Hermes: A Modular Multi-Agent System for Structuring Clinical Text

Aarat Satsangi[1,2*], **Joud El-Shawa**[1*], **Uday Devulapalli**[1,2*], **Apurva Narayan**[1]

[1]Western University, London, ON, Canada
[2]International Centre for Applied Systems Science for Sustainable Development, Cambridge, ON, Canada
{asatsang, jelshawa, udevulap, apurva.narayan}@uwo.ca

Abstract

In today's age of information, unstructured information can become overwhelming and difficult to interpret, particularly in safety critical domains such as healthcare where the volume and complexity of unstructured textual notes is required to be interpretable, insightful, and easily automated for processing. This paper introduces Hermes, a modular agentic system that transforms unstructured clinical text into a modified version of the Subjective-Objective-Assessment-Plan (SOAP) format and generates a knowledge graph offering a high-level, distilled view that facilitates downstream clinical reasoning and decision-making. Hermes employs a multi-agent architecture consisting of four specialized components: Hermes-R (report generation), Hermes-G (knowledge graph generation), Hermes-Q (question-answer pair generation), and Hermes-A (answer generation). These agents operate sequentially with validation to generate structured medical information using iterative refinement. Preliminary evaluations on a few samples demonstrate that Hermes is able to generate structured clinical reports and knowledge graphs according to provided specifications from unstructured discharge summaries with good consistency, accuracy, and reward score. Hermes offers a unified framework that advances clinical natural language processing, bridging structured representation, question answering, and semantic validation.

Code — https://github.com/joudelshawa/hermes
Dataset — https://physionet.org/content/mimic-iv-note/2.2

Introduction

The healthcare industry faces growing challenges in managing the vast volume of clinical documentation generated daily through Electronic Health Records (EHRs), much of which is unstructured or semi-structured. Clinicians now spend over 43% of their time on EHR-related tasks (Sinsky et al. 2016; Raghupathi and Raghupathi 2014), leading to reduced productivity and increased risk of errors. Unstructured notes make it difficult to extract specific information, and the lack of standardized documentation formats across systems hampers interoperability and data sharing. The SOAP (Subjective-Objective-Assessment-Plan) format offers a structured approach (Podder, Lew, and Ghassemzadeh 2023), but converting free-text notes into this format remains labor-intensive and requires preserving the clinical accuracy and context of the original content. Recent advances in natural language processing (NLP) and machine learning have shown promise in addressing these challenges. However, existing solutions often focus on specific aspects of clinical documentation, such as information extraction or question answering, without providing a comprehensive framework that combines structured representation, semantic validation, and knowledge graph generation (Leong et al. 2024). This fragmented approach limits the potential impact of these technologies in improving clinical documentation and decision-making. To address these limitations, we present Hermes, a modular agentic system that transforms unstructured clinical text into a modified SOAP format while generating a comprehensive knowledge graph. The system employs a multi-agent architecture consisting of four specialized components that iteratively process, validate, and structure medical information. This approach not only improves the efficiency of clinical documentation but also enhances the quality and accessibility of patient information for downstream clinical reasoning and decision-making. The main contributions of this paper are:

1. A modular framework composed of four specialized Large Language Model (LLM)-based agents (Hermes-R, Hermes-G, Hermes-Q, Hermes-A) that collaboratively transform unstructured clinical notes into structured SOAP reports, knowledge graphs, and clinically relevant Question-Answer (QA) pairs.
2. A novel feedback-driven refinement loop where QA pairs generated from the knowledge graph are semantically matched with answers derived from the original notes.
3. A knowledge graph generation component that produces color-coded graphs with directed relationships and descriptive metadata, offering interpretable and actionable representations for downstream clinical reasoning and decision-making.

Hermes is a unified clinical documentation system that integrates information extraction, knowledge representation, and validation into a single framework. It processes unstructured clinical notes into interpretable and meaningful structured outputs, supporting healthcare providers, researchers, and administrators. This work presents a proof-of-concept

*These authors contributed equally.

system and does not yet include large-scale evaluation or clinical expert validation.

The *Literature Review* situates Hermes within prior work on clinical text understanding, structured report generation, knowledge graph construction, and multi-agent systems. *Methodology* details the system architecture and key components, followed by *Evaluation* describing the experimental setup. *Results* presents our findings, and *Conclusion* discusses implications and future directions.

