EuroVerdict: A Multilingual Dataset for Verdict Generation Against Misinformation

Daniel Russo[1,2], **Fariba Sadeghi**[1], **Stefano Menini**[1], **Marco Guerini**[1]
[1]Fondazione Bruno Kessler, Italy
[2]University of Trento, Italy
{drusso, fsadeghi, menini, guerini}@fbk.eu

Abstract

Misinformation is a global issue that shapes public discourse, influencing opinions and decision-making across various domains. While automated fact-checking (AFC) has become essential in combating misinformation, most work in multilingual settings has focused on claim verification rather than generating explanatory verdicts (i.e. short texts discussing the veracity of the claim), leaving a gap in AFC resources beyond English. To this end, we introduce EuroVerdict, a multilingual dataset designed for verdict generation, covering eight European languages. Developed in collaboration with professional fact-checkers, the dataset comprises claims, manually written verdicts, and supporting evidence, including fact-checking articles and additional secondary sources. We evaluate EuroVerdict with `Llama-3.1-8B-Instruct` on verdict generation under different settings, varying the prompt language, input article language, and training approach. Our results show that fine-tuning consistently improves performance, with models fine-tuned on original-language articles achieving the highest scores in both automatic and human evaluations. Using articles in a different language from the claim slightly lowers performance; however, pairing them with language-specific prompts improves results. Zero-shot and Chain-of-Thought setups perform worse, reinforcing the benefits of fine-tuning for multilingual verdict generation.

1 Introduction

Misinformation is increasingly pervasive in modern society (Lazer et al., 2018), with significant impacts at both societal and individual levels (Adams et al., 2023). At the individual level, misinformation can strongly influence people's beliefs and behaviours in response to false claims (Lewandowsky et al., 2012). Well-known and evident examples are vaccine hesitancy, religious extremism, and disengagement from political participation (Bronstein et al., 2021; Booth et al., 2024; Ecker et al., 2024). At the societal level, misinformation undermines trust in media (Wagner and Boczkowski, 2019), erodes public understanding of science — particularly on critical issues like health and climate change (Lewandowsky et al., 2017) — destabilizes markets (Petratos, 2021; Kogan et al., 2023), and poses a serious threat to democracy by preventing the electorate from being accurately informed (Kuklinski et al., 2000). The 2024 Global Risks Report[1] identifies misinformation and disinformation as the most severe short-term global risk for fueling polarization, civil unrest, and the erosion of democratic rights.

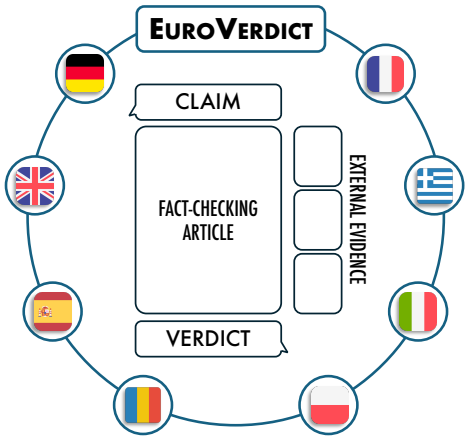


Figure 1: EuroVerdict dataset comprises quadruples of <CLAIM,VERDICT, FC ARTICLE, EXTRA EVIDENCE> in eight European languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian*.

Over the past decade, significant efforts have been made to counter misinformation worldwide (Cazzamatta, 2024; Humprecht, 2019), as evidenced by the proliferation of fact-checking organizations and growing research interest

[1]https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Findings of the Association for Computational Linguistics: ACL 2025, pages 16617–16634

ENGLISH

CLAIM: Donald Trump is the Time Magazine 2023 Person of the Year.
VERDICT: The claim is false. The photo is actually Time Magazine's 2016 cover, as a Time spokesperson confirmed it. In 2023 Taylor Swift was the one named TIME's Person of the Year.
ARTICLE URL: https://checkyourfact.com/2023/12/13/fact-check-no-donald-trump-is-not-time-magazines-2023-person-of-the-year/
EXTERNAL EVIDENCE: https://time.com/6342806/person-of-the-year-2023-taylor-swift/

SPANISH

CLAIM: El censo electoral ha disminuido de 37 a 35 millones de personas en las elecciones del 23 de julio
VERDICT: Los datos oficiales del INE indican que el censo electoral ha aumentado en lugar de disminuir.
ARTICLE URL: https://maldita.es/malditobulo/20230801/censo-electoral-disminuido-elecciones-2023/
EXTERNAL EVIDENCE: https://www.ine.es/prensa/elecgral_nov2019.pdf; https://resultados.generales23j.es/es/resultados/0/0/20#PARTICIPATION

ITALIAN

CLAIM: L'Università di Yale ha sviluppato una tecnologia per vaccinare le persone a loro insaputa
VERDICT: La tecnologia sviluppata dall'università di Yale funziona in modo simile ad altri vaccini nasali già esistenti, e non ci sono prove che questo vaccino possa funzionare se diffuso nell'aria come aerosol o nebulizzato. Inoltre il prodotto non risulta essere stato testato su animali di grandi dimensioni o esseri umani.
ARTICLE URL: https://facta.news/antibufale/2023/10/02/universita-yale-vaccino-covid/
EXTERNAL EVIDENCE: https://www.reuters.com/fact-check/yale-study-did-not-develop-technology-vaccination-without-consent-2023-09-22/

Table 1: Examples of entries of EuroVerdict dataset in English, Spanish, and Italian.

(Lewandowsky and van der Linden, 2021). In particular, the field of computer science has increasingly focused on developing automated systems to detect and fact-check misleading claims online (Guo et al., 2022). Within the automated fact-checking (AFC) domain, two primary approaches have emerged: (1) *verdict prediction*, which involves classifying the truthfulness of a claim using binary or multi-class labels, and (2) *verdict production* (or *generation*), which generates a short explanatory text, i.e, a verdict, detailing *why* a claim is true or false (Kotonya and Toni, 2020a; Atanasova et al., 2020; Russo et al., 2023b).

Although research shows that providing a short verdict with few key arguments is more effective than simply labelling claims as true or false (Lewandowsky et al., 2012; Sanna and Schwarz, 2006; Lombrozo, 2007), multilingual AFC efforts have focused largely on verdict prediction, leaving a notable gap in the verdict production literature. Indeed, the only multilingual resource for verdict production, RU22Fact (Zeng et al., 2024), includes texts in English, Chinese, Russian, and Ukrainian, limited to the Russia-Ukraine conflict.

In this paper, we introduce *EuroVerdict*, a multilingual dataset for verdict generation that includes claims, explanatory verdicts, and their supporting evidence, consisting of fact-checking articles and additional secondary sources.[2] EuroVerdict spans *eight* European languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian* (Figure 1). The dataset was developed in collaboration with European professional fact-checkers, who selected the claims, manually crafted the verdicts, and supplemented existing fact-checking articles with additional evidence. The final dataset comprises 1,642 entries (~200 per language) and covers a wide range of topics, such as *health*, *politics*, and *economy* (examples in Table 1). To evaluate EuroVerdict, we tested the `Llama-3.1-8B-Instruct` language model on the verdict generation task, exploring various input settings (English and language-specific prompts) and training approaches, including in-context learning (zero-shot and Chain-of-Thought) and fine-tuning.

