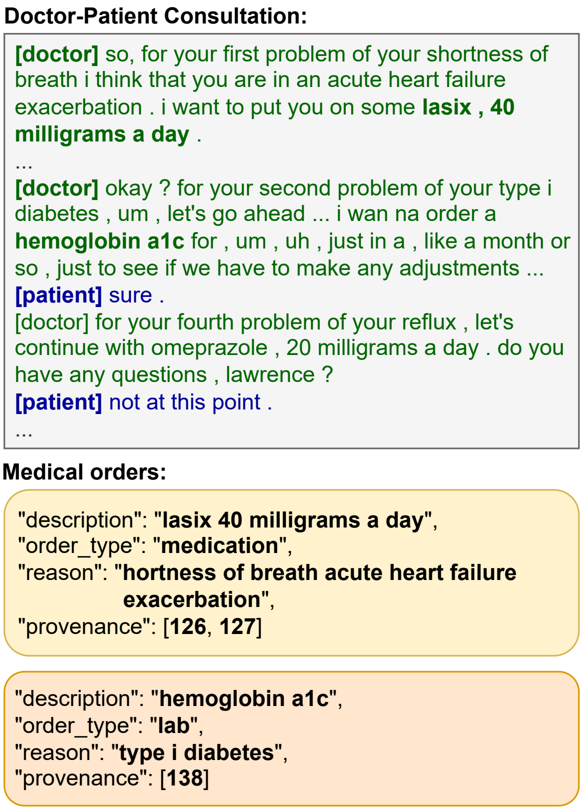
Overview of the MEDIQA-OE 2025 Shared Task on Medical Order Extraction from Doctor-Patient Consultations

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, Francois Beaulieu
Microsoft Healthcare & Life Sciences
Correspondence: {jcorbeil,abenabacha}@microsoft.com

Abstract

Clinical documentation increasingly uses automatic speech recognition and summarization, yet converting conversations into actionable medical orders for Electronic Health Records remains unexplored. A solution to this problem can significantly reduce the documentation burden of clinicians and directly impact downstream patient care. We introduce the MEDIQA-OE 2025 shared task, the first challenge on extracting medical orders from doctor-patient conversations. Six teams participated in the shared task and experimented with a broad range of approaches, and both closed- and open-weight large language models (LLMs). In this paper, we describe the MEDIQA-OE task, dataset, final leaderboard ranking, and participants' solutions.



Doctor-Patient Consultation:

[doctor] so, for your first problem of your shortness of breath i think that you are in an acute heart failure exacerbation . i want to put you on some **lasix , 40 milligrams a day** .
...
[doctor] okay ? for your second problem of your type i diabetes , um , let's go ahead ... i wan na order a **hemoglobin a1c** for , um , uh , just in a , like a month or so , just to see if we have to make any adjustments ...
[patient] sure .
[doctor] for your fourth problem of your reflux , let's continue with omeprazole , 20 milligrams a day . do you have any questions , lawrence ?
[patient] not at this point .
...

Medical orders:

"description": "**lasix 40 milligrams a day**",
"order_type": "**medication**",
"reason": "**hortness of breath acute heart failure exacerbation**",
"provenance": [**126**, **127**]

"description": "**hemoglobin a1c**",
"order_type": "**lab**",
"reason": "**type i diabetes**",
"provenance": [**138**]

Figure 1: The medical order extraction task takes a doctor-patient dialog and extracts a JSON list of orders containing four keys (description, order_type, reason, and provenance). Orders that were previously prescribed but not explicitly renewed should be excluded (e.g. omeprazole in this example).

1 Introduction

In recent years, the burden of clinical documentation has reduced the time clinicians can devote to direct patient care, and ultimately limited the number of patients physicians can help. To mitigate this, many hospitals and clinics now deploy automatic speech recognition and note summarization tools during consultations. A natural next step in this pipeline is medical order extraction (e.g., medications, labs, imaging, follow-ups) from conversation transcripts to directly populate Electronic Health Records (EHRs).

While named entity recognition (NER) and relation extraction (RE) have been extensively studied in clinical NLP[1] (Xu et al., 2010; Doan and Xu, 2010; Yang et al., 2020; Fabacher et al., 2025; Henry et al., 2020; Lybarger et al., 2023), extracting actionable, structured orders from full-length consultations remains underexplored despite its potential impact. The task is challenging: inputs are long, dialogues contain interruptions as well as revisions, and outputs combine schema-constrained fields (e.g., order type) with free-text attributes (e.g., description, reason). These challenges are compounded by distributional shifts, as clinicians adapt their language to patients without medical training during consultations.

In this era of LLMs (Brown et al.; Achiam et al., 2023), new approaches have become feasible for the medical order extraction task — combining

[1] Natural Language Processing.

