Proceedings of the 5th International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/2027.37537

Double-Layer Semantic Disambiguation of Pediatric Medical Spoken Abbreviations Based on 'Lung' and 'Allergy' Examples

Lejia Wang

CDUTCM-KEELE Joint Health and Medical Sciences Institute, Chengdu University of Traditional Chinese Medicine, Chengdu, China
wang.lejia@stu.cdutcm.edu.cn

Abstract. Traditional abbreviation disambiguation only maps abbreviations to their full forms and cannot identify different clinical meanings of the same full form in pediatric colloquial language. This paper constructs a dual-layer disambiguation framework for word meaning and clinical semantics, taking "lung" and "allergy" as examples, fine-tunes Bidirectional Encoder Representations from Transformers (BERT) based on the IMCS-V2 corpus, and compares the zero-shot performance of DeepSeek and Tongyi Qianwen. Experiments show that the disambiguation accuracy of 'lung' is 95.0%, while the identification of the 'allergy' phenotype is only 76.6%. The zero-shot performance of large models is worse than fine-tuned BERT, and window ablation determines that the optimal setting is to consider the previous and next sentences. The results show that lightweight BERT fine-tuning outperforms zero-shot inference with large models, traditional frameworks underestimate the difficulty of this task, and this paper provides a new hierarchical evaluation benchmark for pediatric medical dialogue Natural Language Processing.

Keywords: Medical abbreviation disambiguation, Pediatric doctor-patient dialogue, Dual-layer semantic framework, Large language model, Zero-shot learning

1. Introduction

Medical abbreviations are a frequent phenomenon in clinical texts, but their ambiguity has become an important interfering factor in natural language processing. The problem of one abbreviation corresponding to multiple full forms, known as 'one word with multiple meanings,' can easily lead to misunderstandings or even diagnostic and treatment deviations in a clinical context [1]. Therefore, abbreviation disambiguation is regarded as a fundamental task in clinical natural language understanding. Existing studies mostly focus on standardized written texts, such as electronic medical records and discharge summaries [2]. Methods based on dictionaries, statistical approaches, or pre-trained models have achieved good performance on structured corpora [3]. However, the characteristics of pediatric doctor-patient communication lie in highly fragmented content, the use of colloquial expressions, and incomplete sentence structures. Parents often use everyday, non-standard

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2027.37537

expressions such as 'always allergic' or 'itchy throat,' which are difficult to be directly addressed by traditional diagnostic and treatment methods.

More critically, the existing paradigm of disambiguation tasks has fundamental limitations. Because their goal has always been limited to mapping abbreviations to their corresponding full forms, thereby addressing the superficial lexical ambiguity problem of "one word corresponding to multiple full forms" [4]. However, in real children's oral expressions, the deep semantic issue of 'the same name but different meanings' is also very common and even more concealed. Even if the abbreviation is correctly expanded to its fixed full form, this full form may still refer to entirely different clinical phenotypes in different clinical contexts, which is crucial for diagnostic and treatment decisions. However, this deep-level differentiation has long been systematically ignored by existing evaluation frameworks [5], leading to a serious overestimation of the clinical understanding ability of current models.

Internationally, medical abbreviation disambiguation has evolved from dictionary matching to BERT fine-tuning. Rajkomar and others achieved over 90% accuracy on MIMIC-III [6], and Wei and others achieved high-performance prediction with the help of large language models [7]. However, all of these works are based on English standard texts, and the tasks always stop at 'choosing the correct full form,' without involving semantic differentiation within the full form, nor adapting to pediatric oral scenarios. In China, research has long focused on the standardization of abbreviation terms and dictionary construction. At the technical level, there have been attempts at medical disambiguation based on Bi-LSTM, while the field of pediatric artificial intelligence focuses on symptom recognition and disease classification [8], and resources such as PedCorpus and PediatricsGPT have been constructed. Looking at work in China, there are also two significant gaps: the data sources are all written standardized texts and do not cover the highly colloquial expressions in pediatric dialogues; the disambiguation task is simplified to full-term mapping at the morphological level, ignoring the multiple clinical phenotypic meanings carried in different spoken contexts after the full term is determined. And, this deep semantic disambiguation has not yet been systematically evaluated in domestic pediatric NLP [2].

