# HiTZ-IXA at ArchEHR-QA 2026: Evidence Alignment Through Self-Consistency and Prompt Curation in Memory-Constrained Environments

**Xabier Irastortza-Urbieta, Maite Oronoz, Alicia Pérez**
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{xabier.irastorza, maite.oronoz, alicia.perez}@ehu.eus

**Abstract**

The development of question-answering systems capable of grounding their answers in Electronic Health Records could provide patients with faithful assistance while reducing the clinical workload. The ArchEHR-QA 2026 Shared Task was organized to advance progress in this context. In this paper, we present our strategies for addressing this shared task, which are focused primarily on evidence alignment and, to a lesser extent, on evidence identification. Our approaches rely exclusively on open-source models with up to 8 billion parameters, aiming to produce systems suitable for environments with memory constraints. We experimented with methods based on embedding models, prompt curation, self-consistency, and combination of LLMs. We concluded that prompt curation together with an effective post-processing step was crucial for creating stable systems, while self-consistency yielded considerable gains in performance. The results of our approaches suggest that small LLMs can substantially improve their accuracy in the evidence alignment task via simple and affordable techniques.

**Keywords:** Medical NLP, Clinical LLM, ArchEHR-QA

## 1. Introduction

One of the key challenges in clinical computer science is the creation of conversational agents capable of answering medical questions asked by laypeople. A substantial proportion of the health-related questions posed to medical professionals are already addressed within patients' own medical reports. However, the large volume of clinical data and the complexity of medical language make locating this information challenging for non-expert users. Developing systems capable of automatically and reliably extracting relevant answers from clinical documentation and presenting them to end users could significantly reduce the clinical workload. Large Language Models (LLMs) have shown significant medical knowledge (Singhal et al., 2025), but applying that knowledge in practical situations—such as answering patient questions by grounding responses in an Electronic Health Record (EHR) in a reliable way—remains a challenge.

In this context, the second edition of the *ArchEHR-QA Shared Task: Grounded Question Answering from Electronic Health Records* has been organized (Soni and Demner-Fushman, 2026b). The Shared Task aims to encourage research into creating a system capable of handling a scenario where a patient provides a narrative—containing inaccuracies and an underlying question—along with an EHR related to that narrative, and the system must produce an answer correctly grounded in the EHR. This paper presents a compilation of our approaches to solving the two subtasks related to evidence, identification, and alignment in this Shared Task. All the code implemented and the prompts employed are made available in a repository[1] to aid reproducibility.

## 2. Related Work

Interest in medical question-answering systems dates back a long time, and extensive research has been conducted on the topic (Bardhan et al., 2024). Traditionally, the evaluation of medical systems was carried out using multiple-choice datasets, with questions extracted from medical exams and other sources (Pal et al., 2022). Recently, research has been shifting toward more practical approaches, where the generation of expert-level free-text answers and the grounding of those answers to ensure reliability play a crucial role. In this context, several challenges have emerged, ranging from hallucinations and fairness biases to legal considerations (Yang et al., 2025b). To address these challenges, new systems are being developed, such as Med-PaLM 2 (Singhal et al., 2025), which is a milestone worth mentioning.

Methodologically, LLM-based agent systems represent the primary approach, and research has shown that some systems have achieved expert-level performance (Al Radi et al., 2025). The usual framework involves fine-tuning a backbone LLM on the medical domain. Then, the main challenge lies in correctly managing the vast informa-

[1] Repository containing the code and prompts: https://github.com/XabierIU/HiTZ-IXA-ArchEHR-QA



*Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 434–440
12 May 2026. 

tion and reasoning capabilities of the LLM. To address this, emerging agent systems are being developed based on different methods: tool-augmented agents (Schick et al., 2023), agents with long-term memory modules (Zhang et al., 2024), and multi-agent systems (Kim et al., 2024).

