Fact or Fiction? Can LLMs be Reliable Annotators for Political Truths?

Veronica Chatrath*, Marcelo Lotif, Shaina Raza*
Vector Institute for Artificial Intelligence
Toronto, ON, M5G 0C6, Canada
{veronica.chatrath, marcelo.lotif, shaina.raza}@vectorinstitute.ai

Abstract

Political misinformation poses significant challenges to democratic processes, shaping public opinion and trust in media. Manual fact-checking methods face issues of scalability and annotator bias, while machine learning models require large, costly labelled datasets. This study investigates the use of state-of-the-art large language models (LLMs) as reliable annotators for detecting political factuality in news articles. Using open-source LLMs, we create a politically diverse dataset, labelled for bias through LLM-generated annotations. These annotations are validated by human experts and further evaluated by LLM-based judges to assess the accuracy and reliability of the annotations. Our approach offers a scalable and robust alternative to traditional fact-checking, enhancing transparency and public trust in media.

1 Introduction

Political bias is the tendency to favour or oppose certain political ideologies, parties, or candidates in a way that skews impartiality, posing significant challenges in today's digital landscape [2]. While political bias refers to a large spectrum of topics like disinformation (false information spread intentionally to deceive) or misinformation (false information spread unknowingly) [22], this study is concentrated on distinguishing factual inaccuracies within politics. Traditional fact-checking approaches are not scalable given the vast amount of online information. Manual fact-checking methods, such as expert annotation and crowdsourcing, are viable alternatives but are often costly and time-consuming. Crowdsourcing lacks expert oversight and can lead to inconsistencies due to the varying expertise of contributors [21]. These limitations highlight the need for efficient and scalable fact-checking approaches.

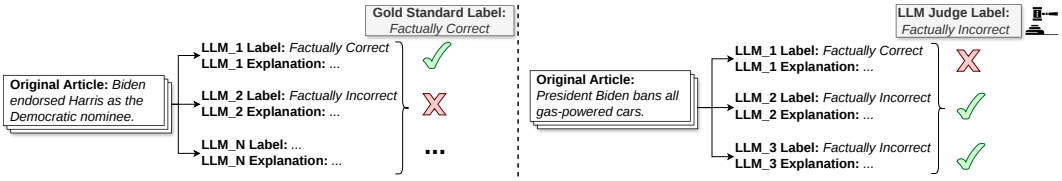

Figure 1: A working instance of our proposed LLM Annotator and Judge approach.

Recent studies have leveraged LLMs as annotators for various natural language processing (NLP) tasks including text classification [11], question answering [8], named entity recognition [37, 31], and other related applications [5, 25, 32, 24]. Newer LLMs like GPT-4, Llama-3.1, Claude 3.5 Sonnet, and Mistral Large 2, updated for 2024, are equipped with extensive knowledge of the global political

*Denotes equal contribution.

Accepted at Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

landscape and are often used for political analysis [15]. Despite advancements, the use of these LLMs as annotators for labelling unstructured data and training machine learning models on political news remains an under explored application. Current annotation methods primarily rely on paid models such as GPT-4, leading to concerns regarding reproducibility, cost, and the quality of annotations.

We propose a scalable and cost-effective framework using open-source LLMs to reduce the effort and cost of annotating political data, as shown in Figure 1. Our framework utilizes LLMs as both annotators and judges (evaluation models), making it adaptable to various LLMs. Our research questions include: *Can LLMs effectively serve as annotators for political misinformation?*, *Does including demonstrations (examples with prompts) improve LLM annotation performance?*, and *How can we evaluate LLM-based annotations using LLMs as judges?*

Our key contributions are:

1. Construction of a high-quality, up-to-date human-annotated dataset.
2. Development of an LLM framework for binary annotation of factually correct and incorrect political news reporting. Our framework utilizes open-source LLMs for annotation, offering reproducible, accessible, and cost-free usage.
3. A reference-based comparison of the LLM annotations against gold (human verified) labels is presented. A human review process has been included to ensure quality and accuracy, and to address any discrepancies identified in the LLM annotation process.

