FactOReS: Fact-checking with an Evidence-based Open Resource in Spanish

Nagore Bravo[1], **Jaione Bengoetxea**[1], **Iker García-Ferrero**[3], **Alba Bonet-Jover**[2], **Estela Saquete**[2], **Rodrigo Agerri**[1]
[1]HiTZ Center - Ixa, University of the Basque Country EHU
[2]Department of Software and Computing Systems, University of Alicante, Spain
[3]Multiverse Computing
{jaione.bengoetxea, rodrigo.agerri}@ehu.eus
{alba.bonet, stela}@dlsi.ua.es

Abstract

Automated Fact-Checking (AFC) has become a popular research area in Natural Language Processing (NLP), intending to support human verification through evidence-based veracity prediction systems that provide transparency at each stage of the process. Despite the global significance of misinformation and the substantial progress made in AFC research, multilingual approaches to evidence-based fact-checking remain inadequately addressed. This work introduces FactOReS, the first publicly available dataset evaluated for evidence-based veracity prediction in Spanish, constructed from real Spanish-language claims and verified fact-checking articles. We establish performance baselines by systematically applying In-Context Learning (ICL) with Large Language Models (LLMs) to both an established English dataset and our novel Spanish dataset. Despite good zero-shot and few-shot performance, results in both languages demonstrate that each step requires further research in order to improve the overall results in the evidence-based veracity prediction task. Finally, we propose a semi-automated methodology that integrates computational processing with human validation, offering a reproducible framework for developing multilingual evidence-based fact-checking resources for the benefit of the NLP research community. Data and code available: https://github.com/hitz-zentroa/AFC_FactOReS.

Keywords: Evidence Retrieval, Veracity Prediction, Multilingualism

1. Introduction

The rapid propagation of fake news through online media poses a growing societal threat, as false or misleading content spreads much faster than it can be manually verified by human fact-checkers (Lazer et al., 2018; Dulhanty et al., 2019). Despite the growing number of fact-checking agencies focusing on detecting misinformation, the exponential growth of digital information and user-generated content has surpassed the capacity of traditional verification efforts, leading to an urgent need for scalable, automated solutions (Panchendrarajan and Zubiaga, 2024).

In this context, Automated Fact-Checking (AFC) has therefore emerged as a promising research field, aiming to assist human experts in evaluating the veracity of claims by leveraging computational methods ranging from rule-based systems to Transformer-based architectures.

Nowadays, AFC is commonly modeled as a multi-step process involving claim check-worthiness detection, evidence retrieval, stance detection and veracity prediction, where retrieved evidence is integrated to produce a factual judgment (Baly et al., 2018; Guo et al., 2022). Benchmarks such as FEVER (Thorne et al., 2018a) and AVeriTeC (Schlichtkrull et al., 2023) have fostered rapid progress by providing large-scale annotated datasets for evidence-based verification. However, these resources are developed only for English, leaving a glaring gap in the research of evidence-based verification systems for other languages.

Despite the global nature of the problem, evidence-based veracity prediction in languages other than English remains an underexplored challenge. AFC presents substantial technical challenges (Augenstein, 2021), and existing systems exhibit limited performance due to linguistic variation and cultural context affecting both claim interpretation and evidence retrieval (Guo et al., 2022). Mitigating this disparity is critical for ensuring that fact-checking technologies support information integrity across diverse linguistic communities. However, this requires the prior development of evidence-based, annotated fact-checking datasets in non-English languages. This study investigates multilingual veracity prediction through the development and evaluation of AFC systems designed for Spanish. The research objectives are twofold: (i) to benchmark the In-Context Learning (ICL) performance of Large Language Models (LLMs) in multilingual fact-checking contexts, and (ii) to construct a high-quality, evidence-based Spanish dataset addressing current resource limitations.

More in detail, we hypothesize that scaling model size enhances robustness in veracity prediction, thereby reducing systematic errors such as the

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5349–5366
11–16 May 2026.

overproduction of *Not Enough Evidence* outputs. Furthermore, we expect that stance information can significantly improve veracity prediction when reliable, but may degrade performance when noisy. Finally, experimental results suggest that semi-automatic dataset construction through Retrieval-Augmented Generation (RAG) can produce a high-quality Spanish benchmark, mitigating the scarcity of non-English fact-checking resources.

To test these hypotheses, we adopt an In-Context Learning (ICL) paradigm, which enables LLMs to adapt to fact-checking tasks with minimal supervision by providing examples within the prompt. Our methodology combines experiments on existing English datasets together with the development of a new Spanish dataset, FactOReS, derived from real social media claims and verified fact-checking articles. The evidence-based dataset is constructed through a semi-automatic pipeline that integrates RAG techniques and LLM-based filtering, followed by human validation to ensure factual reliability. This dual approach enables cross-lingual benchmarking between English and Spanish while fostering research into how multilingual LLMs generalize across languages with different data availability and cultural contexts.

The contributions of this work are threefold. First, we introduce FactOReS, which, to our knowledge and at the time of our study, is the first publicly available evidence-based dataset evaluated for claim verification in Spanish. Second, we conduct an extensive evaluation of AFC for English and Spanish using LLMs under ICL prompting strategies. Third, we propose a semi-automatic methodology for Spanish dataset generation that combines automation with human oversight, establishing a reproducible framework for building multilingual fact-checking resources. Our study advances the development of AFC systems capable of operating beyond English, contributing to a more inclusive information ecosystem.

As misinformation evolves, AFC systems must adapt to accommodate multilingual contexts. Developing evidence-based verification systems across languages is essential for ensuring global access to reliable information. This work hopes to establish that scalable, language-aware AFC approaches can enhance the detection of misleading content across linguistic boundaries.

2. Related Work

AFC aims to assess the truthfulness of textual claims by retrieving and analyzing supporting evidence from trusted sources (Augenstein, 2021). A typical pipeline for claim verification in AFC involves four steps: (i) identifying claims worth verifying, (ii) generating questions to create search queries, (iii) retrieving relevant documents, (iv) determining the stance and veracity of each claim with respect to the retrieved evidence. Recent systems also emphasize explanation generation to improve transparency (Schlichtkrull et al., 2023).

2.1. Datasets

The AFC publicly available datasets that are closer to our interests are FEVER (Thorne et al., 2018a) and AVeriTeC (Schlichtkrull et al., 2023). FEVER introduced large-scale synthetic claims derived from Wikipedia, establishing foundational benchmarks for AFC but lacking linguistic and contextual realism. In contrast, AVeriTeC advanced the field with real-world claims, temporally consistent web evidence and explicit justifications. Both are available in English only.

