Scale or Specialization? Evaluating Language Models for Italian Fact Verification

Alessandro Pucci[1], Danilo Croce[1,*] and Roberto Basili[1]

[1]University of Rome Tor Vergata, Italy

Abstract

Large Language Models have demonstrated strong capabilities across a broad range of Natural Language Processing tasks. However, the selection of a model for a well-defined application cannot be based exclusively on its general capabilities or reported parameter scale. This issue is especially relevant in document-intensive and regulated settings, where predictive performance must be considered together with computational requirements, deployment constraints, data control, and auditability. In this work, we investigate the relationship among model scale, task specialization, and architecture for Italian fact verification. Using the manually validated FEVER-IT test set, we first compare general-purpose decoder-only models with reported parameter counts ranging from 8 billion to 1.6 trillion under a common few-shot prompting protocol. We then evaluate a Llama-3.1-8B-Instruct model adapted through Low-Rank Adaptation and compare it with two substantially smaller encoder-only classifiers fine-tuned on the same training data. The results show that few-shot accuracy improves substantially from the 8B to the 12B configuration, whereas differences among the larger heterogeneous systems remain limited and non-monotonic. LoRA adaptation increases the accuracy of Llama-3.1-8B-Instruct from 0.7067 to 0.8885, yielding the highest observed result in the study. BGE-M3 and mmBERT-base achieve lower accuracy, but use substantially fewer parameters than the adapted decoder. A class-level analysis further shows that the largest improvement produced by adaptation concerns the recognition of cases in which the available evidence is insufficient. These findings indicate that reported model scale alone is not a sufficient criterion for Italian fact verification. For a stable task with substantial annotated data, task-specific adaptation and architectural choice may provide more effective ways of allocating model capacity than selecting the largest available general-purpose model.

Keywords

Italian Fact Verification, Large Language Models, Parameter-Efficient Fine-Tuning, Model Scaling, Efficient NLP

CEUR-WS.org/Vol-4315/paper21.pdf

1. Introduction

Large Language Models (LLMs) can address a broad range of Natural Language Processing tasks through instructions expressed in natural language. Their flexibility is partly enabled by in-context learning, through which a model can perform a task without updating its parameters, using zero-shot, one-shot, or few-shot prompting [2, 3]. This ability makes general-purpose models attractive when the target task changes frequently, annotated examples are scarce, or rapid experimentation is required. When labeled data are available, however, task-specific training provides a different strategy. Instead of expressing the complete task only through the inference prompt, model parameters can be adapted to the target distribution. Parameter-Efficient Fine-Tuning methods, such as Low-Rank Adaptation (LoRA), reduce the number of trainable parameters and allow specialization without modifying the complete model [4]. Quantized adaptation methods can further reduce the memory required during training, making specialization feasible under more limited hardware conditions [5]. The increasing availability of large general-purpose models raises a practical model-selection question. For a narrow and well-defined task, is reported model scale the main factor determining predictive performance, or can specialization provide a more effective alternative? Larger models may offer broader reasoning, stronger instruction following, and greater flexibility across heterogeneous requests, capabilities that are not fully measured by a single closed-label benchmark. At the same time, depending on their architecture and execution setting, they may require more memory, introduce higher inference latency, rely on external providers,

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]

*Corresponding author.

al.pucci@ing.uniroma2.it (A. Pucci); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)

0009-0003-4999-2629 (A. Pucci); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

or generate less predictable operational costs. This question is relevant in document-intensive and regulated domains. In financial and banking scenarios, language technologies can support document validation, information consistency checking, compliance testing, and the verification of claims extracted from internal repositories [6]. In such settings, model selection is influenced not only by maximum benchmark performance, but also by data control, auditability, infrastructure requirements, and the stability of operational costs. The model with the strongest general-purpose capabilities may therefore not always represent the most appropriate component for a stable and clearly specified workflow.

