FACT-GPT: Fact-Checking Augmentation via Claim Matching with LLMs

Eun Cheol Choi
University of Southern California
Annenberg School of Communication
Information Sciences Institute
Los Angeles, CA, United States
euncheol@usc.edu

Emilio Ferrara
University of Southern California
Thomas Lord Department of Computer Science
Information Sciences Institute
Los Angeles, CA, United States
emiliofe@usc.edu

ABSTRACT

Our society is facing rampant misinformation harming public health and trust. To address the societal challenge, we introduce FACT-GPT, a framework leveraging Large Language Models (LLMs) to assist fact-checking. FACT-GPT, trained on a synthetic dataset, identifies social media content that aligns with, contradicts, or is irrelevant to previously debunked claims. Our evaluation shows that our specialized LLMs can match the accuracy of larger models in identifying related claims, closely mirroring human judgment. This research provides a solution for efficient claim matching, demonstrates the potential of LLMs in supporting fact-checkers, and offers valuable resources for further research in the field.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Clustering and classification**; • **Computing methodologies** → **Semi-supervised learning settings**; **Natural language generation**.

KEYWORDS

fact-checking; misinformation; claim matching; large language model; synthetic data

ACM Reference Format:
Eun Cheol Choi and Emilio Ferrara. 2024. FACT-GPT: Fact-Checking Augmentation via Claim Matching with LLMs. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651504

1 INTRODUCTION

The urgent need for extensive fact-checking has been driven by the rapid proliferation of misinformation on digital platforms [29]. The fact-checking process, though complex and labor-intensive encompassing several stages from claim identification to drawing final conclusions, [7, 10] could be made more efficient through AI tools [1]. It is, however, critical to note that a complete automation could undermine journalistic principles and practices [21], thereby indicating the goal lies in enhancing, not replacing, human expertise [6].

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651504

A key element in monitoring the spread of false claims across various communication platforms is claim matching, or detecting new versions of already debunked claims [25]. Claim matching is crucial because false claims often get recycled and repeated in different formats across different communities [8, 21].

Large Language Models (LLMs) present both risks and benefits in combating the spread of misinformation online [2]. This paper presents a concise version of our work on FACT-GPT, a framework originally introduced in [4] which leverages LLMs to identify social media content that aligns with, contradicts, or is irrelevant to previously debunked claims. We provide a more robust comparison with benchmarks, a detailed breakdown of performance metrics, and highlighting the key contributions of our research. Our study reveals that when fine-tuned appropriately, LLMs can effectively match claims. Our framework could benefit fact-checkers by minimizing redundant verification, support online platforms in content moderation, and assist researchers in the extensive analysis of misinformation from a large corpus.

2 RELATED WORK

The Intersection of Fact-checkers and AI Fact-checkers are instrumental in the fight against misinformation, as they have developed reliable practices and principles over time [15]. The integration of AI into the fact-checking process should be conducted with great care, with the goal of enhancing efficiency without undermining established principles [21]. AI models that support rather than replace fact-checkers are more likely to be embraced. Fact-checkers have shown interest in the capabilities of AI tools to identify claims and analyze their popularity [1, 23]. However, at the same time, they are skeptical about the possibility of AI entirely replacing human involvement in the process of fact-checking, indicating the irreplaceable value of human decision-making.

LLMs in Annotation Tasks Large Language Models (LLMs) have garnered significant interest due to their potential to automate diverse annotation tasks. Given their flexible nature, LLMs' performance in various annotation tasks is being scrutinized. Research has evaluated LLMs in contexts such as fact-checking [13], annotating tweets [9], and beyond. Generating synthetic training data to enhance LLMs' performance in classification tasks has also been explored [5]. However, it is crucial to acknowledge LLMs' inherent limitations. Their probabilistic nature implies that their outputs can vary according to prompts and parameters [24]. When compared to task-specific models, ChatGPT often underperform [17], underlining the need for models that are specifically designed and utilized for certain tasks.