Subsequent efforts such as MultiFC (Augenstein et al., 2019), LIAR (Wang, 2017), SciFact (Wadden et al., 2020) and PubHealth (Bayani et al., 2025) expanded AFC into new domains, including political discourse, scientific literature and health communication. However, few of these corpora include explicit evidence linking claims and documents, and most remain monolingual. In many cases, contextual information is added without ensuring it is sufficient to verify the claim.

To address linguistic diversity, XFact (Gupta and Srikumar, 2021) introduced multilingual verification across 25 languages using real-world claims, though coverage and evidence quality vary. Similarly, localized benchmarks such as CHEF (Chinese) (Hu et al., 2022), ViFactCheck (Vietnamese) (Hoa et al., 2025) and HealthFC (health misinformation) (Vladika et al., 2024).

Recent efforts such as EuroVerdict (Russo et al., 2025), MultiClaimNet (Panchendrarajan et al., 2025) and MultiSynFact (Chung et al., 2025) provide datasets with various sources of information complementary to Spanish claims. However, none of these previous publications provide explicit evidence-based evaluation for veracity prediction in Spanish. EuroVerdict's evidence consists of providing a URL related to the claim, but not the explicit piece of evidence required to perform evidence-based veracity prediction. Furthermore, EuroVerdict is focused on verdict generation and does not provide any evaluation related to evidence-based veracity in Spanish. Regarding MutiClaimNet and MultiSynFact, these resources do not provide any retrieved evidence. Furthermore, MultiSynFact's claims are synthetically generated.

Other Spanish resources, such as the Spanish Fake News Corpus (Gómez-Adorno et al., 2021), are primarily designed for fake news detection rather than claim verification.

Therefore, FactOReS stands out as the first specialized, manually annotated resource for Span-

ish that provides explicit, real, and high-quality claim–evidence pairs evaluated in veracity prediction, bridging the gap between general multilingual corpora and the need for dedicated, evidence-based tools for the Spanish language.

2.2. Systems

Veracity prediction in AFC is closely related to Natural Language Inference (NLI) (Radford et al., 2018), and most systems fine-tune pretrained encoders such as BERT or RoBERTa for claim-evidence classification (Gong et al., 2024). Other approaches employ hybrid or graph-based architectures, such as HybridFC (Qudus et al., 2022), to integrate multiple evidence sources.

The FEVER Shared Task 2018 (Thorne et al., 2018b) established the first large-scale benchmark for AFC, introducing a pipeline approach with document retrieval, sentence selection and NLI. Top systems such as UNC-NLP and UCL-MRG demonstrated the effectiveness of combining neural retrieval with entailment models based on contextual embeddings (e.g., ELMo, LSTM), setting the foundation for evidence-based veracity prediction.

More recently, the AVeriTeC Shared Task 2024 (Schlichtkrull et al., 2024) extended this paradigm to real-world claims verified against web evidence. Leading systems like TUDA_MAI (Rothermel et al., 2024), HUMANE (Yoon et al., 2024) and CTU AIC (Ullrich et al., 2024) leveraged LLMs such as GPT-4o and LLaMA for unified pipelines covering question generation, evidence retrieval and verdict prediction. These approaches highlight a growing trend towards RAG and ICL, where LLMs reason over retrieved passages to produce both answers and explanations (Yoon et al., 2024; Rothermel et al., 2024).

Building on these advances, our work contributes to the development of AFC in Spanish by exploring ICL as a flexible paradigm for veracity prediction for multiple languages. We leverage AVeriTeC for benchmarking and introduce a new dataset, FactOReS, designed to capture real-world Spanish claims and fact-checks. Ultimately, our approach aims to advance scalable, evidence-based and explainable fact-checking through the integration of generation, retrieval and inference.

3. Creation of FactOReS Dataset

Motivated by the lack of Spanish resources for evidence-based veracity prediction, we present FactOReS[1], a publicly available, evidence-based dataset built with source data from *Maldita.es*[2]. Below we summarize the initial data structure (Section 3.1), preprocessing to obtain check-worthy claims (Section 3.2), a retrieval pipeline based on RAG (Section 3.3) and the annotation protocol and agreement (Section 3.4).

3.1. Initial Structure of the Data

We begin from a raw collection of 15,305 items, each including: *claim_id*, *title*, *date* (Unix), *status* (*Bulo*/*Qué sabemos*/*No hay pruebas*), *transcription* (original post/statement), *topic*, *article_id*, *url* and article *content*. Some entries include *media* (image/video URL).

3.2. Preprocessing

We retain only textual, verifiable statements. Specifically:

- Elements that require multimedia evidence for verification are excluded.
- Only claims containing concrete, verifiable information, such as specific events, dates, names of people, organizations or places, are retained.

Those claims identified as check-worthy are then standardized; the claims written as questions or negatively phrased are rephrased into affirmative sentences while preserving their factual meaning. After applying these criteria, a total of 8,831 entries are marked as check-worthy. To ensure quality and remove redundancy, duplicate claims are eliminated, resulting in a final set of 1,101 unique and check-worthy claims.

3.3. Information Retrieval

The process follows with the information retrieval step, shown in Figure 1. We adopt a RAG pipeline in order to gather evidence before annotation, which is split into two phases.

Question Generation. For each claim, GPT-4o[3] produces a brief verification plan, ≈6 targeted *questions* and 6 web *searches*. We validate I/O with *Pydantic* schemas to keep a consistent JSON structure.

Evidence Retrieval. Queries are executed via *Serper*[4], retrieving up to five sites per query. Retrieved pages are split by newlines into chunks (≥128 chars), embedded (OpenAI embeddings) and ranked by semantic similarity score against question embeddings.

[1] https://github.com/hitz-zentroa/AFC_FactOReS

[2] https://maldita.es

[3] https://openai.com/api/

[4] https://serper.dev/

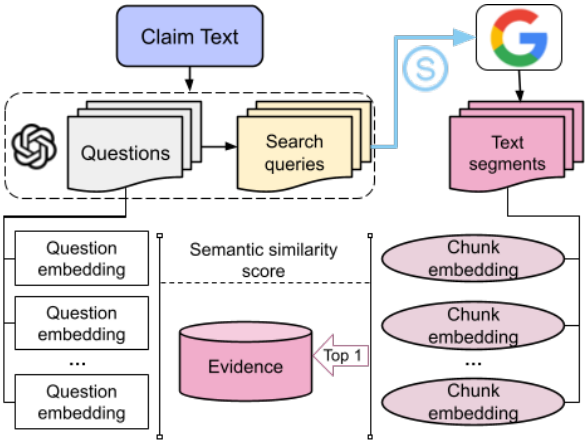


Figure 1: Information retrieval stage: question generation and evidence retrieval.

