# PEI at #SMM4H-HeaRD 2026: Enhancing Patient Metadata Detection via Hypothesis-Conditioned Classification and Paraphrase-Based Data Augmentation

**Farnaz Zeidi, Roman Christof, Farnoush Zeidi, Renate König, Liam Childs**
Host-Pathogen Interactions, Paul-Ehrlich-Institute, Langen, Germany
{farnaz.zeidi, roman.christof, farnoush.zeidikolehparcheh
renate.koenig, liam.childs}@pei.de

## Abstract

This paper presents our approach to Task 5 of the #SMM4H-HeaRD 2026 Workshop, which focuses on detecting patient metadata in SARS-CoV-2 sequencing articles as a binary classification task. We explore both encoder-based and large language model (LLM) approaches, using BioM-BERT as a baseline and Mistral-Nemo as the LLM. To improve performance, we propose a paraphrase-based data augmentation pipeline using Qwen3, where paraphrased training and validation instances are added for fine-tuning. For the LLM, we perform prompt refinement and error analysis, while for the encoder-based model, we reformulate the task as a hypothesis-conditioned classification task inspired by Natural Language Inference (NLI). Our methods improve both models: Mistral-Nemo increases from 0.423 to 0.750 F1, and BioM-BERT from 0.801 to 0.821 on the validation set. Although Mistral-Nemo does not surpass BioM-BERT, our best BioM-BERT model achieves an F1-score of 0.786 on the test set, outperforming the mean and median of competing systems. To support reproducibility, we release our best-performing model on Hugging Face.

## 1 Introduction

Patient metadata is essential for genomic epidemiology, enabling the analysis of virus transmission through demographic, clinical, and geographic information (Grubaugh et al., 2019). However, despite the availability of large-scale SARS-CoV-2 genome sequences in public databases, associated patient metadata are often missing or remain embedded in unstructured scientific text (Chen et al., 2022; O'Connor et al., 2025; Klein et al., 2025). This poses a significant challenge, as manual extraction from large volumes of documents is time-consuming and impractical for healthcare applications (Klein et al., 2025; Zeidi et al., 2025b).

To address this issue, Task 5 of the #SMM4H-HeaRD 2026 Workshop (Lopez-Garcia et al., 2026), focuses on detecting patient metadata in SARS-CoV-2 sequencing articles as a binary classification task using a manually annotated dataset of 22,147 sentences. On the same dataset, Klein et al. (2025) explored BERT-based models with techniques such as under-sampling and class weighting to address class imbalance, achieving an F1-score of 0.776, while their large language model (LLM)-based approach reached 0.558, indicating that LLMs did not outperform encoder-based models. This observation is consistent with prior work showing that LLMs do not consistently surpass encoder-based approaches in classification tasks (Wang et al., 2025). Nevertheless, efforts have been made to improve LLM performance, such as paraphrase-based data augmentation (Ta et al., 2024) and progressive reasoning strategies for text classification (Sun et al., 2023). Similarly, in our previous work on #SMM4H-HeaRD 2025 (Zeidi et al., 2025a), we demonstrated that data augmentation can significantly improve LLM performance, enabling it to outperform a RoBERTa-based baseline, which motivates the use of augmentation strategies in this study.

In our participation in the #SMM4H-HeaRD 2026 Workshop, we leveraged the BERT-based model BioM-BERT (Alrowili and Shanker, 2021) as a baseline, alongside the LLM-based generative model Mistral-Nemo[1]. To enhance performance, we introduce a data augmentation strategy based on paraphrasing, where training and validation samples are rewritten using Qwen3 (Qwen Team, 2025) and incorporated into the training process. For BioM-BERT, we further improve performance using a hypothesis-conditioned formulation inspired by Natural Language Inference

[1] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407



*Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks*, pages 146–153
July 3, 2026 ©2026 Association for Computational Linguistics

(NLI). For Mistral, we investigate multiple prompt formulations, including classification-based and hypothesis-conditioned settings, followed by error analysis and prompt refinement. While the LLM does not outperform the encoder-based model, our approach yields consistent improvements across both model types. To support reproducibility, we release our best-performing model on Hugging Face[2].

## 2 Task and Dataset

Task 5, titled "Detection of Patient Metadata in SARS-CoV-2 Sequencing Articles", focuses on identifying patient-related metadata (e.g., demographics, symptoms, treatments, outcomes, and geographic information) in SARS-CoV-2 sequencing articles. The task is formulated as a binary sentence classification problem, where systems determine whether a sentence contains patient metadata (label "1") or not (label "0").