Empirical analysis indicates that annotations generated by LLMs closely match human annotations, as evidenced by high reference-based scores (Table 1). Additionally, LLM-based evaluations for assessing label quality demonstrate strong performance of LLM annotators (Table 2).

2 Related Works

Political Disinformation Political disinformation involves framing or altering information to make a political position or candidate appear more attractive, or unattractive [3], and poses significant challenges, such as bias or toxicity [30]. In response, LLMs have emerged as a promising method to combat political disinformation by automating the detection and annotation of deceptive content, reducing the reliance on costly and time-consuming manual processes. Recent studies [14, 1, 18] highlight the potential of open-source LLMs in detecting and labelling disinformation. These models are shown to improve the efficiency and accuracy of disinformation detection, strengthening defences within the digital information ecosystem.

LLM-as-an-Annotator Recent advances have demonstrated the potential of LLMs like OpenAI models [2], such as GPT-3.5, GPT-4 and recently GPT-4o as effective annotators for various NLP tasks [32]. These models can annotate data for classification and entity recognition through prompting methods. To maximize their utility, LLMs can be deployed within active learning loops [37], leveraging vast amounts of unlabelled data. Alignment tuning can further refine LLM annotations to match human preferences [39], mitigating potential biases.

Studies have shown that LLM-based annotations can match or exceed human performance across diverse domains, including social media analysis [16], computational science [42], and medical information extraction [13]. Comparative analyses between LLMs and human annotators further highlight their potential [11, 14, 29]. Building on these findings, our work employs LLMs as annotators to address political misinformation, an area yet to be fully explored in this context.

LLM-as-a-Judge Various LLMs such as GPT-3.5 [26] and GPT-4 [27] are increasingly utilized as evaluators, or judges, to ensure outputs align with human values [40]. LLMs can assess the quality of model outputs against specific criteria such as accuracy, toxicity, and relevance [10, 41], leveraging methods like asking for correctness or agreement with human annotations in a controlled experiment. In one LLM evaluation [3], OpenAI's GPT-4o is at the top of the leaderboard, followed by Google's Gemini-1.5-Pro, and Meta's Llama-3.1-470b. Motivated by these results, we employ the newly available GPT-4o-mini [28] as a LLM judge in this work.

[2] `https://openai.com/`

[3] LMSYS Chatbot Arena (lmsys.org)

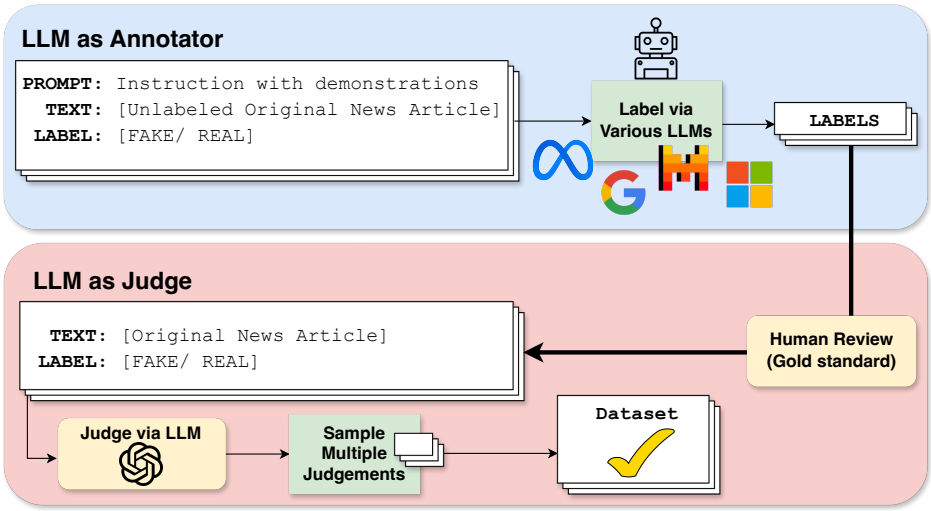

Figure 2: A visualization of our annotator (top) and judge (bottom) pipeline.