3.4. Annotation and Labeling

For each claim-question pair, we select the top-scoring evidence snippet and automatically summarize it with GPT-4o, enforcing schema consistency via *Pydantic*. We then annotate a gold subset of 81 claims (571 question-evidence pairs) following two complementary evaluation schemes: one for the claim-evidence relationship and another for the claim-question relationship.

Claim-Evidence Evaluation. Each claim-evidence pair is scored according to the following dimensions:

- **Relevance (topicality)**: binary indicator that captures if the evidence is related to the claim. If relevance is assigned with 0, all remaining annotation fields are also recorded as 0, *stance* is marked as *Neutral* and *veracity* as *Not Enough Evidence*.
- **Objectivity**: binary indicator which is assigned with 1 if it is factual and 0 if opinionated.
- **Stance**: *Positive*, *Negative* or *Neutral*, indicating the position of the evidence relative to the claim.
- **Veracity**: *Supported*, *Refuted* or *Not Enough Evidence (NEI)*, reflecting the truthfulness of the claim given the evidence.

Claim-Question Evaluation. Independently, each question is assessed for its coverage of key informational aspects. **Critical dimensions** are represented through binary indicators that capture whether the question addresses the essential components of the claim: *what*, *who*, *where*, *when* and *how*.

Figure 2 illustrates this stage. Each case presents a *Claim Text* and a corresponding *Question*. The semantic categories are color-coded as follows: **what**, **who**, **where**, **when** and **how**, whenever these elements appear in the text.

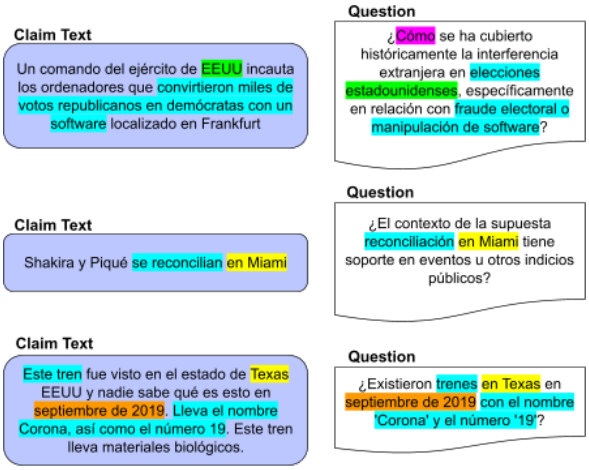


Figure 2: Examples of the claim-question evaluation stage.

Assigning a single veracity label can be challenging when evidence only partially covers a claim. To address this, if most key points are addressed, we avoid labeling it as *Not Enough Evidence* and assign *Supported*/*Refuted*. As an illustrative example, consider Figure 3. Here, while the claim asserts that silver iodide is used to prevent rain, the evidence shows that it is actually used to induce it (highlighted in green). Although it does not mention the use of airplanes to disperse silver iodide (shown in red), which is the unverified component of the claim, the *stance* label assigned is *Negative* and the *veracity* label is *Refuted*.

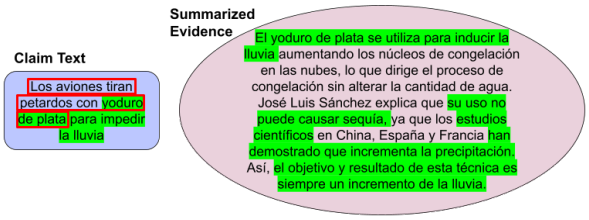


Figure 3: Example of a claim where the evidence refutes only part of the information.

Inter-annotator agreement (Cohen's Kappa) was high overall ($\kappa = 0.7$ on average), with firm consistency for *stance* and *veracity* ($\kappa = 0.88$). The label distribution across the 571 annotated pairs shows an imbalance, with 63% *Not Enough Evidence*, 23% *Refuted* and 14% *Supported* cases.

4. Methodology

This section describes the materials and methods used, including LLMs, datasets, prompting techniques and evaluation methods.

4.1. Large Language Models

We experiment with both closed and open-weight instruction-tuned LLMs: **GPT-4o** (Hurst et al., 2024), **LLaMA 3 Instruct** (8B and 70B) (Grattafiori et al., 2024) and **Qwen 2.5 Instruct** (7B and 72B) (Team, 2024).

All are decoder-only Transformer models optimized for natural language understanding and factual reasoning. We use them without fine-tuning, focusing instead on prompt design and evidence conditioning to perform veracity classification.

4.2. Datasets

Experiments on evidence-based veracity prediction are conducted on the AVeriTeC and our newly released FactOReS datasets (Schlichtkrull et al., 2023). Both contain real-world claims, web-retrieved evidence and question-answer (QA) justifications with veracity labels for *Supported*, *Refuted*, *Not Enough Evidence (NEI)*.

To ensure comparability with previous AVeriTeC results, we rely on the official AVeriTeC baseline outputs for the development split. Each claim is paired with three automatically generated QA pairs used as evidence for veracity prediction. This setup isolates the classification stage, avoiding variability from upstream retrieval or generation components.

4.3. Prompting Techniques

We evaluate several ICL approaches. In the **Zero-Shot** setting, the model receives only the task description and must produce an answer based on its pretrained knowledge, without any examples. In the **Few-Shot** setting, the prompt includes a small number of input-output examples that illustrate how the task should be performed. The **Chain-of-Thought (CoT)** approach adds explicit reasoning steps within the prompt, so the model generates intermediate reasoning before the final answer. Finally, the **Few-Shot CoT** setting provides example demonstrations that include such reasoning traces, allowing the model to follow similar structured reasoning when generating its output.

4.4. Evaluation Methods

Model performance is assessed with three complementary metrics. **AVeriTeC score** (Schlichtkrull et al., 2023) captures the ability of the model to generate appropriate questions, retrieve relevant evidence and correctly assess the veracity of factual claims.

Since exact evidence matching is unreliable due to natural variation in phrasing across different web sources, AVeriTeC adopts an approximate sequence matching approach. This relies on a similarity function $f : S \times S \to \mathbb{R}$, where S is the space of token sequences. The similarity is computed using the Hungarian Algorithm (Kuhn, 1955), with METEOR (Banerjee and Lavie, 2005) as the scoring function (Equation 1). This yields the following function:

$$u_f(\hat{Y}, Y) = \frac{1}{|Y|} \max \sum_{\hat{y} \in \hat{Y}} \sum_{y \in Y} f(\hat{y}, y) X(\hat{y}, y) \quad (1)$$

where $\hat{Y}$ denotes the set of sequences generated by the model, Y represents the corresponding reference set and $X(\hat{y}, y) \in \{0, 1\}$ is a binary assignment matrix that selects the optimal alignment of elements to maximize total similarity.

