Leveraging Large Language Models for Fact Verification in Italian

Antonio Scaiella[1,2], Stefano Costanzo[1], Elisa Passone[1], Danilo Croce[1,*] and Giorgio Gambosi[1]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
[2]Reveal s.r.l.

Abstract

In recent years, Automatic Fact Checking has become a crucial tool for combating fake news by leveraging AI to verify the accuracy of information. Despite significant advancements, most datasets and models are predominantly available in English, posing challenges for other languages. This paper presents an Italian resource based on the dataset made available in the FEVER evaluation campaign, created to train and evaluate fact-checking models in Italian. The dataset comprises approximately 240k examples, with over 2k test examples manually validated. Additionally, we fine-tuned a state-of-the-art LLM, namely LLaMA3, on both the original English and translated Italian datasets, demonstrating that fine-tuning significantly improves model performance. Our results suggest that the fine-tuned models achieve comparable accuracy in both languages, highlighting the value of the proposed resource.

Keywords

Automatic Fact Checking, Fact Checking in Italian, Resource in Italian, Large Language Model for Fact Verification

1. Introduction

In recent years, Automatic Fact Checking (AFC) has assumed a significant role as an instrument to identify fake news. AFC is a process that verifies the truthfulness and accuracy of information, claims, and data contained in a text or speech. The focus is on debunking disinformation and misinformation, intercepting errors, and verifying sources and facts.

Automated fact-checking uses AI tools to identify, verify, and respond to misleading claims, using techniques based on natural language processing, machine learning, knowledge representation, and databases to automatically predict the truthfulness of claims [1]. This is a complex process that involves searching, interpreting, and assessing information. As discussed in [1] a NLP framework for automated fact-checking consists of three stages: claim detection to identify claims that require verification; evidence retrieval to find sources supporting or refuting the claim; and claim verification to assess the truthfulness of the claim based on the retrieved evidence.

At first, automating the fact-checking process has been discussed in the context of computational journalism in works like [2], and has received significant attention in the computational linguistics and, in general, the artificial intelligence communities, surveyed in [1] and more recently in [3] and [4]. In particular, in [1] the authors expose a survey on the topic, describing the early developments that were surveyed in [5], which is an exhaustive overview of the subject.

As with most machine learning paradigms [1], state-of-the-art methods require datasets and benchmarks.

One of the most impactful campaigns for collecting a large-scale benchmark is FEVER (Fact Extraction and VERification) [6]. In this context, fact-checking involves verifying whether a claim is supported by one or more pieces of evidence. FEVER is a publicly available dataset designed for claim verification against textual sources. It comprises about 180K claims generated by altering sentences extracted from Wikipedia. The claims are classified into three categories: SUPPORTED (a piece of evidence exists and it supports the claim), REFUTES (a piece of evidence exists and it contradicts the claim), or NOTENOUGHINFO (there is insufficient evidence to verify the claim). The challenge, therefore, is to retrieve the relevant evidence and verify the accuracy of the claims, categorizing them with the correct label.

Many works like FEVER have recently focused on building data and datasets for the task of Fact Verification, achieving very good results [7, 8, 9, 10, 11, 12]. However, all of these datasets are designed for the English language. Although multilingual models exist (e.g., in [13, 14]), fine-tuning a model on a specific language, pre-training it for a specific task and use case, could lead to a significant decline in quality if applied to another language. Few studies have worked on training models for languages other than English. An example is the work presented in [15], which focuses on developing automated claim detection for Dutch-language fact-checkers.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
scaiella@revealsrl.it (A. Scaiella); stefano.costanzo@students.uniroma2.eu (S. Costanzo); passone@ing.uniroma2.it (E. Passone); croce@info.uniroma2.it (D. Croce); giorgio.gambosi@uniroma2.it (G. Gambosi)
0000-0001-9111-1950 (D. Croce); 0000-0001-9979-6931 (G. Gambosi)

In this work, we propose a FEVER-IT dataset in which the FEVER dataset has been translated into Italian to train the model for the Italian language. Inspired by SQUAD-IT [16] and MSCOCO-IT [17], we worked to obtain quality data. Although the training set may be affected by translation errors, the test set will not, as it is composed of manually validated data. Furthermore, while the original FEVER dataset contained evidence only for SUPPORTS and REFUTES, in this work we have also added and translated examples for the NOTENOUGHINFO category using the heuristics proposed in [18]. This work extends the experience described in [19], where translations were done using Google API, by using publicly available models ([20]) and adding data for the NOTENOUGHINFO category.

The contribution of this work is twofold. Firstly, we release FEVER-IT, a corpus with 228K claims each associated with at least one (possibly useful) piece of evidence, including a test set of 2,000 manually validated claims. In addition, we fine-tuned and validated a state-of-the-art model, LLaMA3 [14], on both the original English dataset and the Italian dataset. While this provides a high-performance model ready for the task in both languages, the primary goal is to assess whether the quality of the Italian data is comparable to the English one. By training the model separately on each dataset, we can evaluate its stability: if the model performs similarly on the manually validated Italian dataset and the English test set, we can conclude that the quality of the Italian data is on par with the English data.

