Explainable Numerical Claim Verification

Peter Røysland Aarnes
peter.r.aarnes@uis.no
University of Stavanger
Stavanger, Norway

Abstract

The rapid proliferation of mis- and disinformation in the digital age highlights the urgent need for scalable, transparent, and trustworthy automated fact-checking systems. Large Language Models (LLMs) offer strong language understanding capabilities but suffer from opacity and brittleness, particularly in reasoning over numerical claims. This work explores how Explainable Artificial Intelligence (XAI)—through the lens of counterfactual explanations and adversarial training—can be used to systematically evaluate and improve the robustness of LLMs against perturbed numerical inputs. We propose a framework that employs counterfactual generation to both probe LLM reliability and generate user-appropriate explanations. Through empirical evaluations using a large-scale numerical fact-checking dataset (QuanTemp), we show that even state-of-the-art LLMs are susceptible to subtle numerical perturbations, impacting verdict accuracy. Our methodology contributes a dual-purpose diagnostic and training strategy that not only bolsters robustness but also enables both global and local interpretability—thereby improving explainability in automated fact-checking systems.

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

XAI; Explanability; LLMs; Automated Fact-Checking; Adversarial Attacks; Adversarial Training; LLM Probing

ACM Reference Format:
Peter Røysland Aarnes. 2025. Explainable Numerical Claim Verification. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3746252.3761666

1 Introduction

The rise of mis- and disinformation in digital media has underscored the urgent need for scalable, reliable, and explainable fact-checking systems. Factual misinformation frequently includes specific numerical values, such as statistics about healthcare costs, economic trends, or climate metrics, that are easy to manipulate and challenging to verify at scale. Despite advances in automated fact-checking powered by Large Language Models (LLMs), a major blind spot persists: these systems often falter when it comes to reasoning over numerical claims, especially when those claims are subtly perturbed [1, 32, 36].

Numerical claim verification poses a unique set of challenges for LLMs. Unlike categorical or semantic claims that can be verified via shallow linguistic cues or retrieved knowledge, numerical claims demand arithmetic operations, comparative understanding, and contextual reasoning across potentially long and complex evidence documents. Our preliminary experiments show that even state-of-the-art models such as GPT-4 and Gemini struggle when minor numerical changes are introduced into otherwise true claims. These perturbations, whether rounding, approximations, or unit changes, can induce drastic failures in otherwise high-performing models.

Equally concerning is the lack of transparency in these models' decision-making processes. When LLMs make fact-checking predictions, especially incorrect ones, users are often left in the dark about why. This "black box" behavior is particularly problematic in contexts requiring high-stakes decision-making or public trust [23, 30]. Explainable Artificial Intelligence (XAI) offers a pathway toward increased transparency and accountability in these systems, but much of the existing XAI literature focuses on classification tasks in vision or tabular domains, and not on the intricacies of natural language fact-checking.

In this project, we propose to bridge this gap by leveraging *counterfactual explanations*, i.e., minimally perturbed versions of input claims, to both evaluate and improve LLM behavior on numerical claims. Counterfactuals serve a dual purpose. First, they act as a diagnostic tool, exposing vulnerabilities in model reasoning through controlled input alterations. Second, they provide a natural foundation for user-facing explanations, showing how slight changes in claim content would alter model predictions. For instance, when a model fails to distinguish between "Unemployment is 5%" and "Unemployment is five percent," a counterfactual explanation can expose its limited capacity to interpret different forms of numerical representation.

To operationalize this idea, we distinguish between two levels of explanation: **global** and **local**. At the global level, explanations summarize model behavior across a wide range of perturbed claims. These are intended for ML engineers or researchers seeking to improve model robustness. At the local level, explanations are instance-specific, helping end-users understand how a particular verdict was reached and whether it can be trusted in light of small input changes.

This dual-explanation framework allows us to diagnose, train, and ultimately explain LLMs in a unified way, grounded in counterfactual reasoning and focused specifically on the neglected but

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761666

critical domain of numerical claim verification. Through rigorous testing on the QuanTemp dataset [32], our preliminary findings reveal significant model brittleness. In response, we propose a full pipeline encompassing counterfactual generation, adversarial fine-tuning, and explanation generation tailored to user needs. The goal is to develop LLM-based fact-checking systems that are not only more robust but also meaningfully interpretable to both technical and non-technical stakeholders.