Our results show that fine-tuning consistently improves performance, while the Chain-of-Thought (CoT) strategy yields the weakest results. Using translated articles[3] slightly lowers performance compared to original-language sources; however, this difference remains small, demonstrating both the robustness of the Llama model in handling multilingual prompts and its ability to generate high-quality verdicts even when the supporting information is not in the same language as the claim. Notably, fine-tuning leads to significant improvements on specific languages, such as Greek,

[2]The dataset is publicly available at github.com/LanD-FBK/EuroVerdict

[3]With "translated article" we refer to the article in one language translated to English to simulate the case where the available evidence is in a different language than the claim.

without degrading performance in other languages, highlighting the robustness of the Llama model in multilingual settings. The human evaluation with professional fact-checkers aligns with these findings: fine-tuned models received the highest scores, and verdicts generated from original-language articles were preferred over those generated starting from translated ones. Additionally, using language-specific prompts with translated articles improves quality compared to using English prompts, while CoT remains the least effective configuration.

2 Related Work

Since the earliest theorization of automated fact-checking (AFC; Thorne and Vlachos, 2018), researchers have devoted considerable attention to the development of systems capable of detecting potentially misleading claims (*claim detection*), retrieving reliable and trustworthy information (*evidence retrieval*), and ultimately determining whether the claims are true or false (*verdict prediction*; Guo et al., 2022). More recently, advancements in large language model (LLM) generation capabilities have enabled researchers to focus on providing explanations for why claims may be true or false. This emerging task, often referred to in the literature as *verdict production/generation* (Guo et al., 2022; Kotonya and Toni, 2020a), is hereafter referred to as *verdict generation* for consistency.

Among AFC tasks, verdict generation is particularly challenging (Atanasova et al., 2020) as it requires systems not only to select and extract the most relevant arguments for fact-checking claims but also to present them in a manner that is coherent, grammatically correct, and faithful to the context. Due to these complexities, early attention-based (Kang et al., 2024) and rule-based (Yang et al., 2019) approaches were quickly superseded by summarization methods. Specifically, end-to-end fine-tuning of transformer models - whether with extractive objectives (Atanasova et al., 2020), abstractive objectives (Kotonya and Toni, 2020a), or a combination of both (Russo et al., 2023b,a) - has proven effective in generating high-quality verdicts by summarizing fact-checking articles.

However, a major limitation of this approach is the tendency of language models to produce factual inaccuracies, commonly referred to as 'hallucinations' (Huang et al., 2025). Recent advances in LLM generation quality have provided a pathway to address this issue by integrating evidence retrieved from curated and reliable knowledge bases (Lewis et al., 2020). The retrieval-augmented generation approach has demonstrated promise in enhancing the factual accuracy of generated verdicts, both in multimodal (Yao et al., 2023) and text-only scenarios (Zeng and Gao, 2024; Russo et al., 2024).

Ad-hoc strategies for collecting verdict generation data have relied on synthetic data generation, such as e-FEVER (Stammbach and Ash, 2020), and journalistic sources, including LIARPLUS (Alhindi et al., 2018), PUBHEALTH (Kotonya and Toni, 2020b), LIAR++, and FullFact (Russo et al., 2023b). For more nuanced and realistic claims in the style of social media posts, datasets such as MisinfoCorrect (He et al., 2023) and VerMouth (Russo et al., 2023a), an extension of FullFact, have incorporated emotional claims and verdicts grounded in trustworthy fact-checking articles.

Despite the global popularity of AFC, research into multilingual AFC has primarily concentrated on the task of verdict prediction (Panchendrarajan and Zubiaga, 2024). Efforts in this area have produced both language-specific datasets, such as those for Danish (Nørregaard and Derczynski, 2021), Chinese (Hu et al., 2022), and Arabic (Baly et al., 2018; Sheikh Ali et al., 2021), as well as multilingual datasets (Gupta and Srikumar, 2021). However, datasets for verdict generation remain predominantly limited to the English language.

To the best of our knowledge, RU22Fact (Zeng et al., 2024) is the only dataset that directly addresses multilingual verdict generation. This dataset includes claims and verdicts in English, Chinese, Russian, and Ukrainian. RU22Fact was created by collecting claims related to the Russia-Ukraine conflict from fact-checking websites and credible news outlets, using fact-checking justifications as explanations for claims from the former, and summarizing and manually verifying news articles for claims from the latter.

In this paper, we take a significant step forward by introducing a novel multilingual dataset for verdict generation, encompassing claims and verdicts written in eight European languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, *Romanian*. The claims, spanning a wide range of topics, were sourced from the Google Fact Check platform. Verdicts were meticulously written by professional fact-checkers, that were actively involved in the dataset's development, ensuring its accuracy and reliability.

3 Dataset Creation

We propose the EuroVerdict, the first multilingual dataset for verdict generation comprising data written by professional fact-checkers around Europe. Hereafter, we detail the dataset creation.

3.1 Data Collection

We began by collecting pairs of claims and corresponding fact-checking (FC) articles in multiple languages from reliable fact-checking sources. Subsequently, professional fact-checkers were tasked with writing a verdict for each claim based on the information presented in the article and potentially additional external resources. We adopted a *nichesourcing* approach for data collection to enhance the quality of our final resource, drawing inspiration from prior works (e.g., Chung et al. (2019)). This section will provide an in-depth description of our data collection procedure. In Appendix A.2 we detailed the guidelines provided to the professional fact-checkers for the creation of the dataset.

Annotators and Languages. The data collection of quadruples <CLAIM,VERDICT, FC ARTICLE, EXTRA EVIDENCE> was performed in eight languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian*. The annotation process involved two professional fact-checkers per language across the eight languages. In order to collect high-quality data, all fact-checkers are native speakers of their respective languages, have a minimum of two years of fact-checking experience, and are members of the European Fact-Checking Standards Network. They were tasked with (i) *selecting claims*, (ii) *writing verdicts* following specific guidelines while grounding the information from FC articles, and (iii) *providing additional material* supporting the verdict. The data collection required roughly two months.

Claim Selection Procedure. This step aimed to collect a balanced dataset of approximately 200 misinformation claims per language. To ensure that our data met fact-checking standards, we started by selecting only reliable and trusted sources. For this purpose, we resorted to gathering claims from Google Fact Check Tools[4] verified websites. We selected the claims through a publisher-based[5] search with the service API. In addition to the claim itself, for each language, we collected other relevant information provided by Google Fact Check Tools, including the date, the URL of the fact-checking article,[6] the information about the source publisher who fact-checked the claim (name and official website), and the claims rating (e.g., true, false). We then applied a filter to the collected claims, by considering only those that were rated as FALSE, PARTLY FALSE or MISSING CONTEXT (or equivalent labels according to the language being used). If different labels (RATINGS) were used, we asked the professional annotator to select only those claims that could be mapped to one of these three categories. Furthermore, we instructed annotators to exclude, as much as possible, claims containing direct quotes from politicians to mitigate potential annotator bias. After the initial filtering and collecting roughly 200 claims per language, we began the annotation procedure.

Verdict Writing Procedure. To ensure consistency across languages and minimize dependence on annotators' personal style or organizational affiliations, we provided annotators with guidelines for the writing of the verdicts. First, we requested that the provided verdict be a concise text of a few sentences explaining why the claim is false/partly false, etc., serving as evidence or a basis for the rating. We also asked the verdict to be as neutral as possible so as to comply with *The European Code of Standards for Independent Fact-Checking Organisations* (EFCSN, 2022). Furthermore, we paid particular attention to multi-modal aspects, i.e. if the verdict was discussing images or video. We required annotators to ensure that the verdict was understandable without viewing accompanying images or videos. In cases where this was not possible, such claims were discarded. Finally, we provided some examples in English as a reference, such as the one below.