Proceedings of the 7th Clinical Natural Language Processing Workshop, pages 11–16
October 30, 2025 ©2025 Association for Computational Linguistics

Table 1: Final Ranking of MEDIQA-OE competition on the test set (100 samples) in which 6 teams participated. Our two baselines (excluded from the ranking) are a simple one-shot prompt for MediPhi-Instruct (1) and GPT-4o-0806 (2) with one example from the training set.

Rank	Team Name	Method	Match	Desc.	Reason	Type	Prov.	AVG
1	WangLab	GPT-4 constrained dec. Detailed instructions	**81.8**	**66.8**	29.5	**81.5**	**63.0**	**60.2**
2	silver_shaw	Gemini 2.5 Pro w/ thinking Detailed plan & instructions	76.4	64.1	**41.3**	74.7	60.4	60.1
3	MISo KeaneBeanz	Qwen3 32B Q4_K_M w/o thinking Instructions w/ 2 shots	73.4	58.0	35.6	71.6	48.4	53.4
4	EXL Health AI Lab	MedGemma 27B One shot (short format)	67.7	54.5	30.5	66.2	52.5	50.9
5	MasonNLP	Llama4 17B 16E Instruct One shot w/o orders	55.5	39.1	19.8	50.9	41.3	37.8
-	`Baseline 2`	GPT-4o Simple prompt w/ one shot	63.6	39.5	20.4	59.3	1.0	30.1
-	`Baseline 1`	MediPhi-Instruct 3.8B Simple prompt w/ one shot	43.3	25.8	19.5	39.6	13.8	24.7
6	HerTrials	Llama3.2 3.2B Instructions w/ one shot	31.2	19.6	9.0	29.6	5.6	15.9

long-context reasoning with schema-aware generation — yet limitations in context length, controllability, and calibration persist. The MEDIQA-OE shared task[2] investigates these challenges and benchmarks solutions to improve EHR clinical documentation, which we believe can reduce the burden on providers while ensuring the accurate capture of critical patient orders.

2 Previous Work

Tasks similar to order extraction in clinical NLP are commonly formulated as NER and RE. Early systems were rule-based (e.g., MedEx by Xu et al. (2010)) or used classical machine learning such as support vector machines (Doan and Xu, 2010). With pretrained contextual encoders, fine-tuned transformer models (e.g., BERT (Devlin et al., 2019), ClinicalBERT (Alsentzer et al., 2019)) became the standard for NER/RE and delivered consistent gains on clinical benchmarks (Yang et al., 2020; Fabacher et al., 2025).

More recently, LLMs enable span-free formulations that cast extraction as reading-comprehension style generation. Prompting methods (Peng et al., 2023; Cui et al., 2023; Peng et al., 2024) have shown strong results on several clinical information extraction tasks, including adverse drug events (Henry et al., 2020) and social determinants of health (Lybarger et al., 2023). However, order extraction from full patient–doctor dialogues remains underexplored, particularly when models must (i) handle long, multi-speaker inputs and (ii) generate outputs that mix schema-constrained fields (e.g., order type, provenance) with free-text attributes (e.g., description, reason).

3 Methodology

3.1 Source Datasets

The long-form doctor-patient conversations used for the order-extraction task are primarily drawn from two datasets: ACI-Bench (Yim et al., 2023) and PriMock57 (Papadopoulos Korfiatis et al., 2022). The ACI-Bench corpus comprises 207 naturalistic conversations between physicians and patients, curated by domain experts to reflect real-world clinical interactions. Similarly, the PriMock57 dataset contains 57 mock doctor-patient dialogues, designed to simulate clinical scenarios in a controlled setting. Recent works such as Notechat (Wang et al., 2024) has introduced large-scale synthetic dialogue datasets. While this corpus is the largest, we excluded it due to the prevalence of low-quality dialogues we observed.

3.2 Annotations

We asked medically trained annotators to produce the gold-standard medical orders for the

[2] https://sites.google.com/view/mediqa-2025

high-quality conversations of Primock57 and ACI-Bench. Annotation guidelines instructed to assess every medical order of type medication, imaging, lab, or follow-up within the conversation the way a doctor would create them in the EHR. This was intended to replicate doctors' current process executed at the end of a patient encounter. We measured an inter-annotator agreement of 0.768. We sampled 100 examples containing 255 medical orders across both data sources as a test set and kept the others as training set (64 samples) used for few-shot prompting, and development set (100 samples) (Corbeil et al., 2025).