The dataset introduced by Klein et al. (2025) consists of 22,147 manually annotated English sentences from 150 COVID-19-related articles in LitCovid (Chen et al., 2022). The dataset is split into training (70%), validation (10%), and test (20%) sets, and is highly imbalanced, with only 13.3% positive samples. Test labels are not publicly available.

## 3 Methodology

To select the baseline encoder model, we used the BioM-BERT-PubMed-PMC-Large model (Alrowili and Shanker, 2021) and selected Mistral-Nemo-Instruct-2407 as the LLM-based model. To enhance model performance, we applied several strategies to improve both BioM-BERT and Mistral. We introduce a paraphrase-based data augmentation pipeline using the Qwen3-8B model (Qwen Team, 2025). After removing duplicate instances, all remaining training and validation samples were paraphrased and added back to the dataset for fine-tuning. The augmented training data (original + paraphrased) are used to train both models, while the augmented validation data (original + paraphrased) are used only for model selection across different checkpoints for BioM-BERT.

For the LLM, we first performed prompt engineering to identify the most effective prompt and evaluated it in a zero-shot Mistral setting. After fine-tuning the model on the original dataset, we experimented with different prompt types (classification-based and hypothesis-conditioned). Based on the best-performing prompt, we then fine-tuned the model on the augmented dataset (original + paraphrased data). Finally, we conducted error analysis and refined the prompts to further improve performance.

For the encoder-based model, we fine-tune a binary classification model using both the original and augmented training data. After that, we reformulate the classification task as a hypothesis-conditioned classification task inspired by NLI by experimenting with different hypotheses. The motivation behind this approach is that NLI-style formulations capture the relationship between an input sentence and a hypothesis (e.g., entailment, contradiction, or neutral), whereas traditional classification assigns a label to a single input without explicitly modeling such relationships (Bowman et al., 2015; Williams et al., 2018; Devlin et al., 2019).

Models were evaluated using precision, recall, and F1-score on the validation set. Since test labels are not publicly available and submissions are limited, we report only the results of the best-performing model on the test set.

## 4 Experiments

To generate paraphrased data, we first removed duplicate samples (80 from the validation set and 1,085 from the training set) and paraphrased the remaining instances (14,419 training and 2,134 validation samples). Each remaining sentence was paraphrased exactly once. The paraphrasing prompt is provided in Appendix A. We used the Qwen3-8B model with a maximum of 32,768 generated tokens, a batch size of 16, and a temperature of 0.6, while keeping the remaining parameters consistent with the official Hugging Face implementation[3].

For the Mistral experiments, we first evaluated multiple prompts in a zero-shot setting and selected the best-performing one. Based on this, we explored two prompting strategies: a hypothesis-conditioned prompt (output: "Entailment" or "Neutral") and a classification-based prompt (output: "1" or "0"). We then fine-tuned the Mistral model and, in the final step, conducted error analysis to

[2] https://huggingface.co/pei-germany/biom-bert-smm4h2026-task5-patient-metadata

[3] https://huggingface.co/Qwen/Qwen3-8B

refine the prompts. The prompts used in all settings are provided in Appendix B. Fine-tuning was performed with a batch size of 1, a learning rate of 6e-5, a sequence length of 4096, and 300 training steps, while keeping the remaining parameters consistent with the official Mistral repository[4].

For the BioM-BERT experiments, we trained both classification and hypothesis-conditioned variants for 7 epochs using a batch size of 4 and a maximum sequence length of 512. We used a learning rate of 1e-5 and otherwise kept the default BERT settings (Devlin et al., 2019). We experimented with three different hypotheses:

H1: This sentence refers to patients or cases, including their clinical, demographic, or sample-related information.

H2: This sentence contains information about patients or cases, such as clinical characteristics, treatments, epidemiological details, or patient-derived samples.

H3: This sentence explicitly describes patients, cases, or patient-related clinical or epidemiological information.

All experiments were conducted on a single NVIDIA A100 GPU (GA100, 80GB HBM2e). The total runtime for paraphrasing was approximately 7,032 minutes (∼4.9 days) for the training set and 957 minutes (∼0.7 days) for the validation set. Fine-tuning on the original dataset took approximately 19 minutes per model for Mistral, and 41 minutes for classification and 34 minutes for hypothesis-conditioned classification using BioM-BERT.

## 5 Results Analysis

Tables 1 and 2 summarize the validation performance of the Mistral and BioM-BERT variants, respectively, while Table 3 presents the test set results of the submitted system in comparison to other participants.

