Comparing LLMs and BERT-based Classifiers for Resource-Sensitive Claim Verification in Social Media

Max Upravitelev[1], Nicolau Duran-Silva[2,3], Christian Woerle[4], Giuseppe Guarino[5], Salar Mohtaj[6], Jing Yang[1,7], Veronika Solopova[7], and Vera Schmitt[1,6,7,8]

[1]Technische Universität Berlin [2]SIRIS Lab, Research Division of SIRIS Academic
[3]LaSTUS Lab, TALN, Universitat Pompeu Fabra [4]Climate+Tech AI Think-tank [5]Data for good
[6]German Research Center for Artificial Intelligence (DFKI)
[7]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[8]Centre for European Research in Trusted AI (CERTAIN)

Abstract

The overwhelming volume of content being published at any given moment poses a significant challenge for the design of automated fact-checking (AFC) systems on social media, requiring an emphasized consideration of efficiency aspects. As in other fields, systems built upon zero-shot LLMs have achieved good results on different AFC benchmarks. The application of LLMs, however, is accompanied by high resource requirements. The energy consumption of LLMs poses a significant challenge from an ecological perspective, while remaining a bottleneck in latency-sensitive scenarios like AFC within social media. Therefore, we propose a system built upon fine-tuned smaller BERT-based models and comprised of components for abstract retrieval and claim verification. When evaluated on the ClimateCheck dataset against decoder-only LLMs, our best fine-tuned model outperforms Phi 4 14B and approaches Qwen3 14B in reasoning mode — while significantly reducing runtime per claim. Our findings demonstrate that small encoder-only models fine-tuned for specific tasks can still provide a substantive alternative to large decoder-only LLMs, especially in efficiency-concerned settings.

1 Introduction

While social media can be a space for public discourse, it can also be a place where misinformation and disinformation claims become dominant. In real-life claim verification, fast response times could be decisive in regard to the impact of harmful claims, such as providing verdicts before the claims start to spread. In the context of climate-related topics, where claims can be verified by a large amount of research, an opportunity is provided to combat misinformation by retrieving relevant research to verify said claims.

Like many other tasks in the natural language processing (NLP) domain, automated fact-checking systems are gaining significant performance boosts with the rise of large language models (LLMs). In the context of social media, however, the application of LLMs for tasks such as claim verification is greatly hindered by their high computational costs and latency. Which, on a large scale, is problematic from an ecological point of view (Jegham et al., 2025), as well as when considered from a latency-sensitive system design perspective (Wang et al., 2025).

Moreover, recent research indicates that BERT-based models fine-tuned for specific tasks can still be competitive with zero-shot LLMs in text classification (Kostina et al., 2025), or even outperform LLMs as shown in Bucher and Martini (2024) while also outperforming other classifiers in related challenging tasks like propaganda detection (Solopova et al., 2024). As discussed in related studies such as Li (2025), many encoder-only BERT-based models like deberta-v3 (He et al., 2023) are accompanied by significantly lower computational costs and therefore have a lower ecological impact due to a smaller number of parameters than many of their recent decoder-only counterparts like Qwen3 (Yang et al., 2025) or Phi 4 (Abdin et al., 2024). Thus, we want to explore how both model classes perform on the ClimateCheck dataset (Abu Ahmad et al., 2025a) – which was released in the context of the ClimateCheck@SDP 2025 Shared Task (Abu Ahmad et al., 2025b) – with respect to veracity prediction. In both cases, the input for prediction is acquired by an abstract retrieval pipeline, which we propose in this paper, and which also does not rely on LLMs.