3.3 Evaluation

We evaluate model performance across four key metrics: description, reason, type, and provenance. Results are reported after performing a matching between reference and hypothesis orders based on description field's word overlap[3]. An intermediary metric, the match score, is computed from this alignment process as the F1 between reference and predicted orders without looking at the content, thus specifically accounting for the amount of fabricated or omitted orders. It represents an upper bound for other metrics that are penalized for empty values for fair comparison. For description and reason metrics, we compute F1 scores of the rouge metric (Lin, 2004) over unigrams. Type is evaluated using accuracy due to its limited label space, and provenance is assessed via F1 score over provenance labels[4]. Finally, the leaderboard ranking is assessed via the average of all four key metrics: description, reason, type, and provenance.

4 Results

4.1 Leaderboard Ranking

We provided in Table 1 the final leaderboard of the MEDIQA-OE along participants' approaches and our two baselines, which were used as reference points while being excluded from the ranking. All solutions are based on prompting language models. While there are two closed-source LLMs at the top of the ranking, the remaining submissions are leveraging open-weight LLMs in few-shot settings. WangLab obtained the 1^{st} rank of the competition by prompting GPT-4 (Achiam et al., 2023) with JSON-constrained decoding and detailed instructions. Following closely by 0.1% on the average score, silver_shaw (Mehta, 2025) achieved the 2^{nd} place by using Gemini 2.5 Pro (Comanici et al., 2025) in thinking mode. The other approaches (Balachandran et al., 2025; Karim and Özlem Uzuner, 2025) leveraged different open-weight models in few-shot settings such as Qwen3 32B (Qwen Team, 2025), MedGemma 27B (Sellergren et al., 2025), Llama4 Scout 17B (Meta AI, 2025) and Llama3.2 3.2B (Meta AI, 2024). Participants only appended one or two shot(s) examples due context limitations from long input-output pairs, and some even reduced examples into shorter formats. Overall, they also wrote simpler prompts compared to the two closed-weight LLM solutions.

4.1.1 WangLab's Approach

WangLab won the competition by prompting GPT-4 (Achiam et al., 2023) in a zero-shot setting. They obtain an average score of 60.2% with JSON-constrained decoding on the order format as well as using very detailed rules in the instructions. They achieved the highest match score at 81.8%, which indicates that a large proportion of reference orders are well matched. The average gains are double digits over the `Baseline 1` based on GPT-4o with improvements on the provenance (+62.0%), description (+27.3%), type (+22.2%) and reason (+9.1%) scores.

Their prompt provides very detailed instructions, and is as follows:

1. Role attribution
2. Transcript definition with example
3. Task definition
4. Type definitions with rules and examples
5. Output JSON key definitions
6. Reason guidelines with specific examples
7. JSON output example
8. Overall guidelines
9. Eliciting JSON output

4.1.2 silver_shaw's Approach

Following closely 0. 1% on the average score, silver_shaw (Mehta, 2025) achieved 2^{nd} position with the highest reason score at 41. 3% by prompting Gemini 2.5 Pro (Comanici et al., 2025) in thinking mode with a detailed reasoning plan and instructions.

[3] Necessary to compare orders with each other.

[4] Turn numbers where the order originates in the transcript.

Their one-call prompting approach asks the model to proceed in three steps aimed at mirroring the clinical reasoning processes: chain-of-thought analysis, self-critique & verification, and deterministic JSON generation.

4.1.3 MISo KeaneBeanz's Approach

MISo KeaneBeanz's approach reached the 3^{rd} rank by prompting the 4-bit quantized open-weight model Gwen3 32B (Qwen Team, 2025) in a two-shot setting.

4.1.4 EXL Health AI Lab's Approach

EXL Health AI Lab achieved the 4^{th} rank at 50.9% leveraging a one-shot solution prompting MedGemma 27B (Sellergren et al., 2025), an open-weight medical LLM. Their experiments covered agentic workflows such as ReAct (Yao et al., 2023) and a four-step multi-agent pipeline. The one-shot method remained more accurate potentially because of the negative impact of noises introduced by multi-step approaches.

4.1.5 MasonNLP's Approach

The 5^{th} rank of the MEDIQA-OE competition was attributed to MasonNLP (Karim and Özlem Uzuner, 2025). They used Llama4 17B (Meta AI, 2025) in a minimal one-shot prompting setting. The authors also reported an experiment with Llama4 8B.

4.1.6 HerTrials' Approach

HerTrials team ranked 6^{th} with a one-shot prompting of the smallest open-weight language models Llama3.2 (Meta AI, 2024).

4.2 Analysis of Open-weight LLMs

We show the correlation between final accuracy and open-weight model sizes in Figure 2. Despite prompt variations, we computed a strong Pearson correlation of 0.981 between leaderboard ranking and model sizes, which is in line with previous work in clinical NLP (Dada et al., 2025).