3 Methodology

3.1 Dataset Construction

We curated a dataset of news articles related to the North American political landscape, covering the period from May 6, 2023, to May 6, 2024. Data collection was conducted using automated scraping tools, including Newspaper and Selenium, with Google RSS as the main source. To ensure comprehensive coverage of electoral topics, a sample of 6,100 articles was selected based on the 26 topics and 16 sources listed in Appendix A.3.

3.2 LLM-as-an-Annotator

We employ a suite of open-source LLMs, recognized for their robustness and accuracy in state-of-the-art applications [19]. These models include Llama-3-8B-Instruct [34], Llama-3.1-8B- Instruct, Mistral-7B-Instruct-v0.3 [17], Gemma-2-9b-Instruct [33], and Phi-3-medium-128k-Instruct [23]. The annotation task involves labeling each news summary article as either *factually correct* or *factually incorrect*. The LLMs are employed in both zero-shot and five-shot settings. The following is a snippet of the prompt used for annotation, with the entire prompt listed in Appendix A.1.

Prompt used for Annotation

role: `user`, **content:** `You are a helpful news fact-checking bot trained to assess the accuracy of information. Analyze the given article and provide your assessment in the following format: ...`

`Classification:` `[Factually Correct/Factually Incorrect]`

`Explanation:` `Provide a concise, evidence-based explanation for your classification. Reference specific examples from the article and contradicting evidence from trusted sources, if applicable ...`

`Article to analyze:` `article`

The LLM-based annotations, determined by LLM majority voting, were reviewed by three human reviewers. The reviewers assessed each piece of data based on majority voting, resolving discrepancies, with unresolved cases escalated to a senior reviewer for the final decision. In this work, *gold data* refers to the dataset portion annotated by an LLM and, subsequently, reviewed by humans.

3.3 LLM-as-a-Judge

We employ LLMs as evaluators (judges) to ensure objective and scalable assessment of our dataset's annotations. Inspired by previous works [35, 12, 7], the evaluation task requires an LLM judge to assess the LLM annotator's response in a binary fashion (yes/no), determining whether it agrees with the annotation. The prompt used for the LLM-as-a-Judge experiments can be seen in Appendix A.2.

GPT-4o-mini is employed as the first judge, as it is quite affordable ($0.150 per 1M input tokens, $0.600 per 1M output tokens [4]). To abide by our commitment to open-source science, Llama-3.1-70B-instruct is also used as a judge. Both models are employed to prevent self-enhancement bias [40] which occurs when a model from the same family serves as both annotator and judge.

4 Experiments

4.1 Experimental Setup

Settings Inference was run on a single A40 GPU with 4 CPU cores. For 6,000 samples across 5 LLMs, the process took about 16.67 hours, consuming 9.34 kWh and emitting 4.2 kgCO2e [9]. This carbon footprint is relatively low compared to fine-tuning. Open-source versions of Llama, Mistral, Gemma, and Phi models were used. GPT-4o-mini was employed as a judge via the OpenAI API, with total approximate experiment costs of $2 USD on sample data.

Evaluation Metrics Two approaches are used to evaluate the annotations: 1) Reference-based evaluation involves employing standard classification metrics by comparing the precision, recall, and accuracy of the LLM-generated annotations with the gold-standard ground truth labels, inspired by similar works [11]. 2) LLM-as-a-judge-based evaluations, inspired by [40], are calculated as the percentage of times the LLM judge's assessment aligns with the LLM-based annotations, also known as Agreement Rate. This experiment is performed on 500 samples.