Fact verification [7] provides a controlled setting in which to study this problem. Given a textual claim and a corresponding evidence passage, a system determines whether the evidence supports the claim, contradicts it, or does not provide sufficient information. Although fact verification has traditionally been investigated in relation to misinformation and computational journalism, the underlying claim–evidence classification problem is also relevant to document validation and information consistency. FEVER [7] introduced a widely adopted English benchmark containing claims and evidence labeled as Supports, Refutes, or NotEnoughInfo. FEVER-IT [8] extended this setting to Italian by translating more than 228,000 training examples and manually validating a test set of 2,063 claim–evidence pairs. The resource makes it possible to evaluate whether results observed on English fact-verification tasks also transfer to Italian, where language coverage in general-purpose pre-training data may be more limited. The original FEVER-IT study primarily investigated the quality and utility of the translated resource. It compared zero-shot, one-shot, and fine-tuned LLaMA3 configurations, showing that supervised adaptation substantially improves performance in Italian. Building on that result, the present work addresses a complementary model-selection question: how do reported scale, parameter-efficient specialization, and architectural choice affect performance on the same task?

We investigate three research questions:

RQ1 How does few-shot fact-verification performance vary across selected general-purpose decoder configurations of different reported scales?

RQ2 Can task-specific LoRA adaptation enable a compact decoder to outperform substantially larger general-purpose few-shot configurations?

RQ3 What performance–parameter trade-off emerges between task-adapted decoder-only and encoder-only models?

The contribution of this work is a comparative analysis of three model-selection strategies. First, we compare general-purpose decoder-only configurations with reported total parameter counts ranging from 8 billion to 1.6 trillion. Second, we compare the use of larger non-adapted models with LoRA specialization of an 8B decoder. Third, we compare the adapted decoder with two substantially smaller encoder-only classifiers trained on the same task. The study does not provide a complete economic or Total Cost of Ownership analysis. Parameter count is used only as an approximate descriptor of model scale and compactness. Direct measurements of inference latency, memory consumption, throughput, energy use, and monetary cost are left to future work. The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the experimental setting, results, and error analysis. Section 4 summarizes the main findings and limitations.

2. Related Work

Automatic Fact Checking commonly involves three main stages: claim detection, evidence retrieval, and claim verification [9, 10]. The present work focuses on the final stage. The evidence passage is already associated with the claim, and the system predicts the semantic relationship between them. Evidence retrieval, ranking, and document selection are therefore outside the scope of the experiments. FEVER [7] established a reference benchmark for evidence-based claim verification. Subsequent work extended this setting in several directions, including adversarial robustness in FEVER 2.0 [11], structured and multi-format evidence in FEVEROUS [12], and real-world web-retrieved evidence in AVeriTeC [13]. The CheckThat! evaluation campaigns have similarly investigated multiple components of automated

fact checking [14, 15, 16]. Broader studies have examined the changing role of LLMs in misinformation detection, both as potential verification tools and as sources of automatically generated misleading information [17]. Despite these developments, large-scale fact-verification resources remain predominantly English. Prior Italian work has addressed disinformation diffusion and the collection of claims for verification [18, 19]. FEVER-IT [8] addresses the resource gap by translating the FEVER training and development data into Italian and providing a manually reviewed test set of 2,063 claim–evidence pairs. Its experiments demonstrate that supervised adaptation produces substantial gains over prompting alone, but do not compare a broad set of general-purpose model configurations or directly investigate the trade-off between adapted decoder-only and encoder-only architectures. Few-shot in-context learning allows an LLM to infer a task from a limited number of demonstrations [2, 3]. Increasing model scale often improves this capability, but downstream gains need not be smooth or monotonic across tasks and evaluation metrics [20]. This issue may be especially relevant in multilingual evaluation, where the representation of each language in the pre-training and post-training data can differ considerably. RQ1 therefore examines the observed relationship between reported scale and Italian fact-verification performance across a heterogeneous selection of contemporary model configurations. Scale is not the only way to allocate model capacity. Parameter-Efficient Fine-Tuning updates a small number of additional parameters instead of the complete model [21]. LoRA [4] represents task-specific updates through low-rank matrices while keeping the original weights frozen. QLoRA [5] combines this mechanism with 4-bit quantization of the frozen base model, reducing the memory required during adaptation. Previous work has applied comparable approaches to Italian decoder models [22]. More generally, parameter-efficient adaptation has been shown to outperform few-shot prompting with substantially larger models on fixed downstream tasks [23]. RQ2 investigates whether this pattern also emerges for Italian fact verification. Encoder-only architectures provide a different alternative. Models derived from BERT [24] produce contextual representations that can be combined with a lightweight classification head. They do not provide the same generative flexibility as decoder-only LLMs, but they remain attractive for tasks with a fixed and closed label set. Their smaller parameter footprint and more predictable inference behaviour can be relevant when deployment cost and latency are limiting factors [25]. Recent comparative studies suggest that fine-tuned encoders can remain competitive with few-shot generative models on structured classification tasks while requiring fewer computational resources [26]. RQ3 examines this trade-off directly on FEVER-IT.