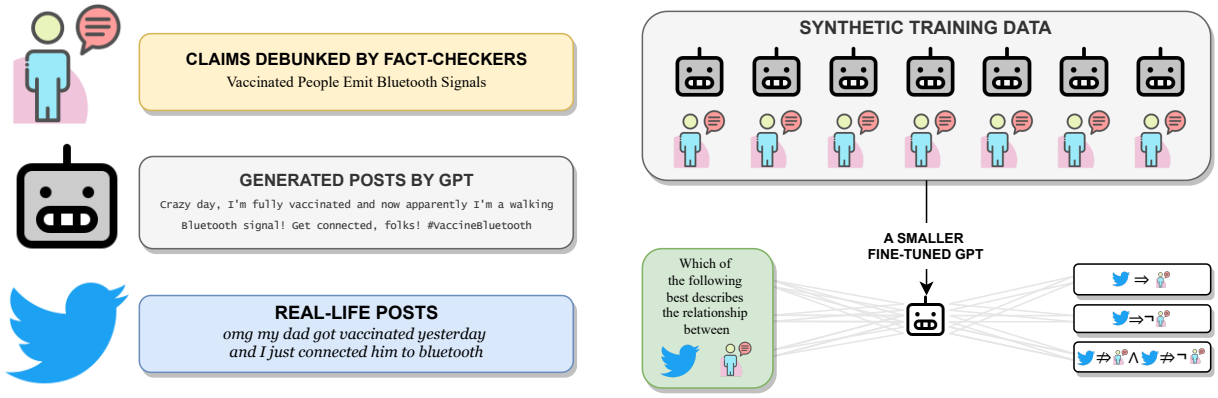

Figure 1: Overview of FACT-GPT, our framework aimed at assisting the claim matching task

3 PROPOSED FRAMEWORK

3.1 Task Description

In our study of Large Language Models' (LLMs) abilities in claim matching, we employ a *textual entailment task*. This task involves distinguishing the associations between two statements into one of three categories: Entailment, Contradiction, or Neutral. The task involves categorizing whether assuming the first statement to be true *implies* that the second statement is true, false, or neither. This approach is based more on practical reasoning than formal logic, with human insight and common sense playing a key role in determining the categorization [20, 22]. This methodology has also proven useful in the past for rumor detection [30].

Claim matching tasks can be configured in various forms including but not limited to textual entailment [19], ranking [18, 26], and binary detection tasks [16]. Defining claim matching as a 3-class entailment task poses both advantages and challenges. Identifying contradicting pairs is important as such rebuttals play a crucial role in mitigating the spread of misinformation [11, 28]. However, it's challenging due to the scarcity of contradiction pairs in real-world instances [20].

3.2 Datasets

In this study, we focus on misinformation relating to public health, specifically COVID-19 related false claims that have been fact-checked. 1,225 False claims debunked by professional fact checkers in 2020 and 2021 were obtained from *Google Fact Check Tools* and *PolitiFact*.

3.2.1 Synthetic Training Datasets Generation. When it comes to detecting misinformation, it is crucial to acquire sufficiently large labeled misinformation datasets, particularly in emerging domains [27]. We utilized Large Language Models (LLMs) to generate synthetic training data, allowing for the creation of a balanced dataset specifically designed for claim matching tasks. Fine-tuning language models on synthetic datasets can enhance their adaptability to specific task nuances, potentially leading to better classification accuracy. In addition, fine-tuning smaller models reduces the computational cost involved in large-scale operations while making it easier to customize these models based on emerging new claims.

We utilized three language models, `GPT-4`, `GPT-3.5-Turbo`, and `Llama-2-70b-chat-hf` (available via the *OpenAI API* or the *HuggingFace Inference API*), for generating training data. We generated tweets based on the false claims debunked by fact checkers. To generate varied styles in the outputs by the language models, we set the temperature parameter at 1. Figure 2 provides an example of a prompt used for data generation. A total of 3,675 synthetic tweets were generated from each model, ensuring an equal distribution across all three categories.