Method	Acc.	Prec.	Rec.	F1
Llama-3-8B-Instruct (zero-shot)	80.4%	82.3%	80.4%	80.7%
Llama-3-8B-Instruct (5-shot)	**89.3%**	**89.3%**	**89.3%**	**89.2%**
Llama-3.1-8B-Instruct (zero-shot)	74.5%	81.0%	74.5%	74.7%
Llama-3.1-8B-Instruct (5-shot)	82.1%	86.4%	82.1%	82.3%
Mistral-7B-Instruct-v0.3 (zero-shot)	80.0%	79.7%	80.0%	79.7%
Mistral-7B-Instruct-v0.3 (5-shot)	82.5%	83.1%	82.5%	81.8%
Gemma-2-9b-Instruct (zero-shot)	77.6%	82.0%	77.6%	77.9%
Gemma-2-9b-Instruct (5-shot)	83.3%	85.4%	83.3%	83.6%
Phi-3-med-128k-Instruct (zero-shot)	80.7%	80.8%	80.7%	80.7%
Phi-3-med-128k-Instruct (5-shot)	86.9%	87.1%	86.9%	86.7%

Table 1: Comparison of different LLM annotation's accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-score (F1) against the gold labels in zero-shot and 5-shot settings. The best results are in **bold**.

4.2 Comparison with Gold Data

We compare various LLMs as annotators against the gold data. This data was formed by a review team consisting of twelve volunteer members representing a diverse range of disciplines and demographics. Additional information on the composition of the gold data can be found in Appendix A.5. Table 1 shows the performance of the LLMs in an annotator role. We observe that Llama-3-8B-Instruct (5-shot) outperforms all other models. Additionally, five-shot experiments outperform all zero-shot experiments, answering our second research question. The inference times for each run can be seen in Appendix A.4.

[4] GPT-4o-mini Pricing

4.3 LLM-based Evaluation Results

Method	AR (GPT-4o-mini)	AR (Llama-3.1-70B)	AR (Ground Truth)
Llama-3-8B-Instruct	76.4%	**79.6%**	87.20%
Llama-3.1-8B-Instruct	70.8%	**73.2%**	78.0%
Mistral-7B-Instruct-v0.3	**74.2%**	71.4%	81.4%
Gemma-2-9b-Instruct	77.2%	**77.6%**	84.0%
Phi-3-med-128k-Instruct	**74.4%**	74.0%	84.0%

Table 2: Comparison of annotation agreement rates (AR) among LLMs, using GPT-4o-mini and Llama-3.1-70B as evaluators, against the ground truth agreement rate. Best judge results are in **bold**.

Table 2 shows the annotation agreement rates (AR) for various LLM annotators, evaluated by GPT-4o-mini and Llama-3.1-70B against ground truth scores. Llama-3-8B-Instruct achieves the highest agreement rates, with 76.4% for GPT-4o-mini and 79.6% for Llama-3.1-70B, closest to the ground truth score of 87.2%. Gemma-2-9b-Instruct also performs well, with consistent rates of 77.2% and 77.6% for the two judges, aligning relatively closely with its ground truth score of 84.0%.

Notably, there are differences in how the two judges evaluate certain annotators. Mistral-7B-Instruct-v0.3 has a higher agreement rate with GPT-4o-mini (74.2%) compared to Llama-3.1-70B (71.4%), yet both are below its ground truth of 81.4%. Phi-3-medium-128k-Instruct shows similar agreement with both judges (around 74%) but falls short of its ground truth (84.0%).

The differences in agreement rates between the judges highlight the importance of using multiple evaluation methods for assessing LLM performance. Judges may have different preferences or criteria, leading to varying assessments. While a ground truth standard is useful, these discrepancies highlight the need for diverse perspectives to fully understand LLM performance.