However, there is a gap in the research concerning the integration of information from medical records when answering patients' clinical questions. To foster research in this area, the first edition of the ArchEHR-QA Shared Task was presented in 2025 (Soni et al., 2025). In that edition, many participants developed systems commonly found in the state of the art, such as multi-agent systems (Hwang et al., 2025) and prompt-optimization techniques (Bogireddy et al., 2025).

## 3. Data and Models

We used exclusively the data provided by the Shared Task organizers (Soni and Demner-Fushman, 2026a), consisting of 167 clinical cases. Of these, the first 20 included answers for development. Cases 121–167 were reserved by them for Subtask 2 evaluation, and cases 21–167 for Subtask 4 evaluation. Each case in the development dataset contains: a free-text patient question, a clinical-interpreted concise question, an EHR with the necessary data to answer the question, and an answer text that addresses the clinical question, along with the reference numbers of the EHR sentences on which each sentence of the answer is grounded.

In order to select backbone language models, we followed two principles. On the one hand, we decided to work with open-source models executed on local GPUs; on the other hand, we used small LLMs of up to 8 billion parameters to avoid excessive memory consumption. We believe that these two principles ensure the applicability of our systems in memory-constrained environments, while avoiding dependence on non-free APIs.

Specifically, we tested the Aloe (Garcia-Gasulla et al., 2025), MedGemma (Sellergren et al., 2025), Qwen3 (Yang et al., 2025a), and Llama Instruct models—all with 8 billion parameters except for MedGemma, which has 4 billion. Regarding the reasons for this selection, we selected the Aloe and Qwen models because the participants in the previous year's edition found them to be successful in addressing this task (Le et al., 2025). In the case of Aloe 8B, it was shown to be even more effective than the 70B version (Cuadron Cortes et al., 2025). MedGemma is claimed to have expert-level capabilities in clinical texts (Al Radi et al., 2025), and Llama was used as a baseline model in ArchEHR-QA 2025 (Soni et al., 2025).

## 4. Tasks Description and Evaluation

The Shared Task is composed of four subtasks: Question Interpretation, Evidence Identification, Answer Generation, and Evidence Alignment. We participated in two of them: the second and the fourth. The aim of the second subtask is, given a patient narrative, a clinical question extracted from that narrative, and an EHR, to identify the essential sentences from that EHR needed to answer the aforementioned clinical question. Meanwhile, in the fourth subtask, the system is provided with the narrative, the question, the EHR, and the answer to the question. For each sentence of the answer text, the objective is to determine which sentences in the EHR it is grounded in.

Regarding evaluation, both tasks are assessed using micro precision, recall, and F1 metrics, and the systems are ranked according to the F1 metric, which is referred to as the overall score.

## 5. Subtask 2: Evidence Identification

In the previous edition of the Shared Task, most participants divided their approaches into two parts (Soni et al., 2025): identifying the essential sentences from the EHR and generating the answer text. The first part corresponds to Subtask 2 of the current edition. Although we focused our work on Subtask 4, we conducted some minor experiments aimed at improving upon the systems from the previous year's edition.

Our approach to this subtask involved employing LLM agents in a zero-shot setting, complemented by prompt refinements. All experiments were made using the Aloe model, described in Section 3.

We observed that some EHRs are considerably longer than others, ranging from fewer than 10 sentences to up to 40 sentences. We hypothesized that longer EHRs might be more challenging due to the potential difficulty small models face in attending to all sentences effectively. To address this, we proposed two approaches:

- To divide the EHR into smaller segments (which we referred to as windows) and input one segment per execution to the agent, thereby reducing the input size. This approach is illustrated in the row *Window* in Table 1.

