DKITNLP at ArchEHR-QA 2025: A Retrieval Augmented LLM Pipeline for Evidence-Based Patient Question Answering

Provia Kadusabe **Abhishek Kaushik** **Fiona Lawless**
Dundalk Institute of Technology
Regulated Software Research Centre
provia.kadusabe@dkit.ie abhishek.kaushik@dkit.ie Fiona.Lawless@dkit.ie

Abstract

This paper describes our submission for the BioNLP ACL 2025 Shared task on grounded Question Answering (QA) from Electronic Health Records (EHRs). The task aims to automatically generate answers to patients' health related questions that are grounded in the evidence from their clinical notes. We propose a two stage retrieval pipeline to identify relevant sentences to guide response generation by a Large Language Model (LLM). Specifically, our approach uses a BioBERT based bi-encoder for initial retrieval, followed by a re-ranking step using a fine-tuned cross-encoder to enhance retrieval precision. The final set of selected sentences serve as an input to Mistral 7B model which generates answers through few-shot prompting. Our approach achieves an overall score of 31.6 on the test set, outperforming a substantially larger baseline model LLaMA 3.3 70B (30.7), which demonstrates the effectiveness of retrieval-augmented generation for grounded QA.

1 Introduction

The widespread adoption of patient portals and digital health platforms has led to a growing volume of patient messages directed to healthcare providers (Martinez et al., 2024; Sieck et al., 2017). Responding to these messages in a timely, accurate, and personalized manner presents a challenge for healthcare providers often contributing to burnout (Stillman, 2023; Shanafelt et al., 2017). The ArchEHR-QA 2025 task aims to develop automated responses to patient messages that are grounded in clinical evidence from their Electronic Health Records (EHRs) (Soni and Demner-Fushman, 2025b).

Large Language Models (LLMs) have recently shown exceptional performance on general domain QA benchmarks (Singhal et al., 2025; Wang et al., 2024). However, directly applying LLMs to clinical EHR-based QA often results in models hallucinating or generating irrelevant details especially if prompted without proper grounding (Jeong et al., 2024; Elgedawy et al., 2024). The key challenge LLMs face is identifying the relevant evidence from patients' lengthy EHRs (Ahsan et al., 2024). To address this, modern QA pipelines often utilize neural retrieval models such as bi and cross-encoders (Karpukhin et al., 2020; Nogueira and Cho, 2019).

Neural retrievers typically serve as the retrieval components in Retrieval Augmented Generation (RAG) frameworks which provide LLMs with grounded document context to mitigate hallucinations and improve factuality (Lewis et al., 2020). Despite their wide adoption in open domain QA, neural retrievers are still underexplored in clinical EHR patient specific QA. A recent review found that most current QA models rely on span extraction methods which are inherently unable to generate coherent answers (Bardhan et al., 2024).

In this work, we propose a two stage retrieval pipeline as shown in figure 1. A bi-encoder first retrieves a broad set of top-K candidate sentences, these sentences are then re-ranked by a fine-tuned cross-encoder to produce top-N sentences. The top-N sentences are ultimately used as context for the LLM response generation.

2 Background & Related work

Previous research in clinical QA has primarily focused on developing datasets that map natural language queries to structured data or extract relevant spans from EHRs (Bardhan et al., 2024). A common approach involves semi-automated template-based generation of QA pairs. For instance, emrQA utilized annotations from i2b2/n2c2 clinical shared tasks to create over 1 million question answer pairs by populating templates with entities from EHRs (Pampari et al., 2018). RxWhyQA focused on extractive QA by leveraging annotated drug–reason relations to produce multi-answer and

BioNLP 2025 Shared Tasks, pages 165–170

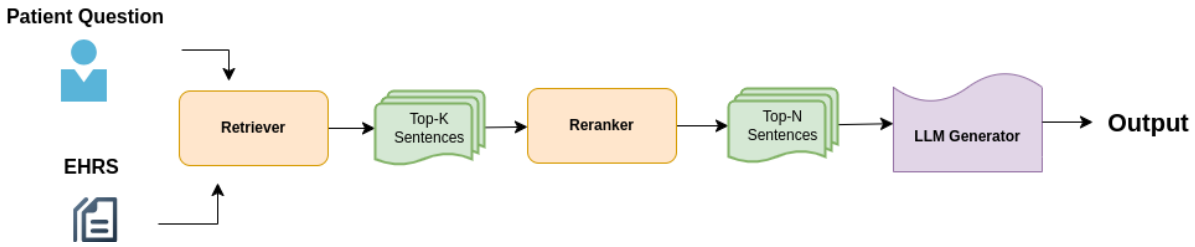

Figure 1: Our retrieval augmented pipeline for patient QA.