Additionally, we want to assess whether using an Italian train dataset, despite the noise from automatic translation, is truly beneficial. LLMs like LLaMA3 can already perform tasks in other languages through zero-shot or few-shot learning, without requiring fine-tuning on a specific dataset, especially if that dataset is noisy. Therefore, we aim to compare the performance on the test set between a LLaMA3 model that hasn't been fine-tuned on the noisy Italian data and one that has been fine-tuned, to determine whether fine-tuning actually improves results or if the model performs on par or better without it.

The experimental results show that the model without fine-tuning achieves an average accuracy of only about 45%. Fine-tuning on the English dataset yields about 90% mean accuracy, while fine-tuning on the Italian dataset results in a percentage quite similar to the fine-tuned English model and much greater than testing without fine-tuning[1].

The remainder of the paper is organized as follows: Section 2 discusses related work, Section 3 presents FEVER-IT, Section 4 details the experimental measures, and Section 5 provides the conclusions.

[1]The resource, fine-tuned models, and code will be released on a dedicated repository: https://github.com/crux82/FEVER-it

2. Related Work

One of the pioneering works in autonomous fact-checking was conducted by [21], which proposed creating publicly available datasets and developing automated systems using natural language processing technologies. Recent challenges such as CheckThat! at CLEF [10, 11, 12] and Fever [7, 8, 9] from 2018 have advanced fact-checking tasks by leveraging advanced approaches and integrating Large Language Models (LLMs) like BERT and GPT. These models represent the current state of the art in many Natural Language Processing tasks, including fact-checking. Notable examples of such technology include FacTeR-Check [22], a multilingual architecture for semi-automated fact-checking and hoax propagation analysis using the XLM-RoBERTa Transformer [13], and FACT-GPT [23], a framework that automates the claim-matching phase of fact-checking using LLMs to identify social media content that supports or contradicts claims previously debunked by fact-checkers.

The success of these systems is largely due to the capabilities of LLMs as summarized in [3], which are neural models based on the Transformer architecture. Specifically, decoder-based architectures, such as GPT [24], GPT-3 [25], and LLaMA [14], generate output sequences in an auto-regressive manner. These models have demonstrated impressive capabilities following pre-training on large collections of documents. One notable outcome is few-shot learning, where models can adapt to new tasks with only a few examples [25], greatly enhancing their flexibility and applicability.

When new annotated data is available, fine-tuning further enhances a model's capabilities. This process involves taking the pre-trained base model and training it on a smaller, specialized dataset relevant to the desired task. Parameter Efficient Fine-Tuning (PEFT) is an optimized technique that involves training only a small portion of the weights, typically by adding a new layer to the model. One widely used technique is LoRA [26], which adds an adapter consisting of two matrices of weights that are relatively small compared to the original model. Extremita [27] is an example of a decoder-based model fine-tuned with LoRA in Italian for multi-task executions.

Several benchmark datasets have been developed to fine-tune and evaluate fact-checking systems, typically collected by organizations like Snopes, FullFact, and PolitiFact. The FEVER challenge has produced four major datasets: FEVER (2018) [6], FEVER 2.0 (2019) [8], FEVEROUS (2021) [9], and AVeriTeC (2024) [28]. These datasets range from labeled claim-evidence associations to verified claims with structured and unstructured evidence. Despite the wealth of resources available, there is a lack of large benchmark datasets in Italian. This work addresses this gap by providing a large-scale Italian resource.

3. Fact Verification in Italian

As in [6], the original FEVER dataset is composed of claims that can potentially be verified against an encyclopedic resource, in this case, Wikipedia. The claims are classified into three categories: SUPPORTED, REFUTES and NOTENOUGHINFO. For the first two categories, each claim is associated with one or more passages from Wikipedia, each specifying the page from which it was extracted. For the NOTENOUGHINFO category, no passages are provided because no information was found on Wikipedia to support or refute the claim. For instance, the sentence "*Dan Brown is illiterate.*" is a claim associated with pieces of evidence such as: "*Angels and Demons is a 2000 best-selling mystery-thriller novel written by American author Dan Brown and published by Pocket Books and then by Corgi Books.*". These pieces of evidence prove that the claim is incorrect, so it can be classified with the label REFUTES. In FEVER, a claim is thus a sentence that expresses information (true or mutated) about a target entity.

To generate the Italian dataset, we started from the dataset version[2] proposed in [29], which consists of 260k claims. This version extends the original FEVER by adding evidence associated with claims justified as NOTENOUGHINFO in FEVER, using the heuristics in [18]. The approach involved using a search engine to retrieve potential evidence and a textual entailment system based on GPT [24]. Claims not judged as SUPPORTS or REFUTES were classified as NOTENOUGHINFO.

This gives us examples of sentences that are closely related to the claim (according to the search engine) but neither support nor refute it. This makes it more straightforward and efficient to train and/or evaluate a classifier, even though some of the derived examples might be somewhat noisy, as they were generated through heuristics.