3. Scale, Specialization, and Model Selection

This study investigates model selection for Italian fact verification along three complementary dimensions: the reported scale of general-purpose models, the effect of task-specific adaptation, and the choice between generative and discriminative architectures. These dimensions are not treated as controlled variations of a single system. Instead, they represent alternative strategies for allocating computational capacity to a stable NLP system for a realistic and complex task.

Task and evaluation setting. Given a claim c and an evidence passage e, the model predicts one of three labels: Supports, Refutes, or NotEnoughInfo. The Supports label indicates that the evidence confirms the claim, whereas Refutes indicates a contradiction. NotEnoughInfo is assigned when the evidence does not establish either relation. For example, for the claim $c =$ *"Dante Alighieri was born in Florence"*, the passages $e_1 =$ *"Dante Alighieri was born in Florence"*, $e_2 =$ *"Dante Alighieri was born in Venice"*, and $e_3 =$ *"Dante Alighieri wrote the Divine Comedy"* lead to Supports, Refutes, and NotEnoughInfo, respectively. The experiments use FEVER-IT [8]. Its training and development partitions consist of machine-translated silver data, while the test set contains 2,063 manually validated gold examples. The training partition contains 114,801 Supports, 47,096 Refutes, and 66,380 NotEnoughInfo examples, for a total of 228,277. The development partition contains 4,638 Supports, 4,887 Refutes, and 6,410 NotEnoughInfo examples, for a total of 15,935. The test set contains 654 Supports, 643 Refutes, and 766 NotEnoughInfo examples. The study considers only claim verification: the relevant evidence has already been associated with the claim, while document retrieval and evidence

selection are outside the experimental scope. All configurations are evaluated on the same manually validated test set. We report accuracy and macro-F1 as complementary metrics. Accuracy measures overall classification performance, whereas macro-F1 assigns equal importance to the three labels.

General-purpose few-shot configurations. We select six decoder-only language models spanning a broad range of reported parameter scales and architectural configurations: Llama-3.1-8B-Instruct, Gemma-4-12B-IT, Gemma-4-31B-IT, GPT-OSS-120B, GLM-5.2, and DeepSeek-V4-Pro. Each model receives the same Italian instruction used by Scaiella et al. [8], followed by the claim and its evidence. The few-shot prompt also reuses their three demonstrations per label. Llama-3.1-8B-Instruct and the Gemma models were evaluated locally through Transformers, with their Hugging Face checkpoints loaded in 4-bit NF4 using bitsandbytes. GPT-OSS-120B, GLM-5.2, and DeepSeek-V4-Pro were accessed through Ollama-packaged variants. The evaluation retained the decoding settings exposed by the corresponding runtimes. Consequently, the comparison concerns complete deployed configurations rather than isolating the causal effect of parameter scale. The prompt, demonstrations, label space, and test set were kept constant across models. Experiments were conducted in June 2026. Table 1 reports the results.