multi-focus questions (Moon et al., 2023). Furthermore, DrugEHRQA compiled over 70,000 medication related QA pairs from structured tables and unstructured notes, aiming to support multimodal QA systems (Bardhan et al., 2022). While these datasets have enabled development of clinical QA methods, they often rely on simple rule based or retrieval only methods that lack the capability to generate coherent and accurate answers. Although LLMs can generate coherent responses, they often struggle to extract relevant information from EHRs, which leads to irrelevant outputs (Huang et al., 2025; Maynez et al., 2020). Retrieval methods, such as RAG, have been explored to guide factual generation (Lewis et al., 2020), but existing studies mainly focus on general biomedical QA rather than patient-specific QA (Elgedawy et al., 2024; Xu et al., 2024; Chung et al., 2025; Jiang et al., 2024).

3 Methodology

In this section, we describe our proposed methodology for the task of grounded QA from EHRs.

3.1 Dataset

The dataset used in this study was provided by the organizers of the ArchEHR-QA shared task. It comprises 120 patient cases (20 development and 100 test). Each case includes a patient question, patient narrative and a clinician rewritten version of the patient question, along with the associated clinical notes with pre-annotated sentence numbers for grounding. The development set has relevance labels indicating whether each sentence is *essential*, *supplementary*, or *not-relevant* for answering the question (Soni and Demner-Fushman, 2025a).

3.2 Problem Formulation

Given a dataset $\mathcal{D}$ of patient questions and expert-annotated clinical note excerpts, the task is to classify whether a sentence $s \in \mathcal{S}$ is *essential* for answering a question $q \in \mathcal{Q}$. Each instance includes a label $y \in 0, 1$, defined as:

$$y = \begin{cases} 1 & \text{if } s \text{ is essential,} \\ 0 & \text{otherwise.} \end{cases}$$

The dataset is $\mathcal{D} = (q_i, s_i, y_i)_{i=1}^{N}$, where N is the total number of question-sentence pairs.

3.3 Model Fine-tuning

We fine-tune three BERT-based cross encoders: BERT-base (uncased) (Devlin et al., 2019), BioBERT (Lee et al., 2020) [1], and BioClinicalBERT (Alsentzer et al., 2019) [2] using the dataset described in section 3.2. For each model, the objective is to predict whether a candidate sentence s from the clinical note is *essential* to answer the patient question q.

Input Representation: Each question-sentence pair (q_i, s_i) is concatenated and tokenized as follows:

$$x_i = [\texttt{[CLS]}\ q_i\ \texttt{[SEP]}\ s_i\ \texttt{[SEP]}]$$

The resulting sequence is tokenized with a maximum length of 512 tokens and fed into the transformer encoder to produce contextualized representations:

$$h_i = \text{Transformer}(x_i)$$

The embedding corresponding to the [CLS] token, denoted $h_i^{\texttt{[CLS]}} \in \mathbb{R}^d$, is used as a joint representation of the question and candidate sentence.

3.3.1 Classification and Training

The joint representation is passed through a linear classification head followed by a sigmoid activation to produce a relevance score $\hat{y}_i$:

$$\hat{y}_i = \sigma(W h_i^{\texttt{[CLS]}} + b)$$

[1] https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

where $W \in \mathbb{R}^{1\times d}$ and $b \in \mathbb{R}$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function. The models are optimized using binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

3.4 Retrieval and re-ranking

3.4.1 Bi-encoder Retrieval

For initial retrieval, we adopt a bi-encoder architecture using BioBERT [3]implemented via Sentence-Transformers (Reimers and Gurevych, 2019).

Given a question q and a set of candidate sentences $\{s_j\}_{j=1}^{M}$, we first encode them independently using a bi-encoder architecture:

$$e_q = \text{BiEncoder}(q)$$
$$e_{s_j} = \text{BiEncoder}(s_j), \quad \forall j = 1, \ldots, M$$

where $e_q, e_{s_j} \in \mathbb{R}^d$ are the resulting dense embeddings. Cosine similarity between the question and each candidate sentence is computed as:

$$\text{Sim}(q, s_j) = \frac{e_q \cdot e_{s_j}}{\|e_q\|\|e_{s_j}\|}$$

The top-K candidates with the highest similarity scores are selected for re-ranking:

$$\mathcal{S}_{\text{top}} = \{s_j \mid \text{rank}(\text{Sim}(q, s_j)) \leq K\}$$

where $\text{rank}(\cdot)$ denotes ranking based on similarity in descending order.