For the automatic translation process, we utilized MADLAD400 [20], a machine translation system based on the Transformer architecture[3], trained on MADLAD, a manually audited, general domain 3T token multilingual dataset based on CommonCrawl, spanning 419 languages. Since the Italian data are obtained through machine translation, and thus potentially incorrect as suggested in [16, 17], we needed validated test data to obtain a realistic benchmark. Our hypothesis is that an LLM is robust enough to generalize from the 228k examples and recognize the relationships involved in FEVER without inheriting translation errors. However, to prevent these errors from being inherited by the model, we manually corrected the translations of the test set.

Out of the approximately 16k available test examples, three annotators were involved in verifying and correcting 2,063 translations from the test set. The annotators focused on correcting mistakes related to the proper sentence structure in Italian, the accurate meaning of specific English words that MADLAD had translated literally, any misunderstandings of the intended meaning in Italian, and a few grammatical errors.

In some cases, translation errors do not completely undermine the examples with respect to the task's purpose. For instance, the English sentence from an evidence, "*he was booked to win a third world championship at a WWE event on the night of his death*" was translated into Italian as "*era stato prenotato per vincere un terzo titolo mondiale in un evento della WWE la notte della sua morte*". A more accurate translation would be "*si pensava avrebbe vinto un terzo titolo mondiale in un evento della WWE la notte della sua morte*", better capturing the verb's meaning. In other, more problematic cases, translation errors, loss of information, or introduction of hallucinations could even change the classification in the fact verification task. For example, in the claim "*The Thin Red Line (1998 film) has an all-British cast.*", the automatic translation was "*La sottile linea rossa (The Thin Red Line) è un film del 1998.*", which is incorrect because it omits the information about the cast. This detail is crucial, as its absence could lead to incorrect labeling.

Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Claim	0,9776	0,9695	0,9623	0,9544
Evidence	0,9529	0,9411	0,9309	0,9207

Table 1
BLEU score metrics of Claim and Evidence manually validated (gold) respect automatic translation version (silver)

	Train (S)	Dev (S)	Test (G)	Total
SUPPORTS	114,801	4,638	654	120,095
REFUTES	47,096	4,887	643	52,626
NEI	66,380	6,410	766	73,556
Total	**228,277**	**15,935**	**2,063**	**246,275**

Table 2
Number of claims and evidence in the Italian dataset. (S) indicates silver data (automatically translated), and (G) indicates gold data (manually validated).

A quantitative analysis of the translation quality suggests that MADLAD performs well in translating simple assertive sentences such as claims. In fact, 91% of the claims were not altered by the validators, who considered them completely correct. This percentage is lower for the Wikipedia passages, dropping to 76%. This discrepancy may be due to the greater complexity of the evidence compared to the simpler sentence structures in the claims. Additionally, we reported the results in terms of BLEU score [30] for the corrected translations compared to the originals, as shown in Table 1. It should be noted that measuring the translation quality after correcting the

[2] https://huggingface.co/datasets/copenlu/fever_gold_evidence
[3] https://github.com/google-research/google-research/tree/master/madlad_400

sentences introduces a strong bias in the measurements; however, it provides a more specific idea of the translation quality, especially in understanding the potential noisiness of the training and development sentences. In this case, results of over 95% for BLEU-1 and over 92% for BLEU-4 suggest that very few terms were altered during validation, and even the grammatical patterns remained largely unchanged. At most, a few mistranslated terms needed updating, as indicated by the qualitative analysis.

Table 2 summarizes the number of examples created for the Italian dataset. In line with the original English material, the dataset is divided into training, development, and test sets, with claims categorized into SUPPORTS, REFUTES, and NOTENOUGHINFO (NEI). The table also distinguishes between silver data (automatically translated) and gold data (manually validated). The training set consists of 228,277 claims, the development set contains 15,935 claims, and the test set has 2,063 claims. Each Italian claim or evidence is aligned with the English counterpart, facilitating future research in cross-lingual fact verification.

Language Models for Fact Verification. For addressing the capabilities of Large Language Models in Fact Verification, they can be utilized through In-Context Learning techniques [31] or by directly fine-tuning the model for specific downstream tasks. In-context learning relies on the model's pre-existing knowledge acquired during pre-training and on instructions provided in natural language at inference time. This method does not involve additional training and can be categorized based on the number of examples provided: *i) 0-shot Learning*, where no examples are given, and the model generates responses based solely on its pre-existing knowledge and the provided instructions; *ii) 1-shot Learning*, where one example per class is added to provide a more precise context, helping the model better understand the task by offering a concrete reference point; *iii) Few-shot Learning*, where more than one example per class is provided to give the model additional contextual information during decision-making. When the model's pre-existing knowledge is insufficient, we can fine-tune it on the downstream task. Fine-tuning involves training the model in a traditional manner using input-output pairs (training data) to adjust its parameters. This process improves the model's performance on specific tasks, allowing it to learn from a more extensive set of examples. As a result, the model becomes more adept at handling similar queries in the future, with a focus on the specific task at hand. We thus evaluated the application of state-of-the-art LLM, namely LLAMA3 [32], by providing just the definition of the task (zero-shot) or adding an example (one-shot) or by performing fine-tuning, to demonstrate the necessity of a training dataset like the one constructed in this work, as discussed in the following section.