The main contributions of this paper can be summarized as follows:

1. Proposing a new pipeline for retrieving abstracts from the ClimateCheck dataset corpus;
2. Exploring the fine-tuning of BERT-based models on the ClimateCheck dataset;

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 281–287
July 31, 2025 ©2025 Association for Computational Linguistics

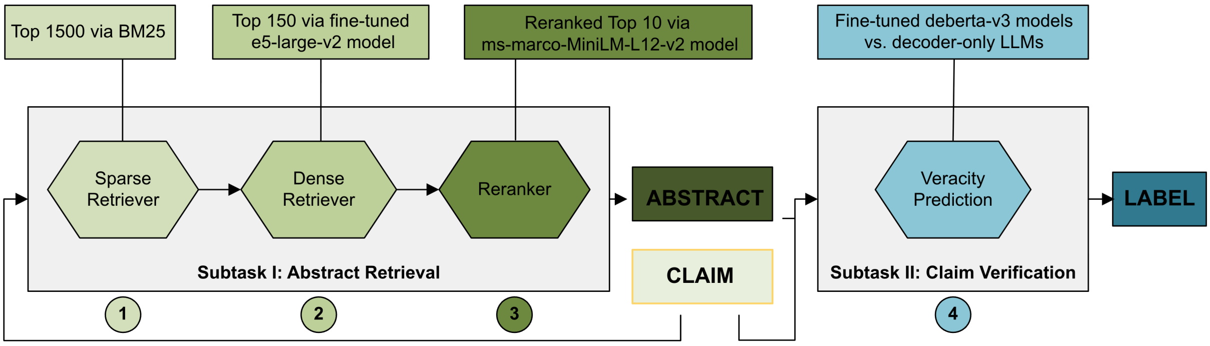
Top 1500 via BM25
Top 150 via fine-tuned e5-large-v2 model
Reranked Top 10 via ms-marco-MiniLM-L12-v2 model
Fine-tuned deberta-v3 models vs. decoder-only LLMs
Sparse Retriever
Dense Retriever
Reranker
ABSTRACT
Veracity Prediction
LABEL
Subtask I: Abstract Retrieval
CLAIM
Subtask II: Claim Verification
1
2
3
4

Figure 1: Architecture of the proposed system

3. Evaluating the claim verification results of fine-tuned BERT-based models against LLMs on runtime and the official ClimateCheck scores.

We have released our code[1] and the models[2,3] we fine-tuned in the context of this paper.

2 Related Work

In recent studies on the verification of climate claims (Leippold et al., 2024), agent-based LLM systems have been shown to achieve promising results when verifying claims based on retrieved knowledge from a corpus such as provided by the Intergovernmental Panel on Climate Change (IPCC). However, in a dynamic situation with a large unfiltered corpus of scientific papers and the frequency of social media claims, the cost and latency may limit the applicability of such a pipeline alone.

At the same time, several datasets were published for the verification of claims outside of the climate domain. For example, PubHealth (Kotonya and Toni, 2020) focuses on public health-related claims, which are accompanied by claims labeled with "true", "false", "mixture" and "unproven". The FEVER (Fact Extraction and VERification) dataset (Thorne et al., 2018) aims at the development of systems for the verification of claims on different topics against textual sources, using the labels "Supported", "Refuted" or "NotEnoughInfo" – a label scheme similar to the labels in ClimateCheck. AVeriTeC (Automated Verification of Textual Claims) (Schlichtkrull et al., 2023) focuses on retrieved evidence from the open web to verify claims, also providing samples with the additional label "Conflicting Evidence/Cherrypicking". In Yang and Rocha (2024), the AVeriTeC task is understood as related to natural language inference (NLI) tasks, which focus on logical inference based on free-text data. In this paper, the authors proposed a label mapping scheme for PubHealth and AVeriTeC and fine-tuned a T5-3B model (Raffel et al., 2023), whose initial training included data from NLI datasets. This strategy inspired us to explore models beyond decoder-only architectures that were fine-tuned on NLI datasets and to fine-tune them further in the context of ClimateCheck.

3 Methodology

Subtask I: Abstract Retrieval The first subtask focuses on the retrieval of relevant abstracts from a corpus of around 400K abstracts of publications from the climate science domains. We propose the following pipeline for this subtask, also illustrated in Figure 1:

1. Sparse retrieval: Get the top 1500 most relevant abstracts from the corpus using each claim as the query via BM25
2. Dense retrieval: Get the most relevant top 150 results from (1)
3. Rerank the results from (2) with a reranking model and return the final top 10 results

The inclusion of step (1) was the result of preliminary experiments, where we first explored the strategy of running dense retrieval on the full set of the embeddings of all 400k abstracts. Since this strategy yielded subpar results, we opted for a hybrid search approach by including sparse retrieval, which is a frequent approach in retrieval tasks to improve retrieval scores (as shown in Sawarkar et al.