After paraphrasing the original dataset, we added 14,419 training and 2,134 validation samples. The paraphrasing process preserves key medical terms while varying sentence structure (see examples in Appendix A.2). To assess augmentation quality, we analyzed semantic and lexical similarity between original and paraphrased sentences using BioM-BERT cosine similarity and ROUGE-L, which measures lexical similarity based on overlapping word sequences. The paraphrases achieved a mean semantic similarity of 0.985 (± 0.033) and a mean ROUGE-L score of 0.637 (± 0.155), indicating strong preservation of meaning while introducing moderate lexical and structural variation. Additional visualizations are provided in Appendix C. As shown in Tables 1 and 2, training on the augmented data (original + paraphrased samples) led to consistent improvements in F1-score: BioM-BERT improved by 0.012 (from 0.801 to 0.813), while Mistral showed a larger gain of 0.029 (from 0.713 to 0.742).

Regarding Mistral, in a zero-shot setting with a classification-based prompt, the model achieved an F1-score of 0.423. After fine-tuning on the original training dataset, performance improved to 0.713. However, switching to a hypothesis-conditioned prompt led to a slight decrease (0.710). Error analysis on the validation set shows that the hypothesis-conditioned model (FP: 101, FN: 75) is more conservative in predicting label "1" compared to the classification-based model (FP: 88, FN: 84). Based on this, we continued with the classification-based prompt. Incorporating paraphrased data further improved performance from 0.713 to 0.742. However, error analysis (FP: 106, FN: 58) indicates that the model tends to produce false positives, particularly for sentences containing location or structural cues (e.g., section titles). To address this, we introduced additional prompt constraints to filter out general or non-informative sentences. This led to an improvement in F1-score from 0.742 to 0.750 and reduced false positives (FP: 91), although false negatives increased slightly (FN: 63). Further prompt refinements did not yield additional gains and occasionally reduced performance.

Regarding BioM-BERT, adding paraphrased data led to an improvement of 0.012 (from 0.801 to 0.813). We then reformulated the task as a hypothesis-conditioned classification task using different hypotheses. While H1 and H2 resulted in slight decreases in F1-score (0.811 and 0.812, respectively), H3 achieved the best performance, improving the F1-score to 0.821. This gain can be attributed to the more explicit and restrictive wording of H3, which helps the model better distinguish patient-related information from general or ambiguous content. Error analysis further supports this observation: the H3-based model (FP: 59, FN: 48) produces fewer false negatives compared to the classification-based model (FP: 58, FN: 53), indicating improved sensitivity to relevant cases.

[4] https://github.com/mistralai/mistral-finetune

| Setting | Prompt Type | Training Dataset | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Zero-shot | Classification | – | 0.284 | 0.830 | 0.423 |
| Fine-tuned | Hypothesis-conditioned | Original | 0.705 | 0.714 | 0.710 |
| Fine-tuned | Classification | Original | 0.684 | 0.745 | 0.713 |
| Fine-tuned | Classification | Original + Paraphrased | 0.690 | 0.803 | 0.742 |
| Fine-tuned | Classification (refined prompt) | Original + Paraphrased | **0.717** | 0.786 | **0.750** |

Table 1: Performance of Mistral variants on the validation set.

| Task Formulation | Hypothesis | Training Dataset | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Classification | – | Original | 0.775 | 0.830 | 0.801 |
| Classification | – | Original + Paraphrased | 0.806 | 0.820 | 0.813 |
| Hypothesis-conditioned | H1 | Original + Paraphrased | 0.786 | 0.837 | 0.811 |
| Hypothesis-conditioned | H2 | Original + Paraphrased | 0.743 | **0.895** | 0.812 |
| Hypothesis-conditioned | H3 | Original + Paraphrased | **0.807** | 0.837 | **0.821** |

Table 2: Performance of BioM-BERT variants on the validation set.

Overall, our best BioM-BERT model (trained on augmented data with a hypothesis-conditioned formulation) achieved an F1 score of 0.821, while the best Mistral variant reached 0.750. Error analysis shows that BioM-BERT (FP: 59, FN: 48) produces fewer false positives and false negatives than Mistral (FP: 91, FN: 63), indicating more stable performance. This difference may be attributed to BioM-BERT's bidirectional encoder architecture, which captures contextual information from both left and right contexts simultaneously, while the autoregressive Mistral-based setup generates text sequentially from left to right. We also observed overlapping errors between the two models, including 31 common false positives and 26 common false negatives. These errors often occur in sentences containing domain-specific or structural terminology (e.g., "Label: Phylogenetic tree of SARS-CoV-2 genomes from CUH."), which may lead both models to over-predict patient-related information. Conversely, both models tend to miss cases where patient metadata is expressed in a more implicit or structural manner (e.g., "Study Design, Patients, and Study Procedures"), suggesting limitations in capturing less explicit cues.