3.4.2 Cross-encoder re-ranking

Each of the top-K candidates is concatenated with the question and scored for relevance using the fine-tuned cross-encoder:

$$x_j = [\texttt{[CLS]}\ q\ \texttt{[SEP]}\ s_j\ \texttt{[SEP]}]$$

$$\hat{y}_j = \sigma(W h_j^{\texttt{[CLS]}} + b)$$

where $h_j^{\texttt{[CLS]}}$ is the contextualized embedding of the input, and $\hat{y}_j \in [0, 1]$ is the predicted relevance score. The top-N candidates with the highest scores are selected as evidence for generation:

$$\mathcal{S}_{\text{evidence}} = \{s_j \mid \text{rank}(\hat{y}_j) \leq N\}$$

[3]https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

3.5 Answer Generation

For the Answer generation stage, we employ Mistral-7B-DPO [4], an instruction tuned causal language model denoted as $G(\cdot; \theta)$. This model is based on the Mistral 7B architecture (Jiang et al., 2023) and has been optimized via Direct Preference Optimization (DPO) to follow instruction and align human preferences (Rafailov et al., 2023).

Given a structured prompt P which includes the patient narrative, patient and clinician questions, and the top-N evidence sentences, the model generates free-text answers in an autoregressive manner:

$$A = G(P;\ \theta)$$

where A denotes the generated response and θ represents the pretrained model parameters. The final output consists of sentences that cite supporting evidence by including sentence identifiers inline using pipe symbols.

4 Experiments

4.1 Experimental Setup

We fine-tune the cross-encoder models on the development set using the patient question. Given the small size of the development set, we performed a fixed split over cases to separate training and validation subsets. Finetuning was conducted with a batch size of 8 for up to 10 epochs with early stopping if there is no improvement for 2 consecutive evaluations. Optimization is performed using AdamW with a weight decay of 0.01 and a learning rate of 2×10^{-5}.

For sentence retrieval, we experimented with different combinations of the number of candidates retrieved by the bi-encoder (K) and re-ranked by the cross-encoder (N). Specifically, we evaluated $(K, N) = (5, 20)$, $(7, 20)$, $(10, 25)$, $(12, 30)$, $(13, 30)$, and $(15, 35)$. The configuration $(13, 30)$ yielded the best performance and was adopted in the final retrieval pipeline.

For answer generation, we used a few-shot prompt (Brown et al., 2020) using the two examples provided in the shared task description (Soni and Demner-Fushman, 2025a). Generation was performed with a sampling temperature of 0.70, a maximum length of 200 tokens, and a target answer length of up to 75 words, as specified by the task organizers. If the model produced no output

[4]https://huggingface.co/NousResearch/Nous-Hermes-2-Mistral-7B-DPO

or generated an answer shorter than 65 words or longer than 75 words, generation was retried up to 10 times.

4.2 Evaluation

For sentence retrieval, we evaluated our models on the development set using precision, recall, and F1-score, comparing the retrieved sentences against the manually annotated ground truth. During fine-tuning, we used the same metrics on the development set to assess sentence-level classification performance. The generated responses were assessed using the official evaluation framework provided by the organizers (Soni and Demner-Fushman, 2025b), which balances two key aspects, Factuality and Relevance. Factuality was measured by calculating Precision, Recall, and F1 Scores between the cited evidence sentences in the generated answer and the manually annotated ground truth evidence sentences. Relevance, on the other hand, was assessed by comparing the generated answers to the ground truth essential note sentences and the questions using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). The overall score was computed as the mean of the Factuality and Relevance scores.

4.3 Experimental Results & Discussion

Experimental results on the development set show that among the fine-tuned models, shown in table 1, BioBERT achieved the best performance and was therefore selected as the cross-encoder re-ranker in the retrieval pipeline.

Model	Precision	Recall	F1-Score
BioClinicalBERT	41.49	72.46	52.77
BERT-base	46.06	80.43	58.58
BioBERT	**51.45**	**89.86**	**65.44**

Table 1: Performance of fine-tuned cross-encoders on the essential sentence prediction task in (%).