- To perform two separate executions of the model: one asking it to identify essential sentences, and the other asking it to identify non-relevant sentences. The results from both executions would then be combined by a third agent. For the first two agents, we requested that they output their reasoning to help guide the final agent in producing the final result. In this way, we explored two complementary

| Set | Approach | Prec. | Rec. | F1 |
|---|---|---|---|---|
| Dev | Aloe-basic | 40.84 | 58.84 | 48.16 |
| | Window | 41.43 | 61.16 | 49.40 |
| | Ess-Nrev | **46.65** | **62.55** | **53.25** |
| Test | Aloe-basic | 38.64 | **54.70** | 45.29 |
| | Ess-Nrev | **44.24** | 47.92 | **46.01** |

Table 1: Micro precision, recall and F1 of our approaches in Subtask 2.

strategies for solving the task, generating additional information to support the final decision. Row *Ess-Nrev* in Table 1.

Table 1 presents the results obtained with the approaches described above, evaluated on both partitions using the official scoring program. To determine the window size for the *Window* approach, we performed a grid search on the development set and found that the best results were achieved with a window size of $W = 15$, meaning that 15 sentences were input per execution to the agent. Note that the sentence sets are disjoint from one another. It can be observed that while our approaches yielded slight improvements on the development set, these gains did not generalize to the test set, likely due to the high variability in EHR characteristics across the two partitions.

## 6. Subtask 4: Evidence Alignment

Our main focus rests on the fourth subtask. The approaches followed are presented in the following subsections.

### 6.1. Avoiding LLMs: Embedding Models

The first step we considered for solving this task was to attempt to avoid computationally costly LLM systems by using embedding models. In order to find which EHR sentences served as the basis for each answer sentence, we followed the procedure outlined as follows: We calculated the embedding vectors for each sentence in the EHR and the answer text using two language models: BERT (Devlin et al., 2019) and MedBERT (Vasantharajan et al., 2022). Then, for each answer sentence, we computed its cosine similarity with all EHR sentences. Finally, we determined that the EHR sentences with a cosine similarity higher than a threshold with value $T$ were considered the basis for the current answer sentence. In practice, the value of the threshold was set using the development set by identifying the value that yielded the best results. After performing a grid search, it was determined that a cosine similarity higher than $T = 0.60$ was the optimal threshold.

Table 2 shows the performance of the embeddings from BERT (trained on generalist texts), MedBERT (trained on clinical texts), and a random baseline. The random system generates random EHR reference numbers for each answer sentence. The substantial influence of the training domain is noteworthy: the model trained on clinical texts, MedBERT, achieves a considerably higher performance, whereas BERT's performance is close to that of the random system. In any case, the performance of the embedding models remained limited and, therefore, we decided to explore alternative approaches.

| Embedding | Prec. | Rec. | F1 |
|---|---|---|---|
| Random | 6.82 | 37.28 | 11.52 |
| BERT | 9.67 | 42.75 | 15.77 |
| MedBERT | **39.07** | **42.75** | **40.83** |

Table 2: Performance in the development set of similarity with different embedding models.

### 6.2. LLM Agents and Prompt Adjustment

Our second approach was to develop an LLM-based agent system using the backbone models mentioned in Section 3. As an initial step, we sought to assess the capabilities of these models in a zero-shot setting.

**Prompt curation** A critical component of such systems is the prompt. Hence, several variants of a crafted prompt structure were explored.

- First: we took this as a prompt template. In this prompt, first, we assign a role to the LLM; then we provide the EHR, followed by the question and the answer text. Finally, we include a block of instructions and constraints to ensure a predictable and structured output. The curated prompt is given in Appendix TBA.
- NoQu: the crafted prompt (i.e. First) without the clinical question;
- CoT: the crafted prompt with a Chain of Thought block (Wei et al., 2022);
- Invr: the inverse prompt presenting the answer first, then the EHR, and finally the question;
- NoIns: the crafted prompt without the instructions block, included as a baseline.

Table 3 presents the impact of each prompt on alternative backbone models (i.e. Aloe, Qwen, Llama, MegGemma abbreviated as MedGm).