[1] https://github.com/XplaiNLP/climatecheck-submission

[2] https://huggingface.co/xplainlp/e5-large-v2-climatecheck

[3] https://huggingface.co/xplainlp/DeBERTa-v3-large-mnli-fever-anli-ling-wanli-climatecheck

#	Embedding Model	Reranking Model	R@2	R@5	R@10	B-Pref	Score
1	e5-large-v2-climatecheck	ms-marco-MiniLM-L12-v2	0.217	0.405	0.574	0.449	0.411
2	e5-large-v2	ms-marco-MiniLM-L12-v2	0.208	0.399	0.560	0.437	0.401
3	e5-large-v2-climatecheck	bge-reranker-large	0.176	0.348	0.502	0.414	0.360
4	e5-large-v2-climatecheck	jina	0.193	0.328	0.464	0.398	0.346
5	#1 w/o bm25	ms-marco-MiniLM-L12-v2	0.197	0.365	0.521	0.397	0.370
6	e5-large-v2-climatecheck	-	0.151	0.257	0.375	0.311	0.273

Table 1: Evaluation on the abstract retrieval subtask. "R" refers to Recall and "Score" to the final ClimateCheck Subtask I Score. "jina" in Configuration #4 refers to jina-reranker-v2-base-multilingual.

(2024), for example). The top k value of 1500 retrieved abstracts was another result of preliminary testing, where we tried different values and chose the one with the best scores on the ClimateCheck dataset.

The results of step (2) are dependent on the embedding model. Here, we experimented with different fine-tuning strategies on e5-large-v2 (introduced in Wang et al. (2022a)). Finally, we fine-tuned the model for three epochs on the entire dataset while incorporating positive and negative examples into the training process. The related claims and abstracts in the ClimateCheck dataset can be seen as sets of positive pairs that map semantically close pairs of texts to each other, which can can be used as positive examples during fine-tuning. As shown in studies like Zhan et al. (2021), the performance in retrieval tasks can be further improved by expending such sets with negative examples. We mined three negative examples by retrieving the three least relevant abstracts via dense retrieval-based ranking.

Finally, we refined the ranking of the result from step (2) with a reranker model in step (3), which was chosen by comparing which model yielded the best results.

Subtask II: Claim Verification The second subtask focuses on the prediction of veracity labels based on the claims and abstracts retrieved in subtask I.

Inspired by Yang and Rocha (2024), our strategy was to fine-tune a BERT-based model previously fine-tuned on related NLI tasks to predict the veracity on the ClimateCheck dataset. This strategy deviates from Yang and Rocha (2024), in which a T5-3B model with an encoder-decoder architecture was used. Since our goal was to achieve good results while minimizing computational inference cost, we opted to work with smaller, encoder-only architectures. Finally, we explored publicly available options of models fine-tuned for NLI tasks and decided to compare two fined-tuned versions of deberta-v3 (He et al., 2023), which allowed for better comparison of the fine-tuning effects due to the same base model:

1. nli-deberta-v3-large from the cross-encoders series by Sentence Transformers[4] fine-tuned on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018).

2. DeBERTa-v3-large-mnli-fever-anli-ling-wanli (Laurer et al., 2022), which was fine-tuned on five NLI-related datasets including MultiNLI, ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), WANLI (Liu et al., 2022) and FEVER NLI, which is a FEVER variant transposed into the NLI schema (Nie et al., 2019). Unlike the model in (1), it is also explicitly not fine-tuned on SNLI.

We fine-tuned the models as follows:

1. Each input consisted of a claim and abstract concatenated with a [SEP] token.

2. Training was stopped when the evaluation metric failed to improve over successive epochs, resulting in 8 epochs in total.

3. We computed class-wise accuracies SUP_{acc}, REF_{acc} and NEI_{acc} and used $Acc_{min} = min(SUP_{acc}, REF_{acc}, NEI_{acc})$ as the optimization target to penalize imbalance.

