

MedVerify: A Severity-Aware System for Verification of Medical Claims Generated by Conversational AI Systems

Aditi Londhe¹, Aarya Badhe², Tanish Charthankar³, Shrujan Kundgulwar⁴, Sachin Shelke⁵

^{1,2,3,4,5}*SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India*

aditivlondhe@gmail.com¹, ryabadhe11@gmail.com², vctanish7@gmail.com³, shrujankundgulwar47@gmail.com⁴, sachindshelke@pict.edu⁵

Abstract.

Large Language Models (LLMs) are increasingly used for health-related queries but may generate hallucinated or medically unsupported information, creating significant risks for patient safety. This paper presents MedVerify, a severity-aware verification tool that identifies, prioritizes, retrieves, and validates medical claims generated by conversational AI. The system combines structured drug-fact verification using FDA drug-label data with evidence-based Natural Language Inference using biomedical text. Implemented as a browser extension, MedVerify verifies chatbot responses against trusted biomedical databases and provides claim-level verification results, severity indicators, and supporting references. This approach enables transparent, evidence-based validation of AI-generated medical information in real time.

Keywords. Medical Claim Verification, Conversational AI, Large Language Models, Medical Hallucination, Severity-Aware Verification.

1. INTRODUCTION

Large Language Models (LLMs) are increasingly used to obtain medical information, increasing reliance on conversational AI in healthcare. Users frequently seek information about symptoms, medications, treatments, and health conditions. Although LLMs can provide fluent and informative responses, they may generate hallucinated content that appears factual but lacks reliable evidence or cannot be verified. Such hallucinations are particularly concerning in healthcare because incorrect dosage instructions, fabricated drug interactions, or unsupported treatment recommendations may cause serious harm. Existing research on hallucination detection and fact-checking primarily focuses on general-domain information rather than medical knowledge. Medical verification is more challenging because biomedical information requires structured knowledge, authoritative evidence, and high reliability. To address this gap, this paper presents MedVerify, a browser-based medical claim verification tool. MedVerify analyzes conversational AI outputs, identifies medical claims, retrieves supporting evidence from trusted biomedical databases, and validates

claims using evidence-based verification methods. The system provides users with reliable, transparent, and evidence-supported verification of AI-generated medical information.

2. RELATED WORK

Research in hallucinations has increased in natural language generation (NLG). Ji et al. [6] gave an extensive review and taxonomy of hallucinations in NLG and created a benchmark for future research in hallucination detection. TruthfulQA, developed by Lin et al. [9] is a dataset that helps in assessing the ability of language models to faithfully emulate human lies and showed that bigger models are no more truthful than smaller models. In the field of medicine, Chelli et al. [4] performed a systematic analysis of hallucinations in medical LLMs.

A Retrieval-Augmented Generation (RAG) model was proposed by Lewis et al. [8], which verifies the generated outputs using evidence retrieval and represents a vital constituent of the evidence retrieval module in MedVerify. There exist pretrained models for biomedical language understanding in the form of BioBERT by Lee et al. [7], PubMedBERT by Gu et al. [5], and BioMedBERT by Chakraborty et al. [3]. In the context of entailment relations for medical QA systems by Abacha and Demner-Fushman [1]. Alsentzer et al. [2] have come up with clinical BERT embeddings for biomedical purposes. Luo et al. [10].

3. KEY CONTRIBUTION

MedVerify introduces risk-based medical claim classification, combines structured drug verification with evidence-based Natural Language Inference, and provides real-time chatbot response verification through a Chrome Extension without modifying underlying AI models.

4. METHOD, EXPERIMENTS AND RESULTS

4.1. System Architecture

The verification pipeline of MedVerify is modular and multi-stage. First, when MedVerify receives an AI response (R), it collects a list of all of the potential claims that were made in that response.

$$C = \{c_1, c_2, \dots, c_n\} \quad (1)$$

Next, all potential claims (C) are run through a sequential process as shown in Table 1:

$$R \rightarrow C \rightarrow S(C) \rightarrow E(C) \rightarrow V(C) \rightarrow Score(C) \quad (2)$$

where:

C = Extracted Claims, S(C) = Severity Classification, E(C) = Retrieved Evidence, V(C) = Verification Outcome, Score(C) = Final Confidence Score.

4.2. Medical Claim Extraction

The primary objective of this stage is to determine whether there are any medical claims (which may include references to medication, suggestions of beneficial results from medications, dosage descriptions, treatment suggestions, side effect disclosures, diagnostic procedures and interactions between medications) in the AI-derived response.