**Post-processing:** It is known that LLMs can sometimes produce unexpected outputs. Our prompts instruct the LLM to output an exact copy of the answer text, divided into sentences, with

the EHR sentence references for each sentence included in brackets after the period. This format was selected after experimenting with other formats that yielded poorer results, such as outputting directly in JSON format or omitting the answer sentences copies. To handle outputs that do not strictly follow this format, we developed a post-processing step with the following capabilities:

- Ability to identify the end of each sentence using different reference characters (e.g., periods, newlines, brackets).
- Ability to match each output sentence with the corresponding answer sentence by finding the longest common character sequence.
- Ability to detect and correct sentence-order inconsistencies between the answer text and the output text.
- Ability to identify and remove duplicate sentences.

This post-process ensures the correct interpretation of the LLM outputs and ultimately enhances the overall performance of the systems.

**Results using Zero-shot:** The results attained by Zero-shot learning approaches with different prompts and followed by the post-processing are shown in Table 3.

| Prompt | Aloe | Qwen | Llama | MedGm |
|---|---|---|---|---|
| First | **54.14** | 54.56 | **53.57** | **52.51** |
| NoQu | 53.90 | 50.48 | 49.20 | 51.87 |
| CoT | 52.01 | 47.91 | 51.10 | 39.21 |
| Invr | 45.10 | **60.56** | 46.55 | 47.57 |
| NoIns | 39.82 | 39.56 | 18.14 | 25.94 |

Table 3: Performance on the development set of Zero-shot of various LLMs using different prompt templates. Results represent the mean micro F1 scores across 10 repeated runs.

From Table 3, it can be observed that all LLMs achieved their best results with the crafted prompt, except for Qwen, which performed best with the inverse prompt. The Chain of Thought block did not lead to performance improvements, a finding consistent with reports from some participants in the previous edition of this Shared Task on other subtasks (Cuadron Cortes et al., 2025). Furthermore, it can be noted that the models whose performance varied the least across different prompts were Aloe and Qwen, and that structuring the prompt did not yield better results.

### 6.3. Self-Consistency Techniques

Self-consistency (Wang et al., 2022) was found to be relevant in related work (Bogireddy et al., 2025) as it tends to improve results for evidence identification and answer generation. In an attempt to explore the benefits of self-consistency, we followed the steps outlined below:

- Selection of the best approach (LLM and prompt). Based on Table 3, we opted for Aloe and Qwen respectively with First and Invr prompts.
- Each approach was executed $N$ times for each example in the dataset. Note that Zero-shot models are not fully-deterministic, instead, the responses might vary.
- Rely on evidence. After the $N$ executions, for each sentence in the answer text, a result was eligible provided that it was present in a percentage ($MV$) of the executions. That is a minimum voting threshold was applied as eligibility criterion.

We conducted a grid search on the development set to determine the minimum threshold percentage $MV$ (with $MV \in \{0, 0.25, 0.5, 0.75, 1\}$) and number of executions $N \in \{3, 5, 7, 9\}$. For both models and for all the values tested for $N$, the optimal value for $MV$ was $0.75$. Each evaluation was conducted 5 times to gain insights about the stability of each approach.

| N | LLM | Prompt | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| 3 | Aloe | First | 50.00 | 65.87 | 56.80 |
| | Qwen | Invr | 60.58 | **70.63** | **65.14** |
| 5 | Aloe | First | 57.10 | 62.60 | 59.70 |
| | Qwen | Invr | **62.61** | 65.74 | 64.12 |
| 7 | Aloe | First | 53.19 | 66.42 | 59.06 |
| | Qwen | Invr | 58.98 | 66.62 | 62.40 |
| 9 | Aloe | First | 56.09 | 65.55 | 60.43 |
| | Qwen | Invr | 60.00 | 68.82 | 64.01 |

Table 4: Performance on the development set of LLMs enhanced with self-consistency techniques varying the number of executions (N) for Aloe and Qwen. Results represent the mean micro scores across 5 repeated runs of each system.

Table 4 presents the results of our agents incorporating the self-consistency criterion. It can be observed that this approach leads to a significant improvement in the results. The number of iterations does not appear to be highly critical; performing just three executions is sufficient to achieve notable improvements. From five or seven executions onward, the gains become less substantial compared to those obtained with fewer executions.