We formulate the following research questions to address these challenges:

- **RQ1**: How brittle are state-of-the-art LLMs to perturbation of numerical values?
- **RQ2**: How to make LLMs robust to adversarial numerical perturbations?
- **RQ3** What constitutes insightful fact-checking explanation for different users and how to generate them?

2 Background

2.1 Explainable Artificial Intelligence

XAI aims to improve user understanding and trust in AI and ML outputs, especially since models can sometimes arrive at correct predictions for incorrect reasons, risking failures under changing contexts. With the growing sophistication of AI, especially LLMs, and tightening regulations[1], interpretability techniques have become increasingly critical.

Earlier post-hoc XAI methods like LIME [28] and SHAP [25] have declined in relevance with the rise of LLMs, which are often closed-source and comprise billions of parameters [6, 12]. This project focuses on the computationally efficient *Counterfactual Explanations*, which reveal model sensitivity to input changes [34], helping uncover potential vulnerabilities [35].

Counterfactuals also relate to *adversarial attacks*, where minor input perturbations can drastically affect model outputs [31]. In open-source models, word embedding perturbation using top-k nearest neighbors often yields nonsensical text [20, 37]. Closed-source models are typically attacked via brute-force methods like modifying keywords or token-level transformations, typos, character edits, or common misspellings, though such methods can be mitigated by spell-checking defenses [9, 19, 22, 29].

For transformer models [33], adversarial techniques include paraphrasing, back-translation, and synonym substitutions to preserve semantics [7, 9, 14, 21, 37]. To build robustness, adversarial training is widely used, blending clean and perturbed inputs during fine-tuning [8, 26].

Despite these defenses, LLMs struggle with numerical reasoning [2, 15, 17, 27], particularly under numerical perturbations [1, 36]. This project proposes using counterfactual perturbation of numerical values in factual claims to assess AI models' ability to reason over quantitative information in claims and their evidence.

2.2 The Automated Fact-checking Pipeline

An automated fact-checking framework is typically a multifaceted task, split into different stages. First, *claim detection* is used to identify statements that need verification. However, not every claim is worth using resources on for further verification [4]. That is why it is common to classify a claim for check-worthiness, where the most important factors are whether a claim is factually verifiable and in the public's interest. [3, 11, 16].

Thereafter, evidence is retrieved, and the relevant, trusted information is categorized using *stance detection* to determine whether it supports or refutes a particular claim [10].

Verdict prediction, often accompanied by *justification production* encompass the last stages. The verdict is typically classified as support or refuted, something in-between (i.e. "mostly supported/refuted") or "not enough evidence". Whereas justification production should clearly outline how a verdict was reached given the evidence, what assumptions or commonsense knowledge were involved, and the reasoning that connects them–all in a way that is easily understandable to the user in order to persuade them. A challenging task, as challenging existing beliefs, users often reject the new information outright, and backfire–further strengthening existing beliefs [18]

In the following section, we outline the proposed methodology and present the areas of interest that will guide this project.

3 Methodology and Research Questions

This project investigates the intersection of XAI and automated fact-checking, focusing on leveraging counterfactual explanations for numerical claims. To this end, we utilize the QuanTemp dataset [32], where we implement numerical perturbation methods for creating counterfactual examples. Our project is structured around three key areas of interest:

3.1 Perturbation of numerical values

This research investigates the vulnerabilities of LLMs when exposed to perturbed inputs though adversarial examples and counterfactual claims.

We have conducted preliminary experiments using the QuanTemp dataset [32], which provides numerical claims paired with extensive long-form evidence, typically ranging from one to several thousand words per document. Our perturbation methodology is inspired by prior work from Xu et al. [36] and Akhtar et al. [1], enabling a targeted evaluation of LLM performance under controlled semantic shifts in numerical representation within claims. Notably, Akhtar et al. [1] explored models' numerical reason capabilities are on tabular numerical data. We employ similar perturbation techniques, such as converting numerical figures into their word-based form (e.g., "1,000$" would be transformed into "one thousand dollars").

