Less Can be More: An Empirical Evaluation of Small and Large Language Models for Sentence-level Claim Detection

Andrew Bell
Verita Research
New York, New York
andrew@veritaresearch.org

Abstract

Sentence-level claim detection is a critical first step in the fact-checking process. While Large Language Models (LLMs) seem well-suited for claim detection, their computational cost poses challenges for real-world deployment. This paper investigates the effectiveness of both small and large pretrained Language Models for the task of claim detection. We conduct a comprehensive empirical evaluation using BERT, ModernBERT, RoBERTa, Llama, and ChatGPT-based models. Our results reveal that smaller models, when finetuned appropriately, can achieve competitive performance with significantly lower computational overhead on *in-domain* tasks. Notably, we also find that BERT-based models transfer poorly on sentence-level claim detection in *out-of-domain* tasks, often over-predicting the positive outcome. We discuss the implications of these findings for practitioners and highlight directions for future research.

1 Introduction

Due to the increasing flow of global information, distinguishing factual content from opinion, speculation, or misinformation through fact-checking has become critically important. A foundational step in the fact-checking process is *sentence-level*[1] *claim detection*, or identifying whether a given sentence contains a factual claim or assertion. Without accurate claim detection, fact-checking efforts risk wasting resources.

Tools and approaches for automatically detecting factual claims from text have been developed in tandem with advances in Natural Language Processing. As of the writing of this paper, approaches using BERT-based models (Ni et al., 2024; Soleimani et al., 2020) are being replaced with those using state-of-the-art Large Language Models (LLMs) (Wang et al., 2024; Metropolitansky and Larson, 2025). Yet, there are inherent disadvantages to using LLMs for sentence-level claim detection: the computational demands of these models — both during training and inference — pose significant barriers to their deployment in real-time or in resource-constrained environments.

To this end, in this short paper, we present an empirical evaluation of the performance of different-sized Language Models on sentence-level claim detection. Specifically, we evaluate BERT, ModernBERT, RoBERTa, Llama, and ChatGPT-based models on a composite dataset constructed from three publicly available, human-curated datasets. we also test the ability of finetuned Language Models to generalize on "out-of-domain" data.

Contributions. First, we conduct experiments evaluating both the *in-domain* (Section 3) and *out-of-domain* (Section 4) performance of six models on sentence-level claim detection tasks. Second, we release four artifacts:[2] a composite dataset of approximately 13,000 sentences containing sentences and a binary label for whether or not the sentence is a claim, a finetuned BERT model, a finetuned ModernBERT model, and a finetuned Llama-3.2-1B-Instruct model. Third, in Section 5, we offer recommendations to practitioners on when it is worthwhile to use BERT-based language models as opposed to LLMs. Fourth, also in Section 5, we identify several research gaps and future research directions.

Findings. This paper finds that smaller, fine-tuned BERT-based models outperform LLMs on *in-domain* sentence-level claim detection tasks, making them a practical choice for resource-constrained settings. However—and perhaps expectedly—LLMs generalize better to *out-of-domain* data *without the need for fine-tuning*. In

[1] Sentence-level claim detection is distinguished from *document-level claim extraction* (Deng et al., 2024).

[2] https://github.com/VeritaResearch/claim-extraction

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 85–90
July 31, 2025 ©2025 Association for Computational Linguistics

Source	# of records	% positive
Claimbuster (Hassan et al., 2017)	7,976	25.00
PoliClaim Gold (Ni et al., 2024)	1,953	59.09
AVeriTeC (Schlichtkrull et al., 2023)	3,068	100.00
Total	12,997	47.83

Table 1: Composite dataset used for training and testing claim detection models.

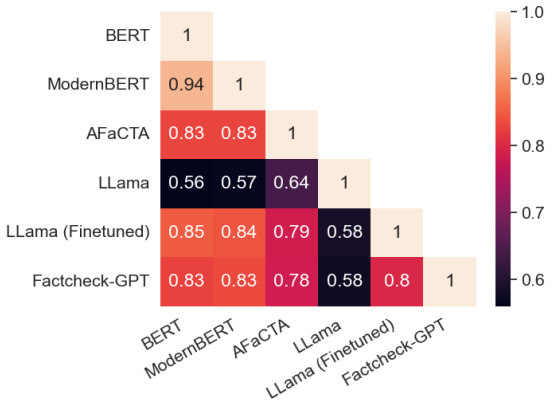

Figure 1: Overlap in positively predicted labels corresponding to Table 2.

some cases, finetuning can even harm LLM performance. Therefore, our results suggest that specialized models are more suitable for narrow domains, while LLMs are more effective for broad, diverse domains.

