WisPerMed at ArchEHR-QA 2025: A Modular, Relevance-First Approach for Grounded Question Answering on Eletronic Health Records

Jan-Henning Büns[1], Hendrik Damm[1,3], Tabea M. G. Pakull[1,2], Felix Nensa[4,5], Elisabeth Livingstone[2]

[1]Department of Computer Science, University of Applied Sciences and Arts Dortmund
[2]Institute for Transfusion Medicine, University Hospital Essen
[3]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen
[4]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen
[5]Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen

Correspondence: jan-henning.buens@fh-dortmund.de

Abstract

Automatically answering patient questions based on electronic health records (EHRs) requires systems that both identify relevant evidence and generate accurate, grounded responses. We present a three-part pipeline developed by WisPerMed for the ArchEHR-QA 2025 shared task. First, a fine-tuned BioClinicalBERT model classifies note sentences by their relevance using synonym-based and paraphrased data augmentation. Second, a constrained generation step uses DistilBART-Med-Summary to produce faithful answers strictly limited to top-ranked evidence. Third, we align each answer sentence to its supporting evidence via BiomedBERT embeddings and ROUGE-based similarity scoring to ensure citation transparency. Our system achieved a 35.0% overall score on the hidden test set, outperforming the organizer's baseline by 4.3 percentage points. Gains in BERTScore (+44%) and SARI (+119%) highlight substantial improvements in semantic accuracy and relevance. This modular approach demonstrates that enforcing evidence-awareness and citation grounding enhances both answer quality and trustworthiness in clinical QA systems.

1 Introduction

As patient–portal adoption accelerates, message volume now exceeds pre-pandemic levels; a longitudinal study found a 55% rise in medical-advice requests and 24% increase in daily inbox time for physicians between 2019–2023 (Arndt et al., 2024). Large Language Models (LLMs) can draft fluent replies, yet uncontrolled hallucinations threaten patient safety (Nov et al., 2023; Biro et al., 2025). The ArchEHR-QA 2025 shared task (Soni and Demner-Fushman, 2025b) extends this trajectory by pairing genuine portal questions with sentence-level evidence annotations and requiring grounded answers.

This paper presents the submission by WisPerMed, a three-part pipeline:

- BioClinicalBERT (Lee et al., 2019a) classifies note sentences as *essential*, *supplementary*, or *not-relevant*, with robustness improved via synonym and paraphrase augmentation;
- DistilBART-Med-Summary (Lewis et al., 2019) generates an answer conditioned solely on the top-ranked evidence and
- BiomedBERT (Gu et al., 2021a) embeddings align each answer sentence to its most similar evidence, yielding explicit citations.

2 Related Work

This section establishes the context for our multi-component system that combines evidence classification, answer generation, and citation alignment.

Electronic Health Record Question Answering. Electronic Health Records (EHRs) contain valuable patient information that can benefit both healthcare providers and patients. Giving patients access to their EHRs can increase patient and physician trust, improve communication, strengthen the physician–patient relationship, increase medication adherence, and improve patient outcomes (Tapuria et al., 2021). Question Answering (QA) systems on patient-related data can assist clinicians in decision-making and enable patients to better understand their medical history (Bardhan et al., 2024). Unlike general medical QA tasks that rely on curated knowledge sources (e.g., PubMed or medical websites), EHR QA requires answer generation grounded in patient-specific records. This introduces challenges in interpreting both informal patient queries and domain-specific clinical text.

BioNLP 2025 Shared Tasks, pages 41–49

Datasets. non Early progress relied on synthetic corpora such as EMRQA (Pampari et al., 2018), which repurposed i2b2 annotations (Özlem Uzuner et al., 2011) to create ~ 0.4 M evidence–answer pairs. Work on structured records introduced MIMICSQL for question-to-SQL generation on MIMIC-III tables (Wang et al., 2020). To improve realism and coverage, consumer-health resources like MEDIQA-ANS (Savery et al., 2020) added question-driven answer summaries, while MEDIQA-CHAT captured full doctor–patient dialogues (Ben Abacha et al., 2023). Recent benchmarks push modality boundaries: EHRXQA integrates tabular EHR data with chest-X-ray images for cross-modal reasoning (Bae et al., 2023). The ArchEHR-QA dataset (Soni and Demner-Fushman, 2025a) extends this trajectory by pairing genuine portal questions with sentence-level evidence annotations and enforcing grounded answers. The dataset is derived from the MIMIC-III dataset (Johnson et al., 2016) and comprises 120 patient cases (20 development, 100 test). Every case consists of a realistic patient question, corresponding clinician-rewritten questions, and annotated clinical note excerpts. Each clinical note excerpt is segmented into sentences, which are manually annotated as "essential", "supplementary", or "not-relevant" for answering the question.