4. Experimental Evaluation

The goal of our experimentation is to assess the performance of a state-of-the-art LLM applied to Fact Verification. Specifically, we aim to determine whether a multilingual model maintains consistent quality when applied to both the English FEVER dataset and our Italian dataset. We utilize LLaMA3-Instruct[4], an instruction-tuned generative text model from META with 8 billion parameters, released in April 2024. This model is trained to execute specific instructions or prompts across various tasks. To ensure alignment, we evaluate the systems on the manually validated Italian test set and the same subset of 2,063 claims in the English counterpart. The model is evaluated in 0-shot and 1-shot settings to assess its capability without fine-tuning. The prompts used in English and Italian are provided in Appendix A. Additionally, we fine-tuned LLaMA3 on the English datasets from [29] and separately on the Italian datasets obtained via machine translation. Fine-tuning was conducted on an NVIDIA A100 using the LoRA technique[5].

In FEVER, the title of the document associated with each claim often provides crucial context. For example, the claim "*The University of Leicester discovered and identified the remains of a king.*" relies on the document titled "*University of Leicester*" to correctly classify the claim as SUPPORTS. To ensure the model's generalization, we will evaluate the impact of including document titles in prompts. The metrics used to analyze the results are recall, precision, accuracy, and F1 score, calculated globally and for each label (SUPPORTS, REFUTES, NOTENOUGHINFO).

The results are reported in Tables 3 and 4 for the English and Italian datasets, respectively. Each table shows whether the model underwent fine-tuning (column FT), whether a prompt without examples (0-shot) or with one example per class (1-shot) was used (column Prompt), and whether the document title was included (column Doc). Notably, if no fine-tuning was performed, the original LLaMA3-Instruct model was used. Given that the system's response can consist of multiple words, we search the output for the mention of one of the classes and associate the example with that class. If no class is identified, the result is classified as NOTENOUGHINFO. In general, the fine-tuned model is extremely stable, consistently outputting one of the three categories for every request. The non-fine-tuned model, on rare occasions—just a few dozen times out of 2000—produces responses that do not correspond to any of the required classes. This highlights the inherent stability of LLaMA3 while also supporting

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] The following hyperparameters were used: a learning rate of 0.0001, two epochs, LoRA_R set to 8, LoRA_alpha set to 16, and LoRA_dropout at 0.05. The micro-batch size was 2, and gradient accumulation steps were set to 8.

FT	Prompt	Doc	Acc	Support			Refutes			Not enough info			Macro Average		
				P	R	F1	P	R	F1	P	R	F1	P	R	F1
No	0-shot	No	0.449	0.784	0.161	0.267	0.647	0.236	0.346	0.395	0.873	0.544	0.609	0.423	0.386
		Yes	0.374	0.343	**0.976**	0.507	0.763	0.160	0.265	0.477	0.041	0.075	0.528	0.392	0.282
	1-shot	No	0.591	0.555	0.864	0.675	0.699	0.415	0.521	0.586	0.507	0.543	0.613	0.595	0.580
		Yes	0.383	0.929	0.020	0.039	0.867	0.020	0.040	0.376	**0.999**	0.546	0.724	0.346	0.208
Yes	0-shot	No	0.917	0.932	0.947	0.939	0.924	0.888	0.906	0.899	0.916	0.908	0.918	0.917	0.918
		Yes	**0.922**	**0.938**	0.953	**0.945**	**0.929**	**0.896**	**0.912**	0.902	0.918	0.910	**0.923**	**0.922**	**0.923**
	1-shot	No	0.914	0.928	0.948	0.938	0.927	0.883	0.905	0.893	0.911	0.902	0.916	0.914	0.915
		Yes	0.921	0.931	0.956	0.943	0.927	0.891	0.909	**0.907**	0.916	**0.912**	0.922	0.921	0.921

Table 3
Performance in terms of Accuracy, Precision, Recall and F1-measure of our systems on Fever-EN dataset