4. To account for randomized factors (data split, model initialization), we ran the training procedure multiple times and selected the model with the highest Acc_{min} score.

[4] https://huggingface.co/cross-encoder/nli-deberta-v3-large

#	Model	s/claim	Precision	Recall	F1	Score
1	DeBERTa-v3-large-climatecheck	**0.032**	0.686	0.683	0.683	1.257
2	DeBERTa-v3-large-mnli-fever-anli-ling-wanli	0.032	0.261	0.154	0.104	0.678
3	nli-deberta-v3-large-climatecheck	0.032	0.604	0.607	0.602	1.176
4	nli-deberta-v3-large	0.032	0.413	0.418	0.289	0.863
5	Phi 4 14B	0.729	0.668	0.662	0.660	1.234
6	Qwen3 14B	12.229	**0.716**	**0.717**	**0.716**	**1.291**
7	Qwen3 14B w/o reasoning	0.363	0.690	0.629	0.597	1.171
8	Qwen3 1.7B	9.176	0.697	0.661	0.646	1.242

Table 2: Evaluation of subtask II concerning claim verification. The full name of our fine-tuned model in #1 is "DeBERTa-v3-large-mnli-fever-anli-ling-wanli-climatecheck". "Score" refers to the final ClimateCheck Subtask II score.

4 Evaluation

Subtask I The first subtask is evaluated on Recall@k, where $k = [2, 5, 10]$, and Binary Preference (B-Pref). All 4 scores are averaged into a final Subtask I score. Our pipeline achieved 4th place out of 10 on the subtask. Our evaluation results are documented in Table 1.

The first two results highlight the influence of our fine-tuning strategy by ablating it, resulting in worse performance. Next, we evaluate the influence of the reranking model by running bge-large-rerank (Xiao et al., 2023), a jina model[5], and a model from the Sentence Transformers Cross-Encoder series[6] against each other. For our final pipeline, we choose the highest scoring model, which was also explicitly fine-tuned on the information retrieval MS MARCO dataset (Bajaj et al., 2018). In the last section of Table 1 we assess the influence of retrieval components by ablating them. Setting (S) #5 documents our best performing configuration from S#1 without the BM25 step, indicating its importance due to a performance drop. Similarly, another drop is shown by S#6, where reranking was removed from the pipeline.

Subtask II The second subtask is evaluated on Precision, Recall, and the weighted F1-score. The final Subtask II score is the F1-score scaled by the number of claim-abstract pairs that were retrieved correctly, represented by the Recall@10 score of Subtask I. Since runtime was an important factor in our system design, we also included the processing time per claim in our evaluation. All experiments were run on a system with one NVIDIA H100 80 GB GPU. Table 2 documents our results.

S#1 achieves competitive results against our LLM configurations, while processing claims at only 0.032 seconds on average, outperforming LLMs on this metric by a margin. The other NLI-fine-tuned model in S#3 performed worse, which could be related to the selection of the datasets both were fine-tuned on, respectively. Both models perform worse without our fine-tuning strategy, as documented by S#2 and S#4. Surprisingly, there is also a large performance gap between both, where S#4 outperforms S#2 despite S#2 being more successful with our fine-tuning strategy.

For the comparison with current decoder-only LLMs, we start by evaluating against Phi 4 (Abdin et al., 2024), which is a recent model with 14B parameters and good performance results on many benchmarks. It is outperformed by S#1 across all metrics, most notably on the runtime. For better comparison, we also evaluate against members of the Qwen3 (Yang et al., 2025) series. S#6 was our final submission in the shared task, achieving 3rd place in the Subtask II score and 2nd place in Recall, Precision and F1.

Compared to our other settings, it has the best results in all metrics – except on runtime, yielding 12.229 seconds per claim. Turning off the reasoning in S#7 greatly improved the runtime while still achieving competitive results. However, this configuration was outperformed by S#1 and S#2 on the final Subtask II score while being around 14.4 times slower. In S#8 we replaced the Qwen 14B model with the 1.7B variant. Although still slower compared to S#1, it outperformed Phi 4 and S#7 on the Subtask II metric.