Biomedical Language Models. Domain-specific language models have revolutionized biomedical NLP applications (Yang et al., 2023). While early approaches fine-tuned general-domain models like BERT (Devlin et al., 2019) on biomedical corpora, research has demonstrated that pre-training language models from scratch on biomedical text yields substantial performance gains across various tasks (Gu et al., 2021b). In the realm of medical text summarization, models like DistilBART-Med-Summary[1] have been developed to condense clinical documents into concise summaries while preserving essential information. These models are trained on large-scale medical datasets and fine-tuned to capture the specific linguistic characteristics of clinical narratives.

BioBERT (Lee et al., 2019b), a domain-specific model pretrained on large-scale biomedical corpora, significantly outperforms general-domain BERT on biomedical text mining tasks. Building on this foundation, BiomedBERT (Gu et al., 2021a) was trained solely on biomedical text from scratch and achieved excellent results across multiple biomedical NLP benchmarks. Bio_ClinicalBERT (Alsentzer et al., 2019) specializes further in clinical text by initializing from BioBERT and training on MIMIC notes, a database containing electronic health records from ICU patients.

Data Augmentation In the medical domain, the scarcity of annotated datasets poses a challenge to the development of robust models. To address this, data augmentation techniques have been employed to artificially expand training datasets, thereby enhancing model generalizability and mitigating overfitting. In clinical contexts, leveraging domain-specific resources such as the Unified Medical Language System (UMLS) (Bodenreider, 2004a) and WordNet (Miller, 1994) for synonym replacement has proven effective in maintaining the integrity of medical terminology during augmentation (Kang et al., 2020; Shorten et al., 2021). Furthermore, the use of LLMs like Gemini (Hoffmann et al., 2023) to generate synthetic data have shown promise in producing high-quality, diverse clinical text (Wang et al., 2024), which is particularly beneficial for tasks in low-resource settings.

3 Methods

WisPerMed adopts a three-part pipeline summarized in Figure 1.

Sentence-level relevance classification. Each clinical note sentence is encoded with BIOCLINICALBERT (Lee et al., 2019a). The model is fine-tuned on the ArchEHR-QA development split (batch size 8, 5 epochs, initial learning rate set to 2×10^{-6} according to the default learning rate scheduler from the transformers library (Wolf et al., 2020)) to predict *essential*, *supplementary*, or *irrelevant* labels. The training data are expanded by 500%, to 100 cases, using: (1) synonym substitution derived from UMLS (Bodenreider, 2004b) and WordNet (Miller, 1994) and (2) paraphrase generation with Gemini.

Answer generation. The evidence set, clinician-rewritten question, and a fixed instruction prompt are concatenated and passed to DistilBART-Med-Summary. The prompt (refer to Listing 2 in Appendix 6) instructs the model to (i) restrict content to the provided evidence. Decoding employs beam search (Meister et al., 2020) (beam size 5, repetition penalty 1.2) and truncates the output to ≤ 75 tokens, as required by the task limit. Only the first

[1] https://huggingface.co/Mahalingam/DistilBart-Med-Summary, Last Accessed: 30.04.2025

75 tokens are included in the performance evaluation.

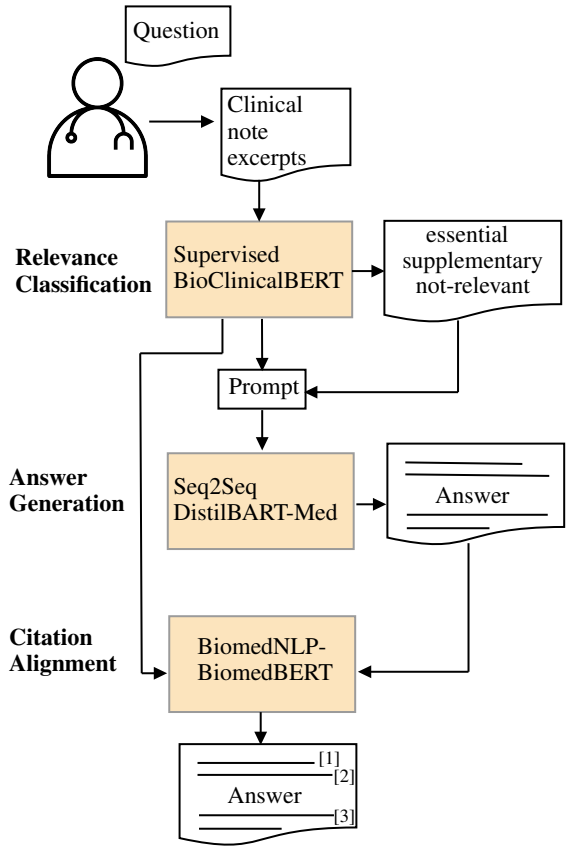

Figure 1: Workflow of the three-part pipeline. The first stage performs relevance classification, identifying sentences as essential, supplementary, or not-relevant for answer generation. The second stage generates an answer using the prioritized evidence. The final stage adds explicit citations by linking each answer sentence to its supporting evidence.