FT	Prompt	Doc	Acc	Support			Refutes			Not enough info			Macro Average		
				P	R	F1	P	R	F1	P	R	F1	P	R	F1
No	0-shot	No	0.462	0.411	0.951	0.574	0.607	0.457	0.522	0.585	0.050	0.092	0.534	0.486	0.396
		Yes	0.507	0.463	0.942	0.620	0.587	0.663	0.622	0.800	0.005	0.010	0.617	0.537	0.418
	1-shot	No	0.425	0.376	0.963	0.541	0.671	0.333	0.445	0.478	0.043	0.079	0.508	0.446	0.355
		Yes	0.462	0.403	**0.968**	0.569	0.632	0.361	0.459	0.698	0.115	0.197	0.578	0.481	0.409
Yes	0-shot	No	0.897	0.897	0.940	0.918	0.924	0.845	0.882	0.877	0.903	0.890	0.899	0.896	0.897
		Yes	0.901	0.899	0.936	0.917	0.923	**0.855**	0.888	**0.887**	0.910	0.898	0.903	0.900	0.901
	1-shot	No	0.895	0.891	0.947	0.918	0.919	0.843	0.879	0.881	0.894	0.887	0.897	0.895	0.895
		Yes	**0.905**	**0.913**	0.942	**0.927**	**0.924**	0.854	**0.888**	0.883	**0.915**	**0.899**	**0.907**	**0.904**	**0.905**

Table 4
Performance in terms of Accuracy, Precision, Recall and F1-measure of our systems on Fever-IT dataset

the soundness of the results achieved.

A key finding is that the multilingual model generally achieves similar, though modest, results on English and Italian datasets without fine-tuning, with accuracy values around 0.40-0.50 and average F1 scores in the range of 0.35-0.55. This performance is relatively unstable, and the addition of an example in the prompt does not lead to significant improvements. In English, there are some improvements, but in Italian, there are fewer. We believe this is because, although LLaMA is multilingual, the percentage of Italian examples observed during training is less than 1%, making it less performant and less stable in this language.

However, when fine-tuning is applied, the results improve dramatically, with accuracy exceeding 90% in both languages. This demonstrates the utility of the translated dataset, even if it contains some noise. In this scenario, adding an example in the prompt leads to negligible but consistent improvements. Additionally, the inclusion of the document title, while sometimes causing inconsistencies in zero-shot learning, is better utilized by the fine-tuned model, leading to slight but not significant improvements. This is interesting because it suggests that the model not relying on document titles is more broadly applicable. Overall, the fine-tuned models perform significantly better, highlighting the importance of the translated dataset for achieving high accuracy in fact verification tasks in both English and Italian.

The error analysis suggests that the model sometimes inherits the mathematical reasoning limitations of the LLM. For example, the claim "*Il Castello di Praga attira oltre 18 milioni di visitatori ogni anno.*[6]" was given the evidence "*Il castello è tra le attrazioni turistiche più visitate di Praga che attira oltre 1,8 milioni di visitatori all'anno.*[7]" The model's predicted label was Refutes, while the true label was Supports. Here, the true label should be Supports since 18 million is indeed greater than 1.8 million, but the model found the numbers inconsistent. In another case, the claim "*Ned Stark è stato introdotto nel 1996 in Tempesta di spade.*[8]" was paired with the evidence "*Introdotto nel 1996 in Il Trono di Spade, Ned è l'onorevole signore di Winterfell, un'antica fortezza nel nord del continente immaginario di Westeros.*[9]" The model predicted Refutes, although the true label was Supports. The confusion here is due to the difference in the book titles, which are from the same series but are distinct works. The error analysis revealed that the model occasionally struggled with mathematical reasoning and contextual understanding, highlighting areas for future enhancement. Larger models and further fine-tuning could potentially address these issues, which remain open questions for future research.

[6] In English: "*The Prague Castle attracts over 18 million visitors every year.*"
[7] In English: "*The castle is among the most visited tourist attractions in Prague, attracting over 1.8 million visitors every year.*"
[8] In English: "*Ned Stark was introduced in 1996 in A Storm of Swords.*"
[9] In English: "*Introduced in 1996 in A Game of Thrones, Ned is the honorable lord of Winterfell, an ancient fortress in the north of the imaginary continent of Westeros.*"

5. Conclusion

In this work, we have introduced FEVER-IT, an Italian version of the FEVER dataset, designed to improve the training and evaluation of models for fact verification in the Italian language. Using a machine translation system, we translated a large-scale dataset of 228,000 claims/-pieces of evidence pairs and manually validated 2,000 test instances to ensure meaningful evaluations. This enabled us to fine-tune a state-of-the-art LLM, specifically LLaMA3, and assess its performance in both English and Italian.

Our experiments demonstrated that the multilingual model, without fine-tuning, performed similarly on both English and Italian datasets, though the accuracy and stability were limited. Fine-tuning significantly improved the model's performance, achieving over 90% accuracy in both languages. This underscores the importance and effectiveness of the translated dataset, even if it contains some noise.

Future work will explore the performance of larger models and further refinement of the dataset to enhance accuracy and generalization capabilities or explore more complex settings such as those described in [9].

Acknowledgments

The team would like to thank Monika Kakol for her invaluable support in the validation of the translations. This work was supported by Project ECS 0000024 Rome Technopole, - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

References

[1] Z. Guo, M. S. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Trans. Assoc. Comput. Linguistics 10 (2022) 178–206.

[2] A. D. Terry Flew, Christina Spurgeon, A. Swift, The promise of computational journalism, Journalism Practice 6 (2012) 157–171.