> **CLAIM**: The Pope Francis has been seen partying and drinking alcohol
>
> **VERDICT**: Details in the images (that circulated online) may indicate that they have been created with AI tools. Also, it was proved that the social media profiles disseminating the content were defined as 'meme account'.

[4] https://toolbox.google.com/factcheck/

[5] More details on the full list of publishers selected for collecting the claims can be found in Table 6 in Appendix A.1.

[6] Articles were scraped using either ad-hoc scrapers or the `Newspaper` library: https://github.com/codelucas/newspaper

External Evidence Annotation Procedure. Along with the original fact-checking article link, we asked annotators to provide a list of links to resources comprising relevant information supporting the verdict. These links could be extracted directly from the references used in the fact-checking article itself or were provided by fact-checkers. We further specified that the relevant evidence should be presented in a text-based format, meaning that direct links to videos or images are excluded.

3.2 Dataset Description

The final dataset consists of 1,642 entries, with nearly 200 per language (see Table 2). Each entry includes the claim, the fact-checking article (along with details about its publisher), the gold verdict written by professional fact-checkers, and links to additional evidence providing relevant information for verifying the claim. Examples from EuroVerdict can be found in Table 1.

In Table 2 we provide the average length of the main components of EuroVerdict dataset. In particular, we report the average number of sentences and words for the claims, the verdicts, and the fact-checking articles

Additionally, in order to have an overview of the topic covered by EuroVerdict dataset, we performed topic modeling on the claims annotated by the fact-checkers during data collection. Using the BERTtopic strategy (Grootendorst, 2022), we identified underlying themes within the data (see Figure 2 for a visualization of the resulting topics on the whole dataset). Our implementation incorporated `bge-M3` (Chen et al., 2024) as a multilingual embedded model, *Uniform Manifold Approximation and Projection* (UMAP; McInnes et al., 2018) for dimensionality reduction, and `HDBSCAN` algorithm from (Campello et al., 2013) as hierarchical clustering strategy. To automatically assign cluster-specific labels to the topics, we used `Llama-3.1-8B-Instruct` for one-shot classification as our representation model. For technical details, as well as for language-specific topic analysis, please refer to Appendix B.

	# Items	# Ext. Ev.	Article		Claim		Verdict	
			Sent.	Words	Sent.	Words	Sent.	Words
EuroVerdict	1642	2	41	904	1	16	2	35
German	201	1	54	802	1	19	2	15
Greek	195	7	50	968	1	21	2	45
English	195	1	29	567	1	14	2	30
Spanish	263	2	19	563	1	15	1	23
French	190	1	55	1.585	1	17	2	44
Italian	202	1	17	428	1	16	2	41
Polish	204	2	64	1.112	1	10	2	32
Romanian	192	1	49	1.357	1	14	2	52

Table 2: EuroVerdict statistics. We report the total number of items (#Items) and the average count of external evidence (#Ext. Ev.), words, and sentences.

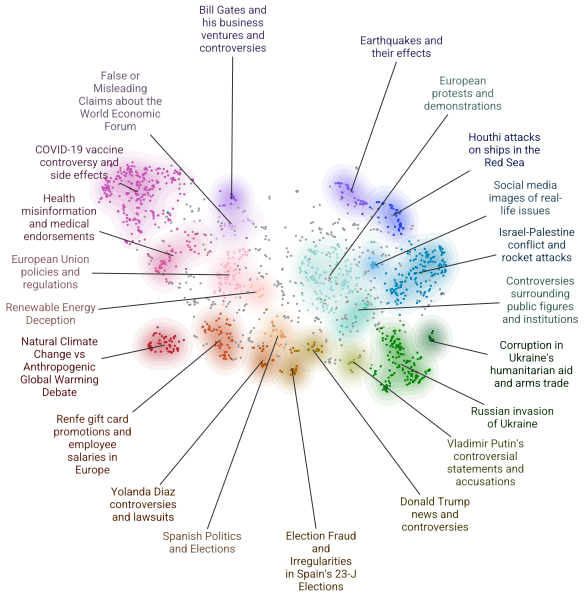


Figure 2: EuroVerdict topics distribution obtained with the BERTopic topic modeling technique.

The topic modeling analysis identified several macro-topics related to misinformation and public discourse (Figure 2). **Political discussions** were prominent, with topics such as *"Spanish Politics and Elections"*, *"Election Fraud and Irregularities in Spain's 23-J Elections"*, and *"European Union policies and regulations"*. **Global conflicts** were also well-represented, including *"Russian invasion of Ukraine"*, *"Israel-Palestine conflict and rocket attacks"*, and *"Houthi attacks on ships in the Red Sea"*. **Health-related misinformation** emerged in topics like *"COVID-19 vaccine controversy and side effects"* and *"Health misinformation and medical endorsements"*. Additionally, the analysis captured **economic and environmental themes**, such as *"Renewable Energy Deception"*, and *"Natural Climate Change vs Anthropogenic Global Warming Debate"*. Finally, **public political figures** and institutions were frequently discussed, with topics including *"Donald Trump news and controversies"*, *"Bill Gates and his business ventures and controversies"*, *"Vladimir Putin's controversial statements and accusations"*, and *"Yolanda Diaz controversies and lawsuits"*.

4 Experimental Design

We tested the `Llama-3.1-8B-Instruct` multilingual model[7] on EuroVerdict for the task of verdict generation, exploring various configurations.[8] Our experimental design considered three key aspects: *prompt configuration*, *article configuration*, and *training setup*. The first two aspects address the multilingual setting by varying the language of the prompt and input, while the third pertains to the training setup.

Prompt Configurations. We investigated whether using English prompts versus language-specific prompts would influence the model's performance, regardless of the language of the claims and the evidence articles. In particular, we started with the following English prompt.

> **SYSTEM:** You are an expert fact-checker. Your task is to evaluate a claim based solely on the provided context. You must strictly adhere to the following rules:
> 1. Base your response only on the given context; do not bring in external knowledge.
> 2. Respond concisely in no more than three sentences.
> 3. Do not reference the context or mention that you had a context in your response.
> 4. Match the emotional tone and communication style of the claim.
> 5. Respond entirely in `{language}`.
> **USER:** Evaluate the following claim based on the given information.
> <context> `{article}` </context>
> Claim: `"{claim}"`

Subsequently, we automatically translated with *Google Translate* the prompt into the eight languages of EuroVerdict.

Article Configurations. Regarding the article configuration, we compared the model's performance when given fact-checking articles in their original language versus their English translations. Specifically, the articles were automatically translated using *Google Translate*. For the purposes of this study, we opted to use Google Translate to automatically translate the documents. This decision was primarily driven by practical considerations: Google Translate is one of the few freely available tools capable of handling the automatic translation of long documents. While we acknowledge the potential for translation errors, this approach represented one of the few viable and cost-effective solutions available to us. Moreover, our methodological assumption aligns with a realistic deployment scenario, in which translated documents are typically provided in advance and assumed to be of sufficient quality for downstream processing. This configuration serves a dual purpose: (i) to assess the model's ability to generalize across languages and evaluate its performance on non-English content, and (ii) to address scenarios where the only available information for a given claim is in a different language.