5 Discussion

Practical Impact Our framework significantly improves the labeling process in NLP. By using LLMs for initial annotation, incorporating human oversight, and leveraging stronger LLMs for evaluation, we achieve a balanced solution that combines AI efficiency with human accuracy. This approach reduces costs and time compared to fully manual labeling, while maintaining high quality. It is practical for real-world applications like labeling news articles, analyzing social media, and processing customer feedback. For elections, this method can help analyze political content, fact-check claims, and track misinformation trends at scale.

Limitations and Future Work Our findings demonstrate LLMs' effectiveness as annotators, though potential biases in the annotation remain a concern, warranting further investigation [6]. Research shows LLM judges exhibit their own biases [4], impacting assessment reliability. They often favor their own responses [38] and are susceptible to order and length biases [20]. Prompt adjustments can alter LLM judgments, necessitating the use of multiple judges, including humans and various LLM models.

Despite controlled temperatures and carefully designed prompts, annotation responses and LLM judgments exhibit inherent stochasticity. These models rely on internal confidence mechanisms rather than explicit probability scores [36]. Potential mitigation methods include majority voting or aggregating multiple LLM responses. Future work will focus on incorporating associated images collected during initial data scraping, with the aim of developing a multi-modal model that leverages both textual and visual information for more robust analysis.

6 Conclusion

In conclusion, this study highlights the potential of open-source LLMs as effective annotators for political misinformation detection, demonstrating their ability to produce annotations that align closely

with human judgments. The variations in evaluation between different LLM judges highlight the need for multiple evaluation methods to comprehensively assess performance. While LLMs provide a promising and cost-effective approach for annotation, integrating them with human oversight enhances the reliability of the results. This framework has significant implications for improving labeling processes in NLP applications, especially in analyzing political content for sentiment analysis and misinformation detection. Future work should explore multi-modal approaches that incorporate visual data to further enhance political misinformation annotation and combat misinformation effectively.

7 Social Impacts Statement

Bias is inherently subjective, and its definitions widely vary across different disciplines. While our work addresses misinformation for news media, it is not intended to cover the full range of biases that exist globally. We have made possible efforts to mitigate bias, but acknowledge that our approach may have its limitations. Additionally, while our methods aim to identify bias and ensure accurate fact-checking, it is important that the data and techniques we release are used responsibly and ethically. The same data, if misused, could unintentionally fuel the spread of hate, which is not our intention. We have employed LLMs as annotators and evaluators, with human oversight included in the process. However, there is still a possibility of inherent biases in LLMs, and we recommend cautious use of both the data and methods developed. Lastly, while we focus on addressing misinformation that affects certain groups, our intent is not to exacerbate the sentiments of those already impacted by systematic inequalities.

References

[1] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, M. Zahedivafa, J. D. Bermeo, M. Korobeynikova, and F. Gilardi. Open-source llms for text annotation: A practical guide for model setting and fine-tuning, 2024.

[2] Y. Bang, D. Chen, N. Lee, and P. Fun. Measuring political bias in large language models: What is said and how it is said. In *62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[3] D. Barman, Z. Guo, and O. Conlan. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16, 2024.

[4] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, P. Zhou, Y. Wan, and L. Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

[5] J. Choi, J. Yun, K. Jin, and Y. Kim. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. *ArXiv*, abs/2404.09682, 2024.

[6] A. Das, Z. Zhang, F. Jamshidi, V. Jain, A. Chadha, N. Raychawdhary, M. Sandage, L. Pope, G. Dozier, and C. Seals. Investigating annotator bias in large language models for hate speech detection. *arXiv preprint arXiv:2406.11109*, 2024.

[7] B. Ding, C. Qin, L. Liu, L. Bing, S. R. Joty, and B. A. Li. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[8] B. Ding, C. Qin, L. Liu, Y. K. Chia, S. Joty, B. Li, and L. Bing. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.

[9] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, and et al. Measuring the carbon intensity of AI in cloud instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894, 2022.

[10] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, and et al. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[11] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[12] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), 2023.

[13] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, and et al. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, 2023.