[3] C. Chen, K. Shu, Combating misinformation in the age of llms: Opportunities and challenges, 2023. URL: https://arxiv.org/abs/2311.05656. arXiv:2311.05656.

[4] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, A. Vlachos, Multimodal automated fact-checking: A survey, 2023. URL: https://arxiv.org/abs/2305.13507. arXiv:2305.13507.

[5] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3346–3359. URL: https://aclanthology.org/C18-1283.

[6] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[7] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501. doi:10.18653/v1/W18-5501.

[8] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The FEVER2.0 shared task, in: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–6. URL: https://aclanthology.org/D19-6601. doi:10.18653/v1/D19-6601.

[9] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Dominican Republic, 2021, pp. 1–13. URL: https://aclanthology.org/2021.fever-1.1. doi:10.18653/v1/2021.fever-1.1.

[10] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Proceedings of the 43rd European Conference on Information Retrieval, ECIR '21, Lucca, Italy, 2021, pp. 639–649. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1_75.

[11] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting

the covid-19 infodemic and fake news detection, in: Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[12] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[15] B. Berendt, P. Burger, R. Hautekiet, J. Jagers, A. Pleijter, P. Van Aelst, Factrank: Developing automated claim detection for dutch-language fact-checkers, Online Social Networks and Media 22 (2021) 100113. doi:https://doi.org/10.1016/j.osnem.2020.100113.

[16] D. Croce, A. Zelenanska, R. Basili, Enabling deep learning for large scale question answering in italian, Intelligenza Artificiale 13 (2019) 49–61. URL: https://doi.org/10.3233/IA-190018. doi:10.3233/IA-190018.

[17] A. Scaiella, D. Croce, R. Basili, Large scale datasets for image and video captioning in italian, Italian Journal of Computational Linguistics 2 (2019) 49–60. URL: http://www.ai-lc.it/IJCoL/v5n2/IJCOL_5_2_3___scaiella_et_al.pdf.

[18] C. Malon, Team papelo: Transformer networks at FEVER, in: J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal (Eds.), Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 109–113. URL: https://aclanthology.org/W18-5517. doi:10.18653/v1/W18-5517.

[19] L. Canale, A. Messina, Experimenting ai technologies for disinformation combat: the idmo project, 2023. URL: https://arxiv.org/abs/2310.11097. arXiv:2310.11097.

[20] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, O. Firat, Madlad-400: A multilingual and document-level large audited dataset, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 67284–67296.

[21] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, N. A. Smith (Eds.), Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508. doi:10.3115/v1/W14-2508.

[22] A. Martín, J. Huertas-Tato, Álvaro Huertas-García, G. Villar-Rodríguez, D. Camacho, Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference, Knowledge-Based Systems 251 (2022) 109265. doi:https://doi.org/10.1016/j.knosys.2022.109265.

[23] E. C. Choi, E. Ferrara, Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation, in: Companion Proceedings of the ACM on Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1441–1449. URL: https://doi.org/10.1145/3589335.3651910. doi:10.1145/3589335.3651910.

[24] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, CoRR abs/1801.06146 (2018). URL: http://arxiv.org/abs/1801.06146. arXiv:1801.06146.

[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December, 2020, pp. 6–12.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[27] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR*

Workshop Proceedings, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper13.pdf.

[28] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 65128–65167.

[29] P. Atanasova, D. Wright, I. Augenstein, Generating label cohesive and well-formed adversarial claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3168–3177. URL: https://aclanthology.org/2020.emnlp-main.256. doi:10.18653/v1/2020.emnlp-main.256.

[30] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[31] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, 2024. URL: https://arxiv.org/abs/2301.00234. arXiv:2301.00234.

[32] AI@Meta, Llama 3 model card, 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

A. Prompting Engineering

This appendix contains the prompts used in the experiments. The prompts are provided in both Italian and English, reflecting the task-specific nature of the experiments. Each prompt begins with an explanation of the task and the meaning of the classes. In the different variants, the 0-shot setting does not include any examples, unlike the 1-shot setting. Where necessary, the name of the document from which the evidence is taken is also specified.

A.1. Prompts in English

A.1.1. 0-shot Setting

The following prompt is used for 0-shot learning, where the task and classes are presented without additional information.

```
### Instruction
Evaluate if the claim is supported by the
    evidence provided. Definitions for key
    terms used in this task are:
- Claim: A statement or assertion under
    examination.
- Evidence: Information that either supports
    or opposes the claim.

Answer with one of the following judgments
    based on the evidence provided:
- SUPPORTS: if the evidence substantiates the
     claim.
- REFUTES: if the evidence directly
    contradicts the claim.
- NOT ENOUGH INFO: if there is insufficient
    evidence to determine the claim's
    validity
### Input
- Claim: [CLAIM HERE]
- Evidence: [EVIDENCE HERE]
### Answer: [ANSWER HERE]
```