3.2.2 Ground Truth Dataset. Our method for creating a ground truth dataset is illustrated in Figure 3. Initially, we paired tweets from the publicly available *Coronavirus Twitter Dataset* [3] with

`System`	Generate TWEET so that if TWEET is true, then CLAIM is also true. Be brief. Do not start a sentence with 'Just'.
`Input`	Vaccininated people emit Bluetooth signals.
`Output`	Crazy day. I'm fully vaccinated and now apparently I'm a walking Bluetooth signal! Get connected, folks! #VaccineBluetooth

Figure 2: Example of synthetic tweet generation prompts

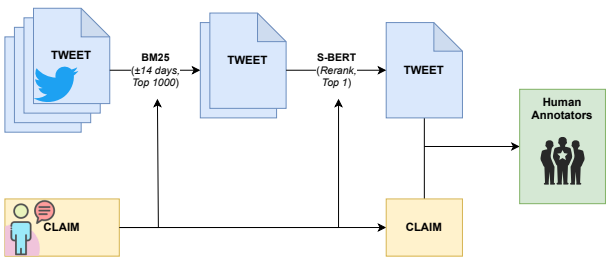

Figure 3: Workflow of test data construction

Table 1: Descriptive statistics for test data.

Label	Count	Percentage
ENTAILMENT	647	52.8%
NEUTRAL	433	35.3%
CONTRADICTION	90	7.3%
(Two-way ties)	55	4.5%
TOTAL	1225	100%

`System`	Which of the following best describes the relationship between TWEET and CLAIM? You must choose from ENTAILMENT, NEUTRAL, or CONTRADICTION. If TWEET is true: (ENTAILMENT) then CLAIM is also true. (NEUTRAL) CLAIM cannot be said to be true or false. (CONTRADICTION) then CLAIM is false.
`Input`	TWEET: omg my dad got vaccinated yesterday and I just connected him to bluetooth CLAIM: Vaccininated people emit Bluetooth signals.
`Output`	ENTAILMENT

Figure 4: Example of an entailment task prompt

debunked false claims, considering both token and semantic similarities. This process generated a unique set of 1,225 pairs consisting of tweets and claims. Experienced annotators on Amazon Mechanical Turk then classed each of these pairs into one of the three categories. The final categorization was based on which class received the majority of votes, creating a fully annotated test dataset, as illustrated in Table 1. As for dealing with two-way ties, we created 1,000 different scenarios of test data with random tie-breakers, and calculated the average performance across every scenario.

3.3 Experiments

3.3.1 Baselines. We established comparison benchmarks by assessing the performance of several pre-trained Large Language Models (LLMs), including `GPT-4`, `GPT-3.5-Turbo`, `Llama-2-13b`, and `Llama-2-7b`, against human annotations. We then presented prompts as illustrated in Figure 4 to each LLM and collected their responses at the temperature of 0 (OpenAI models) or 0.01 (Llama models).

3.3.2 Fine-tuning. Our assessment of FACT-GPT's effectiveness involved fine-tuning `GPT-3.5-Turbo`, `Llama-2-13b`, and `Llama-2-7b` with the synthetic training dataset outlined in 3.2.1. We allocated 80% of the data for training and the remaining 20% for validation. `GPT-3.5-Turbo` underwent fine-tuning using *OpenAI's Fine-tuning API*. Meanwhile, for the LLaMa models, we applied LoRA (Low-Rank Adaptation, [14]) in *LLaMa-Factory* [12], which is an efficient tuning framework for LLMs. `BERT-base` model was fine-tuned on `GPT-4`-generated train set to provide an additional benchmark. Each model went through three epochs (five for `BERT-base`) of fine-tuning on a single A100 GPU.