[14] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, and et al. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190. Association for Computational Linguistics, 2024.

[15] R. Hernandes. Using gpt to assess the political bias of news outlets. Leverhulme Centre for the Future of Intelligence Blog, 2024. Accessed September 3, 2024.

[16] F. Huang, H. Kwak, and J. An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297, 2023.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[18] H. Kim, K. Mitra, R. L. Chen, S. Rahman, and D. Zhang. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*, 2024.

[19] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, and et al. AlpacaEval: An Automatic Evaluator of Instruction-following Models, May 2023.

[20] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, and C. Tao. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*, 2024.

[21] H. Liu, J. Thekinen, S. Mollaoglu, D. Tang, J. Yang, Y. Cheng, and et al. Toward annotator group bias in crowdsourcing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1797–1806, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[22] N. Micallef, V. Armacost, N. Memon, and S. Patil. True or false: Studying the work practices of professional fact-checkers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[23] Microsoft. Phi-3-medium-128k-instruct. 2024.

[24] A. H. Nasution and A. Onan. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900, 2024.

[25] J. Ni, M. Shi, D. Stammbach, M. Sachan, E. Ash, and M. Leippold. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. *Association for Computational Linguistics*, 2024.

[26] OpenAI. Introducing chatgpt, 2022.

[27] OpenAI. Gpt-4 technical report, 2023.

[28] OpenAI. Hello gpt-4o, 2024.

[29] M. Pavlovic and M. Poesio. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110. ELRA and ICCL, May 2024.

[30] S. Raza, O. Bamgbose, V. Chatrath, S. Ghuge, Y. Sidyakin, and A. Y. M. Muaad. Unlocking bias detection: Leveraging transformer-based models for content analysis. *IEEE Transactions on Computational Social Systems*, 2024.

[31] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding. Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, 237:121542, 2024.

[32] Z. Tan, A. Beigi, S. Wang, R. Guo, A. Bhattacharjee, B. Jiang, and et al. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

[33] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[35] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, and et al. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024.

[36] Z. Yan. Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). *eugeneyan.com*, Aug 2024.

[37] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103. Association for Computational Linguistics, 2023.

[38] T. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

[39] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, and et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[40] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, and et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

[41] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, and et al. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[42] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

A Appendix

A.1 Prompt for Annotation

The full prompt used for the annotation phase, as described in Section 3.2.

Prompt used for Annotation

role: user,

content: You are a helpful news fact-checking bot trained to assess the accuracy of information. Your task is to analyze the given article and determine whether it is *Factually Correct* or *Factually Incorrect*.

Fact-checking is the methodical process of verifying claims in public discourse or media reports. It is vital for countering misinformation and disinformation, thereby enhancing public knowledge and trust. Consider the following in your evaluation:

Misinformation: Incorrect or misleading information shared without intent to harm.

Disinformation: Information that is knowingly false, often prejudiced, and disseminated with the intent to mislead.

Your analysis should include:

Verification of key claims against multiple reliable sources. Identification of logical fallacies or statements that may mislead readers. Assessment of the context in which the information was presented, including the source's history and potential motivations. Evaluation for any presence of hate speech, linguistic harm, or intent to spread prejudice.

Provide your assessment in the following format:

Classification: [Factually Correct/Factually Incorrect]

Explanation: Provide a concise, evidence-based explanation for your classification. Reference specific examples from the article and contradicting evidence from trusted sources, if applicable.

Ensure to remain objective, basing your assessment strictly on facts and evidence rather than opinions or biases.