A.1.2. 1-shot Setting

The following prompt is used for 1-shot learning, where the task and classes are explained, and one example per class is provided. Notice that only the evidence is reported without the title of the original document.

```
### Instruction
Evaluate if the claim is supported by the
    evidence provided. Definitions for key
    terms used in this task are:
- Claim: A statement or assertion under
    examination.
- Evidence: Information that either supports
    or opposes the claim.

Answer with one of the following judgments
    based on the evidence provided:
- SUPPORTS: if the evidence substantiates the
     claim.
- REFUTES: if the evidence directly
    contradicts the claim.
- NOT ENOUGH INFO: if there is insufficient
    evidence to determine the claim's
    validity

### Examples
These examples demonstrate how to apply the
    evaluation criteria:
- Claim: The Germanic peoples are also called
     Gothic.
- Evidence: The Germanic peoples (also
    referred to as Teutonic, Suebian, or
    Gothic in older literature) are an Indo-
    European ethno-linguistic group of
    Northern European origin.
- Answer: SUPPORTS

- Claim: Tennis is not a sport.
- Evidence: Tennis is played by millions of
    recreational players and is also a
    popular worldwide spectator sport.
```

```
- Answer: REFUTES

- Claim: Kick-Ass is a horror film.
- Evidence: Kick-Ass is a 2010 British-
    American film based on the comic book of
     the same name by Mark Millar and John
    Romita, Jr.
- Answer: NOT ENOUGH INFO
### Input
- Claim: [CLAIM HERE]
- Evidence: [EVIDENCE HERE]
### Answer: [ANSWER HERE]
```

A.1.3. 0-shot Setting with Document Title

The following prompt is used for 0-shot learning, where the task and classes are explained without additional information. Each input evidence is provided with the title of its original document.

```
### Instruction
Evaluate if the claim is supported by the
    evidence provided. Definitions for key
    terms used in this task are:
- Claim: A statement or assertion under
    examination.
- Evidence: Information that either supports
    or opposes the claim.
- Document: denotes the source document for
    the evidence.

Answer with one of the following judgments
    based on the evidence provided:
- SUPPORTS: if the evidence substantiates the
     claim.
- REFUTES: if the evidence directly
    contradicts the claim.
- NOT ENOUGH INFO: if there is insufficient
    evidence to determine the claim's
    validity
### Input
- Claim: [CLAIM HERE]
- Evidence: [EVIDENCE HERE]
- Document: [DOCUMENT HERE]
### Answer: [ANSWER HERE]
```

A.1.4. 1-shot Setting with Document Title

The following prompt is used for 1-shot learning, where the task and classes are explained, and one example per class is provided. Each input evidence is provided with the title of its original document.

```
### Instruction
Evaluate if the claim is supported by the
    evidence provided. Definitions for key
    terms used in this task are:
- Claim: A statement or assertion under
    examination.
- Evidence: Information that either supports
    or opposes the claim.
- Document: denotes the source document for
    the evidence.

Answer with one of the following judgments
    based on the evidence provided:
- SUPPORTS: if the evidence substantiates the
     claim.
- REFUTES: if the evidence directly
    contradicts the claim.
- NOT ENOUGH INFO: if there is insufficient
    evidence to determine the claim's
    validity

### Examples
These examples demonstrate how to apply the
    evaluation criteria:
- Claim: The Germanic peoples are also called
     Gothic.
- Evidence: The Germanic peoples (also
    referred to as Teutonic, Suebian, or
    Gothic in older literature) are an Indo-
    European ethno-linguistic group of
    Northern European origin.
- Document: Germanic peoples
- Answer: SUPPORTS

- Claim: Tennis is not a sport.
- Evidence: Tennis is played by millions of
    recreational players and is also a
    popular worldwide spectator sport.
- Document: Tennis
- Answer: REFUTES

- Claim: Kick-Ass is a horror film.
- Evidence: Kick-Ass is a 2010 British-
    American film based on the comic book of
     the same name by Mark Millar and John
    Romita, Jr.
- Document: Kick-Ass (film)
- Answer: NOT ENOUGH INFO
### Input
- Claim: [CLAIM HERE]
- Evidence: [EVIDENCE HERE]
- Document: [DOCUMENT HERE]
### Answer: [ANSWER HERE]
```

A.2. Prompts in Italian

A.2.1. 0-shot Setting

The following prompt is used for 0-shot learning, where the task and classes are presented without additional information.

```
### Istruzioni
Valuta se l'affermazione è supportata dalle
    prove fornite. Le definizioni dei
    termini chiave utilizzati in questo
    compito sono:
- Affermazione: Una dichiarazione o
    asserzione sotto esame.
- Prova: Informazioni che supportano o
    contraddicono l'affermazione.
```

```
Rispondi con uno dei seguenti giudizi basati
    sulle prove fornite:
- SUPPORTS: se le prove confermano l'
    affermazione.
- REFUTES: se le prove contraddicono
    direttamente l'affermazione.
- NOT ENOUGH INFO: se le prove non sono
    sufficienti per determinare la validità
    dell'affermazione.
### Input
- Affermazione: [CLAIM HERE]
- Prova: [EVIDENCE HERE]
### Risposta: [ANSWER HERE]
```