Citation alignment. Each answer sentence is embedded with BIOMEDBERT (Gu et al., 2021a). Using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score (Lin, 2004) calculations, we link sentences in the answer to the most similar sentences of the clinical notes. A similarity threshold of 0.30 ensures that lower-scoring sentences are tagged as unsupported. The selected citations are then being added to the corresponding sentences in the answer to maintain the task's citation format.

Implementation. Models are trained and executed with PyTorch 2.6.0 (Paszke et al., 2019) using Python 3.12.9 on a single Nvidia RTX 4080 Super (16GB). Source code is released under MIT License.[2]

[2] https://github.com/rtg-wispermed/ArchEHR-QA, Last Accessed: 09.05.2025

4 Evaluation

The metrics for evaluation are divided into factuality and relevance metrics. Factuality metrics include Precision, Recall, and F1-score (Powers, 2020) in both micro and macro variations. In addition, all scores are measured in a strict (including only sentences classified as "essential") and a lenient (including sentences classified as "essential" and "supplementary") variation. The mean of all factuality scores (Strict Citation F1 scores) is the Overall Factuality score. Relevance metrics include Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), ROUGE (Lin, 2004) and System output Against References and against the Input (SARI) (Xu et al., 2016). Semantic similarity is measured with BERTScore (Zhang et al., 2019). AlignScore (Zha et al., 2023) provides task-agnostic factual consistency, and MEDCON (Medical Concept Overlap (Yim et al., 2023)) captures clinical concept agreement. The mean of all surface metrics is the overall relevance score. Lastly, the overall score is calculated by the mean of the Overall Factuality score and the overall relevance score.

5 Results and Discussion

Table 1 presents overall scores on the ArchEHR-QA hidden test set. The approach by WisPerMed improved upon the organizer's baseline by $\approx 4.3\%$. Both the Overall Factuality and the overall relevance improved by $\approx 2.6\%$ and $\approx 6.1\%$ respectively.

Metric	WisPerMed	Baseline	DMIS Lab
Overall	35.0	30.7	53.7
OF	36.2	33.6	58.6
OR	33.9	27.8	48.8

Table 1: Comparison of Overall, Overall Factuality (OF), and Overall Relevance (OR) scores for WisPerMed, the organizer's baseline and DMIS Lab

The three-part pipeline demonstrates consistent improvements over the organizer's baseline across key relevance and factual accuracy metrics, as shown in Table 2. Notably, it achieves a 44% relative improvement in BERTScore (29.5 vs. 20.5), indicating superior semantic alignment with reference texts through contextual embeddings. The 119% improvement in SARI (61.0 vs. 27.8) highlights enhanced content preservation during text simplification or rewriting tasks, even compared

to DMIS Lab (36.7). While both systems show comparable performance in UMLS-based concept recognition (MEDCON), WisPerMed's 3.5-point gain in AlignScore (62.3 vs. 57.7) suggests better factual consistency in clinical narratives. The first-place team, DMIS Lab, achieved significantly higher overall scores, indicating that there is still headroom for improving our approach.

Metric	WisPerMed	Baseline	DMIS Lab
BLEU	2.0	0.1	14.3
ROUGE-LSum	22.6	33.6	46.5
SARI	61.0	27.8	36.7
BERTScore	29.5	20.5	53.9
AlignScore	62.3	57.7	92.4
MEDCON	25.9	25.6	49.3

Table 2: Comparison of relevance metrics between WisPerMed, organizers-baseline and DMIS Lab

Table 3 shows that our approach achieves consistently higher recall and F1 scores than the organizer's baseline across both strict and lenient, micro-averaged settings, with strict recall (micro) improving from 21.9 to 26.9 and strict F1 (micro) from 33.6 to 36.2. These gains indicate a higher ability to identify a greater proportion of relevant information, reducing false negatives. On the other hand the organizer's baseline demonstrates higher precision, indicating that our approach contains more false positives. Overall, the metrics demonstrate the focus on maximizing relevant coverage.