2 Methods

2.1 Dataset

We construct and release a composite dataset for training and testing sentence-level claim detection models made from three high-quality, publicly available datasets, summarized in Table 1. Importantly, all three of these datasets are *human-curated*. Claimbuster (Hassan et al., 2017) and PoliClaim (Ni et al., 2024) were collected specifically to train and test machine learning models for sentence-level claim detection, and contain sentences from US political speeches and debates. AVeriTeC (Schlichtkrull et al., 2023) contains claims published by over 50 different organizations, including fact-checking organizations like FullFact and Snopes.[3]

For finetuning and evaluation, wedivided this composite dataset into a *training* set (via random sampling of 80% of the samples), and a *testing* set (the remaining 20%). These samples were frozen throughout the finetuning and evaluation procedure, and can be found in the accompanying GitHub repository.[4]

2.2 Models

We evaluated the efficacy of using six models for sentence-level claim extraction, which are listed in Table 2 (and 3). The first two models were a finetuned BERT model[5] and a finetuned ModernBERT model,[6] an updated version of its namesake (Warner et al., 2024). These models have 110 million and 150 million parameters, respectively. The third model, known as AFaCTA, comes from Ni et al. (2024), and is a fine-tuned RoBERTa model (Zhuang et al., 2021) containing 125 million parameters.[7] The fourth and fifth models were a base Meta-Llama-3.2-1B-Instruct model[8] and a finetuned version of that same model. The sixth model used was Factcheck-GPT (Wang et al., 2024), for which the underlying model is OpenAI's ChatGPT-3.5 Turbo model.[9] The system and user prompts used for Factcheck-GPT and the Llama-3.2-1B-Instruct model can be found in Appendix Section A.

2.3 Finetuning details

As described in Section 2.1, 80% of the composite dataset constructed and released with this work was reserved for finetuning (training). Full parameter finetuning was used for the BERT and ModernBERT models, while LoRA was used to finetune the Llama-3.2-1B-Instruct model. Finetuning was carried out using the Huggingface `Trainer` class, and exact implementation details can be found in the GitHub repository accompanying this work.

[3]The full AVeriTeC dataset contains 4,568 real-world claims, but weonly include the publicly released training dataset which contains 3,068 claims.

[4]https://github.com/VeritaResearch/claim-extraction

[5]https://huggingface.co/google-bert/bert-base-uncased

[6]https://huggingface.co/answerdotai/ModernBERT-base

[7]https://huggingface.co/JingweiNi/roberta-base-afacta

[8]https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[9]https://platform.openai.com/docs/models/gpt-3.5-turbo

Table 2: Sentence-level claim detection results (speeches & fact-checks, composite dataset described in Section 2.1).

Model	Accuracy	Precision	Recall	F1 Score
BERT (Finetuned)	**0.917**	**0.918**	0.904	**0.911**
ModernBERT (Finetuned)	0.911	0.907	0.902	0.904
AFaCTA (Ni et al., 2024)	0.831	0.755	**0.945**	0.839
LLama-3.2-1B-Instruct	0.571	0.526	0.816	0.640
LLama-3.2-1B-Instruct (Finetuned)	0.850	0.844	0.834	0.839
Factcheck-GPT (Wang et al., 2024)	0.824	0.802	0.829	0.815

All finetuning was performed on an NVIDIA GTX 4060 Ti with 8GB of VRAM and took 24 hours or less to complete for each model. The training loss can be seen in Figure 2. The BERT and ModernBERT models were trained over 5 epochs which was sufficient for training loss to converge to close to 0. Due to resource constraints, the Llama-3.2-1B-Instruct model was only trained for 30 epochs and training loss was reduced from 2.3932 to 0.8570 (a 64.2% decrease).

3 Results

The performance for all six models on sentence-level claim extraction can be found in Table 2. The best performing model with respect to accuracy, precision, and F1 score was the finetuned BERT model at 91.7%, 91.8%, and 91.1%, respectively. We found that the AFaCTA model had the best recall at 94.5% — although, the model has a relatively low precision. As we will discuss in Section 4, we observed a tendency of BERT-based models to over-predict claims (the positive outcome) when used on "out-of-domain" data.