3.3.3 Results. The overall performance of FACT-GPTs are summarized in Table 2. Notably, models fine-tuned on synthetic datasets exhibited superior performance in comparison to the pre-trained versions. There was a consistent pattern in the performance among the fine-tuned models, with all models exhibiting improved outcomes when fine-tuned using training data generated by `GPT-4` as opposed to those generated by `GPT-3.5-Turbo` or `Llama-2-70b`. This trend emphasizes the significance of the quality of training data in determining the effectiveness of the resulting models.

Table 2 also reveals that our top-performing models are more adept at classifying *Entailment* and *Neutral* labels, but face challenges with *Contradiction* labels. This suggests that our FACT-GPTs are proficient in determining the relevance or irrelevance of social media posts to the original debunked claims. However, given that rebuttals to false claims play a crucial role in preventing the spread of misinformation [11, 28], future work should focus on improving the detection of contradictory posts.

4 DISCUSSION

This paper proposes a framework for assisting fact-checkers with LLMs by performing the claim matching task. Our research demonstrates that LLMs have the capacity to discern entailment relationships between social media posts and debunked claims. Importantly, our study reveals that appropriately fine-tuned, smaller LLMs can yield a performance comparable to larger models, thereby offering a more accessible and cost-effective AI solution without compromising quality. However, while our models excel in detecting whether social media content is relevant to or irrelevant from debunked claims, they show struggles with categorizing posts that contradict these claims. This is an area that requires further refinement, given the importance of rebuttals in curbing the spread of misinformation.

Looking forward, it is crucial to encourage ongoing collaborations among researchers, developers, and fact-checkers to fully exploit AI benefits while mitigating its potential drawbacks. The importance of human expertise and supervision in this context cannot be overstated. Completely automating fact-checking procedures using AI carries certain risks and limitations, such as the perpetuation of biases intrinsic to models and inherent inconsistencies due to their probabilistic nature. However, with thoughtful incorporation, technologies could substantially augment the capabilities of fact-checkers to detect and debunk misinformation.

Future studies should focus on discovering different methods for further optimizing FACT-GPT in terms of performance, inference time, memory allocation, continuous learning, and so on. Additionally, examining the models' generalizability in different contexts, as well as their capabilities in generating explanations for classification results, are other important next steps. This research adds substantively to a growing body of work examining the use of LLMs in support of human fact-checkers, offering a foundation for continued studies and the responsible advancement of AI tools to effectively combat the rapid spread of misinformation.

Table 2: Performance Metrics of Pre-trained and Fine-tuned Models.

Model	Train Set From	Precision	Recall	Accuracy	$F1_{Ent}$	$F1_{Neu}$	$F1_{Con}$
BERT-base	*GPT-4*	.44	.46	.49	.66	.33	.21
GPT-4	—	.64	.70	.63	.62	.66	.51
GPT-3.5-Turbo	*GPT-4*	.64	.68	.73	.83	.67	.44
	GPT-3.5-Turbo	.51	.59	.57	.76	.35	.32
	Llama-2-70b	.57	.65	.60	.73	.57	.34
	—	.56	.61	.58	.58	.64	.39
Llama-2-13b	*GPT-4*	.63	.69	.71	.79	.69	.45
	GPT-3.5-Turbo	.52	.57	.60	.73	.50	.34
	Llama-2-70b	.56	.64	.63	.74	.57	.39
	—	.51	.47	.30	.37	.36	.19
Llama-2-7b	*GPT-4*	.64	.70	.73	.79	.72	.46
	GPT-3.5-Turbo	.48	.52	.56	.69	.41	.29
	Llama-2-70b	.60	.60	.68	.74	.65	.40
	—	.42	.46	.40	.63	.02	.23

ACKNOWLEDGMENTS

This work was supported in part by DARPA (contract no. HR001121-C0169). The code and datasets used in this study are available at *https://doi.org/10.5281/zenodo.10807885*

REFERENCES

[1] P. Arnold. The challenges of online fact checking, 2020. URL https://fullfact.org/media/uploads/coof-2020.pdf.