Article to analyze: article

A.2 LLM-as-a-Judge Prompt

Prompt used by the LLM Judge

Prompt: `Determine which annotation model better aligns with the given label and text.`

Text: `[Insert Text Here]`
Label: `[Insert Label Here]`

Annotation from all Models: `[Insert annotation from Model A, B, C, etc. here]`

Evaluation Criteria:
- `Alignment Assessment: Evaluate which model's annotation aligns better with both the text and the label.`
- `Instruction: Indicate whether Model A, B, C, etc. provides an annotation that better fits the given label and text context.`

`Provide your evaluation and a brief justification.`

A.3 Dataset News Sources and Topics

Topics	Candidates	Party Affiliations
• Healthcare • Climate Change • Education • Foreign Policy • Tax Reforms • Social Justice • Racial Equality • Gender Equality • Economic Inequality • Immigration • Gun Control • Culture-war • Abortion • Equality • Democracy • Environmental Policy • Technology and Innovation • Veterans Affairs • Public Safety • Mental Health • Drug Policy • Employment • Trade Policies • International Relations • Judicial Appointments • International Relation	• Presidential Candidates • Congressional Elections • House of Representatives • State Governors • State Legislature • Local Government • Mayor • City Council • County Officials • Judicial • Federal • Prime Ministerial • Parliamentary • Premier • Provincial Legislature • Municipal • Territorial	• Democratic • Republican • Independent • Libertarian Party • Green Party • Constitution Party • Liberal Party of Canada • Conservative Party of Canada • New Democratic Party • Bloc Québécois • Green Party of Canada • People's Party of Canada • Tories

Table 3: Topics, Candidates, and Party Affiliations scraped for the dataset.

• CNN • BBC • The New York Times • The Guardian	• CBS News • ABC News • Fox News • Al Jazeera	• Reuters • Associated Press • Bloomberg • USA Today	• Real Clear Politics • PEW Research • CBC • Global News

Table 4: News Sources scraped for the dataset.

A.4 Performance and Inference Time Comparison of LLM Annotation Methods

Method	Acc.	Prec.	Rec.	F1	Time
Llama-3-8B (zero-shot)	80.4%	82.3%	80.4%	80.7%	3 hr. 23 min.
Llama-3-8B (5-shot)	**89.3%**	**89.3%**	**89.3%**	**89.2%**	5 hr. 5 min.
Llama-3.1-8B (zero-shot)	74.5%	81.0%	74.5%	74.7%	3 hr. 58 min.
Llama-3.1-8B (5-shot)	82.1%	86.4%	82.1%	82.3%	4 hr. 15 min.
Mistral-7B (zero-shot)	80.0%	79.7%	80.0%	79.7%	3 hr. 10 min.
Mistral-7B (5-shot)	82.5%	83.1%	82.5%	81.8%	4 hr. 5 min.
Gemma-2-9b (zero-shot)	77.6%	82.0%	77.6%	77.9%	3 hr. 5 min.
Gemma-2-9b (5-shot)	83.3%	85.4%	83.3%	83.6%	5 hr. 20 min.
Phi-3-med (zero-shot)	80.7%	80.8%	80.7%	80.7%	2 hr. 40 min.
Phi-3-med (5-shot)	86.9%	87.1%	86.9%	86.7%	3 hr. 52 min.

Table 5: Comparison of different LLM annotation's accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-score (F1) against the gold labels in zero-shot and 5-shot settings. The best results are in **bold**.

This table, an extension of Table 1 from the main paper, includes an additional column indicating the inference time for each model. To conserve space, the model names in Column 1 have been abbreviated.

A.5 Gold Data Composition Process

Our gold data review team consists of twelve volunteer members representing a diverse range of disciplines and demographics. The team includes three undergraduate students, four master's students, and two professionals with doctoral degrees, spanning the fields of computer science, media studies, and political science. We ensured balanced gender representation, with six female and six male members from diverse cultural and ethnic backgrounds. Two subject matter experts in computer science and linguistics led the annotation efforts. The team followed detailed guidelines for identifying disinformation, ensuring consistent criteria across all annotations. After completing the annotation process, a sample of annotations was evaluated using Label Studio. Disagreements in labeling were resolved through consensus discussions, ensuring high agreement and reliability in the final annotations.