In pediatric terminology, although the full term 'allergy' is fixed, it often refers to different phenotypes such as allergic rhinitis, atopic dermatitis, food allergies, and allergic asthma. The word 'lung' also has the classic ambiguity between 'organ' and 'pneumonia.' These two happen to form a layered example, going from 'lexical ambiguity' to 'clinical semantic ambiguity,' making it a perfect entry point to reveal the shortcomings of the current framework. To address the gaps mentioned above, this article builds a dual-layer word meaning-clinical semantic disambiguation evaluation framework. Taking 'lung' and 'allergy' as examples, annotation and system experiments were carried out based on the IMCS-V2 pediatric dialogue corpus. Specifically, it includes: designing a two-level task with word meaning binary classification and clinical semantics four-class classification; fine-tuning BERT with 10-fold cross-validation to establish a baseline; comparing the zero-shot performance of DeepSeek and Tongyi Qianwen; and using window ablation to determine the optimal context length. This article aims to reveal how traditional single-layer evaluation frameworks seriously underestimate the difficulty of clinical semantic understanding, and to provide a new hierarchical evaluation benchmark for natural language understanding in pediatric medical dialogues, pushing the field from a 'word sense disambiguation' paradigm toward 'semantic understanding'.

DOI: 10.54254/2753-8818/2027.37537

2. Method

2.1. Data sources and selection

The original dialogue text comes from the public dataset IMCS-V2, and Python was used for data filtering, cleaning, and preprocessing. By screening pediatric consultation cases using codes, this study ultimately included 320 cases and 985 dialogue records, covering common pediatric diseases such as pneumonia, bronchitis, and diarrhea. Use Python to filter formal written texts such as physical examination reports and medical records, retaining only fragmented doctor-patient conversations, and completely preserving informal and incomplete expressions.

2.2. Annotation standards and dual-layer task definition

This article builds a five-dimensional annotation system for abbreviations, full names, standardized symptoms, conversational behaviors, and diagnoses, using Python to store and manage all the annotation data. A two-level classification task based on the ambiguity differences of two types of vocabulary: Word Sense Disambiguation (binary classification) for the polysemous term 'lung', distinguishing whether the context refers to pneumonia as a disease or the lung as a human organ.