Model	Total parameters	Active parameters	Accuracy	Macro-F1
Llama-3.1-8B-Instruct	8B	8B	0.7067	0.7026
Gemma-4-12B-IT	12B	12B	0.8076	0.8078
Gemma-4-31B-IT	31B	31B	0.8236	0.8255
GPT-OSS-120B	117B	5.1B	**0.8434**	**0.8447**
GLM-5.2	753B	40B	0.8425	0.8444
DeepSeek-V4-Pro	1.6T	49B	0.8371	0.8385

Table 1
Few-shot accuracy and macro-F1 on FEVER-IT. Reported total parameter counts and, for Mixture-of-Experts models, active parameters per token are taken from the corresponding model documentation.

The largest improvement is observed between the 8B and 12B configurations, with accuracy increasing by more than ten percentage points. Performance then improves more gradually up to GPT-OSS-120B, which obtains the numerically highest few-shot accuracy. Differences among GPT-OSS-120B, GLM-5.2, and DeepSeek-V4-Pro are small, despite the substantial differences in their reported total parameter counts. The configurations with the largest reported total parameter counts therefore do not achieve further improvements in this evaluation. This result should not be interpreted as a general limitation of larger models. The evaluated systems differ in architecture, training data, post-training procedures, active parameters, quantization, runtime, and inference configuration. Within this heterogeneous set, however, reported parameter count is not associated with a monotonic increase in performance on the considered Italian classification task. Reported total parameter count is also only an approximate descriptor of scale. Dense and Mixture-of-Experts models activate different proportions of their parameters during inference, and similar active parameter counts do not necessarily imply similar latency, memory usage, or serving cost. Table 1 therefore supports a comparison of model configurations, not a direct comparison of computational or monetary efficiency. A supplementary evaluation using the same prompting protocol on the original English FEVER data produces a similar descriptive pattern: the three largest configurations achieve closely comparable results, with accuracy scores ranging from 0.857 to 0.865. Although this is not intended as a controlled cross-lingual analysis, it suggests that the limited differences observed among the larger configurations are not exclusively associated with Italian.

Task-specific adaptation. To compare supervised adaptation with larger non-adapted models, we select Llama-3.1-8B-Instruct as the smallest evaluated decoder, retaining continuity with the LLaMA-based FEVER-IT study. It is adapted on the FEVER-IT training partition using a QLoRA-style setup on a single NVIDIA A100. The base model is loaded in 4-bit NF4 precision, while LoRA is configured with rank $r = 8$, scaling factor $\alpha = 16$, dropout rate 0.05, and learning rate 1×10^{-4}. Training is performed

for five epochs with an effective batch size of 32. The best checkpoint is selected on the development partition and evaluated once on the manually validated test set. Table 2 reports the comparison with the few-shot configuration of the same base model.

Model configuration	Accuracy	Macro-F1
Llama-3.1-8B-Instruct few-shot	0.7067	0.7026
Llama-3.1-8B-Instruct + LoRA	**0.8885**	**0.8881**

Table 2
Effect of task-specific LoRA adaptation on Llama-3.1-8B-Instruct.

LoRA adaptation improves accuracy by approximately 0.18 and macro-F1 by approximately 0.19. The adapted model achieves a higher observed accuracy than all the larger general-purpose few-shot configurations reported in Table 1. This comparison does not demonstrate that the 8B architecture is intrinsically more capable than the larger models. The adapted configuration has access to more than 228,000 labeled training examples from the silver training set, whereas the general-purpose systems are all evaluated through in-context learning without parameter updates. The result instead compares two different model-selection strategies: increasing general-purpose scale or using the available task supervision to specialize a smaller model. Under the present conditions, specialization produces the stronger observed result. Our 0.8885 accuracy approaches the 0.905 fine-tuned FEVER-IT reference [8]. The remaining gap may reflect several factors, including base-model quantization and differences in adaptation and training configurations; nevertheless, comparable performance with a 4-bit base model remains relevant for memory-efficient adaptation.

Class-level error analysis. Aggregate accuracy and macro-F1 do not reveal whether the adaptation gain is distributed uniformly across labels. Table 3 reports row-normalized confusion matrices for the Llama-3.1-8B-Instruct few-shot baseline, the corresponding LoRA-adapted model, and DeepSeek-V4-Pro, the configuration with the largest reported total parameter count.