5 Discussion and Limitations

In spite of top-ranking solutions achieving considerable scores with zero- and few-shot prompting and reasoning, significant gaps remain to push further the performance of the medical order-extraction task.

First, we notice that the maximum match F1 score is of 81.8%, which means that there is still room of nearly 20% to match the number of orders in the reference.

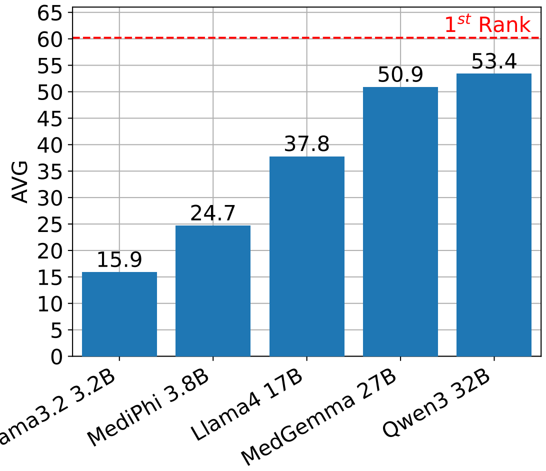


Figure 2: Open-weight models ranking obtained with few shots correlates with parameter count.

Second, description and provenance are both lagging behind the raw match score (i.e., their upperbound) by approximately 15-18%. Provenance was considerably improved in this challenge by using: larger models, constrained decoding, and specific instructions. Future work could explore embedding-based and hybrid systems.

Third, order types are all very close to match scores, which highlights how such classification tasks are well suited for LLMs.

Fourth, we observe low performances on the reason field which might come from the dispersion of reasons across the conversation and the fact that it is an optional field with scarcer annotations.

One of the main limitations of the current task is the small dataset sizes. The current trainset size of 64 samples limits the ability to use it for finetuning — which could particularly make open-weight small language models more competitive. Future work might produce larger datasets or leverage synthetic ones. While the inter-annotator agreement is considerably high, annotations might also present noises (e.g., span boundaries, non-expert conversational style instead of formal writing, etc.) which limit the maximum score below 100%.

6 Conclusion

To conclude, the medical order-extraction task was tackled by a variety of zero- and few-shot approaches using open- and closed-weight LLMs. Closed-weight models such as GPT-4 and Gemini 2.5 Pro in zero-shot setting dominated the top ranks, leveraging detailed instructions, constrained decoding and reasoning. We observed a significant correlation of 0.981 between open-weight model sizes in few-shot settings and final accuracy. Although final scores considerably improved over the baselines especially in the match and provenance metrics, we still observe a significant gap in total extracted orders performance from the match score of 81.8% as well as in performances on the description and the reason free-form fields. We believe future works include synthetic data generation, model fine-tuning, hybrid systems, and focus on improving small language models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Abhinand Balachandran, Bavana Durgapraveen, Gowsikkan Sikkan Sudhagar, VIDHYA VARSHANY J S, and Sriram Rajkumar. 2025. Exl health ai lab at mediqa-oe 2025: Evaluating prompting strategies with medgemma for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. Language models are few-shot learners.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel A. del Agua, Jérôme Tremblay, Akila Jeeson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, François Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. *CoRR*, abs/2507.05517.

Yang Cui, Lifeng Han, and Goran Nenadic. 2023. Medtem2. 0: Prompt-based temporal classification of treatment events from discharge summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Julian.friedrich@uk-essen.de Julian.friedrich@uk-essen.de, and Jens Kleesiek. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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.

Son Doan and Hua Xu. 2010. Recognizing medication related entities in hospital discharge summaries using support vector machine. In *Proceedings of COLING. International conference on computational linguistics*, volume 2010, page 259.

Thibaut Fabacher, Erik-André Sauleau, Emmanuelle Arcay, Bineta Faye, Maxime Alter, Archia Chahard, Nathan Miraillet, Adrien Coulet, and Aurélie Névéol. 2025. Efficient extraction of medication information from clinical notes: an evaluation in two languages. *Preprint*, arXiv:2502.03257.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

A H M Rezaul Karim and Özlem Uzuner. 2025. Masonnlp at mediqa-oe 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.

Parth Mehta. 2025. silver_shaw at mediqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical*

Natural Language Processing Workshop. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile\-devices/. Blog post announcing Llama 3.2 models with lightweight text and vision capabilities, consulted on 2025-08-18.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Blog post on Meta AI's launch of Llama 4 models, consulted on 2025-08-18.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Cheng Peng, Xi Yang, Kaleb E Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, 153:104630.

Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.

Qwen Team. 2025. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/. Blog post, published by the Qwen Team, consulted on 2025-08-18.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.