### 6.4. Cross-Consistency

While in the previous section we looked for self-consistency of a single approach, next we com-

bined both Aloe and Qwen as a committee of experts or ensemble LLMs. Considered together all the executions, we applied the minimum threshold voting heuristic. Also, we used the best working prompts for each model (*First* for Aloe and *Invr* for Qwen) in all executions.

| $N_A$ | $N_Q$ | MV | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| 2 | 3 | 0.50 | **67.68** | 63.03 | 65.22 |
| | | 0.75 | 56.09 | 76.08 | 64.52 |
| 3 | 2 | 0.50 | 62.46 | 61.30 | 61.82 |
| | | 0.75 | 51.74 | **78.43** | 62.30 |
| 4 | 5 | 0.50 | 65.94 | 65.09 | **65.50** |
| | | 0.75 | 53.19 | 74.28 | 61.77 |

Table 5: Cross-consistency experiments on the development set. The first and second columns indicate, respectively, the number of Aloe and Qwen executions in each committee of experts system. Results represent the mean micro scores across 5 repeated runs of each system.

Table 5 presents the most relevant systems from our grid search. The results suggest that the improvements achieved through cross-consistency are bottlenecked by the performance of the strongest individual model—in this case, Qwen. Committees comprising a majority of Aloe executions showed improved performance compared to Aloe-only systems using self-consistency, with scores approaching those of Qwen-only systems that employ self-consistency (see Table 4). In contrast, committees with a majority of Qwen executions were unable to substantially surpass the performance of Qwen-only models using self-consistency. These findings indicate that while self-consistency was worthy, cross-consistency did not result in synergistic.

### 6.5. Results on Test Partition

The settings for the three runs submitted to the Subtask 4 are as follows:

- **Run 1**. The Aloe model using the *First* prompt described in Subsection 6.2, without using self-consistency techniques. The results in the development set are shown in Table 3. This is our baseline run.
- **Run 2**. The Qwen model using the *Inverse* prompt described in Subsection 6.2, enhanced with the self-consistency approach employing the most successful hyperparameters in the development set ($N = 3$ and $MV = 0.75$, see Table 4), as described in Subsection 6.3.
- **Run 3**. The best cross-consistency approach described in Subsection 6.4. That is, the committee of experts composed of 4 Aloe executions with the *First* prompt and 5 Qwen executions with the *Invr* prompt, using $MV = 0.50$.

| Run | Prec. | Rec. | F1 |
|---|---|---|---|
| Run 1 | 58.19 | 56.80 | 57.49 |
| Run 2 | **74.72** | **60.33** | **66.76** |
| Run 3 | 74.44 | 59.43 | 66.09 |

Table 6: Results on the test partition. Micro precision, recall and F1.

Table 6 shows that the trends observed in the development partition are also reflected in the test partition. Our best-performing approach is the second one, which uses a Qwen model enhanced with self-consistency and the hyperparameters tuned on the development set. It is worth noting that this approach outperforms our first submission by more than 9 points in Micro F1, illustrating the importance of selecting the most effective model, carefully tailoring prompts to it, and repeating and combining executions to produce more robust outputs.

## 7. Conclusions

We conclude that the performance of small backbone LLMs can be considerably improved for clinical evidence grounding through simple techniques, such as prompt adjustment or the use of self-consistency. Our best approach for Subtask 4 of the Shared Task involved using a Qwen model with a carefully curated prompt and merging the outputs from three executions to increase the robustness of the model. This approach improved the results by more than 9 Micro F1 points compared to our initial submission. In addition, we illustrated that the performance of small LLMs in this task is often highly variable and tends to promote error accumulation in approaches that combine outputs from different models. Consequently, post-processing becomes a core component of the overall pipeline.