2.3. Fine-tuning the BERT model construction

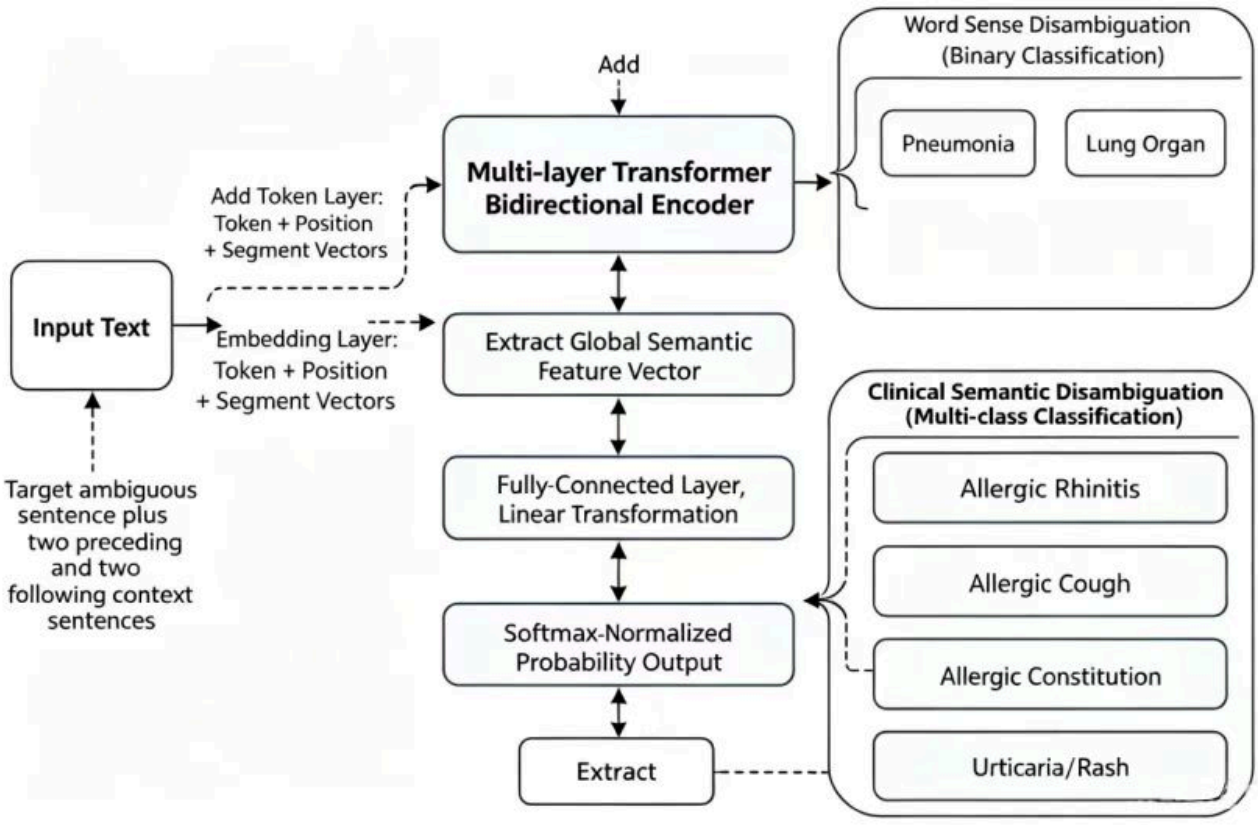

Figure 1. Fine-tuning the BERT model (picture credit: original)

As shown in Figure 1, the two layers of tasks are uniformly modeled as text multi-classification, with the basic pre-trained model chosen as Chinese BERT-wwm. Concatenate the input text with

DOI: 10.54254/2753-8818/2027.37537

two sentences of context before and after the ambiguous sentence to expand semantic cues, extract the [CLS] feature vector, and connect it to a fully connected layer with Softmax to output prediction probabilities. All experiments employed 10-fold cross-validation to avoid single-split errors, and the significance of performance differences was determined using paired sample t-tests, with the threshold set at $p<0.05$.

2.4. Zero-shot experimental setup for large models

Select the two domestic large models, DeepSeek-V3-0324 and Qwen2.5-72B-Instruct, and perform inference through OpenAI-compatible API calls. To minimize output randomness as much as possible, set temperature=0, top_p=1, max_tokens=10 uniformly, and fix the random seed at seed=42 to stabilize the generation process. Simultaneously design standardized prompts: the system prompt is fixed as 'You are a medical text classification assistant; only output the classification name without any additional explanations,' and the user prompt is 'Please determine which category of clinical phenotype the term "allergy" belongs to in the following sentence: {sentence}.' Each sample is reasoned individually following the same template. Using 73 "allergy" samples, with classification accuracy as the core evaluation metric, a comparison was made between fine-tuning BERT and the majority class baseline. Considering uncontrollable factors such as underlying server operators and caching, the average of multiple repeated experiments is taken to enhance robustness and ensure that the results are reproducible and comparable.

3. Experimental results and analysis

3.1. Dataset category distribution

All 171 ambiguous samples contain only the two high-frequency terms "lung" and "allergy." Among the 98 "lung" samples, 62.6% refer to pneumonia, and 37.4% refer to the lung organ itself, with the majority class baseline at the semantic level being 62.6%. The 73 "allergy" samples are severely imbalanced in distribution: constitutions prone to allergies account for 74.0%, allergic rhinitis 13.7%, allergic cough 6.8%, and rashes 5.5%, with the clinical semantic level majority class baseline at 74.0%. Category imbalance is a key factor limiting recognition accuracy.

