower accuracy than $BERT - B_{ft}$. $PPL_{GPT2\text{-}XL}$ achieves 73.67% Accuracy and a 71.71% F1-macro score, significantly improving over the two fine-tuned models due to its large model size and the different approach.

4.5 Ablation Study

The validation of the proposed approach passed through additional experimentation in which verdicts are extracted by leveraging summaries without generating relations. In other words, the *Claim Verification* phase has also been carried out considering only summaries for a given claim. Therefore, each partial verdict represents the output provided by the LLM, considering a given claim and a summary. Thus, a set of partial verdicts is created, and the most frequent of them is the final outcome of the claim.

Table 2 shows that simplifying the process by eliminating the relation extraction step reduces performance considerably. Accuracy and F1-macro grow when the LLM receives a more structured context, as in the case of relation extraction.

Table 2. Evaluation metrics of the proposed approach compared with results given by considering only summaries, without extracting relations.

Approach	**Accuracy (%)**	**F1-macro (%)**
Without relation extraction	77.33	73.02
With relation extraction	**84.23**	**84.23**

5 Discussion & Limitations

The approach proposed in this study demonstrates robust and promising performance in Claim Verification. Its performance exceeds compared approaches by also guaranteeing the feasibility of the approaches in different domains and the reliability of results. However, it is important to highlight some limitations which could be explored in future research.

5.1 Evidence Retrieval Without Temporal Consideration

The *Evidence Retrieval* phase does not account for temporal information when searching for claim-related web pages. This oversight can pose issues in real-time scenarios. When an influential individual makes a claim, it can quickly become viral, leading to numerous web pages reporting it. Consequently, the proposed approach might erroneously consider it well-supported if used during the spreading of the claim, regardless of its veracity. This limitation underscores the necessity for incorporating temporal data to more accurately evaluate the validity and reliability of the information retrieved in future research.

5.2 Summaries Without Relevant Relations

During *Claim-focused Relation Extraction*, the Large Language Model (LLM) identifies relations from a summary that are relevant to assessing a claim's truthfulness. If no relevant relations are found, the LLM outputs *No Relevant relations* *for the given claim*. This outcome is recorded but not used in determining the most frequent verdict, as only *Supports* and *Refutes* are valid partial verdicts. However, if all summaries for a claim lack relevant relations, the claim is excluded from the final evaluation.

Additionally, for a specific claim, having an equal number of *Supports* and *Refutes* is possible, making it difficult to determine the correct label. Therefore, claims with this ambiguity are also excluded from the final evaluation. Finally, around 16.96% of the considered subset is excluded from the final evaluation due to those circumstances.

The issues described above are related, in the majority of cases, to the scraping process, where at times scraping is not possible, and the snippet is added. As a result, the snippet may be too short to contain relevant information to evaluate the veracity of the claim.

5.3 Closed LLM

Although the Large Language Model, Gemini 1.5 Pro, achieves state-of-the-art performance on many benchmarks, it is proprietary and requires specific API for access, which introduces several limitations. Firstly, the use of API needs increasing financial expenditures. Also, reliance on external API means any changes or discontinuities in this service can disrupt the research continuity. Additionally, the closed nature of the model restricts transparency and flexibility, preventing the research from inspecting or modifying the model as needed.

6 Conclusion

This manuscript presents an innovative approach for claim verification employing a RAG-based architecture that leverages and enhances the capabilities of Large Language Models (LLMs) for the claim verification task. First, evidence are retrieved from the open Web, then, following the proposed RAG phases, they are provided to the LLM to create summaries, extract claim-focused relations, and provide a final verdict for the claim. Experimentation on a shared dataset (i.e., FEVER) highlights its superiority with respect to the state-of-the-art in terms of Accuracy and F1-macro. Moreover, the approach guarantees the reliability of the results and a broader applicability by its nature.

In the future, as described in the previous section, it could be relevant to refine the approach to filter out evidence following the claim publication to simulate real-time situations. Moreover, to broaden the experimentation, the approach may consider claims labeled as *NOT ENOUGH INFO* and be compared with additional existing approaches.