Metric	WisPerMed	Baseline	DMIS Lab
Strict Precision (mic)	55.4	71.6	57.9
Strict Recall (mic)	26.9	21.9	59.3
Strict F1 (mic)	36.2	33.6	58.6
Lenient Precision (mic)	59.1	77.0	61.2
Lenient Recall (mic)	27.1	22.3	59.2
Lenient F1 (mic)	37.1	34.6	60.2
Strict Precision (mac)	54.0	77.4	62.1
Strict Recall (mac)	34.0	31.5	69.0
Strict F1 (mac)	37.7	39.0	61.2
Lenient Precision (mac)	59.5	83.0	66.6
Lenient Recall (mac)	33.9	30.8	67.1
Lenient F1 (mac)	39.9	39.9	63.2

Table 3: Comparison of strict and lenient (micro/-macro) precision, recall, and F1 scores for WisPerMed, organizers-baseline and DMIS Lab

Further experiments on the ArchEHR-QA development set have been conducted to compare three different sequence-to-sequence text generation models. Specifically, we chose three models from huggingface: (1) Flan-T5 (Chung et al., 2022), (2) BART-Large-CNN (Lewis et al., 2019) and (3) DistilBART-Med-Summary. The results (refer to Table 4) indicate that both BART-models capture medical concepts in their generated answer more precisely compared to Flan-T5. While DistilBART-Med-Summary achieves the highest Overall Factuality score due to its finetuning on medical data, BART-Large-CNN can capture the relevance of information with a higher precision. Another finding is that Flan-T5 requires a detailed and specific prompt to generate answers that adhere to task requirements (see Listing 1). Both BART models, on the other hand, perform well with a much simpler prompt.

Model	OF	OR	Overall
Flan-T5	54.92	29.63	42.27
BART-Large-CNN	64.04	52.69	58.36
DistilBART-M-S	70.42	49.33	59.87

Table 4: Overall score, Overall Factuality (OF) and Overall Relevance (OR), for each model

The impact of data augmentation was evaluated on the ArchEHR-QA development set. The results (refer to Table 5 in Appendix) demonstrate that synonym augmentation can greatly improve the model's performance in every metric. Including synthetic data generated by Gemini on the other hand has minor impact on the performance metrics.

6 Conclusion

The three-part pipeline proposed by WisPerMed system demonstrates that a modular, relevance-first approach can deliver competitive performance on ArchEHR-QA 2025 while retaining transparency. The combination of BioClinicalBERT-based (Lee et al., 2019a) sentence selection, answer generation with DistilBART-Med-Summary, and Biomed-BERT citation alignment (Gu et al., 2021a) yielded results that surpassed the organizers' baseline and maintained strong precision across strict and lenient settings. We demonstrated that models based on BART (Lewis et al., 2019) are better suited for grounded answer generation for EHR questions compared to Flan-T5 (Chung et al., 2022) variants. We further conclude that synonym augmentation based on UMLS (Bodenreider, 2004a), and WordNet (Miller, 1994) can greatly improve the performance of relevance classification.

Limitations

While the WisPerMed pipeline achieves a strong improvement in the relevance metrics, several weaknesses remain. Reliance on hard probability thresholds in the relevance classifier caps citation recall at roughly 27%. Synthetic training data generated via Gemini paraphrasing occasionally alters medical meaning, introducing label noise that propagates downstream. Because all models are tuned on MIMIC style documentation, performance may degrade when confronted with different institutional note formats or specialty-specific jargon. The ROUGE-score-based similarity method for citation alignment may misassign identifiers when multiple sentences are semantically similar. The decision to use BERT-based sequence-to-sequence (seq2seq) models was made to minimize hardware requirements, enabling the three-step pipeline to be trained on a single consumer GPU, such as the Nvidia RTX 4080 Super (16GB). However, our three-part pipeline could be outperformed by more demanding Retrieval-Augmented Generation (RAG) approaches, which jointly optimize retrieval and generation while explicitly linking answers to sources, reducing citation errors.

Acknowledgments

The work of Jan-Henning Büns, Hendrik Damm and Tabea M. G. Pakull was funded by a PhD grant from the DFG Research Training Group 2535 *Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed)*.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proc. ClinicalNLP*, pages 72–78.

Brian G. Arndt, Mark A. Micek, Adam Rule, Christina M. Shafer, Jeffrey J. Baltus, and Christine A. Sinsky. 2024. More tethered to the EHR: EHR workload trends among academic primary care physicians, 2019–2023. *Annals of Family Medicine*, 22(1):12–18.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, et al. 2023. EHRXQA: A multimodal question answering dataset for electronic health records with chest x-ray images. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question answering for electronic health records: Scoping review of datasets and models. *J Med Internet Res*, 26:e53636.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor–patient conversations. In *Proc. ClinicalNLP*, pages 503–513.