3.2. BERT two-layer disambiguation experiment results

The results of 10-fold cross-validation show that the disambiguation accuracy of the lexical-semantic level for the word 'lung' is 95.0%, representing an improvement of 32.4 percentage points compared to the majority class baseline. The model can reliably distinguish between the two meanings based on surface words such as chest X-ray, inflammation, and cough medicine. The accuracy of identifying 'allergy' at the clinical semantic level is only 76.6%, with an 18.4 percentage point difference in precision between the two levels, and the difference is significant. The model accuracy is only slightly higher than the category baseline, indicating that fine-tuning BERT did not learn the detailed clinical features. It relies solely on sample frequency to preferentially predict the high-frequency category 'allergic constitution,' showing a clear issue of 'pseudo-learning'.

DOI: 10.54254/2753-8818/2027.37537

3.3. Zero-shot large model comparative experiment

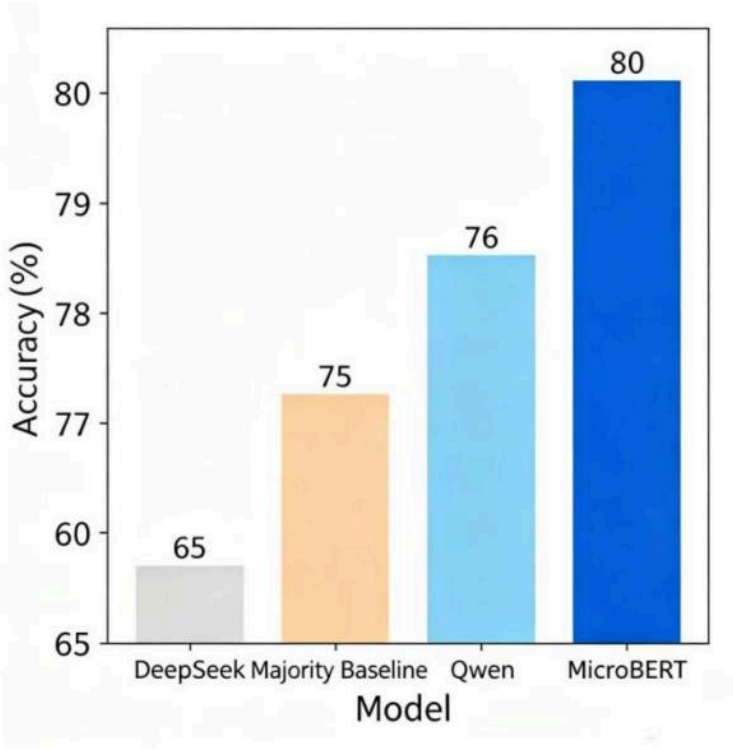

Figure 2. Comparison experiment of zero-shot large models (picture credit: original)

As shown in Figure 2, the test results for 73 "allergy" samples indicate that the performance of the two general-purpose large models is inferior to that of the fine-tuned BERT. Tongyi Qianwen has an accuracy rate of 74.0%, which is completely on par with most baseline models, and it cannot rely on context to extract fine-grained semantic features; DeepSeek's accuracy is only 69.9%, lower than the baseline, proving that the general medical knowledge built into large models is insufficient to support the fine-grained discrimination of pediatric oral language. BERT fine-tuned with only a small amount of labeled data achieves optimal results, demonstrating the irreplaceable value of domain-specific fine-tuning for specialized clinical tasks.

3.4. Statistics of model error types

All erroneous samples of BERT have been manually compiled, clearly showing the proportion of the three main causes of errors. Insufficient contextual information accounts for 42%, mainly manifesting as a lack of key diagnostic clues in short dialogue sentences; unclear clinical semantic boundaries account for 31%, and due to the presence of multiple overlapping allergic features within a single conversation segment, the distinction of phenotypes becomes ambiguous; non-standard colloquial expressions caused interference in 27% of cases; everyday terms such as 'child with lung issues' or 'lung discomfort' can interfere with the feature extraction process and reduce discrimination accuracy.