This metric $u_f(\hat{Y}, Y)$ is evaluated only if the retrieved QA pairs surpass a threshold λ, with a typical value of $\lambda = 0.25$. Specifically, if $u_f(\hat{Y}, Y) \geq \lambda$ the prediction is evaluated based on both the veracity label and the quality of the QA evidence, and 0 otherwise.

In addition to the above strategy, we also report the **macro F1 score** and **per-label F1 scores** to handle class imbalance and provide detailed insight for each label.

5. Experimental Setup

This section describes the experimental design used to evaluate veracity prediction across datasets, models and prompting strategies. It focuses on assessing the capacity of state-of-the-art LLMs to perform fact-checking tasks in multiple languages within an ICL framework, without requiring additional training. This design enables the evaluation of both the predictive performance of the models and their ability to generalize to unseen claims across different linguistic contexts.

Experiments are conducted in two phases. We first use the English-only AVeriTeC dataset to evaluate LLM performance on the veracity prediction task. We then extend the evaluation to a multilingual setting by including the Spanish evidence-based dataset FactOReS, where we assess both stance prediction and veracity prediction. Both phases compare the performance of several LLMs under the same ICL configurations, described in Section 4, and examine the effects of prompt structure and the inclusion of intermediate reasoning steps (stance prediction) on overall AFC performance. In addition, the complete set of prompts is available in the Appendix.

5.1. Experiments on AVeriTeC

In the first phase, each LLM performs stance detection and veracity prediction using the AVeriTeC development split. We test the selected LLMs by comparing performance across model scales and prompting styles.

These experiments aim to assess performance on the final task of **veracity prediction**. Therefore, while stance labels are generated as part of the process, they are not evaluated, as the primary focus of AVeriTeC is on veracity classification. However, the predicted stance labels are used as intermediate indicators, which allow us to explore whether incorporating this information improves, worsens or has no effect on the final predictions. To that end, we define two experimental conditions:

- **Direct:** the model predicts veracity using only the claim, question and evidence.
- **Stance-informed:** the model additionally receives the predicted stance labels as context.

All prompts are written in English and use a structured JSON output parsing process using the *Pydantic* validation library in Python, ensuring a consistent format and facilitating integration with the rest of the fact-checking pipeline. In addition to the predicted label, models are required to provide a *reasoning* field that explains the reason behind their decision. This reasoning, while not quantitatively evaluated in the present work, (i) offers transparency for the decision-making process and (ii) enables future work on the assessment of explanation quality and correctness. Predictions are scored using the metrics presented in Section 4.4.

5.2. Experiments on FactOReS

After completing all experiments on AVeriTeC, we apply the same configurations to the FactOReS dataset. In this phase, we evaluate both **stance detection** and **veracity prediction**, following the same processing pipeline described for AVeriTeC. Prompts remain in English, as prior work shows LLMs often achieve higher accuracy when prompted in English, even on non-English content (Dey et al., 2024; Vadlapati, 2023).

As in the AVeriTeC setup, we leverage a structured output parsing pipeline based on the *Pydantic* validation library. Consistent with the AVeriTeC configuration, models are also required to provide a *reasoning* field alongside the predicted label, which are not evaluated in this work, but are included to enhance explainability and enable future research on the generation of explanations task. In addition, since FactOReS excludes the *Conflicting/Cherrypicking* label, the veracity space is reduced to *Supported*, *Refuted* and *Not Enough Evidence*. Regarding prediction evaluation, we also report macro F1, per-label F1 and AVeriTeC score presented in Section 4.4.

6. Results

This section presents the results of the experiments conducted to evaluate model performance in the previously defined settings, divided into two parts. In the first part, we evaluate the models using the AVeriTeC English-language dataset 6.1. In the second part, we assess model performance on our own Spanish-language dataset 6.2, FactOReS, which introduces additional linguistic and domain-specific challenges. Additionally, the Appendix contains the complete result tables for all experimental configurations.

6.1. Results on AVeriTeC

This evaluation helps to identify the strengths and weaknesses of each approach before introducing the additional linguistic and domain-specific challenges posed by the FactOReS dataset.

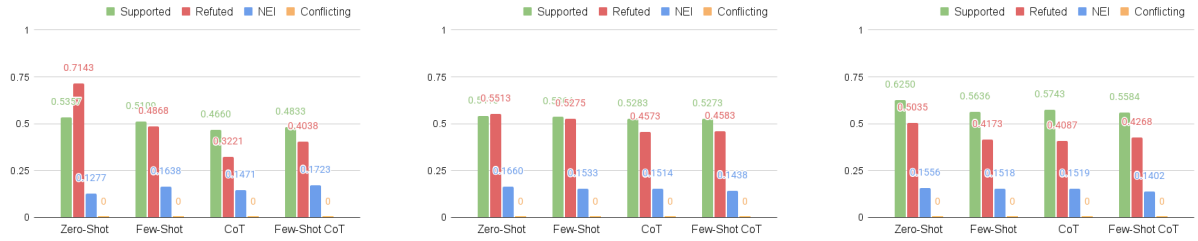
The evaluation task is situated within a fact-checking pipeline involving real-world claims, where the questions and answers have been automatically generated by the baseline system AVeriTeC. During its evidence retrieval phase, this system achieved a HU-METEOR score of 0.240 for question-only inputs (*Q only* metric) and 0.185 for the QA pair (*Q + A* metric). It indicates a noisy informational basis that is only partially aligned with the facts to be verified, so we assess robustness rather than absolute accuracy.

To contextualize the performance of our models, it is useful to examine how other systems have performed on the same development split in the AVeriTeC Shared Task 2024, as they serve as a benchmark for assessing the relative performance of participating systems under comparable conditions. Table 1 presents the results on the development split for the top four systems (based on test split performance) and the baseline, sorted by the values of AVeriTeC score metric at a threshold of 0.25 (@.25).

Rank	Team Name	AVeriTeC Score @.25
1	TUDA_MAI	**0.60**
2	HUMANE	0.58
3	Dunamu-ml	0.48
6	CTU AIC	0.42
17	Baseline	0.09

Table 1: Development phase leaderboard of the AVeriTeC Shared Task 2024, sorted by AVeriTeC score.

Overall, the results reveal substantial differences



(a) LLaMA 3 70B Instruct (b) Qwen 2.5 72B Instruct (c) GPT-4o

Figure 4: Per-label F1 scores on veracity prediction (no-stance) for AVeriTeC dataset.

in system performance, with top-ranked teams significantly outperforming the baseline.

Regarding the structure of the analysis, we start by examining smaller models, followed by the study of larger models and closed models. Our results show that veracity prediction under noisy evidence is primarily driven by model scale and prompt simplicity. First, small-scale instruct-tuned models such as LLaMA 3 8B and Qwen 2.5 7B exhibit strong biases toward the *Not Enough Evidence* class, which substantially limits their veracity prediction performance. While stance integration provides slight improvements for specific labels (notably *Refuted*), the overall benefits remain limited and inconsistent.