Finally, we submitted our H3-based hypothesis-conditioned BioM-BERT model for the test set evaluation, achieving an F1 score of 0.786, outperforming the baseline of 0.776 reported by (Klein et al., 2025). Table 3 shows that our model also surpasses both the mean (0.729) and median (0.754) F1-scores of competing systems, indicating strong and competitive performance on the task.

## 6 Conclusion

In this paper, we presented our approach to Task 5 of the #SMM4H-HeaRD 2026 Workshop, which focuses on binary classification of patient metadata in SARS-CoV-2 sequencing articles. We explored both an encoder-based model (BioM-BERT) and a large language model (Mistral-Nemo), and introduced a paraphrase-based data augmentation pipeline using Qwen3 to enrich the training data. For the LLM, we applied prompt engineering, fine-tuning, and error-driven refinement, while for BioM-BERT, we adopted a hypothesis-conditioned formulation with different hypothesis designs. These strategies led to consistent improvements across both models, with Mistral-Nemo improving from an F1 score of 0.423 to 0.750 and BioM-BERT from 0.801 to 0.821. Despite these gains, the encoder-based model remained more effective, particularly in handling class imbalance and reducing both false positives and false negatives. Our best hypothesis-conditioned BioM-BERT model achieved an F1 score of 0.786 on the test set, outperforming the mean and median scores of competing systems. These results demonstrate the effectiveness of hypothesis-conditioned classification and paraphrase-based data augmentation, while future work will explore hybrid encoder–generative approaches.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| Our BioM-BERT (Hypothesis-conditioned H3) | **0.797** | **0.774** | **0.786** |
| Mean (all teams) | 0.712 | 0.756 | 0.729 |
| Median (all teams) | 0.741 | 0.772 | 0.754 |

Table 3: Test set performance of the submitted system compared to other participants.

## Limitations

This work has several limitations. First, evaluation is primarily conducted on the validation set due to unavailable test labels and limited submission runs, restricting a comprehensive assessment of model variants. Second, the dataset is highly imbalanced, which may limit generalization, and we do not explicitly address this issue with techniques such as resampling or class weighting. In addition, paraphrase augmentation was applied uniformly to all samples, preserving the original class imbalance rather than explicitly addressing it. Third, although paraphrase quality analysis indicates strong semantic preservation overall, the augmentation process relies on a single LLM and may still introduce occasional noise or semantic drift. Furthermore, some implicit patient-related information may require broader contextual understanding beyond individual sentences. Finally, both the LLM and hypothesis-conditioned approaches depend on prompt design and hypothesis formulation, which require manual tuning and may not generalize across different contexts.

## Acknowledgments

This work was carried out by the Artificial Intelligence Working Group, led by Dr. Liam Childs within the FoG3 Research Group, under the leadership of Dr. Renate König at the Paul Ehrlich Institute (PEI).

## References

Sultan Alrowili and Vijay Shanker. 2021. BioM-Transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Zhiyuan Chen, Andrew S. Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, and Xiangyanyu Xu. 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics*, 54(4):499–507.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nathan D. Grubaugh, Jason T. Ladner, Philippe Lemey, Andrew Rambaut, Oliver G. Pybus, Edward C. Holmes, and Kristian G. Andersen. 2019. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*, 4(1):10–19.

Ari Z. Klein, Davy Weissenbacher, Karen O'Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models. *medRxiv*.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z. Klein, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, and 9 others. 2026. Overview of the 11th social media mining for health (#smm4h) and health real-world data (heard) shared tasks at acl 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Karen O'Connor, Davy Weissenbacher, Amir Elyaderani, Ebbing Lautenbach, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Patient-related metadata reported in sequencing studies of SARS-CoV-2: Protocol for a scoping review and bibliometric analysis. *JMIR Research Protocols*, 14(1):e58567.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005.

Thang Ta, Abu Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2024. ThangDLU at #SMM4H 2024: Encoder-decoder models for classifying text data on social disorders in children and adolescents. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 1–4.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Farnaz Zeidi, Roman Christof, Renate König, and Liam Childs. 2025a. PEI at #SMM4H-HeaRD 2025: Enhancing adverse event detection via error-driven data augmentation. In *Proceedings of the 19th International AAAI Conference on Web and Social Media (ICWSM) – SMM4H-HeaRD 2025 Workshop and Shared Tasks*.