DOI: 10.54254/2753-8818/2027.37537

3.5. Context window ablation experiment

To verify the impact of context window length on the two-level disambiguation task, this study designed four control experiments based on the lexical semantic layer of the word 'lung' and the clinical semantic layer of 'allergy'. The experimental content includes:"a single sentence," "one sentence before and after," "two sentences before and after," and "three sentences before and after". The results are shown in Figure 3.

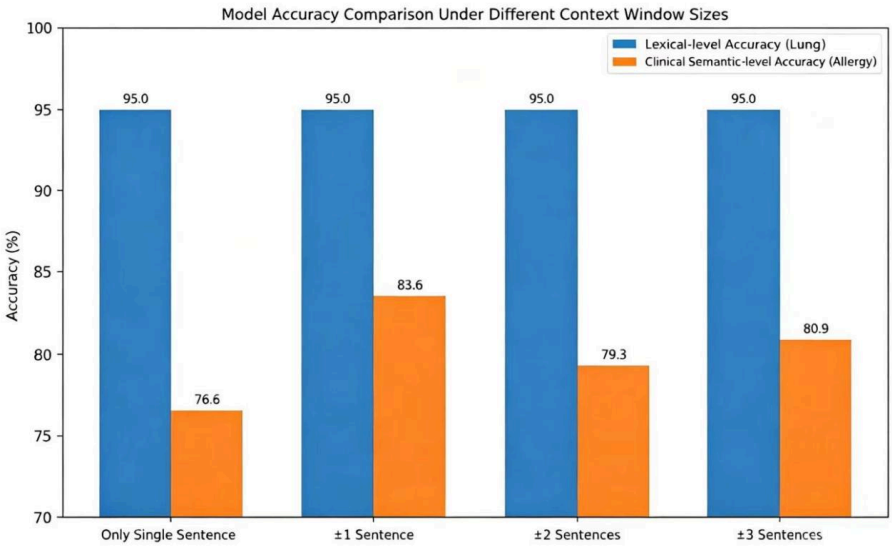

Figure 3. Disambiguation accuracy under different context windows (picture credit: original)

The word meaning recognition task achieved an accuracy of 95.0% across all four window groups, with no fluctuations. Indicating 'lung' pneumonia, organ binary classification relies on strong local feature words (such as 'chest X-ray' and 'anti-inflammatory'), and a single sentence context is sufficient to support stable discrimination. In contrast, the clinical semantic level is highly sensitive to window length: the accuracy is lowest when only a single sentence is used (76.6%), after adding one sentence before and after, it significantly increased to 83.6% (an increase of 7.0 percentage points). This indicates that the chief complaints, medication history, and physical examination descriptions in the neighboring dialogue provide key clues for distinguishing allergic phenotypes. However, when the window was further expanded to include the two preceding and succeeding sentences and the three preceding and succeeding sentences, the accuracy actually dropped to 79.3% and 80.9%, respectively, indicating that excessively long context introduced irrelevant diagnostic information (such as repeated descriptions of past medical history and examination request forms), which interfered with the model's focus on the core meaning of the current sentence. This result clearly supports the setting of 'one sentence before and after as the optimal window' in Section 2.3, while also revealing the essential differences in contextual requirements between the lexical-semantic layer and the clinical-semantic layer, further corroborating the dual-layer task difficulty stratification.

4. Discussion

Word sense disambiguation can be accomplished solely by relying on surface-level lexical co-occurrence features, and it belongs to a simple shallow matching task. Distinguishing the four types of allergic phenotypes requires professional pediatric knowledge and the ability to discern the