4.3. Severity Classification

MedVerify categorizes every claim c_i with one of four distinct severity labels:

$$S(c_i) \in \{Critical, Serious, Moderate, Low\} \quad (3)$$

The categorization is made using a rules-based classification technique which relies on three primary factors to classify whether a claim should be assigned a severity label: (i) dosage expressions (e.g. mg, frequency) present (ii) drug-related entities such as medications, interactions (iii) clinical action verbs such as prescribe, diagnose and treat. Severity weights are assigned to each of the four severity levels to be used when calculating the total score for each claim as follows:

$$SeverityWeight(c) = 1.5 (Critical), 1.2 (Serious), 1.0 (Moderate), 0.8 (Low) \quad (4)$$

4.4. Evidence Retrieval

The evidence retrieval component finds trustworthy biomedical literature (peer-reviewed), trustworthy labels & safety information for medications (OpenFDA), trustworthy consumer health information (MedlinePlus), and Wikipedia (general context and fallback information) to support or not support statements made by individual claims. Each claim c has an associated query that is generated using the keywords associated with the entities connected with the claim:

$$Q(c) = keywords(entities(c)) \quad (5)$$

The system obtains the top k document(s) ($k = 3-5$) whose keyword relevance score is the highest and has the most shared entity types and then pulls out the relevant sentence(s) needed to verify that individual claim(s) c are true:

$$E(c) = \{e_1, e_2, \dots, e_k\} \quad (6)$$

4.5. Hybrid Verification Framework

The MedVerify Tool uses a Hybrid Verification framework which includes two ways to verify claims: Structured Drug Fact Verification: Uses OpenFDA databases for verification against FDA approved drug label dosage, contraindications, & interaction data structured in the database, allowing for precise comparison of the attributes at the attribute level. Evidence-Based NLI Verification: Utilizes a biomedical NLI to verify general medical claims retrieved from electronic sources such that each claim is classified as either Entailment, Contradiction, or Neutral. Any claim that has been classified as a contradiction will be flagged as a disputed claim.

4.6. Medical Confidence Scoring

Verification results, source authority, and severity weight are combined into a final confidence score.

$$Score(c) = V(c) \times Authority(e) \times SeverityWeight(c) \quad (7)$$

Claims lacking sufficient evidence are labelled as Cannot Verify.

4.7. System Implementation

MedVerify, built as a Manifest V3 Chrome extension, monitors chatbot responses using PubMed, OpenFDA, and MedlinePlus APIs for evidence-based verification.

4.8. Evaluation

MedVerify analysed 100 medical claims, identifying various hallucination severities and verifying supporting evidence within an average response time of 2.3 seconds.

Table 1. Performance of MedVerify.

Category	Precision	Recall	F1	Samples
Critical	0.91	0.88	0.89	25
Serious	0.87	0.85	0.86	25
Moderate	0.82	0.80	0.81	25
Low	0.78	0.76	0.77	25
Overall	0.85	0.82	0.83	100

5. DISCUSSION

MedVerify is most accurate for claims of critical severity, while moderate and low-severity claims tend to be less accurate owing to complicated medical data. The limitations involve reliance on medical evidence that may be available, limited coverage of rare illnesses and recently approved medicines, and complications with understanding medical language. MedVerify facilitates medical decision-making but cannot substitute professional doctors.

6. CONCLUSION

MedVerify integrates biomedical NER, severity classification, multi-source evidence retrieval, and hybrid NLI-based verification. Testing on 100 claims achieved an overall F1-score of 0.83, with 0.89 for critical claims and 2.3-second processing. Future work includes advanced NLI models, broader databases, multilingual support, and user studies.

7. REFERENCES

- [1] Abacha, S.B. & Demner-Fushman, D. 2020. A question-entailment approach to question answering. *CEUR Workshop Proc.* 2696: 1–10.
- [2] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T. & McDermott, M. 2019. Publicly available clinical BERT embeddings. In *Proc. Workshop Biomed. Natural Language Process. (BioNLP)*, Florence, Italy: 72–78.
- [3] Chakraborty, S., Saha, S. & Bandyopadhyay, S. 2020. BioMedBERT: A pre-trained biomedical language model for question answering and information retrieval. In *Proc. Int. Conf. Comput. Linguistics (COLING)*, Barcelona, Spain: 669–679.
- [4] Chelli, M., Descamps, J., Lavoué, A., Trojani, M., Azar, P., Deckert, J., Raynier, A., Clowez, J.-B., Kün, S. & Ruetsch-Chelli, A. 2024. Hallucination in medical LLMs: A systematic evaluation. *JMIR Med. Informatics* 12(1): e53164.

- [5] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* 3(1): 1–23.
- [6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. & Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55(12): 1–38.
- [7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. & Kang, J. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4): 1234–1240.
- [8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Event, vol. 33: 9459–9474.
- [9] Lin, S., Hilton, J. & Evans, O. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland: 3214–3252.
- [10] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H. & Liu, T.-Y. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinformatics* 23(6): 1–10.

Biographies



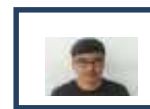
Aditi Londhe. Student, SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India.



Aarya Badhe. Student, SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India.



Tanish Charthankar. Student, SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India.



Shrujan Kundgulwar. Student, SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India



Sachin Shelke. Assistant Professor, SCTR's Pune Institute of Computer Technology, Department of Information Technology, Pune, India.