In contrast, large-scale models demonstrate better robustness. Both LLaMA 3 70B (see Figure 4a) and Qwen 2.5 72B (see Figure 4b) consistently surpass the AVeriTeC baseline (0.3211 macro F1, 0.092 @.25), with simple Zero-Shot prompting emerging as the most reliable configuration (LLaMA 3 70B: 0.3444 macro F1 and 0.115 AVeriTeC score). When aggregating results without including stance, Qwen 2.5 72B achieves the highest mean performance (0.3193 ± 0.0160 macro F1), confirming its robustness compared to LLaMA 3 70B (0.2832 ± 0.0470). GPT-4o reaches a very similar average performance (0.2871 ± 0.0166 macro F1) to Qwen 2.5 72B, while maintaining high consistency across prompting styles (Figure 4c).

Taken together, these results indicate that scaling enhances robustness and alignment between balanced classification (macro F1) and threshold-based evaluation (AVeriTeC score), while stance integration yields at best marginal and uneven improvements.

6.2. Results on FactOReS

The results of the experiments aim to determine whether the models generalize effectively beyond English and how well they perform when confronted with domain-specific terminology. The analysis is structured by task: we examine model behavior first on stance detection (Section 6.2.1) and then on veracity prediction (Section 6.2.2). As in AVeriTeC, we follow the same structure for each task.

6.2.1. Stance Detection

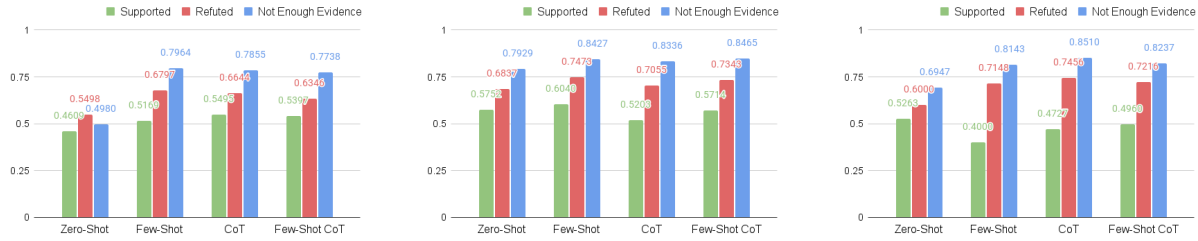
On the stance detection task with the FactOReS dataset, small-scale models reveal a clear hierarchy: Qwen 2.5 7B (0.6094 ± 0.0336) consistently outperforms LLaMA 3 8B (0.4302 ± 0.0954) across prompting strategies, showing stronger robustness and better balance across stance labels. Among small models, CoT prompting enhances the detection of *Neutral* and *Negative* classes, while *Positive* remains the most challenging.

Scaling up further mitigates inter-class disparities: LLaMA 3 70B improves over its smaller counterpart (0.5714 ± 0.0491), and Qwen 2.5 72B achieves the highest open-weight stance score (0.6239 ± 0.0345), exhibiting stable behavior across prompting setups. Finally, GPT-4o slightly outperforms the Qwen models (0.6560 ± 0.0085), which may suggest that closed-source instruction tuning yields robustness across stance categories without relying on prompt complexity.

6.2.2. Veracity Prediction

The veracity prediction task in the FactOReS dataset reveals further divergences between model families and scales. Small-scale models again expose weaknesses: LLaMA 3 8B remains limited by poor stance detection and excessive confusion between *Supported* and *Not Enough Evidence*, while Qwen 2.5 7B achieves more balanced results and benefits substantially from stance integration in Few-Shot CoT settings.

With scaling, both families improve: LLaMA 3 70B achieves its best results under CoT prompting (see Figure 5a), but is more sensitive to stance inclusion, which occasionally reintroduces noise. In contrast, Qwen 2.5 72B emerges as the strongest open-weight model, consistently surpassing all others across configurations (Figure 5b) and achieving the highest overall macro F1 (0.7114). Its robustness to stance inclusion and prompting confirms that it integrates stance reasoning more effectively than LLaMA. GPT-4o shows relatively stable be-



(a) LLaMA 3 70B Instruct (b) Qwen 2.5 72B Instruct (c) GPT-4o

Figure 5: Per-label F1 scores on veracity prediction (no-stance) for FactOReS dataset.

havior (Figure 5c), with competitive performance across all prompting strategies and limited dependence on stance integration. While this consistency is noteworthy, it is important to consider the potential influence of using the same model family during the semi-automatic construction of FactOReS (Section 3). Although the final dataset was strictly validated by human annotators to ensure factual reliability, the alignment between the model used for evidence processing and the evaluator could contribute to the observed stability. Future work will involve testing these benchmarks with a broader range of independent annotator models to further decouple the resource from specific architectures.

In general, stance inclusion proves beneficial primarily for small models and in Qwen architectures, but offers marginal or even detrimental effects for large LLaMA and closed models. Thus, while scale and architecture drive the main performance differences, stance integration only plays a secondary, model-dependent role.

7. Conclusions

This work advances evidence-based AFC beyond English by introducing FactOReS, the first publicly available evidence-based Spanish dataset evaluated for veracity prediction built from real claims. We present a semi-automated pipeline that combines RAG-driven retrieval and human annotation, and we benchmark state-of-the-art LLMs under ICL on both AVeriTeC (English) and FactOReS (Spanish).

Our findings lead to three main conclusions. First, scale matters: larger models exhibit markedly better robustness to noisy or partially aligned evidence and reduce systematic biases, such as overproducing *NEI*. Among open-weight models, Qwen 2.5 72B is the strongest on average, while GPT-4o is consistently stable across prompting styles. Second, experimental results show that simpler prompts are often more reliable than more elaborate ICL variants in the presence of noisy evidence, and the alignment between macro F1 and AVeriTeC score (@0.25) improves with model scale. Third, experiments with stance information confirmed its conditional contribution: high-quality stance signals improved predictions, whereas noisy stance detection degraded them. Regarding the methodology, while the use of GPT-4o enabled a scalable semi-automatic pipeline for question generation and evidence summarization, we acknowledge that this may introduce a specific model bias in the veracity prediction results. However, the high inter-annotator agreement achieved during the human validation phase ($\kappa = 0.88$ for veracity) suggests that FactOReS remains a reliable, human-centered benchmark. Future evaluations should include a broader range of open-weight models to further assess the independence of the resource from the underlying construction tools. In conclusion, this work provides resources and insights that address a key gap in multilingual AFC, laying the foundation for future systems that are multilingual and transparent.