Another method involves perturbing numerical claims by replacing exact figures with a range. For instance, an unperturbed true claim such as "80 percent of healthcare dollars are spent by 20 percent of the population" could be transformed into a perturbed, counterfactual false claim: "Between 50 and 60 percent of healthcare dollars are spent by 20 percent of the population." Additionally, we investigate model behavior when the claim remains true, but the explicit figure is replaced with a numerical range, e.g., "between 75 and 85 percent" instead of "80 percent". We also plan to deploy perturbation transformations by rounding numbers to approximation

[1] https://artificialintelligenceact.eu/

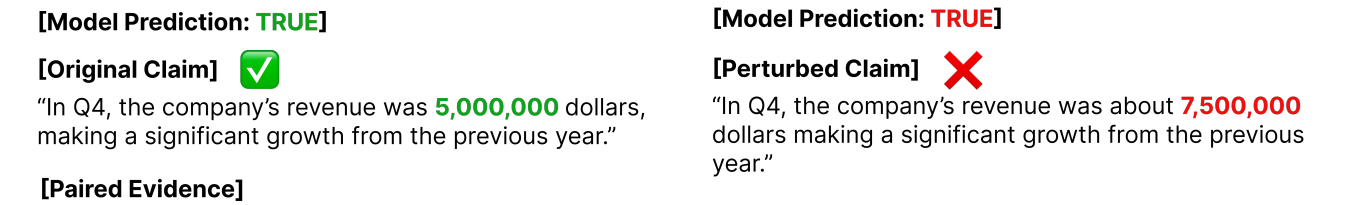

Figure 1: Impact of Approximation Perturbation on Model Predictions: The figure contrasts an original claim (left) with its approximation counterpart (right), which has a flipped label. The model misclassifies the perturbed claim as true, despite it being false, illustrating the challenge perturbations pose in maintaining accuracy. The model is paired with the evidence as a part of the input.

and adding the prefix "about" before the given value to indicate numerical approximation.

Our experiments will specifically focus on claims that are originally labeled as true and supported by their accompanying evidence. By systematically perturbing these numerical values while preserving the original semantic context, we generate two variations of any given claim: one that remains supported by the evidence, and one that is not. These perturbed claims, paired with long-form evidence, allow us to assess how effectively LLMs maintain accurate veracity predictions. This requires both numerical comprehension–inferring numerical values despite varied representations in claim and evidence–and the ability to make these inferences over extended contexts. Figure 1 illustrates how a model could misclassify a numerical claim where the numerical value in the claim does not comply with the paired evidence.

Preliminary Evaluation: The preliminary results suggest that even state-of-the-art LLMs, such as GPT-4o and Gemini 1.5 Pro, alongside smaller open-weight models, remain highly susceptible to counterfactual claim-evidence pairs across various perturbation settings. For example, the most robust model tested (GPT-4o) suffers a 20-40% decrease in accuracy for certain perturbation settings. In light of findings by Liu et al. [24], which indicate that LLMs struggle to identify relevant information in long contexts, investigating how these models reason about numerical values over extensive evidence documents offers particularly valuable insights

3.2 Adverserial fine-tuning of LLMs

Expanding on the aforementioned area of interest, and preliminary insights, we want to investigate if adversarial training can mitigate vulnerabilities of numerical counterfactual claims.

We plan to explore fine-tuning techniques such as Low-Rank Adaptation (LoRA) [13], which has shown strong performance in adapting open-weight LLMs to specific tasks with minimal computational overhead—in addition to fine-tuning proprietary closed-sourced models with platform specific tools. Nevertheless, adversarial training poses a high risk of degrading the generalization abilities of fine-tuned models [5]. This, however, may be a worthwhile trade-off, provided the performance decline is modest and the gain in robustness is substantial.

3.3 Generating user-friendly explanations

To make the output of LLMs more transparent, the explanation must serve different user groups. Our goal is to generate explanations based on the findings of robustness tests and diagnostic benchmarks, constructed using counterfactual examples. We differentiate *Global Explanation* and *Local Explanation*

For the data scientists or machine learning engineers training the LLMs, it would be useful to provide the summary of the vulnerabilities of the LLMs so that they can take steps to enhance the reliability of the models at the *Global Explanation* level. For example, given the preliminary results, described at the beginning of Section 3, would provide how certain counterfactual perturbations impact the model in general.