Farnaz Zeidi, Manuela Messelhäusser, Roman Christof, Xing David Wang, Ulf Leser, Dirk Mentzer, Renate König, and Liam Childs. 2025b. MEDNER.DE: Medicinal product entity recognition in german-specific contexts. *IEEE Access*, 13:159961–159978.

# A Paraphrasing Prompt and Examples

## A.1 Paraphrasing Prompt

Rewrite the following sentence with the same meaning using different wording.

**Guidelines:**

- Preserve the original meaning
- Use different phrasing and structure
- Keep key medical terms unchanged

**Sentence:** `<input sentence>`
**Paraphrased sentence:**

## A.2 Paraphrasing Examples

The paraphrasing process aims to preserve the original meaning while varying wording and sentence structure. Examples are shown below:

**Example 1:** The virus was originally introduced into Houston many times independently.
**Paraphrased:** The virus was first introduced into Houston on multiple separate occasions.

**Example 2:** Study sample descriptions and sequencing results.
**Paraphrased:** Examine the sample data and sequencing results.

# B Mistral Prompts

## B.1 Classification-based Prompt

**Task:**
Determine whether the sentence is related to patients, cases, or patient-derived samples.

**Definition:**
Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

**Guidelines:**
Answer "1" if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer "0" if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases

**Output:**
Answer with only "1" or "0".

**Sentence:** `<input sentence>`

### B.2 Hypothesis-Conditioned Prompt

**Task:**
Determine whether the sentence is related to patients, cases, or patient-derived samples.

**Definition:**
Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

**Guidelines:**
Answer "Entailment" if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer "Neutral" if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases

**Output:**
Answer with only "Entailment" or "Neutral".

**Sentence:** `<input sentence>`

### B.3 Final Refined Prompt

**Task:**
Determine whether the sentence is related to patients, cases, or patient-derived samples.

**Definition:**
Patient metadata includes:

- Clinical information (symptoms, severity, inflammation, outcomes)
- Treatments or medical decisions
- Laboratory or biological properties linked to cases (e.g., viral load)
- Epidemiological information (travel history, outbreak context, location, time)
- Statements about groups of patients or cases (not only individuals)
- Sample-related information

**Guidelines:**
Answer "1" if the sentence refers to patients, cases, or patient-derived data/samples (even indirectly or at group level).

Answer "0" if:

- No patients/cases are mentioned
- The sentence is only a title, number, or section header
- It is general knowledge or purely methodological without reference to specific cases
- The sentence is a fragment, abbreviation, reference, statistical output, or definition
- The sentence contains only generic or implicit subjects without explicit patient or case context
- The sentence contains non-informative mentions (e.g., "a passenger", "a child", "the patient was admitted") without specific clinical or epidemiological details

**Output:**
Answer with only "1" or "0".

**Sentence:** `<input sentence>`

## C Paraphrase Quality Analysis

To further evaluate the quality of paraphrase-based augmentation, we analyzed semantic similarity and lexical overlap between original and paraphrased sentences. Semantic similarity was computed using BioM-BERT embeddings with cosine similarity, while ROUGE-L measures lexical similarity based on overlapping word sequences. As shown in Figure 1, most paraphrases achieve very high

semantic similarity scores (primarily between 0.9 and 1.0) despite varying levels of lexical overlap, indicating strong preservation of semantic content while introducing lexical and structural variation. Most ROUGE-L scores fall between 0.4 and 0.8, suggesting moderate rewording rather than simple lexical copying. In addition, a smaller number of low-overlap outliers likely correspond to heavily restructured sentences, such as section titles or short structural phrases, which may introduce stronger lexical variation while preserving partial semantic meaning.

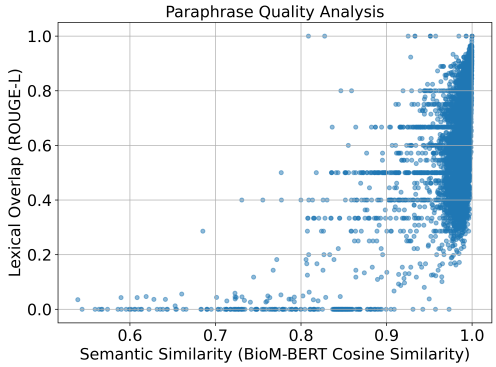


Figure 1: Relationship between semantic similarity (BioM-BERT cosine similarity) and lexical overlap (ROUGE-L) for original and paraphrased sentence pairs.