While the findings of our work represent meaningful progress, they also open several avenues for future research. First, results highlight the conditional utility of stance information. Accordingly, future work should focus on developing more accurate stance detection models, particularly in multilingual contexts, and on investigating methods to integrate stance into veracity prediction in ways that reduce noise rather than amplify it. In addition, another promising direction lies in the generation of explanations. Beyond providing only veracity labels, AFC systems should be able to offer human-understandable justifications that connect claims with supporting or refuting evidence. Advancing in this direction would not only improve system transparency but also foster greater trust in the deployment of AFC tools in real-world settings.

8. Limitations

Our work presents several limitations that should be considered when interpreting the results. First, FactOReS is derived exclusively from Maldita.es, which may introduce source-specific topical and editorial biases and does not fully cover the broader

Spanish- and Latin American–language misinformation landscape. The preprocessing pipeline filters out non-textual and non-check-worthy content. Furthermore, our semi-automatic construction relies on a single proprietary LLM for question generation, evidence summarization and intermediate filtering, so model-specific biases may be reflected in the style and distribution of evidence; this is particularly relevant when related model families are later evaluated on the same benchmark. Despite these limitations, we believe that the resource and analysis provided constitute a substantive step towards a robust, evidence-based, and multilingual veracity-prediction method in Spanish, which can be applied to mitigate the lack of resources in other languages.

9. Ethics Statement

From an ethical perspective, FactOReS is built from publicly available fact-checking articles, but it still contains real-world claims and evidence, including potentially sensitive topics such as health, migration or political discourse. We do not attempt to deanonymize entities beyond what is already present in the publicly available source material, yet releasing structured claim-evidence pairs and model outputs can facilitate large-scale analysis of narratives about specific actors, with possible implications for privacy and stigmatization. Moreover, while our main goal is to support human fact-checkers and foster transparent, evidence-based AFC, the same methods could be repurposed for large-scale monitoring or profiling of online content. We therefore stress that the dataset and accompanying code are intended for research purposes, that AFC systems built on top of them should be deployed only with appropriate human oversight, and that future work should continue to monitor biases, disparate error rates and potential downstream harms, in line with LREC's ethical guidelines.

10. Acknowledgements

This research has been partially funded by the following MCIN/AEI/10.13039/501100011033 projects: (i) COOLANG.CONSENSO/TRIVIAL (PID2021-122263OBC21/PID2021-122263OB-C22) and by ERDF, A way of making Europe; (ii) HEART-NLP (PID2024-156263OB-C22) and by ERDF/EU; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR and (iv) DeepThought (PID2024-159202OB-C21) funded by ERDF, EU.

11. Bibliographical References

Isabelle Augenstein. 2021. Towards explainable fact checking. *arXiv preprint arXiv:2108.10274*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Azadeh Bayani, Alexandre Ayotte, Jean Noel Nikiema, et al. 2025. Transformer-based tool for automated fact-checking of online health information: Development study. *JMIR Infodemiology*, 5(1):e56831.

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. *arXiv preprint arXiv:2502.15419*.

Krishno Dey, Prerona Tarannum, Md Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *arXiv preprint arXiv:2410.13153*.

Chris Dulhanty, Jason L Deglint, Ibrahim Ben Daya, and Alexander Wong. 2019. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *arXiv preprint arXiv:1911.11951*.

Helena Gómez-Adorno, Juan Pablo Posadas-Durán, Gemma Bel Enguix, and Claudia Porto Capetillo. 2021. Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task. *Procesamiento del lenguaje natural*, 67:223–231.

Haisong Gong, Weizhi Xu, Shu Wu, Qiang Liu, and Liang Wang. 2024. Heterogeneous graph reasoning for fact checking over texts and tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 100–108.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. Vifactcheck: A new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 308–316.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Rrubaa Panchendrarajan, Rubén Míguez, and Arkaitz Zubiaga. 2025. Multiclaimnet: a massively multilingual dataset of fact-checked claim clusters. *arXiv preprint arXiv:2503.22280*.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Umair Qudus, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2022. Hybridfc: A hybrid fact-checking approach for knowledge graphs. In *International Semantic Web Conference*, pages 462–480. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. Infact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112.

Daniel Russo, Fariba Sadeghi, Stefano Menini, and Marco Guerini. 2025. Euroverdict: A multilingual dataset for verdict generation against misinformation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16617–16634.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. Aic ctu system at averitec: Re-framing automated fact-checking as a simple rag task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 137–150.

Praneeth Vadlapati. 2023. Multilingual prompting in llms: Investigating the accuracy and performance. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 7(02):1–7.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.

A. Appendix

This appendix provides complementary material that supports the main text. It includes the full prompts used across the system (Section A.1) and the complete result tables (Section A.2).

A.1. Prompts

This section of the appendix presents the different prompts used throughout the system for the following tasks: question generation (Section A.1.1), evidence summarization (Section A.1.2) and veracity prediction (Section A.1.3).

A.1.1. Question Generation

The prompt in Figure 6 is designed to guide the model in generating critical questions and relevant Google search queries based on a given statement.

Input

You are tasked with generating a counter-narrative for a given statement. The statement can also be formulated as a question. This involves creating a plan, generating critical questions and formulating Google searches to gather information that may challenge or verify the statement.
Here is the statement to analyze: `FACT_CHECKING_TOPIC`.
First, create a plan to approach this task. Your plan should outline the steps you'll take to gather information and analyze the statement. You should gather information and compare relevant data.
Next, generate a set of critical questions that would help gather information about the topic. These questions should be critical and aimed at uncovering various aspects of the statement. You can generate a maximum of 6 questions, so be concise and specific. Questions should start with "¿" and end with "?".
Finally, generate a set of Google searches that would help retrieve the necessary information to evaluate the statement as well as the questions you generated. You can generate a maximum of 6 searches, so be concise and specific.
Present your response in the following JSON format. The JSON should include three main keys: "plan", "questions" and "searches". "plan" should be a string in Markdown formatting. "questions" and "searches" should be arrays containing the items you generated in each step. The current year is `CURRENT_YEAR`.
Your answer should be in `'Spanish'if lang == 'es'else 'English'`.

Figure 6: Prompt used to generate questions.

A.1.2. Evidence Summarization

The prompt in Figure 7 is used to create a short and clear summary of a given piece of evidence. The goal is to distill the main idea into a concise, three-sentence maximum summary.

Input

You are tasked with generating a short and clear summary of the following evidence. Here is the evidence: `TOP_EVIDENCE`. Write a short summary in no more than three sentences. Respond ONLY with a JSON object in this format:

```
{
"summarized_evidence":
"your summary here"
}
```

Your answer must be in `'Spanish'if lang == 'es'else 'English'`.

Figure 7: Prompt used to summarize evidence.