At the local explanation level, we focus on the end-user. Here, instance-specific explanations help users gain confidence in an LLM's prediction for a particular claim. For example, if we perturb a claim's critical details, like changing "Unemployment in Norway is **5%**" to "Unemployment in Norway is **50%**," yet the model's prediction remains unchanged, it suggests the model isn't paying attention to those crucial parts of the claim. This type of local insight helps reveal what specific claim features the model is truly sensitive to.

Therefore, a core objective of this research is to establish robust methodologies for generating these insightful explanations, directly utilizing the findings from our robustness tests.

4 Conclusion

In this paper. we have outlined the methodological framework for advancing the automated fact-checking by focusing making LLMs more robust against perturbed numerical values in claims. Central to this framework is the systematic generation and application of counterfactual examples. We leverage them as a diagnostic tool to rigorously test model robustness against different numerical perturbations techniques, as a training mechanism through adversarial fine-tuning to enhance robustness, and as a basis for generating

global and local explanations to potential model vulnerabilities. By exploring counterfactuals, their creation to their application in both training and evaluation–this project aims to improve the reliability and transparency of LLM-based fact-checking systems.

5 Acknowledgments

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

References

[1] Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (ACL '23)*. 15391–15405.

[2] Hadeel Al-Negheimish, Pranava Madhyastha, and Alessandra Russo. 2021. Numerical reasoning in machine reading comprehension tasks: are we there yet?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*. 9643–9649.

[3] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *Findings of the Association for Computational Linguistics: EMNLP 2021 (EMNLP'2021)*. 611–649.

[4] Isabelle Augenstein. 2021. Towards Explainable Fact Checking. arXiv:2108.10274 [cs]

[5] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '21)*. 4312–4321.

[6] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *International Conference on Machine Learning (ICML '18)*. 883–892. arXiv:1802.07814 [cs]

[7] Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers) (ACL '18)*. 650–655.

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2014). arXiv:1412.6572 [cs, stat]

[9] Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A Survey of Adversarial Defences and Robustness in NLP. *Acm Computing Surveys* 55 (2023). Issue 14s.

[10] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[11] 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)*. 1803–1812.

[12] Henning Heyen, Amy Widdicombe, Noah Y. Siegel, Maria Perez-Ortiz, and Philip Treleaven. 2024. The Effect of Model Size on LLM Post-Hoc Explainability via LIME. arXiv:2405.05348 [cs]

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs]

[14] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (2020), 8018–8025.

[15] Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang, and Sung-Hyon Myaeng. 2021. Have You Seen That Number? Investigating Extrapolation in Question Answering Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*. 7031–7037.

[16] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice* 2, 2 (2021), 1–16.

[17] Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. 2021. Adversarial Examples for Evaluating Math Word Problem Solvers. In *Findings of the association for computational linguistics: EMNLP 2021 (EMNLP '21)*. 2705–2712.

[18] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131.

[19] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized Perturbation for Textual Adversarial Attack. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (NAACL '21)*. 5053–5069.

[20] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *The Internet Society (NDSS '19, Vol. abs/1812.05271)*.

[21] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP '2020)*. 6193–6202.

[22] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep Text Classification Can be Fooled. In *Proceedings of the 27th international joint conference on artificial intelligence (IJCAI '18)*. 4208–4215.

[23] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[24] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[25] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*. 4768–4777.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations (ICLR '18)*. arXiv:1706.06083 [cs, stat]

[27] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL '21)*. 2080–2094.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 1135–1144.

[29] Tom Roth, Yansong Gao, Sharif Abuadbba, Surya Nepal, and Wei Liu. 2021. Token-modification Adversarial Attacks for Natural Language Processing: A survey. *AI Communications* (2021), 1–22.

[30] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Publishing Group UK London* 1, 5 (2019), 206–215. arXiv:1811.10154 [cs, stat]

[31] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. arXiv:2310.10844 [cs]

[32] Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A Real-World Open-Domain Benchmark for Fact-Checking Numerical Claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Sigir '24)*. 650–660.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS' 17, Vol. 30)*.

[34] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017).

[35] Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A Survey on Natural Language Counterfactual Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024 (EMNLP '24)*. 4798–4818.

[36] Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han, and Dongmei Zhang. 2022. Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*. 7950–7966.

[37] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. *ACM Transactions on Intelligent Systems and Technology* 11, 3, Article 24 (2020).