Training Setups. Furthermore, we explored three different training setups for `Llama-3.1-8B-Instruct`: in-context learning through *zero-shot* or *Chain-of-Thought (CoT)* prompting, as well as fine-tuning the model on our EuroVerdict dataset. We employed `Llama-3.1-8B-Instruct` based on preliminary experiments with professional fact-checkers. Details are provided in Appendix C.2.

For the zero-shot experiments, we used the aforementioned prompt in both English and the respective languages. For the CoT reasoning approach, we required the model to follow a structured analytical process, beginning with identifying the specific assertion made in the claim, extracting relevant contextual information from the article, evaluating supporting or contradicting evidence, and considering potential exceptions. This step-by-step reasoning framework ensured a thorough examination before reaching a final classification by requiring the model to assign a veracity label (TRUE or FALSE) to the claim and generate a concise justification explaining its decision. CoT was only tested with an English prompt, which is provided in Appendix C.3. As the final training setup, we evaluated the model after fine-tuning it on the EuroVerdict dataset. Specifically, we fine-tuned four different versions of the Llama model while keeping the training parameters constant: two models for the prompt configurations and two for the context input configurations. Details of the fine-tuning process are provided in Appendix C.4.

5 Results and Discussion

To automatically assess model performance, we compared the generated verdicts with the gold verdicts in terms of both lexical and semantic similarity. Lexical similarity was evaluated using ROUGE metrics (ROUGE-1 and ROUGE-L), while semantic similarity was assessed with BERTScore and cosine similarity. ROUGE metrics and BERTScore

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[8] Dataset partition details are provided in Appendix C.1.

		ROUGE-1		ROUGE-L		BERTScore		Cosine Similarity	
		EL	LS	EL	LS	EL	LS	EL	LS
original articles	Zero-shot	0.325	0.311	0.238	0.229	0.734	0.727	0.669	0.642
	Fine-tune	**0.387**	**0.393**	**0.312**	**0.315**	**0.760**	**0.761**	**0.698**	**0.707**
	CoT	0.270	-	0.190	-	0.710	-	0.644	-
translated articles	Zero-shot	0.302	0.296	0.218	0.216	0.724	0.721	0.664	0.657
	Fine-tune	**0.336**	**0.342**	**0.261**	**0.264**	**0.741**	**0.741**	**0.682**	**0.680**
	CoT	0.258	-	0.183	-	0.702	-	0.616	-

Table 3: Averaged experimental results across all languages, presented for all experimental design configurations. Prompt types include EL (english-language prompts) and LS (language-specific prompts).

were computed using the `Evaluate` library;[9] for cosine similarity, embeddings were generated using the `paraphrase-multilingual-MiniLM-L12-v2` model from Sentence Transformers (Reimers and Gurevych, 2019).[10]

In Table 3, we present the results averaged across all languages, providing an overall assessment of model performance. Language-specific results are detailed in Tables 9 and 10 (Appendix D.1). In Table 5, we provide examples of the generations. The results demonstrate that fine-tuning consistently improves performance across all experimental settings. In contrast, the CoT strategy yields the weakest results, suggesting that step-by-step reasoning does not necessarily enhance fact-checking accuracy in this context. In both CoT setups the model was able to correctly determine the veracity of the claims, as shown in the table 4. Given that the model in both CoT configurations was able to assign the correct label in more than 95% of the cases, we believe this step had minimal impact on the final results and did not significantly affect verdict generation. This can be attributed to cases where the model accumulated errors throughout the reasoning steps, selected arguments different from the gold verdict, or responded in the incorrect language. Additionally, using articles in a different language from the claim results in lower scores compared to using articles in their original language. However, this performance drop is not substantial, suggesting that the model is sufficiently robust to retrieve information in one language and transfer it to another for verdict generation. Indeed, the slight decline can be attributed to the additional reasoning step required for the model, which must not only extract relevant information but also effectively "translate" the knowledge into a different language.

[9] https://huggingface.co/docs/evaluate/

[10] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Approach	TRUE	FALSE
CoT original articles	8 (3.27%)	237 (96.73%)
CoT translated articles	10 (4.08%)	235 (95.92%)

Table 4: CoT veracity prediction results.

Analyzing performance across specific languages (Tables 9 and 10 in Appendix D.1), we observe that fine-tuning leads to significant improvements for certain languages, such as Greek, where the increase is particularly pronounced. Notably, this enhancement does not negatively impact performance in other languages, even when training across multiple languages with distinct alphabets. This finding highlights the robustness of the Llama 3.1 model, demonstrating its ability to adapt effectively to multilingual fine-tuning without degradation in overall performance.

5.1 Human Evaluation

To better assess the quality of the generations produced , we conducted a human evaluation. For this, we enlisted a pool of professional fact-checkers who were native speakers of the languages included in EuroVerdict. Evaluators were asked to rate the quality of the generated verdicts on a scale from 1 to 5, with respect to the gold verdict, across three main dimensions: **alignment** (*"How much is the current verdict similar to the gold verdict?"*), **understandability** (*"How much is this verdict written in an understandable way?"*), **soundness** (*"How much does the verdict sound like a well-formed verdict?"*).[11]

Specifically, we provided the evaluators with the claim and gold verdict, along with five different generated verdicts, one for each configuration (always presented in random order). These included verdicts produced in zero-shot, fine-tuning with En-

[11] Details in Appendix D.2.

	Claim	Gold Verdict	Generated Verdict
EN	Israeli PM Benjamin Netanyahu is contracted with a mutated strain of Covid	The claim is false. The image was manipulated from a photo of his wife Sara visiting the victim of a 2015 attack in Jerusalem. As of March 8, 2024, there have been no official reports that Netanyahu was being treated in hospital for a Covid-19 infection.	The claim is false. The picture shown of Netanyahu is digitally altered from a photo of his wife visiting the victim of a 2015 attack in Jerusalem.
			The claim is false. The photo of Netanyahu hospitalised is actually a digital alteration of a photo from 2015 where his wife was visiting a hospital patient. There have been no official reports of Netanyahu's hospitalisation for a Covid-19 infection.
ES	Tom Hanks aparece en la lista de pedófilos de Jeffrey Epstein	El nombre de Tom Hanks no aparece ni en la documentación judicial desclasificada relacionada con Jeffrey Epstein ni en el registro de vuelos a la isla de Epstein.	El nombre de Tom Hanks no aparece en los documentos desclasificados ni en los registros de vuelos a la isla de Epstein.
			El nombre de Tom Hanks no aparece en los documentos desclasificados relacionados con Jeffrey Epstein, por lo que la afirmación es falsa.
DE	Moderne Verbrenner ziehen CO2 aus der Luft.	Die Behauptung ist falsch. Ein Vebrennungsmotor stößt immer CO2 aus.	Die Behauptung ist falsch. Ein mit Diesel oder Benzin betriebener Verbrennermotor strömt im Betrieb stets CO2 aus.
			Die Behauptung ist falsch. In modernen Fahrzeugen stößt der Verbrennungsmotor immer CO2 aus.

Table 5: Examples of generated verdicts in English, French, and German using the fine-tuned Llama model with articles in their original language. Each language includes two generations: one using the English prompt (first example) and one using the language-specific prompt (second example).

glish and language-specific prompts, and CoT with the English prompt. The two context input configurations were not evaluated simultaneously, as assessing ten items at once could have increased task complexity and potentially impacted the results. To mitigate this, we alternated between verdicts generated using original-language articles as context and those generated using their English translations.