Llama few-shot				**Llama + LoRA**				**DeepSeek few-shot**			
Gold \ Pred.	S	R	NEI	Gold \ Pred.	S	R	NEI	Gold \ Pred.	S	R	NEI
S	0.91	0.05	0.04	S	0.93	0.02	0.05	S	0.88	0.03	0.10
R	0.16	0.71	0.12	R	0.06	0.82	0.12	R	0.02	0.74	0.24
NEI	0.26	0.21	0.53	NEI	0.04	0.04	0.91	NEI	0.04	0.08	0.88

Table 3
Row-normalized confusion matrices. Rows indicate gold labels and columns indicate predicted labels.

The Llama-3.1-8B-Instruct few-shot configuration performs relatively well on Supports, correctly identifying 91% of the corresponding examples. Its main weakness concerns NotEnoughInfo: only 53% of these cases are correctly classified, while 26% are predicted as Supports and 21% as Refutes. The model therefore frequently assigns a definite verdict when the evidence does not establish either support or contradiction. LoRA adaptation improves performance across all three classes. Recall rises from 0.91 to 0.93 for Supports, from 0.71 to 0.82 for Refutes, and from 0.53 to 0.91 for NotEnoughInfo. The largest change concerns the ability to recognize insufficient evidence, indicating that task supervision substantially reduces over-commitment in the base few-shot configuration. DeepSeek-V4-Pro presents a different error distribution. It correctly identifies 88% of Supports examples, 74% of Refutes, and 88% of NotEnoughInfo. Its overall performance is therefore considerably stronger than the Llama few-shot baseline, but the matrix still reveals class-specific weaknesses that are not visible from accuracy alone. In particular, 24% of Refutes instances are assigned to NotEnoughInfo, while 10% of Supports cases are misclassified as NotEnoughInfo.

Encoder-only alternatives. Model scale and adaptation are not the only relevant design choices. Fact verification, as evaluated here, is a fixed three-class classification task. It is therefore useful to

compare the adapted decoder with architectures designed specifically for discriminative prediction. We evaluate two recent multilingual encoder-only models: mmBERT-base (approximately 307M parameters) and BGE-M3 (approximately 560M). Both models are fine-tuned end-to-end on the FEVER-IT training partition using a three-class classification head. Training is performed for four epochs. The learning rate is 1×10^{-4} for mmBERT-base and 1×10^{-5} for BGE-M3. Table 4 reports the comparison. The adapted decoder achieves the highest accuracy, exceeding BGE-M3 by 4.60 percentage points and mmBERT-base by 5.99 percentage points. The two encoder-only models therefore do not match the maximum predictive performance obtained through LoRA adaptation. Nevertheless, both encoders remain competitive while using substantially fewer base-model parameters. BGE-M3 contains approximately fourteen times fewer parameters than Llama-3.1-8B-Instruct, while mmBERT-base contains approximately twenty-six times fewer. This identifies a distinct form of efficiency: the encoder-only systems sacrifice part of the observed accuracy in exchange for a considerably smaller architecture. BGE-M3 achieves the stronger result among the two encoders. The lower macro-F1 of mmBERT-base relative to its accuracy further indicates a less balanced per-class behavior. This difference cannot be attributed to parameter count alone, since the models also differ in architecture, pre-training objectives, data, tokenization, and representation mechanisms. The comparison should therefore be interpreted as an evaluation of two complete encoder configurations, rather than as evidence for a general scaling relationship among encoder-only models. The decoder may provide capabilities beyond fixed-label prediction, including natural-language generation, interactive explanations, and integration into workflows that require multiple output formats. These capabilities are not measured by the current evaluation and should not be inferred from classification accuracy. Conversely, when the required output is limited to a stable verdict from a closed label set, the additional generative flexibility may not be necessary.

Model configuration	Parameters	Accuracy	Macro-F1
Llama-3.1-8B-Instruct + LoRA	8B	**0.8885**	**0.8881**
BGE-M3	0.560B	0.8425	0.8427
mmBERT-base	0.307B	0.8286	0.8010

Table 4
Comparison of the LoRA-adapted decoder with end-to-end fine-tuned encoder-only models.