## 8. Limitations

Regarding the limitations of our approaches, we observed that the results of some of the LLMs we used were considerably unstable. In the zero-shot setting, the standard deviation of the micro F1 scores reached 3.27 points for the Qwen models, primarily due to high variability in micro recall across executions. In contrast, the Aloe models exhibited greater stability in the same setting, with a standard deviation of 1.39 micro F1 points. The application of self-consistency techniques proved helpful in reducing this variability. Nonetheless, this instability remains an issue that should be addressed in future work. Furthermore, the effectiveness of fine-tuning techniques could lead to greater improvements for

the small LLMs we employed, particularly if synthetic data is generated in an efficient and task-appropriate manner.

## 9. Ethics Statement

The data used in this paper were processed in compliance with the Shared Task organizers' requirements, following the completion of the required training for accessing MIMIC-III and MIMIC-IV, the sources from which the dataset is curated. All experiments were performed on local GPUs, ensuring that no data were transmitted to external destinations and that the integrity of the data was maintained.

## 10. Acknowledgments

Xabier Irastortza-Urbieta is supported by a doctoral grant from the Basque Government (PRE_2025_1_0026). This work has been partially supported by the HiTZ Center and the following MCIN/AEI/10.13039/501100011033 projects: i) HumanAIze (AIA2025-163322-C61) and, (ii) EDHIA (PID2022-136522OB-C22).

## 11. Bibliographical References

Abdul Mohaimen Al Radi, Xu Cao, Fanyang Yu, Yuyuan Liu, Fengbei Liu, Chong Wang, Yuanhong Chen, Jintai Chen, Hu Wang, Yanda Meng, et al. 2025. Agentic large-language-model systems in medicine: A systematic review and taxonomy. *Authorea Preprints*.

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.

Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.

Adrian Cuadron Cortes, Aimar Sagasti, Maitane Urruela, Iker De La Iglesia, Ane García Domingo-aldama, Aitziber Atutxa Salazar, Josu Goikoetxea, and Ander Barrena. 2025. ArgHiTZ at ArchEHR-QA 2025: A two-step divide and conquer approach to patient question answering for top factuality. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 1–10, Vienna, Austria. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. 2025. The aloe family recipe for open and specialized healthcare llms.

Hyeon Hwang, Hyeongsoon Hwang, Jongmyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang, and Jaewoo Kang. 2025. DMIS lab at ArchEHR-QA 2025: Evidence-grounded answer generation for EHR-based QA via a multi-agent framework. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 118–125, Vienna, Austria. Association for Computational Linguistics.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

Tuan Dung Le, Thanh Duong, Shohreh Haddadan, Behzad Jazayeri, Brandon Manley, and Thanh Thieu. 2025. LAILab at ArchEHR-QA 2025: Test-time scaling for evidence selection in grounded question answering from electronic health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 75–80, Vienna, Austria. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Sarvesh Soni and Dina Demner-Fushman. 2026a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *Scientific Data*.

Sarvesh Soni and Dina Demner-Fushman. 2026b. Overview of the ArchEHR-QA 2026 Shared Task on Grounded Question Answering from Electronic Health Records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

Sarvesh Soni, Soumya Gayen, and Dina Demner-Fushman. 2025. Overview of the ArchEHR-QA 2025 shared task on grounded question answering from electronic health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 396–405, Viena, Austria. Association for Computational Linguistics.

Charangan Vasantharajan, Kyaw Zin Tun, Ho Thi-Nga, Sparsh Jain, Tong Rong, and Chng Eng Siong. 2022. Medbert: A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1482–1488.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and Chenxu Lv et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yifan Yang, Qiao Jin, Qingqing Zhu, Zhizheng Wang, Francisco Erramuspe Álvarez, Nicholas Wan, Benjamin Hou, and Zhiyong Lu. 2025b. Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare. *Annual Review of Biomedical Data Science*, 8(Volume 8, 2025):305–316.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. LLM-based medical assistant personalization with short- and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398, Mexico City, Mexico. Association for Computational Linguistics.