[2] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, et al. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*, 2023.

[3] E. Chen, K. Lerman, E. Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273, 2020.

[4] E. C. Choi and E. Ferrara. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. *arXiv preprint arXiv:2310.09223*, 2023.

[5] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li. Auggpt: Leveraging chatgpt for text data augmentation, 2023.

[6] S. Dégallier-Rochat, M. Kurpicz-Briki, N. Endrissat, and O. Yatsenko. Human augmentation, not replacement: A research agenda for ai and robotics in the industry. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144. doi: 10.3389/frobt.2022.997386.

[7] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, and P. Atanasova. Checkthat! at clef 2019: Automatic identification and verification of claims. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, editors, *Advances in Information Retrieval*, pages 309–315, Cham, 2019. Springer International Publishing. ISBN 978-3-030-15719-7.

[8] T. Fornaciari, L. Luceri, E. Ferrara, and D. Hovy. Leveraging social interactions to detect misinformation on social media. *arXiv preprint arXiv:2304.02983*, 2023.

[9] F. Gilardi, M. Alizadeh, and M. Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), jul 2023. doi: 10.1073/pnas.2305016120.

[10] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, and M. Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10 (12):1945–1948, aug 2017. ISSN 2150-8097. doi: 10.14778/3137765.3137815.

[11] B. He, M. Ahamad, and S. Kumar. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709, 2023.

[12] hiyouga. Llama-factory. https://github.com/hiyouga/LLaMA-Factory, 2023.

[13] E. Hoes, S. Altay, and J. Bermeo. Using chatgpt to fight misinformation: Chatgpt nails 72% of 12,000 verified claims. 2023.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. 2021.

[15] IFCN. Code of principles, 2023. URL https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles.

[16] Y. Jin, X. Wang, R. Yang, Y. Sun, W. Wang, H. Liao, and X. Xie. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754, 2022.

[17] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Łukasz Radliński, K. Wojtasik, S. Woźniak, and P. Kazienko. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101861.

[18] V. La Gatta, C. Wei, L. Luceri, F. Pierri, E. Ferrara, et al. Retrieving false claims on twitter during the russia-ukraine conflict. In *WWW'23 Companion: Companion Proceedings of the ACM Web Conference 2023*, pages 1317–1323, 2023.

[19] J. Ma, W. Gao, S. Joty, and K.-F. Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics, 2019.

[20] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation*, 2014. URL https://api.semanticscholar.org/CorpusID:762228.

[21] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino. Automated fact-checking for assisting human fact-checkers, 2021.

[22] S. Padó and I. Dagan. Textual Entailment. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 06 2022. ISBN 9780199573691. doi: 10.1093/oxfordhb/9780199573691.013.024.

[23] I. V. Pasquetto, B. Swire-Thompson, M. A. Amazeen, F. Benevenuto, N. M. Brashier, R. M. Bond, L. C. Bozarth, C. Budak, U. K. Ecker, L. K. Fazio, et al. Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*, 1(8):01–14, 2020.

[24] M. V. Reiss. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark, 2023.

[25] S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.332.

[26] S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, and P. Nakov. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, 2022.

[27] K. Sharma, E. Ferrara, and Y. Liu. Construction of large-scale misinformation labeled datasets from social media discourse using label refinement. In *Proceedings of the ACM Web Conference 2022*, pages 3755–3764, 2022.

[28] M. Tambuscio and G. Ruffo. Fact-checking strategies to limit urban legends spreading in a segregated society. *Applied Network Science*, 4:1–19, 2019.

[29] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[30] A. Yavary, H. Sajedi, and M. S. Abadeh. Information verification improvement by textual entailment methods. *SN Applied Sciences*, 1:1–6, 2019.