Literature Review

LLMs in Clinical Text Understanding

LLMs have demonstrated strong capabilities in processing unstructured clinical text, enabling a wide range of applications such as named entity recognition, clinical concept extraction, summarization, and question answering. Early work in domain adaptation of language models led to the development of BioBERT, which adapted BERT to the biomedical domain using PubMed abstracts and PMC full-text articles, improving performance on biomedical Named Entity Recognition and QA tasks (Lee et al. 2020). ClinicalBERT built on this foundation by pretraining on MIMIC-III clinical notes (Johnson et al. 2016), achieving improvements in clinical concept recognition and classification (Alsentzer et al. 2019). More recently, general-purpose LLMs such as GPT-3 and GPT-4 have shown promise in zero- and few-shot medical reasoning tasks, particularly for QA and report summarization (Brown et al. 2020). Specialized medical LLMs like Med-PaLM and MedAlpaca have achieved state-of-the-art performance on benchmark medical QA datasets through instruction tuning and human feedback (Singhal et al. 2023; Han et al. 2023). These advancements set the stage for using LLMs not only as powerful interpreters of clinical text, but also as core components in multi-agent systems for structured data generation and validation.

Structured Report Generation from Clinical Notes

Following early rule-based systems by Gunter et al. (2022) and Mykowiecka, Marciniak, and Kupć (2009), machine learning models such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) were introduced to enable data-driven extraction of medications, diagnoses, and other clinical concepts (Meystre et al. 2007; Patrick and Li 2010) so that these models can achieve better generalization across institutions and note types. Hybrid approaches have since emerged, combining the interpretability of rules with the adaptability of learning-based models. For example, Sohn et al. (2014) integrated CRFs with rule-based components for medication extraction, and Uzuner et al. (2012) proposed a hybrid system for parsing discharge summaries, both improving performance and robustness. More recently, LLMs have demonstrated considerable promise in automating structured clinical report generation (Grothey et al. 2025). Fine-tuned LLaMA-2 models have achieved high F1 scores in synoptic reporting of cancer pathology, aligning closely with expert annotations (Rajaganapathy et al. 2025), while ChatGPT has shown competitive accuracy in extracting structured data from clinical notes (Huang et al. 2024). Veen et al. (2024) further reported that LLMs can match, and in some cases exceed, the performance of medical experts in clinical text summarization. These findings indicate that LLMs have the potential to significantly improve the efficiency and accuracy of structured report generation in clinical settings.

Knowledge Graph Construction from Clinical Text

Constructing knowledge graphs (KGs) from clinical text has become increasingly important for enhancing clinical decision support, disease understanding, and healthcare analytics. Recent work has explored the integration of LLMs into KG construction pipelines. For example, Xu et al. (2024) proposed a heart failure–specific KG built through prompt-engineered LLM extraction and combined with expert refinement to ensure both accuracy and efficiency. Lyu et al. (2023) developed a causal KG for diabetic nephropathy by extracting triples from multiple knowledge sources and applying pruning strategies to reduce noise, improving diagnostic decision support. Similarly, Arsenyan et al. (2024) compared multiple LLM architectures for biomedical KG construction from electronic medical records, demonstrating that LLMs can effectively capture complex medical entities and relationships.

Agentic AI Frameworks and Multi-Agent Systems

Agentic Artificial Intelligence (AI) frameworks and multi-agent systems (MAS) are increasingly used to coordinate complex, collaborative tasks across various domains. Incorporating LLMs into MAS has expanded agents' abilities to perceive, reason, and work together effectively. Tran et al. (2025) survey LLM-based MAS, introducing a framework that categorizes collaboration mechanisms by actors, interaction types, structures, strategies, and coordination protocols. Ramachandran (2024) explores the architectures and applications of agentic AI and multimodal frameworks, showing how technologies such as Reinforcement Learning, Neuro-Symbolic AI, and Graph Neural Networks can address challenges in healthcare, finance, and disaster management. In software engineering, He, Treude, and Lo (2025) discuss how LLM-based MAS enable autonomous problem-solving and scalability across the software development lifecycle. Focusing on medicine, Wang et al. (2025) review LLM-based agents, examining their architectures, applications, and challenges, and analyzing key components such as system profiles, clinical planning mechanisms, and medical reasoning frameworks. Finally, Chen et al. (2025) demonstrate how multi-agent conversational LLMs can improve clinical decision support. Collectively, these works show how agentic AI and MAS are evolving into powerful tools for building intelligent, cooperative systems that can address real-world challenges.