In Figure 3 we report the human evaluation results averaged across all languages for *Alignment*, *Understandability*, and *Soundness* as well as the overall scores. Results show that verdicts generated by the fine-tuned model using the original article and the English prompt received the highest scores. Across all training setups, generations based on the original article were rated higher than those using translated articles, confirming that maintaining the original language improves overall quality. These preferences align with the automatic evaluation results reported in Table 3, where fine-tuning outperforms zero-shot, and models fine-tuned on original articles achieve better performance than those fine-tuned on translated articles.

Interestingly, while using translated articles led to lower scores, pairing them with language-specific prompts resulted in better evaluations compared to using them with the English prompt. Finally, the CoT strategy received the lowest scores across all configurations.

6 Conclusion

In this work, we introduced EuroVerdict, a multilingual dataset for verdict generation, covering eight European languages and developed in collaboration with professional fact-checkers. The dataset

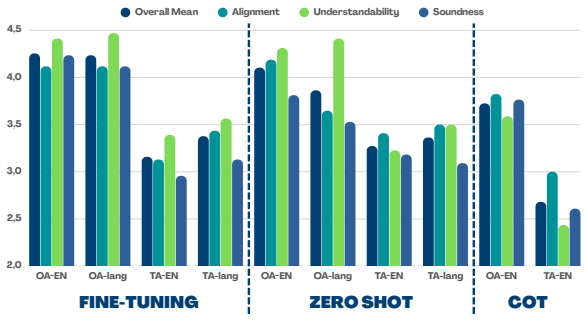


Figure 3: Human evaluation results averaged across all languages. We report the overall score (*Overall Mean*) and scores for *Alignment*, *Understandability*, and *Soundness* across configurations: fine-tuning, zero-shot, CoT, original (*OA*) or translated articles (*TA*), and English (*EN*) or language-specific prompts (*lang*).

includes claims, verdicts, and supporting evidence, consisting of fact-checking articles and additional secondary sources.

`Llama-3.1-8B-Instruct` was employed to test EuroVerdict over different configurations by varying the prompt language, input article language, and training setup. Our results show that fine-tuning consistently improves performance across all settings, while the Chain-of-Thought strategy yields the weakest results. Using articles written in a different language than the claim slightly reduces performance compared to original-language sources; however, the drop remains minimal, showcasing the model's robustness in processing multilingual inputs.

Human evaluation aligns with these findings, with fine-tuned models achieving the highest scores and verdicts generated using original-language articles being preferred over those using translations. Additionally, combining translated articles with language-specific prompts improves quality compared to using English prompts. These results highlight the effectiveness of fine-tuning and the viability of multilingual fact-checking approaches, demonstrating that high-quality verdicts can be generated even when supporting evidence is in a different language than the claim.

Limitations

Despite the remarkable capabilities of large language models, they remain prone to generating factual inaccuracies. While constraining generation by conditioning the model on fact-checking articles helps mitigate this issue, it does not eliminate it entirely. Ensuring that generated verdicts strictly adhere to the provided evidence remains an open challenge.

We acknowledge that the size of our proposed dataset is relatively small. Nevertheless, this limitation is offset by the high quality of its content: the dataset has been manually curated by professional fact-checkers from various European countries. This ensures a high degree of reliability and credibility, as each entry has been rigorously verified by domain experts.

Additionally, we recognize that the use of Google Translate for document and prompt translation may introduce inaccuracies. As previously discussed, this choice was driven by practical constraints and the assumption that, in real-world applications, reliable translations are typically provided.

Additionally, although the EuroVerdict dataset includes extra evidence beyond fact-checking articles, this work does not incorporate it into the experimental design. Generating verdicts using non-fact-checking sources presents additional complexities, as highlighted by Russo et al. (2024), and we believe this should be explored as a separate task. However, with advancements in retrieval-augmented generation (Lewis et al., 2020), this dataset represents a valuable resource for future research in this direction.

Like previous work in the field, the generative system presented in this study relies heavily on the availability of fact-checking articles, assuming that relevant knowledge is readily accessible. This dependency limits its applicability in scenarios where fact-checked information is scarce or unavailable. However, we believe that the insights provided in this work can be easily transferred to more advanced retrieval-augmented generation systems, where knowledge is not assumed to be given but is dynamically retrieved before generation.

Ethical Statement

The automation of the fact-checking process is essential in today's society, where misinformation spreads rapidly, and manual fact-checking alone cannot keep up with the sheer volume of false or misleading claims circulating online (Guo et al., 2022). However, automatic systems are not infallible: detection models may mislabel claims, and generative models can produce factual inaccuracies. Thus, we emphasize that the systems proposed in this work are not intended to function as standalone tools or replace professional fact-checkers. Instead, we view them as supportive tools that can assist fact-checkers by accelerating the verification process and enhancing their efficiency.

Acknowledgments

This work was partly supported by: the AI4TRUST project - AI-based-technologies for trustworthy solutions against disinformation (ID: 101070190), the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS), the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437 (AI-CODE).

We gratefully acknowledge the contribution of the fact-checking organizations and media partners that took part in the data collection:

ADB (Asociația Digital Bridge), *DEMAGOG* (Stowarzyszenie Demagog), *ELLINIKA* (Astiki Mi Kerdoskopiki Etairia Kentro Katapolemisis Tis Parapliroforisis / Civil Non-Profit Company Kentro Katapolemisis Tis Parapliroforisis), *EMS* (Europejskie Media Sp. zoo), *EURACTIV* (Euractiv Media Network B.V.), *MALDITA* (Fundación Maldita.es contra la desinformación: periodismo, educación investigación y datos en nuevos formatos), *SkyTG24* (Sky Italia S.r.l.)

References

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463. PMID: 36795592.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. 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.

Emily Booth, Jooyoung Lee, Marian-Andrei Rizoiu, and Hany Farid. 2024. Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*, 60(2):440–457.

Michael Bronstein, Erich Kummerfeld, Angus MacDonald III, and Sophia Vinogradov. 2021. Investigating the impact of anti-vaccine news on sars-cov-2 vaccine intentions. *Available at SSRN 3936927*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Regina Cazzamatta. 2024. Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries. *New Media & Society*, 0(0):14614448241268896.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. 2024. Misinformation poses a bigger threat to democracy than you might think. *Nature*, 630(8015):29–32.

EFCSN EFCSN. 2022. European code of standards for independent fact-checking organisations.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2698–2709, New York, NY, USA. Association for Computing Machinery.

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.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Edda Humprecht. 2019. Where 'fake news' flourishes: a comparison across four western democracies. *Information, Communication & Society*, 22(13):1973–1988.

Wan Ju Kang, Jiyoung Han, Jaemin Jung, and James Thorne. 2024. XFACT team0331 at PerspectiveArg2024: Sampling from bounded clusters for diverse relevant argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 182–188, Bangkok, Thailand. Association for Computational Linguistics.

Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2023. Social media and financial news manipulation. *Review of Finance*, 27(4):1229–1268.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

James H Kuklinski, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. 2000. Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369.

Stephan Lewandowsky and Sander van der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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.