[5] https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

[6] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L12-v2

To further evaluate runtime differences, we perform a paired t-test over the test set (N = 1760) on per-claim runtimes. The BERT-based model in S#1 (mean = 0.032 s, std = 0.002 s) is significantly faster than the fastest decoder-only LLM in S#7 (mean = 0.363 s, std = 0.133 s) with $t(1759) = -104.541$, $p < 0.001$, Cohen's $d = 2.49$.

5 Discussion

Our results indicate that while recent decoder-only zero-shot LLMs such as Qwen3 are able to receive impressive results on datasets like ClimateCheck just by prompting them without applying any fine-tuning strategies, fine-tuned encoder-only BERT-based models can achieve comparable results at a fraction of the runtime. In conclusion, the smaller model class can still be a valid choice, particularly in scenarios where low latency is a critical factor.

Limitations

This study focuses on the comparison between fine-tuned encoder-only BERT models and decoder-only zero-shot LLMs in task-specific performance and runtime. While our results align with prior work (e.g., Bucher and Martini (2024)), they are limited to the described settings and the dataset used. Our system is tailored to the current iteration of the ClimateCheck dataset, and evaluating it on other datasets is necessary to assess generalizability. This is particularly relevant for the comparison of the two model families: Studies such as Wang et al. (2022b) indicate that decoder-only zero-shot LLMs generalize better than their fine-tuned encoder-only counterparts and therefore are less sensitive to changes in data.

The competitive results of BERT-based models as shown here are limited to the comparison against LLMs in a zero-shot setting. The performance of decoder-only LLMs could be further improved, for example, by prompting strategies such as few-shot learning (adding examples to prompts). Although this could further slow down the inference time due to increased length of input context that needs to be processed, it could also lead to a more consequential performance gap.

Finally, while the reported runtime performance at 0.032 seconds per claim on average can be considered as approaching real-time latency requirements, this results was achieved on a high-end GPU (NVIDIA H100). For real-life deployment, more optimization like quantization and parallelization techniques are needed to enable similar runtime on lower-end devices.

Acknowledgments

The work on this paper is performed in the scope of the projects "VeraExtract" (01IS24066) and "news-polygraph" (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 Technical Report. *arXiv preprint*. ArXiv:2412.08905 [cs].

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025a. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025b. The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Nidhal Jegham, Marwen Abdelatti, Lassad Elmoubarki, and Abdeltawab Hendawi. 2025. How hungry is ai?

benchmarking energy, water, and carbon footprint of llm inference. *Preprint*, arXiv:2505.09598.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arXiv:2501.08457.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. BERT-NLI-transfer-learn-laurer.pdf. Publisher: Open Science Framework.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *Preprint*, arXiv:2401.12566.

Andrew Li. 2025. A Case Study of Sentiment Analysis on Survey Data Using LLMs versus Dedicated Neural Networks.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? *Preprint*, arXiv:2104.07179.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, volume 24, page 155–161. IEEE.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Thirty-thh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Veronika Solopova, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf. 2024. Check news in one click: NLP-empowered pro-kremlin propaganda detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 44–51, St. Julians, Malta. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Rui Wang, Zhiyong Gao, Liuyang Zhang, Shuaibing Yue, and Ziyi Gao. 2025. Empowering large language models to edge intelligence: A survey of edge efficient llms and techniques. *Computer Science Review*, 57:100755.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022b. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jing Yang and Anderson Rocha. 2024. Take it easy: Label-adaptive self-rationalization for fact verification and explanation generation. *Preprint*, arXiv:2410.04002.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’21, page 1503–1512, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Prompts Collection

For predicting veracity labels with LLMs, we used:

> f“<sys>You are a professional fact checker. You get a claim and an abstract of a scientific paper. Assess if the claim is supported or refuted by the abstract! Return only your verdict! Either ’Supports’, ’Refutes’ or ’Not Enough Information’.</sys><user>The claim: {claim}\n {abstract}\n Your verdict: ”</user>

In all cases, the task description was used as the system prompt (indicated by the <sys>-tags), while the actual values of the variables where used within user prompts (indicated by the <user>-tags).