DOI: 10.54254/2753-8818/2027.37537

fundamental differences between disease diagnosis and descriptions of innate constitution. It is a highly challenging and deep clinical reasoning task, representing a common difficulty for both BERT and general large models. Sample class imbalance forces the model to favor high-frequency samples in order to reduce loss. The traditional single-layer evaluation framework fails to reveal this flaw, thereby significantly overestimating the model's true semantic understanding ability. This is also the core reason and innovative value for proposing the dual-layer hierarchical evaluation framework in this paper. Error type statistics and window ablation experiments further reveal the key bottlenecks in clinical semantic tasks. More than 40% of errors stem from insufficient contextual information, indicating that in pediatric short-sentence dialogues, the model finds it difficult to capture phenotypic differences when key clues such as chief complaints or medication history are lacking. At the same time, 31% of the errors involving semantic boundary ambiguities indicate that in some sentences, 'allergy' may refer to multiple phenotypes simultaneously, and the actual annotation also faces differences in subjective judgment. This is not only a problem with the model, but also a reflection of the complexity inherent in the task definition itself. The results of the window ablation indicate that the preceding and following sentences contain the optimal amount of information; overly long windows introduce irrelevant noise such as medical history or examination descriptions, which in turn weakens the model's focus on the core semantics of the current sentence. This finding has direct implications for practical system design; the selection of context is not better with more, and should be dynamically trimmed according to the actual complexity of the task [9].

The pre-training corpus for general large models primarily consists of academic literature and standard medical records, lacking fragmented, conversational dialogues from parents during medical consultations.

Also, zero-shot large model learning relies solely on basic prompts to issue instructions and lacks specific discriminative rule constraints tailored to the pediatric field. The model generates outputs solely based on the statistical frequencies of the training set and lacks sophisticated clinical reasoning capabilities. In pediatric sub-scenarios with limited annotation and computing resources, a lightweight BERT fine-tuning approach has lower costs and stronger recognition stability, making it more suitable for the implementation of grassroots intelligent consultation systems [10].

From a clinical application perspective, misdiagnosing "allergic constitution" as "allergic rhinitis" may lead to unnecessary use of anti-allergic medications, while misinterpreting disease phenotypes as constitution may delay treatment. The dual-layer framework used in this article, by distinguishing between the semantic layer and the clinical layer, can effectively reduce the risk of such decisions [11]. Although BERT fine-tuning outperforms zero-shot large models in terms of accuracy, it should be noted that its performance is highly dependent on the quality of annotated data. Moreover, it remains powerless against rare phenotypes.

This study has three limitations: first, the dataset is small, with only 73 samples labeled as 'allergy', and the class imbalance is pronounced; second, the vocabulary validation is insufficient, as only the ambiguous words 'lung' and 'allergy' were selected, and the generalizability of the dual-layer framework requires validation with a broader range of vocabulary; Third, the large model only conducts zero-shot testing and has not explored few-shot or instruction fine-tuning optimization schemes. Future research will expand the annotated data for various types of pediatric ambiguous terms and increase the sample size and introduce oversampling and loss weighting to mitigate the problem of class imbalance. In terms of improving large models, future exploration will focus on fine-tuning in the field of large models and integrating solutions combined with medical knowledge graphs.

DOI: 10.54254/2753-8818/2027.37537

5. Conclusion

This paper selects 'lung' and 'allergy' in pediatric doctor-patient conversations as the research subjects, and constructs a two-tier disambiguation framework based on lexical semantics and clinical semantics. Leveraging the IMCS-V2 pediatric dialogue annotation dataset, multiple controlled experiments are conducted to systematically compare the disambiguation performance of fine-tuned BERT, DeepSeek, and Tongyi Qianwen. The experimental results indicate that the disambiguation accuracy for the lexical item 'lung' can reach 95.0%, whereas the recognition of fine-grained clinical phenotypes for 'allergy' is only 76.6%, with an 18.4 percentage point difference in accuracy between the two levels of tasks. The zero-shot inference performance of the two general-purpose large models is inferior to that of fine-tuned BERT. The accuracy of Tongyi Qianwen is on par with most class baselines, while DeepSeek is below the baseline. The window ablation experiment further confirms that one sentence before and after is the optimal context, while an excessively wide window introduces noise that weakens the recognition of clinical semantics. This result fully demonstrates that the traditional 'abbreviation-full form' single-layer evaluation standard cannot measure a model's deep clinical semantic understanding ability, and distinguishing fine-grained semantics that rely on professional medical knowledge, such as diseases and body constitutions, is a core challenge faced by both lightweight pre-trained models and general large language models. This article builds a hierarchical assessment and recognition system based on high-frequency ambiguous vocabulary in pediatrics, extending abbreviation disambiguation research from simple lexical matching to deep clinical semantic reasoning of symptoms. This provides a new research approach and a standardized experimental benchmark for natural language understanding in Chinese pediatric doctor-patient dialogues.