A.1.3. Veracity Prediction

This section describes the different prompting configurations used for the veracity prediction task. The objective of the task is to determine the truthfulness of a claim based on all provided evidence pieces, returning a single label and a justification. Configurations where stance prediction is included or not included are illustrated in the prompts, with gray-colored text indicating that stance inputs are optional.

Figure 8 presents the Zero-Shot configuration, where the model is given a claim and several question-evidence pairs. It must analyze all the provided evidence and return an overall veracity prediction without any prior examples.

Figure 9 shows the Few-Shot configuration, which extends the Zero-Shot setup by including illustrative examples that demonstrate how to handle claims, questions and evidence.

Figure 10 presents the CoT configuration, where the model is guided through a logical, step-by-step reasoning process.

Finally, Figure 11 shows the Few-Shot CoT configuration, which combines example-based guidance with explicit reasoning steps.

Input

Given a claim, you are tasked with establishing whether the evidence given *Supported*, *Refuted*, *Not Enough Evidence* or *Conflicting/Cherrypicking* the claim.
For every claim, there is one `evidence` attribute which includes three `answer` elements and their respective `stance` elements. Write one overall prediction for the claim in the `pred_label` attribute in the JSON and provide the `reasoning`.

Analyze the following:

Claim: `CLAIM`
Question 1: `QUESTION_1`
Evidence 1: `EVIDENCE_1`
Stance 1: `STANCE_PREDICTION_1`

Question 2: `QUESTION_2`
Evidence 2: `EVIDENCE_2`
Stance 2: `STANCE_PREDICTION_2`

Question 3: `QUESTION_3`
Evidence 3: `EVIDENCE_3`
Stance 3: `STANCE_PREDICTION_3`

The JSON should include two main keys: `"pred_label"` and `"reasoning"`.
Your answer must be in `'Spanish' if lang == 'es' else 'English'`.

Figure 8: Zero-Shot prompt used to perform veracity prediction. Text that appears in gray indicates optional stance prediction input, based on configuration.

A.2. Tables

This section of the appendix presents tables that complement the results discussed in the main text.

Table 2 presents the results obtained on the AVeriTeC dataset for the veracity prediction task. The table reports the Macro F1 and AVeriTeC score (column *@.25*) across the different model sizes, prompting techniques and stance information. The AVeriTeC baseline performance is also included for comparison and configurations that surpass the baseline in both metrics are shown in blue.

Table 3 presents the results on the FactOReS dataset for both the stance detection and veracity prediction tasks. The table reports the Macro F1 scores for stance (column *S. Macro F1*) and veracity (column *V. Macro F1*) under the same experimental setup as AVeriTeC. Here, the best result for each model is shown in bold, and the overall best-performing configurations across all models are highlighted in blue.

Input

Given a claim, you are tasked with establishing whether the evidence given *Supported*, *Refuted*, *Not Enough Evidence* or *Conflicting/Cherrypicking* the claim.
For every claim, there is one `evidence` attribute which includes three `answer` elements and their respective `stance` elements. Write one overall prediction for the claim in the `pred_label` attribute in the JSON and provide the `reasoning`.

You will be given some examples first.

Examples:
Claim 1: Wearing face masks will stop the spread of COVID-19.
Question 1: Does a face mask prevent the spread of COVID-19?
Evidence 1: Cloth face coverings, even homemade ones, are effective in reducing the spread of COVID-19, according to Oxford's Leverhulme Centre.
Stance 1: `Positive`
Veracity Prediction 1: `Supported`
Reasoning 1: The evidence confirms the claim through a credible study.

Claim 2: Trump Administration claimed Billie Eilish is destroying the country.
Question 2: Has the Trump administration made that claim?
Evidence 2: A Washington Post article wrongly stated this; no official documents confirm the claim.
Stance 2: `Negative`
Veracity Prediction 2: `Refuted`
Reasoning 2: The claim is explicitly debunked by the evidence.

Claim 3: It makes no sense for oil to be cheaper in Nigeria than in Saudi Arabia.
Question 3: Why do fuel prices differ by country?
Evidence 3: Fuel prices vary due to taxes, refining costs and other components.
Stance 3: `Neutral`
Veracity Prediction 3: `Not Enough Evidence`
Reasoning 3: The evidence provides general background but does not evaluate the specific comparison in the claim.

Analyze the following:
Claim: `CLAIM`

Question 1: `QUESTION_1`
Evidence 1: `EVIDENCE_1`
Stance 1: `STANCE_PREDICTION_1`

Question 2: `QUESTION_2`
Evidence 2: `EVIDENCE_2`
Stance 2: `STANCE_PREDICTION_2`

Question 3: `QUESTION_3`
Evidence 3: `EVIDENCE_3`
Stance 3: `STANCE_PREDICTION_3`

The JSON should include two main keys: `"pred_label"` and `"reasoning"`.

Your answer must be in `'Spanish' if lang == 'es' else 'English'`.

Figure 9: Few-Shot prompt used to perform veracity prediction. Text that appears in gray indicates optional stance prediction input, based on configuration.

Input

Given a claim, you are tasked with establishing whether the evidence given *Supported*, *Refuted*, *Not Enough Evidence* or *Conflicting/Cherrypicking*.
For every claim, there is one `evidence` attribute which includes multiple `answer` elements and their respective `stance` elements. You must provide one overall prediction for each claim in the `pred_label` attribute in the JSON.

Reasoning process - Follow these steps systematically:

1. Identify the key assertions and components of the claim that need to be verified.
2. For each piece of evidence provided: summarize what the evidence states, assess the quality and reliability of the evidence, determine how directly it relates to the claim.
3. For each evidence piece, determine if it supports the claim (and to what degree), refutes the claim (and to what degree), is neutral/irrelevant to the claim, contains conflicting information.
4. Look across all evidence pieces for consistent patterns of support or refutation, contradictions between different evidence sources, gaps in information needed to verify the claim, signs of selective evidence presentation (cherry-picking).
5. Combine your analysis of all evidence to determine the overall relationship between the evidence set and the claim.
6. Based on your analysis, classify as *Supported* (evidence consistently and reliably supports the claim), *Refuted* (evidence consistently and reliably contradicts the claim), *Not Enough Evidence* (insufficient reliable evidence to decide), *Conflicting/Cherrypicking* (evidence presents contradictory information or appears selectively chosen).

Analyze the following:
Claim: `CLAIM`

Question 1: `QUESTION_1`
Evidence 1: `EVIDENCE_1`
Stance 1: `STANCE_1`

Question 2: `QUESTION_2`
Evidence 2: `EVIDENCE_2`
Stance 2: `STANCE_2`

Question 3: `QUESTION_3`
Evidence 3: `EVIDENCE_3`
Stance 3: `STANCE_3`

Instructions for your response: Please work through each step of the reasoning process outlined above. Show your thinking clearly for each step before providing your final answer.
The JSON should include two main keys: `"pred_label"` and `"reasoning"`.
Your answer must be in `'Spanish' if lang == 'es' else 'English'`.