Joshua M. Biro, Jessica L. Handley, J. Malcolm McCurry, Adam Visconti, Jeffrey Weinfeld, J. Gregory Trafton, and Raj M. Ratwani. 2025. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digital Medicine*, 8:222.

Olivier Bodenreider. 2004a. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270.

Olivier Bodenreider. 2004b. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. 2022. Scaling instruction-finetuned language models. In *Proc. EMNLP*, pages 277–294.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Caleb Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021a. PubMedBERT: Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint*, arXiv:2007.15779.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021b. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Jordan Hoffmann, Jeffrey Dean, Slav Petrov, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint*, arXiv:2312.11805.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony G. Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. 2020. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019a. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019b. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL Workshop on Text Summarization Branches Out*, pages 74–81.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Oded Nov, Nina Singh, and Devin Mann. 2023. Putting ChatGPT's medical advice to the (turing) test: Survey study. *JMIR Medical Education*, 9:e46939.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proc. EMNLP*, pages 2357–2368.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

David M. W. Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Preprint*, arXiv:2010.16061.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7:322.

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Archana Tapuria, Talya Porat, Dipak Kalra, Glen Dsouza, Sun Xiaohui, and Vasa Curcin and. 2021. Impact of patient access to their electronic health record: systematic review. *Informatics for Health and Social Care*, 46(2):194–206. PMID: 33840342.

Hanyin Wang, Chufan Gao, Bolun Liu, Qiping Xu, Guleid Hussein, Mohamad El Labban, Kingsley Iheasirim, Hariprasad Reddy Korsapati, Chuck Outcalt, and Jimeng Sun. 2024. Adapting open-source large language models for cost-effective, expert-level clinical note generation with on-policy reinforcement learning. *ArXiv*, abs/2405.00715.

Ping Wang, Tian Shi, and Chandan K. Reddy. 2020. Text-to-SQL generation for question answering on electronic medical records. In *Proc. WWW*, pages 350–361.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. In *Trans. ACL*, volume 4, pages 401–415.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. In *Scientific Data*, volume 10, page 586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proc. ACL*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *Proc. ICLR*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Listing 1: Flan-T5 Prompt

```
f"""Question: {question}

Context: {context}
Instructions:
1. Create a comprehensive, narrative
    answer in paragraph form to the
    question based STRICTLY on the
    provided context sentences
2. Use complete sentences. Do NOT use
    lists
3. Every sentence in your answer MUST be
     directly supported by evidence from
     the context
4. Minimize paraphrasing. Prefer using
    exact phrases from the context for
    medical terms, findings, and actions
5. The answer must not exceed 75 words
6. Preserve all medical terminology
    exactly as it appears. Do not
    simplify
7. Ensure clinical accuracy and a
    professional tone

Answer:
"""
```

Listing 2: BART-Large-CNN / DistilBART-Med-Summary Prompt

```
(f"{context} Based on the text above,
    answer the question: {question}\n"
f"Answer:")
```

Metric	No Aug.	Synonym Aug.	Synonym Aug. + Synth. Data
Strict Macro Precision	70.17	100.00	100.00
Strict Macro Recall	40.18	65.15	65.15
Strict Macro F1	49.08	75.84	75.84
Strict Micro Precision	71.64	100.00	100.00
Strict Micro Recall	34.78	53.62	53.62
Strict Micro F1	46.83	69.81	69.81
Lenient Macro Precision	75.17	100.00	100.00
Lenient Macro Recall	34.87	50.60	50.60
Lenient Macro F1	45.09	63.65	63.65
Lenient Micro Precision	77.61	100.00	100.00
Lenient Micro Recall	27.51	39.15	39.15
Lenient Micro F1	40.62	56.27	56.27
Overall Factuality Score	46.83	69.81	69.81
SARI	66.94	73.46	73.56
BLEU	2.74	3.81	3.85
BERTScore	36.06	43.96	43.68
ROUGE-1	30.88	36.89	36.89
ROUGE-2	23.48	31.45	31.67
ROUGE-L	22.65	25.57	28.99
ROUGE-Lsum	29.76	36.31	36.31
AlignScore	64.37	87.05	89.17
MedCon	38.81	49.85	49.85
Overall Relevance Score	33.87	39.38	39.35
Overall Score	40.35	54.60	54.58

Table 5: Scores for each augmentation type: No Augmentation, Synonym Augmentation, and Synonym Augmentation + Synthetic Data.