References

[1] Coghlan, A., Turner, S., & Coverdale, S. (2021). Danger in discharge summaries: Abbreviations create confusion for both author and recipient. *Internal Medicine Journal, 53*(4). https://doi.org/10.1111/imj.15582.

[2] Kugic, A., Martin, I., Modersohn, L., Pallaoro, P., Kreuzthaler, M., Schulz, S., & Boeker, M. (2024). Processing of short-form content in clinical narratives: Systematic scoping review. *Journal of Medical Internet Research, 26*. https://doi.org/10.2196/57852.

[3] Gibson, D., Jackson, S., Ramesh Shanmugasundaram, Seth, I., Siu, A., Ahmadi, N., Kam, J., Mehan, N., Ruban Thanigasalam, Jeffery, N., Patel, M. I., & Leslie, S. (2024). Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: A multi-metric assessment (preprint). *Journal of Medical Internet Research, 26*, e55939–e55939. https://doi.org/10.2196/55939.

[4] Chen, L., Varoquaux, G., & Suchanek, F. M. (2023). GLADIS: A general and large acronym disambiguation benchmark. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2073–2088. https://doi.org/10.18653/v1/2023.eacl-main.152.

[5] Shor, J., Bi, R. A., Venugopalan, S., Ibara, S., Goldenberg, R., & Rivlin, E. (2023). Clinical bertscore: An improved measure of automatic speech recognition performance in clinical settings. *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 1–7. https://doi.org/10.18653/v1/2023.clinicalnlp-1.1.

[6] Rajkomar, A., Loreaux, E., Liu, Y., Kemp, J., Li, B., Chen, M.-J., Zhang, Y., Mohiuddin, A., & Gottweis, J. (2022). Deciphering clinical abbreviations with a privacy protecting machine learning system. *Nature Communications, 13*(1). https://doi.org/10.1038/s41467-022-35007-9.

[7] Wei, S., Liu, X., & Shi, Y. (2026). Selection-based prompting with synthesized candidates unlocks high-performance clinical abbreviation disambiguation via open-source LLMs. *Journal of Biomedical Informatics, 181*, 105085. https://doi.org/10.1016/j.jbi.2026.105085.

[8] Bhargava, H., Salomon, C., Suresh, S., Chang, A., Kilian, R., Stijn, D. V., Oriol, A., Low, D., Knebel, A., & Taraman, S. (2024). Promises, pitfalls, and clinical applications of artificial intelligence in pediatrics. *Journal of Medical Internet Research, 26*, e49022. https://doi.org/10.2196/49022.

DOI: 10.54254/2753-8818/2027.37537

[9] Wu, J., Ruan, X., McNeer, E., Rossow, K. M., & Choi, L. (2026). Developing a natural language processing system using transformer-based models for adverse drug event detection in electronic health records. *PLOS ONE, 21*(7), e0350516. https://doi.org/10.1371/journal.pone.0350516.

[10] Chaichulee, S., Promchai, C., Kaewkomon, T., Kongkamol, C., Ingviya, T., & Sangsupawanich, P. (2022). Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. *PLOS ONE, 17*(8), e0270595. https://doi.org/10.1371/journal.pone.0270595.

[11] Vassileva, S., Grazhdanski, G., Boytcheva, S., & Koychev, I. (2023). FMI-SU at SemEval-2023 task 7: Two-level entailment classification of clinical trials enhanced by contextual data augmentation. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1454–1462. https://doi.org/10.18653/v1/2023.semeval-1.200.