Pythagoras N. Petratos. 2021. Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, 64(6):763–774. CIBER SPECIAL ISSUE: CYBERSECURITY IN CRISIS.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023a. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2024. Face the facts! evaluating rag-based fact-checking pipelines in realistic settings. *Preprint*, arXiv:2412.15189.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Lawrence J. Sanna and Norbert Schwarz. 2006. Metacognitive experiences and human judgment: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science*, 15(4):172–176.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large

Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries forautomated fact checking. In *Conference for Truth and Trust Online*.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

María Celeste Wagner and Pablo J. Boczkowski. 2019. The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital Journalism*, 7(7):870–885.

Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, WWW '19, page 3600–3604, New York, NY, USA. Association for Computing Machinery.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.

Appendix

A Dataset Creation

A.1 Claim Selection

Table 6 lists the reliable publisher websites used for collecting EuroVerdict's claims and their related fact-checking articles for each language.

A.2 Annotation Guidelines

Hereafter we provide the annotation guidelines:

Please open the dataset you have been provided with according to the language you are working with and follow these instructions.

1. The first goal is to reduce the database to at least 200 claims. For this, first apply a filter on Rating and select only: false, partly false and missing context (or equivalent in the language you are working with).

2. Please keep in mind that the fact-checkers providing content to the Google Fact Check Explorer have different ways of naming the ratings but you should find those referring to false, partly false and missing context.

3. Then, discard those claims that are Political fact-checks, meaning those claims attributed directly to what a politician said. You can do this by erasing those claims of political fact-checks directly in the database. For example, these of Kamala Harris and Donald Trump:

4. If this did not reduce the database to at least 200 claims, please let us know through email.

5. Once you have a dataset of between 100-200 claims you can proceed to fill in the fields of Verdict and Relevant Links.

   - **Verdict**: a few sentences long text explaining why the claim is false/partly false, etc. This works as evidence or as the basis for the rating. It should be as neutral as possible. For example, in this claim: "The Pope Francis has been seen partying and drinking alcohol", the verdict would be: "Details in the images (that were circulated online) that may indicate that they have been created with AI tools have been detected. Also it was proved that the social media profiles disseminating the content were defined as 'meme accounts'". When discussing images or video the verdict should be comprehensible even without viewing the media as in the provided example.

   - **Relevant Links**: provide a list of links on which the verdict is based (i.e. containing relevant evidence, they can be information outside of your organization). The relevant evidence should be text based (i.e. no direct link to video or images). -

6. Please let us know whenever you complete the task or if you have any questions.

B Dataset Analysis

B.1 Topic Modeling

For topic modeling, we implemented the *BERTopic* strategy using `BAAI/bge-m3` as the embedding model. Dimensionality reduction was performed with *UMAP* (n_neighbors=15, n_components=20, min_dist=0.0, metric="cosine", random_state=42), while clustering was carried out using HDBSCAN min_cluster_size=10, metric="euclidean", cluster_selection_method="eom", prediction_data=True).

To automatically assign cluster-specific labels, we employed `Llama-3.1-8B-Instruct` for one-shot classification, using the following prompt.[12]

> **SYSTEM:** You are a helpful, respectful and honest assistant for labeling topics.
> **EXAMPLE PROMPT:** I have a topic that contains the following multilingual documents: - Traditional diets in most cultures were primarily plant-based with a little meat on top, but with the rise of industrial style meat production and factory farming, meat has become a staple food.
> - Meat, but especially beef, is the word food in terms of emissions.
> - Eating meat doesn't make you a bad person, not eating meat doesn't make you a good one.
> The topic is described by the following multilingual keywords: 'meat, beef, eat, eating, emissions, steak, food, health, processed, chicken'.
> Based on the information about the topic above, please create a short English label for this topic. Make sure you only return the label and nothing more. """
> **EXAMPLE PROMPT REPLY:** Environmental impacts of eating meat
> **MAIN PROMPT:** I have a topic that contains the following multilingual documents: [DOCUMENTS]
> The topic is described by the following multilingual keywords: '[KEYWORDS]'.
> Based on the information about the topic above, please create a short English label for this topic. Make sure you only return the label and nothing more.

For visualization, we further reduced the embeddings to two components using *UMAP*

[12] We adapted the prompt provided on the BERTopic documentation https://maartengr.github.io/BERTopic/getting_started/representation/llm.html#llama-2

Language	Publishers
RO	verificat.afp.com, brodhub.eu, factual.ro
EL	ellinikahoaxes.gr, factcheckgreek.afp.com, factreview.gr
IT	facta.news, bufale.net, pagellapolitica.it
EN	cjp.org.in, logicallyfacts.com, newschecker.in, rappler.com, thequint.com, factcheck.afp.com, actcheck.org, checkyourfact.com, thip.media, snopes.com, newsweek.com, politifact.com, polygraph.info, fullfact.org
FR	francetvinfo.fr, factuel.afp.com, dpa-factchecking.com, guineecheck.org, tf1info.fr, observers.france24.com, francetvinfo.fr, defacto-observatoire.fr
PL	demagog.org.pl, sprawdzam.afp.com, oko.press
ES	verifica.efe.com, fastcheck.cl, newtral.es, factual.afp.com
DE	dpa-factchecking.com, correctiv.org, faktencheck.afp.com, br.de

Table 6: reliable publisher websites

n_neighbors=15, n_components=2, min_dist=0.0, metric="cosine", random_state=42). In Figure 2 we present the topic for all EuroVerdict data. Language-specific topic modling visualisation are presented in Figure 4.

C Experimental Design Details

C.1 Dataset Partition

For the experiments, we divided EuroVerdict into three subsets: train, evaluation, and test sets. In Table 7, we report the number of items in each set, both for the entire dataset and for each individual language.

	Train	Eval	Test
All	1152	245	245
DE	142	31	31
EL	138	30	30
EN	138	30	30
ES	186	40	40
FR	135	29	29
IT	143	31	31
PL	143	32	32
RO	135	30	30

Table 7: Data distribution across train, eval, and test sets.

C.2 Expert-Based LLM selection

To assess whether `Llama-31-8B-Instruct` could not only generate grammatically correct text in all eight languages of EuroVerdict but also produce plausible verdicts, we conducted a preliminary zero-shot verdict generation experiment followed by a human evaluation. Similar to the main experiments, we provided the model with a subset of claims in all eight languages along with their corresponding fact-checking articles and instructed it to generate verdicts based on the provided information. The generated verdicts were then evaluated by expert annotators, who were native speakers of the respective languages, and rated on a scale from 1 to 5 for grammatical correctness and soundness (i.e., *how closely the generated verdict resembles a real verdict*). Results showed that across all languages, the generated verdicts were considered both grammatically correct and plausible. Based on these findings, we proceeded to employ the LLM throughout the experiments presented in this work.