4. Conclusion

We compared three model-selection strategies for Italian fact verification: scaling general-purpose few-shot models, specializing an 8B decoder through LoRA, and fine-tuning smaller encoder-only architectures. On FEVER-IT, few-shot performance improves substantially among the smallest configurations, while differences among the larger heterogeneous models remain limited and non-monotonic. Task-specific adaptation produces the strongest observed result. LoRA increases the accuracy of Llama-3.1-8B-Instruct from 0.7067 to 0.8885 and substantially improves the recognition of NotEnoughInfo instances. This does not imply that the 8B model is intrinsically superior, since it has access to the complete labeled training set. Rather, it shows that, when substantial supervision is available, specialization can be more effective than selecting a much larger general-purpose few-shot model. BGE-M3 and mmBERT-base achieve lower accuracy, but use approximately fourteen and twenty-six times fewer base-model parameters, respectively. Model selection should therefore consider supervision, adaptation strategy, architectural design, and required output capabilities, rather than reported parameter count alone. The primary evaluation focuses on FEVER-IT, assumes that evidence has already been retrieved, and does not directly measure latency, memory, energy consumption, or monetary cost. The evaluated systems also differ in architecture, training, runtime, and inference configuration. Since FEVER and FEVER-IT are publicly available and the pre-training data of several recent models are not fully disclosed, contamination from prior benchmark exposure cannot be ruled out; no financial or banking documents are evaluated.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] 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, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1107–1128. doi:10.18653/v1/2024.emnlp-main.64.

[4] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[5] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[6] K. Du, Y. Zhao, R. Mao, F. Xing, E. Cambria, Natural language processing in finance: A survey, Information Fusion 115 (2025) 102755. doi:10.1016/j.inffus.2024.102755.

[7] 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.

[8] A. Scaiella, S. Costanzo, E. Passone, D. Croce, G. Gambosi, Leveraging large language models for fact verification in Italian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 898–908. URL: https://aclanthology.org/2024.clicit-1.97/.

[9] 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.

[10] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. doi:10.1162/tacl_a_00454.

[11] 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. doi:10.18653/v1/D19-6601.

[12] 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. doi:10.18653/v1/2021.fever-1.1.

[13] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, 2023. arXiv:2305.13117.

[14] 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.

[15] 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: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2022, p. 416–428. URL: https://doi.org/10.1007/978-3-030-99739-7_52.

[16] 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: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 449–458. URL: https://doi.org/10.1007/978-3-031-56069-9_62.

[17] C. Chen, K. Shu, Combating misinformation in the age of llms: Opportunities and challenges, AI Magazine 45 (2024). doi:10.1002/aaai.12188, first circulated as arXiv:2311.05656 (2023).

[18] F. Pierri, A. Artoni, S. Ceri, Hoaxitaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019, 2020. arXiv:2001.10926.

[19] J. Gili, L. Passaro, T. Caselli, Check-IT!: A corpus of expert fact-checked claims for Italian, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 227–235. URL: https://aclanthology.org/2023.clicit-1.29/.

[20] R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage?, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[21] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. arXiv:2403.14608.

[22] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, LLaMAntino: LLaMA 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[23] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022. arXiv:2205.05638.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[25] M. Treviso, J.-U. Lee, T. Ji, B. van Aken, Q. Cao, M. R. Ciosici, M. Hassid, K. Heafield, S. Hooker, C. Raffel, P. H. Martins, A. F. T. Martins, J. Z. Forde, P. Milder, E. Simpson, N. Slonim, J. Dodge, E. Strubell, N. Balasubramanian, L. Derczynski, I. Gurevych, R. Schwartz, Efficient methods for natural language processing: A survey, Transactions of the Association for Computational Linguistics 11 (2023) 826–860. doi:10.1162/tacl_a_00577.

[26] A. A. V. Gonzalez, Cost-aware model selection for text classification: Multi-objective trade-offs between fine-tuned encoders and llm prompting in production, 2026. arXiv:2602.06370.