We observe a significant performance difference between the two "base" LLMs used in our experiments: the Llama-3.2-1B-Instruct model and

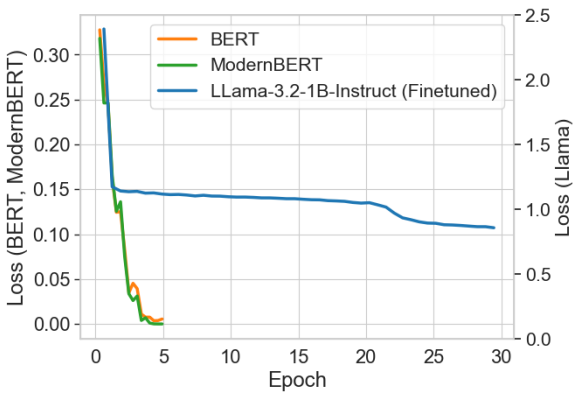

Figure 2: Finetuning training loss.

Factcheck-GPT, which uses OpenAI's ChatGPT-3.5 Turbo model. While the size of ChatGPT-3.5 Turbo is not publicly available, it is believed to be significantly larger than 1B parameters, indicating that larger LLMs may be better suited for claim detection tasks. Significantly, fine-tuning the Llama-3.2-1B-Instruct model results in strong performance improvements: the F1 score increased from 64.0% to 83.9%.

Figure 1 shows the overlap in positive predictions between the six models. In general, overlap closely follows from similarities between model precisions. Perhaps unsurprisingly, BERT and ModernBERT share 94% of their positively predicted labels, indicating that the two models are likely learning the same underlying semantic structures of claims.

4 Transfer and out-of-domain performance

We also evaluated the performance of the six models on *out-of-domain* claims to test how well performance generalizes. In this short paper use "out-of-domain" to refer to a domain of claims that may have a different underlying semantic structure as compared to the domain a model was trained on.

To carry out this evaluation, we obtained a dataset released by CheckThat[10] containing Tweets posted to X.com in the English language, where each Tweet has a label indicating whether or not it contains verifiable, factual claims (Nakov et al., 2022). The dataset contains 911 human-labeled Tweets, where 63.0% are labeled as claims (the positive outcome). Results for sentence-level claim detection among the six models studied can be seen seen in Table 3. Importantly, Tweets have a different semantic character than political speeches: many examples in CheckThat contain internet

[10] https://gitlab.com/checkthat_lab/clef2022-checkthat-lab/clef2022-checkthat-lab

Table 3: Sentence-level claim detection results on **out-of-domain** samples (Tweets posted to X.com, described in 4). For clarity: we did not re-finetune the BERT, ModernBERT, and Llama-3.2-1B-Instruct model on these out-of-domain samples—models were finetuned using the composite dataset in Table 1 as described in Section 2.1.

Model	Accuracy	Precision	Recall	F1 Score
BERT (Finetuned)	0.633	0.632	0.998	0.774
ModernBERT (Finetuned)	0.637	0.634	**1.000**	0.776
AFaCTA (Ni et al., 2024)	0.633	0.642	0.940	0.763
LLama-3.2-1B-Instruct	0.607	0.663	0.765	0.710
LLama-3.2-1B-Instruct (Finetuned)	0.634	0.633	0.996	0.774
Factcheck-GPT (Wang et al., 2024)	**0.680**	**0.676**	0.944	**0.788**

slang, leet-speak, hashtags and other emojis.

Performance varies *significantly* on out-of-domain samples. In all cases and across all models, F1 scores dropped between 2.7% and 13.7%. Accuracy only increased for the Llama-3.2-1B-Instruct, and otherwise saw similar drops as observed with F1 score. The model with the best accuracy, precision, and F1 score was Factcheck-GPT with 68.0%, 67.6% and 78.7%, respectively. The ModernBERT model had the highest recall at 100.00%.

Importantly, we observe another salient finding: all models reported high recall on out-of-domain samples, yet relatively low precision scores close to the number of positive samples in the dataset (63.0%). This indicates a bias of all models—but particularly by those finetuned on in-domain samples—to over-predict the positive label (*i.e.,* that a sentence is a claim). Perhaps most surprisingly, the finetuned Llama-3.2-1B-Instruct model actually performed *worse* than the non-finetuned base version.