A.2.2. 1-shot Setting

The following prompt is used for 1-shot learning, where the task and classes are explained, and one example per class is provided. Notice that only the evidence is reported without the title of the original document.

```
### Istruzioni
Valuta se l'affermazione è supportata dalle
    prove fornite. Le definizioni dei
    termini chiave utilizzati in questo
    compito sono:
- Affermazione: Una dichiarazione o
    asserzione sotto esame.
- Prova: Informazioni che supportano o
    contraddicono l'affermazione.

Rispondi con uno dei seguenti giudizi basati
    sulle prove fornite:
- SUPPORTS: se le prove confermano l'
    affermazione.
- REFUTES: se le prove contraddicono
    direttamente l'affermazione.
- NOT ENOUGH INFO: se le prove non sono
    sufficienti per determinare la validità
    dell'affermazione.

### Esempi
Questi esempi dimostrano come applicare i
    criteri di valutazione:
- Affermazione: I popoli germanici sono
    chiamati anche gotici.
- Prova: I popoli germanici (anche chiamati
    Teutoni, Suebi o Goti nella letteratura
    più antica) sono un gruppo etno-
    linguistico indoeuropeo di origine nord
    europea.
- Risposta: SUPPORTS

- Affermazione: Il tennis non è uno sport.
- Prova: Il tennis è praticato da milioni di
    giocatori amatoriali ed è anche uno
    sport popolare a livello mondiale.
- Risposta: REFUTES

- Affermazione: Kick-Ass è un film horror.
- Prova: Kick-Ass è un film britannico-
    americano del 2010 basato sul fumetto
    omonimo di Mark Millar e John Romita Jr.
- Risposta: NOT ENOUGH INFO

### Input
- Affermazione: [CLAIM HERE]
- Prova: [EVIDENCE HERE]
### Risposta: [ANSWER HERE]
```

A.2.3. 0-shot Setting with Document Title

The following prompt is used for 0-shot learning, where the task and classes are explained without additional information. Each input evidence is provided with the title of its original document.

```
### Istruzioni
Valuta se l'affermazione è supportata dalle
    prove fornite. Le definizioni dei
    termini chiave utilizzati in questo
    compito sono:
- Affermazione: Una dichiarazione o
    asserzione sotto esame.
- Prova: Informazioni che supportano o
    contraddicono l'affermazione.
- Documento: indica la fonte da cui è stata
    estratta la prova.

Rispondi con uno dei seguenti giudizi basati
    sulle prove fornite:
- SUPPORTS: se le prove confermano l'
    affermazione.
- REFUTES: se le prove contraddicono
    direttamente l'affermazione.
- NOT ENOUGH INFO: se le prove non sono
    sufficienti per determinare la validità
    dell'affermazione.
### Input
- Affermazione: [CLAIM HERE]
- Prova: [EVIDENCE HERE]
- Documento: [DOCUMENT HERE]
### Risposta: [ANSWER HERE]
```

A.2.4. 1-shot Setting with Document Title

The following prompt is used for 1-shot learning, where the task and classes are explained, and one example per class is provided. Each input evidence is provided with the title of its original document.

```
### Istruzioni
Valuta se l'affermazione è supportata dalle
    prove fornite. Le definizioni dei
    termini chiave utilizzati in questo
    compito sono:
- Affermazione: Una dichiarazione o
    asserzione sotto esame.
- Prova: Informazioni che supportano o
    contraddicono l'affermazione.
- Documento: indica la fonte da cui è stata
    estratta la prova.

Rispondi con uno dei seguenti giudizi basati
    sulle prove fornite:
- SUPPORTS: se le prove confermano l'
    affermazione.
```

```
- REFUTES: se le prove contraddicono
    direttamente l'affermazione.
- NOT ENOUGH INFO: se le prove non sono
    sufficienti per determinare la validità
    dell'affermazione.

### Esempi
Questi esempi dimostrano come applicare i
    criteri di valutazione:
- Affermazione: I popoli germanici sono
    chiamati anche gotici.
- Prova: I popoli germanici (anche chiamati
    Teutoni, Suebi o Goti nella letteratura
    più antica) sono un gruppo etno-
    linguistico indoeuropeo di origine nord
    europea.
- Documento: Popoli germanici
- Risposta: SUPPORTS

- Affermazione: Il tennis non è uno sport.
- Prova: Il tennis è praticato da milioni di
    giocatori amatoriali ed è anche uno
    sport popolare a livello mondiale.
- Documento: Tennis
- Risposta: REFUTES

- Affermazione: Kick-Ass è un film horror.
- Prova: Kick-Ass è un film britannico-
    americano del 2010 basato sul fumetto
    omonimo di Mark Millar e John Romita Jr.
- Documento: Kick-Ass (film)
- Risposta: NOT ENOUGH INFO
### Input
- Affermazione: [CLAIM HERE]
- Prova: [EVIDENCE HERE]
- Documento: [DOCUMENT HERE]
### Risposta: [ANSWER HERE]
```