We also compare our system with using only few-shot prompting as shown in table 2.

Few-shot prompting achieved a slightly higher overall factuality score (47.90 vs. 45.45), however, our system outperformed it in overall relevance (35.71 vs. 31.08) and overall score (40.58 vs. 39.49). Based on these results, we selected the RAG system for testing.

Metric	RAG	Few-Shot Only
Overall Factuality Score	45.45	47.90
Overall Relevance Score	35.71	31.08
Overall Score	**40.58**	39.49

Table 2: Comparison of our system (RAG) with few-shot prompting only (no retrieval). Both methods use the Mistral 7B model.

Metric	RAG	Baseline
Overall Factuality Score	32.70	33.60
Overall Relevance Score	30.50	27.80
Overall Score	**31.6**	30.70

Table 3: Performance of our system (RAG) on the test set.

Evaluation on the test set in table 3 showed that our system achieved an overall relevance score of 30.50, outperforming the baseline score of 27.80. This suggests that our system's generated answers were more aligned to the ground-truth essential note sentences. However, it slightly underperformed in the overall factuality with a score of 32.70 compared to the baseline score of 33.60. Despite this, our system achieved a higher overall score of 31.6, surpassing the baseline score of 30.7, which was based on LLaMA 3.3 70B. While our model (Mistral 7B parameters) is significantly smaller than the LLaMa 70B model used in the baseline system, it still delivers competitive results which shows the effectiveness of retrieval augmented generation for grounded clinical question answering.

5 Conclusion & Future Work

In this work, we introduced our approach for the grounded patient QA task using EHRs. Our method uses a two stage retrieval pipeline using a BioBERT based bi-encorder for initial relevant sentence retrieval and a fine-tuned cross-encoder for re-ranking to identify the most relevant sentences for LLM (Mistral 7B) generation. Experimental results show that our proposed approach improves performance over the baseline in terms of overall score (31.6 versus 30.70).
Future work should investigate alternative model architectures and evaluate the performance of smaller LLMs on larger datasets.

6 Limitation

Our study was constrained by several factors. First, the development set used for fine-tuning was relatively small thus using a larger dataset could yield better performance. Second, our fine-tuning experiments utilized smaller pretrained language models due to resource constraints, exploring larger LLMs could further improve performance.

7 Acknowledgments

This research was funded through the CREATE-DkIT project, supported by the HEA TU-Rise program and co-financed by the Government of Ireland and the European Union through the Southern, Eastern & Midland Regional Program of the ERDF 2021-27 and the Northern & Western Regional Programme 2021–27.

References

Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C Wallace. 2024. Retrieving evidence from ehrs with llms: possibilities and challenges. *Proceedings of machine learning research*, 248:489.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries. *arXiv preprint arXiv:2205.01290*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

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.

Ran Elgedawy, Ioana Danciu, Maria Mahbub, and Sudarshan Srinivasan. 2024. Dynamic q&a of clinical documents with large language models. *arXiv preprint arXiv:2401.10733*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Emily Jiang, Alice Chen, Irene Tenison, and Lalana Kagal. 2024. Medirag: Secure question answering for healthcare data. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6476–6485. IEEE.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kathryn A Martinez, Rebecca Schulte, Michael B Rothberg, Maria Charmaine Tang, and Elizabeth R Pfoh. 2024. Patient portal message volume and time spent on the ehr: an observational study of primary care clinicians. *Journal of General Internal Medicine*, 39(4):566–572.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, Jungwei Wilfred Fan, and 1 others. 2023. Extractive clinical question-answering with multianswer and multifocus questions: data set development and evaluation study. *JMIR AI*, 2(1):e41818.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Tait D Shanafelt, Lotte N Dyrbye, and Colin P West. 2017. Addressing physician burnout: the way forward. *Jama*, 317(9):901–902.

Cynthia J Sieck, Jennifer L Hefner, Jeanette Schnierle, Hannah Florian, Aradhna Agarwal, Kristen Rundell, and Ann Scheck McAlearney. 2017. The rules of engagement: perspectives on secure messaging from experienced ambulatory patient portal users. *JMIR medical informatics*, 5(3):e7516.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

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.

Michael Stillman. 2023. Death by patient portal. *JAMA*, 330(3):223–224.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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. *Scientific data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.