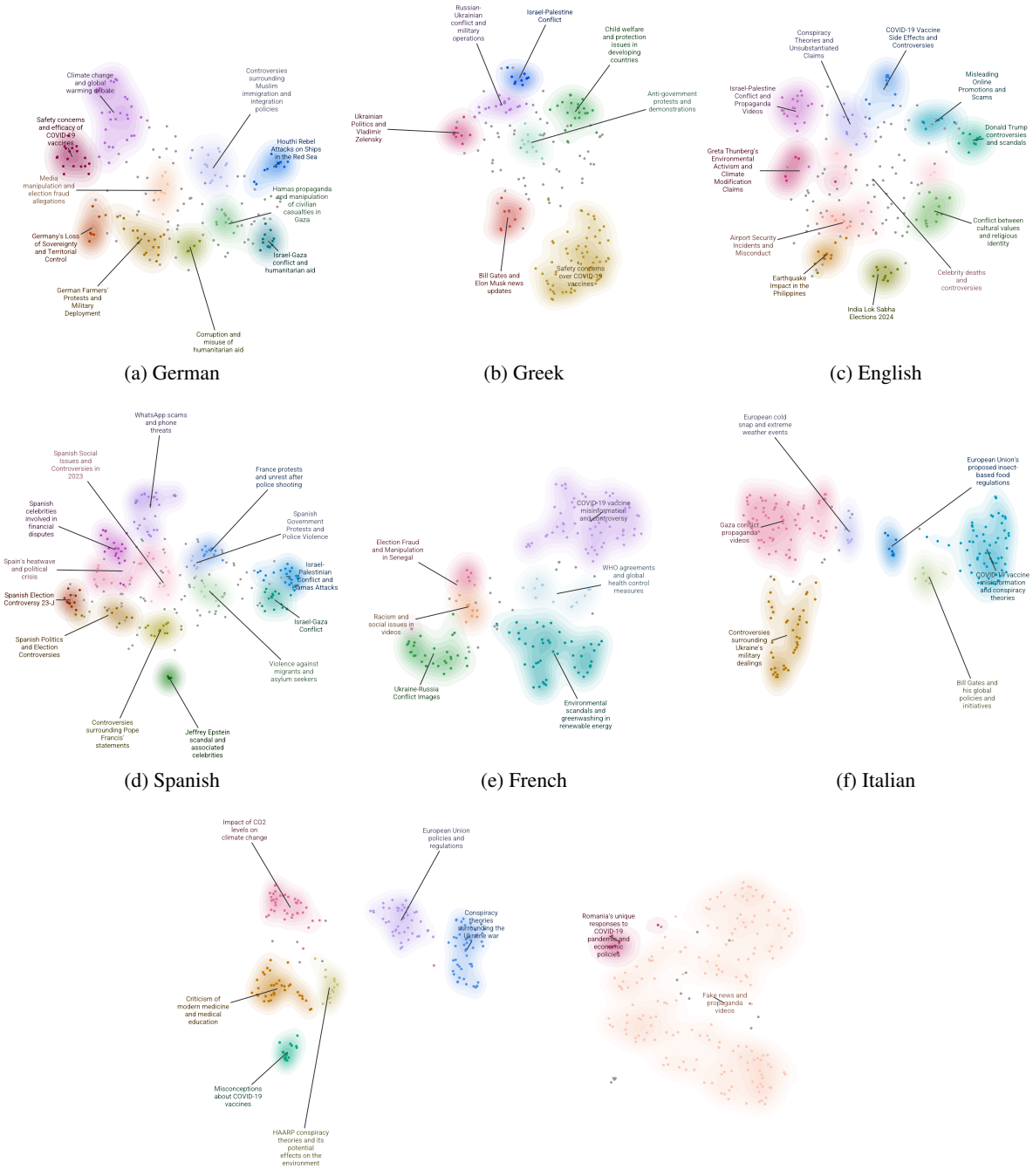
C.3 Chain-of-Thought Prompt

In the following, we provide the prompt employed in the Chain-of-Thought configuration.

SYSTEM: You will be provided with a misleading claim. Evaluate this claim step-by-step using the given article. To do so, answer to the following questions (reasoning). Finally state whether the claim is true or false (veracity label), and provide a short reply to the claim providing the reasons why the claim is true or false (justification).
1. What does the claim specifically state?
2. What context or background information in the article is relevant to evaluating the claim?
3. What evidence supports or contradicts the claim?
4. Are there exceptions or scenarios where the claim doesn't apply?
5. To what extent is the claim accurate or misleading?
Provide a conclusion with key caveats or nuances.
USER: <claim> "`{claim}`" <claim>
<article> "`{article}`" <article>
Reply in `{language}`. Return the response in JSON format: {"claim": "claim", "reasoning": [list of the replies to each question], "veracity_label": "TRUE or FALSE", "justification": "why the claim is true or false"} Leave the JSON names in English. Ensure the structure remains consistent throughout.

C.4 Fine-Tuning Details

We utilized `Llama-3.1-8B-Instruct` for verdict generation. Four versions of the LMM were fine-



(a) German (b) Greek (c) English

(d) Spanish (e) French (f) Italian

(g) Polish (h) Romanian

Figure 4: Topic analysis for each language in EuroVerdict dataset.

tuned, combining English and language specific prompts with the original fact-checking articles and the English translated ones. The same hyperparameters were used across all the fine-tuning, with the only variation being the training data. In particual, we employed the prompt presented in Paragraph 4, adding the gold verdict as for the assistant role. The training was performed on an `NVIDIA Ampere A40` GPU with 48GB of memory, applying a 4bit quantization. Low-Rank Adaptation (LoRA) was utilized with a rank of 16, an α value of 16, and a dropout rate of 0. Training parameters included a learning rate of 5×10^{-5}, a training and evaluation batch size of 6, and gradient accumulation steps of 4. The model was trained for 5 epochs, with a weight decay of 0.01, and a warm-up ratio of 0.03. For inference, we employed the checkpoint with the lowest evaluation loss.

		Overall Mean	Alignment	Understandability	Soundness
Fine-Tuning	OA-EN	4.255	4.118	4.412	4.235
	OA-lang	4.235	4.118	4.471	4.118
	TA-EN	3.159	3.13	3.391	2.957
	TA-lang	3.377	3.435	3.565	3.13
Zero-shot	OA-EN	4.104	4.188	4.312	3.812
	OA-lang	3.863	3.647	4.412	3.529
	TA-EN	3.273	3.409	3.227	3.182
	TA-lang	3.364	3.500	3.500	3.091
CoT	OA-EN	3.725	3.824	3.588	3.765
	TA-EN	2.681	3.000	2.435	2.609

Table 8: Human evaluation results averaged across all languages. We report the overall score (*Overall Mean*) and scores for *Alignment*, *Understandability*, and *Soundness* across configurations: fine-tuning, zero-shot, CoT, original (*OA*) or translated articles (*TA*), and English (*EN*) or language-specific prompts (*lang*).

D Experimental Results Details

D.1 Verdict Generation Results

In Tables In Tables 9 and 10, we present the results for all the experiments. For clarity, experiments with articles in the original language are reported in Table 9, while those with articles automatically translated into English are shown in Table 10.

D.2 Human Evaluation Instructions

Hereafter, we report the task proposed to the human evaluators.

> We prepared a list of Claims and Gold Verdicts.
>
> For each of them we provide a list of 5 additional verdicts, that we ask you to evaluate according to:
>
> - **Alignment**: How much is the current verdict similar to the gold verdict?
> - **Grammaticality**: How much is this verdict written in an understandable way (regardless of being a proper verdict)?
> - **Soundness**: How much does the verdict sounds like a well formed verdict?
>
> Score each of the provided verdicts on a scale from 1 to 5, according to the following guidelines:
>
> **Alignment**
>
> 5. Arguments and conclusions are semantically the same
> 4. Arguments and conclusions are very similar
> 3. Arguments and conclusions have some discrepancies
> 2. The conclusion is similar but arguments are completely different (or vice versa)
> 1. Arguments and conclusions are completely different
>
> **Grammaticality**
>
> 5. Understandable and grammatical
> 4. Understandable but with few grammatical errors
> 3. Understandable but with several grammatical errors
> 2. Difficult to understand with severe grammatical errors
> 1. Not understandable"
>
> **Soundness**
>
> 5. Very convincing verdict
> 4. Proper verdict
> 3. Acceptable Verdict
> 2. Not a proper verdict
> 1. Does not seem like a legit verdict"

D.3 Human Evaluation Results Details

In Table 8 we report the scores obtained from the human evaluation of the generated verdicts. Details are provided in Section 5.1.