5 Discussion

Takeaways. The findings of this short paper suggest two key takeaways: first, when restricted to in-domain data, **less can be more**. we found that smaller, finetuned BERT-based models outperformed LLMs. This is good news for practitioners who are resource constrained: we found that BERT-based models can easily be finetuned over a small number of epochs (we use 5 in this paper) and with a small GPU having only 8GB of VRAM.

Our second key takeaway is a drawback of using finetuned, BERT-based models for claim detection: LLMs perform better on out-of-domain problems **without finetuning**. In fact, we present one example where finetuning actually *worsened* out-of-domain performance for an LLM. Overall, this finding is consistent with current literature which shows that LLMs perform well on zero- and few-shot learning (Kojima et al., 2022). It's also worth noting that because LLMs are trained on such large and diverse sets of data, the out-of-domain data may actually be "in-domain" for an LLM.

Our takeaway for those building claim detection models can be summarized in the following way: if one is detecting claims in a restricted domain (*e.g.,* political speeches), we recommend training a small, specialized model. However, if one will be detecting claims from an *un*restricted domain, LLMs will likely yield better performance over the long run.

Research gaps. We leave several important research gaps for future researchers working on claim detection: first, what constitutes a *domain* in claim-detection? In this paper, we separate domain by speeches and fact-checks versus Tweets posted to X.com. However, if one adopted a speech-like pattern to writing their Tweets, or wrote the sentences of their speech in the style of Tweets, would domain transfer be possible? One could explore the boundary of domains in claim detection, perhaps relating measures of semantic structures or word distributions to define a domain distance. Second, how many training samples are required to ensure that a claim detection model performs well on a domain (Kocielnik et al., 2023)? Third, is it possible to generalize a BERT or ModernBERT model to multiple domains, and is there a limit to that generalization? These research questions generally drive at fundamental questions of transfer learning in Natural Language Processing (Wang and Chen, 2022).[11]

[11] See the *Transfer Learning for Natural Language Processing* workshop co-located with NeurIPS in 2022 at `https://tl4nlp.github.io/`.

6 Limitations

This short paper has several limitations. First, we only explore two different domains, giving a limited insight on when domain transfer may (or may not) be possible. Instead, this paper only serves as a single "counterexample" demonstrating the difficulty of generalizing BERT-based models tuned for sentence-level claim detection across domains. Second, we only explore two LLM choices in this paper, one which is known to be small (1B parameters) and another which is thought to be very large. Future work could include adding "medium"-sized LLMs (like those with 7B or 8B parameters) and mixture-of-experts type models (Jiang et al., 2024).

References

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Rafal Kocielnik, Sara Kangaslahti, Shrimai Prabhumoye, Meena Hari, Michael Alvarez, and Anima Anandkumar. 2023. Can you label less by using out-of-domain data? active amp; transfer learning with few-shot instructions. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 22–32. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. *arXiv preprint arXiv:2502.10855*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *European conference on information retrieval*, pages 416–428. Springer.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912, Bangkok, Thailand. Association for Computational Linguistics.

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

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359.

Jindong Wang and Yiqiang Chen. 2022. Transfer learning for natural language processing. In *Introduction to transfer learning: Algorithms and Practice*, pages 275–279. Springer.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Prompts

Factcheck-GPT System Prompt

```
You are a helpful factchecker
    assistant.
```

Factcheck-GPT User Prompt

```
Your task is to identify whether
    texts are checkworthy in the
    context of fact-checking.
Let's define a function named
    checkworthy(input: List[str]).
The return value should be a list
    of strings, where each string
    selects from ["Yes", "No"].
"Yes" means the text is a factual
    checkworthy statement.
"No" means that the text is not
    checkworthy, it might be an
    opinion, a question, or others.
For example, if a user call
    checkworthy(["I think Apple is
    a good company.", "Friends is a
     great TV series.", "Are you
    sure Preslav is a professor in
    MBZUAI?", "The Stanford Prison
    Experiment was conducted in the
     basement of Encina Hall.", "As
     a language model, I can't
    provide these info."])
You should return a python list
    without any other words,
["No", "Yes", "No", "Yes", "No"]
Note that your response will be
    passed to the python
    interpreter, SO NO OTHER WORDS!

checkworthy({texts})
```

Llama-3.2-1B-Instruct System Prompt

```
Only answer with Yes or No
```

Llama-3.2-1B-Instruct User Prompt

```
SENTENCE: {texts}

Is the sentence a factual claim
    that could be verified by a
    factchecker? Yes or No
```