Acknowledgments. This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alhindi, T., McManus, B., Muresan, S.: What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure. In: Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 380–391 (2021)
3. Barbaresi, A.: Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 122–131 (2021)
4. Chen, J., Kim, G., Sriram, A., Durrett, G., Choi, E.: Complex claim verification with evidence retrieved in the wild. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3569–3587 (2024)
5. Cheung, T.H., Lam, K.M.: Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 846–853. IEEE (2023)
6. Dammu, P.P.S., Naidu, H., Dewan, M., Kim, Y., Roosta, T., Chadha, A., Shah, C.: Claimver: Explainable claim-level verification and evidence attribution of text through knowledge graphs. arXiv preprint arXiv:2403.09724 (2024)
7. Das, A., Liu, H., Kovatchev, V., Lease, M.: The state of human-centered nlp technology for fact-checking. Information processing & management **60**(2), 103219 (2023)
8. Du, S., Gollapalli, S.D., Ng, S.K.: Nus-ids at checkthat! 2022: identifying checkworthiness of tweets using checkthat5. Working Notes of CLEF (2022)
9. Fenza, G., Gallo, M., Loia, V., Petrone, A., Stanzione, C.: Concept-drift detection index based on fuzzy formal concept analysis for fake news classifiers. Technological Forecasting and Social Change **194**, 122640 (2023)
10. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Transactions of the Association for Computational Linguistics **10**, 178–206 (2022)
11. He, Z., Zheng, Y., Ng, P.: Rethinking few-shot learning in language models. In: Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3452–3463 (2023)
12. Huang, K., Sun, Y., Li, Y., Zhou, X., Li, J., Sun, J.: A survey on knowledge-enhanced pre-trained language models. Neurocomputing **517**, 231–250 (2023)
13. Jiang, Z., Xu, F.F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., Neubig, G.: Active retrieval augmented generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7969–7992 (2023)
14. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
15. Lee, J., Yoon, J., Kang, J.: Generating knowledge by driving generative language models with discriminative rewards. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4674–4687 (2022)

16. Lee, N., Bang, Y., Madotto, A., Fung, P.: Towards few-shot fact-checking via perplexity. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1971–1981 (2021)
17. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
18. Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., Choi, Y., Hajishirzi, H.: Generated knowledge prompting for commonsense reasoning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3154–3169 (2022)
19. Liu, Z., Xiong, C., Sun, M., Liu, Z.: Fine-grained fact verification with kernel graph attention network. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7342–7351 (2020)
20. Naderi, N., Hirst, G.: Automated fact-checking of claims in argumentative parliamentary debates. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER). pp. 60–65 (2018)
21. Pankovska, E., Schulz, K., Rehm, G.: Suspicious sentence detection and claim verification in the covid-19 domain. In: Proceedings of the Workshop Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022), CEUR-WS, Stavanger (2022)
22. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
23. Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P.: That is a known lie: Detecting previously fact-checked claims. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3607–3618 (2020)
24. Soleimani, A., Monz, C., Worring, M.: Bert for evidence retrieval and claim verification. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42. pp. 359–366. Springer (2020)
25. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: NAACL-HLT (2018)
26. Vaghefi, S., Muccione, V., Huggel, C., Khashehchi, H., Leippold, M.: Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks. In: NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning (2022)
27. Wadhwa, S., Amir, S., Wallace, B.C.: Revisiting relation extraction in the era of large language models. In: Proceedings of the conference. Association for Computational Linguistics. Meeting. vol. 2023, p. 15566. NIH Public Access (2023)
28. Wang, G., Chillrud, L., McKeown, K.: Evidence based automatic fact-checking for climate change misinformation. In: International Workshop on Social Sensing on The International AAAI Conference on Web and Social Media (2021)
29. Zeng, W., Wang, M., Liu, K., Xiao, X., Sun, R., Han, X., Li, H., Li, H., Wu, H., Chen, E., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1224–1246 (2023)

30. Zhang, X., Gao, W.: Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 996–1011 (2023)
31. Zhu, B., Zhang, X., Gu, M., Deng, Y.: Knowledge enhanced fact checking and verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3132–3143 (2021)