Figure 10: CoT prompt used to perform veracity prediction. Text that appears in gray indicates optional stance prediction input, based on configuration.

Input

Given a claim, you are tasked with establishing whether the evidence given *Supported*, *Refuted*, *Not Enough Evidence* or *Conflicting/Cherrypicking*.
For every claim, there is one `evidence` attribute which includes multiple `answer` elements and their respective `stance` elements. You must provide one overall prediction for each claim in the `pred_label` attribute in the JSON.

Examples with detailed reasoning:

Example 1 shown in Few-Shot...
Chain of Thought Analysis: 1. Claim asserts that masks will stop COVID-19 spread. 2. Evidence from Oxford study shows masks reduce spread. Positive stance aligns with claim. 3. Evidence strongly supports claim. 4. Consistent positive stance from reliable source. 5. Oxford study provides strong backing. 6. Final classification: Supported.

Example 2 shown in Few-Shot...
Chain of Thought Analysis: 1. Claim asserts Trump administration made statements about Billie Eilish. 2. Evidence shows correction from Washington Post; claim is false. 3. Evidence refutes claim, negative stance aligns. 4. Consistent negative stance from reliable source. 5. Correction confirms claim unfounded. 6. Final classification: Refuted.

Example 3 shown in Few-Shot...
Chain of Thought Analysis: 1. Claim questions logic of oil price differences. 2. Evidence gives general factors but not specific comparison. 3. Evidence is relevant but inconclusive, neutral stance aligns. 4. No specific analysis on Nigeria vs Saudi Arabia. 5. Background info insufficient for claim verification. 6. Final classification: Not Enough Evidence.

Reasoning process - Follow these steps systematically:
Reasoning process presented in CoT ...

Analyze the following:
Claim: `CLAIM`
Question 1: `QUESTION_1`
Evidence 1: `EVIDENCE_1`
Stance 1: `STANCE_1`

Question 2: `QUESTION_2`
Evidence 2: `EVIDENCE_2`
Stance 2: `STANCE_2`

Question 3: `QUESTION_3`
Evidence 3: `EVIDENCE_3`
Stance 3: `STANCE_3`

Instructions for your response:
Please work through each step of the reasoning process outlined above. Show your thinking clearly for each step before providing your final answer.

The JSON should include two main keys: `"pred_label"` and `"reasoning"`.
Your answer must be in `'Spanish' if lang == 'es' else 'English'`.

Figure 11: Few-Shot CoT prompt used to perform veracity prediction. Text that appears in gray indicates optional stance prediction input, based on configuration.

Dataset	Type	Model	Technique	Stance	Macro F1	@.25
AVeriTeC	Small	LLaMA3 8B	Zero-Shot	✓	0.2920	0.106
				✗	0.2775	0.116
			Few-Shot	✓	0.320	0.11
				✗	0.317	0.112
			CoT	✓	0.2328	0.086
				✗	0.2098	0.078
			Few-Shot CoT	✓	0.382	0.092
				✗	0.2777	0.09
		Qwen2.5 7B	Zero-Shot	✓	0.2673	0.094
				✗	0.2430	0.088
			Few-Shot	✓	0.2685	0.068
				✗	0.2662	0.088
			CoT	✓	0.2504	0.088
				✗	0.2486	0.082
			Few-Shot CoT	✓	0.2586	0.082
				✗	0.2390	0.084
	Large	LLaMA3 70B	Zero-Shot	✓	0.3317	0.112
				✗	0.3444	0.115
			Few-Shot	✓	0.2762	0.1
				✗	0.2904	0.102
			CoT	✓	0.2015	0.058
				✗	0.2334	0.072
			Few-Shot CoT	✓	0.2633	0.098
				✗	0.2648	0.1
		Qwen2.5 72B	Zero-Shot	✓	0.3004	0.102
				✗	0.3147	0.11
			Few-Shot	✓	0.3057	0.102
				✗	0.3043	0.106
			CoT	✓	0.3021	0.1
				✗	0.342	0.096
			Few-Shot CoT	✓	0.2989	0.1
				✗	0.316	0.1
	Closed	GPT-4o	Zero-Shot	✓	0.3068	0.106
				✗	0.3114	0.108
			Few-Shot	✓	0.3119	0.112
				✗	0.2770	0.094
			CoT	✓	0.2781	0.094
				✗	0.2837	0.092
			Few-Shot CoT	✓	0.2759	0.088
				✗	0.2762	0.096
	AVeriTeC Baseline				0.3211	0.092

Table 2: Macro F1 and AVeriTeC score (@.25) results on AVeriTeC dataset by model, prompting technique and stance inclusion. Veracity prediction task.

Dataset	Type	Model	Technique	S. Macro F1	Stance	V. Macro F1
FactOReS	Small	LLaMA3 8B	Zero-Shot	0.3471	✓	0.5044
					✗	0.5156
			Few-Shot	0.3713	✓	0.4540
					✗	0.4846
			CoT	0.5597	✓	**0.5466**
					✗	0.5340
			Few-Shot CoT	0.4426	✓	0.4843
					✗	0.5328
		Qwen2.5 7B	Zero-Shot	0.5718	✓	0.6374
					✗	0.6290
			Few-Shot	0.5908	✓	0.5918
					✗	0.5950
			CoT	0.6326	✓	0.6519
					✗	0.6528
			Few-Shot CoT	0.6425	✓	**0.6563**
					✗	0.6249
	Large	LLaMA3 70B	Zero-Shot	0.5019	✓	0.4870
					✗	0.4821
			Few-Shot	0.5776	✓	0.6440
					✗	0.6296
			CoT	0.6164	✓	0.6580
					✗	**0.6665**
			Few-Shot CoT	0.5896	✓	0.6257
					✗	0.6494
		Qwen2.5 72B	Zero-Shot	0.5831	✓	0.6665
					✗	0.6743
			Few-Shot	0.6635	✓	0.6871
					✗	**0.7114**
			CoT	0.6374	✓	0.6646
					✗	0.6698
			Few-Shot CoT	0.6115	✓	0.6518
					✗	0.6940
	Closed	GPT-4o	Zero-Shot	0.6536	✓	0.6172
					✗	0.6029
			Few-Shot	0.6517	✓	0.6476
					✗	0.6465
			CoT	0.6500	✓	0.6416
					✗	**0.6949**
			Few-Shot CoT	0.6686	✓	0.6231
					✗	0.6804

Table 3: Macro F1 results on FactOReS dataset by model, prompting technique and stance inclusion. Stance detection and veracity prediction tasks.