		ROUGE-1		ROUGE-L		BertScore		Cosine Similarity	
		EL	**LS**	**EL**	**LS**	**EL**	**LS**	**EL**	**LS**
DE	Zero-shot	0.293	0.288	0.243	0.248	0.752	0.750	0.650	0.654
	Fine-tune	**0.378**	**0.393**	**0.353**	**0.362**	**0.786**	**0.792**	**0.704**	**0.724**
	CoT	0.221	-	0.176	-	0.713	-	0.652	-
EL	Zero-shot	0.294	0.347	0.217	0.257	0.721	0.729	0.599	0.657
	Fine-tune	**0.562**	**0.662**	**0.507**	**0.628**	**0.821**	**0.860**	**0.778**	**0.820**
	CoT	0.267	-	0.180	-	0.694	-	0.626	-
EN	Zero-shot	0.429	0.419	0.305	0.287	0.769	0.758	**0.760**	**0.739**
	Fine-tune	**0.470**	**0.494**	**0.373**	**0.383**	**0.784**	**0.786**	0.723	0.733
	CoT	0.391	-	0.263	-	0.741	-	0.730	-
ES	Zero-shot	0.388	0.340	0.301	0.257	0.758	0.743	0.633	0.580
	Fine-tune	**0.474**	**0.438**	**0.406**	**0.346**	**0.799**	**0.779**	**0.716**	**0.696**
	CoT	0.356	-	0.269	-	0.747	-	0.666	-
FR	Zero-shot	0.307	0.290	0.209	**0.193**	0.724	**0.717**	**0.648**	**0.645**
	Fine-tune	**0.327**	**0.299**	**0.220**	**0.193**	**0.726**	0.714	0.616	0.601
	CoT	0.273	-	0.169	-	0.706	-	0.611	-
IT	Zero-shot	0.288	0.273	0.185	**0.191**	0.722	0.713	**0.694**	0.607
	Fine-tune	**0.293**	**0.298**	**0.201**	0.189	**0.731**	**0.723**	0.671	**0.674**
	CoT	0.250	-	0.166	-	0.708	-	0.611	-
PL	Zero-shot	0.291	0.238	**0.228**	0.176	0.706	0.690	**0.696**	0.609
	Fine-tune	**0.278**	**0.267**	0.210	**0.214**	**0.709**	**0.717**	0.692	**0.725**
	CoT	0.176	-	0.133	-	0.674	-	0.596	-
RO	Zero-shot	0.290	**0.292**	0.198	**0.213**	0.710	0.712	0.640	0.670
	Fine-tune	**0.292**	0.285	**0.201**	0.200	**0.712**	0.712	**0.675**	**0.681**
	CoT	0.202	-	0.139	-	0.685	-	0.657	-
Mean	Zero-shot	0.325	0.311	0.238	0.229	0.734	0.727	0.669	0.642
	Fine-tune	**0.387**	**0.393**	**0.312**	**0.315**	**0.760**	**0.761**	**0.698**	**0.707**
	CoT	0.270	-	0.190	-	0.710	-	0.644	-

Table 9: Experimental results of `Llama-3.1-8B-Instruct` on the EuroVerdict dataset. We report performance across Zero-Shot, Fine-Tuning (using *articles in the original language*), and Chain-of-Thought (CoT) settings. Evaluation metrics include ROUGE-1, ROUGE-L, BERTScore, and Cosine Similarity. For each language, results are presented for two prompt types: EL (English-language prompts) and LS (language-specific prompts).

		ROUGE-1		ROUGE-L		BertScore		Cosine Similarity	
		EL	**LS**	**EL**	**LS**	**EL**	**LS**	**EL**	**LS**
DE	Zero-shot	0.264	0.281	0.211	0.240	0.732	0.745	0.641	0.670
	Fine-tune	**0.422**	**0.351**	**0.381**	**0.314**	**0.805**	**0.777**	**0.722**	**0.681**
	CoT	0.200	-	0.159	-	0.701	-	0.634	-
EL	Zero-shot	0.265	0.296	0.190	0.196	0.698	0.701	0.608	0.647
	Fine-tune	**0.343**	**0.349**	**0.280**	**0.284**	**0.731**	**0.730**	**0.695**	**0.744**
	CoT	0.226	-	0.162	-	0.675	-	0.566	-
EN	Zero-shot	0.437	0.422	0.298	0.299	0.766	0.766	0.718	0.742
	Fine-tune	**0.472**	**0.515**	**0.367**	**0.407**	**0.782**	**0.798**	**0.737**	**0.754**
	CoT	0.381	-	0.255	-	0.739	-	0.710	-
ES	Zero-shot	0.345	0.340	0.256	0.257	0.751	0.743	0.675	0.645
	Fine-tune	**0.395**	**0.405**	**0.294**	**0.329**	**0.770**	**0.770**	**0.869**	**0.676**
	CoT	0.316	-	0.232	-	0.734	-	0.603	-
FR	Zero-shot	**0.303**	0.270	**0.213**	0.182	**0.719**	0.710	**0.632**	0.614
	Fine-tune	0.275	**0.307**	0.181	**0.195**	0.712	**0.718**	0.605	**0.637**
	CoT	0.262	-	0.159	-	0.697	-	0.601	-
IT	Zero-shot	0.261	0.289	0.177	0.202	0.710	0.721	0.652	0.652
	Fine-tune	**0.277**	**0.293**	**0.196**	**0.207**	**0.717**	**0.725**	**0.666**	**0.662**
	CoT	0.233	-	0.158	-	0.700	-	0.606	-
PL	Zero-shot	**0.270**	0.209	**0.205**	0.156	**0.705**	0.674	**0.732**	0.637
	Fine-tune	0.232	**0.232**	0.180	**0.178**	0.701	**0.697**	0.703	**0.660**
	CoT	0.207	-	0.171	-	0.672	-	0.589	-
RO	Zero-shot	**0.264**	0.255	0.180	**0.181**	**0.706**	0.699	**0.641**	**0.652**
	Fine-tune	0.260	**0.272**	**0.195**	0.178	0.699	**0.703**	0.636	0.628
	CoT	0.226	-	0.152	-	0.686	-	0.621	-
Mean	Zero-shot	0.302	0.296	0.218	0.216	0.724	0.721	0.664	0.657
	Fine-tune	**0.336**	**0.342**	**0.261**	**0.264**	**0.741**	**0.741**	**0.682**	**0.680**
	CoT	0.258	-	0.183	-	0.702	-	0.616	-

Table 10: Experimental results of `Llama-3.1-8B-Instruct` on the EuroVerdict dataset. We report performance across Zero-Shot, Fine-Tuning (using *articles translated in English*), and Chain-of-Thought (CoT) settings. Evaluation metrics include ROUGE-1, ROUGE-L, BERTScore, and Cosine Similarity. For each language, results are presented for two prompt types: EL (English-language prompts) and LS (language-specific prompts).