Methodology

Overview of Hermes

The modular system, shown in Figure 1, comprises four specialized agents with distinct yet complementary roles that work together to process unstructured clinical text

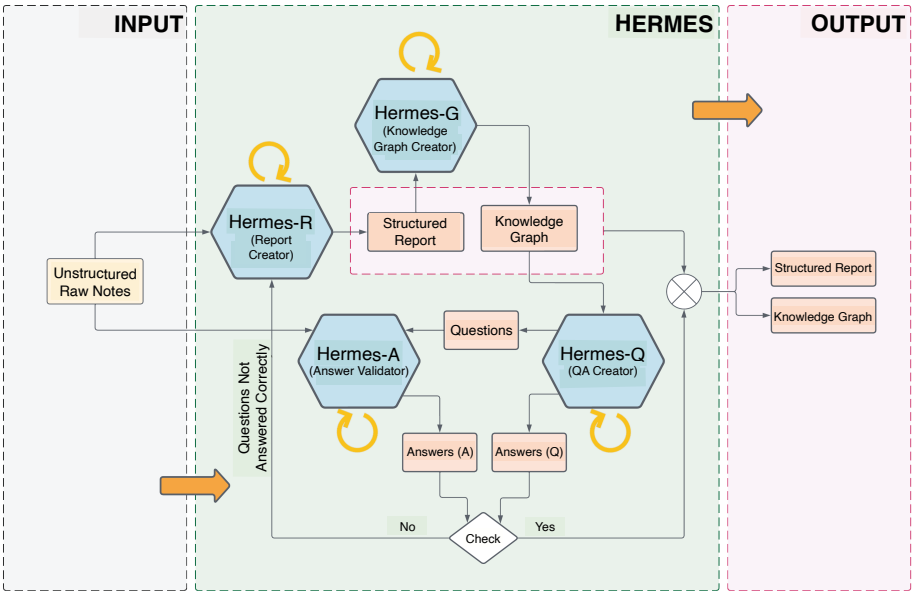

Figure 1: Hermes pipeline

through iterative refinement: Hermes-R, the report creator; Hermes-G, the knowledge graph constructor; Hermes-Q, the question-answer pair generator; and Hermes-A, the answer finder.

Generation. First, the unstructured notes are passed to Hermes-R, which generates a structured report in a modified SOAP format[1] by extracting and organizing all relevant medical details into clearly delineated Subjective, Objective, Assessment, and Plan sections. This report is then handed off to Hermes-G, which focuses on identifying medical entities and their relationships, color-coding each node, and linking them with directed edges to build a knowledge graph. This graph then flows into Hermes-Q, where clinically pertinent question-answer pairs are formulated that cover the full breadth of the graph's content. The questions are sent to Hermes-A, which tries to answer each question using the original unstructured notes.

Validation. For validation, each answer from Hermes-Q and Hermes-A is compared using semantic similarity matching. If any answer falls below the similarity threshold, feedback in the form of incorrectly answered questions is routed back to Hermes-R to refine the structured report and reprocess the pipeline until consistency is achieved or a predefined iteration limit is reached. This feedback loop not only improves accuracy but also helps contain and prevent the propagation of potential hallucinations across agents. Upon successful validation, Hermes produces the final structured report and the color-coded knowledge graph. The iterative nature of the process enables automated correction of errors in both structured reports and knowledge graphs, ensuring higher reliability and trustworthiness of the outputs.

[1]Format is available in system-prompt and an example is available in user-prompt of Hermes-R on GitHub.

Agents

Hermes-R. Hermes-R transforms the unstructured clinical notes into a well-organized report according to the modified SOAP format in markdown. If the format is violated, an error describing the issue is propagated back along with the initial prompts so that it can try to avoid making the same mistake again. The generated report is also verified to ensure all numerical values in the report match those in the source clinical notes.

Hermes-G. Hermes-G generates a knowledge graph using the structured report, providing a high-level, distilled view that can help in downstream clinical reasoning and decision-making. The generated graph represents various medical entities as nodes and the relationships between them as directed edges. It also stores both relationship descriptions and associated factual information as separate fields. The medical entities are color-coded for intuitive visualization (see Figure 2). Hermes-G is constrained to output a valid, non-empty knowledge graph.

Hermes-Q. Hermes-Q formulates clinically relevant question-answer pairs to validate the generated knowledge graph. It must generate at least as many questions as there are nodes in the graph. If it fails, an error describing this shortfall is propagated back along with the response and initial prompts so that it can add more questions.

Hermes-A. Hermes-A answers the questions posed by Hermes-Q using the unstructured clinical notes, providing each answer in a single sentence. Once the answers are generated, they are compared to Hermes-Q's generated answers

for semantic similarity, as described in the *Validation* section.

Modified SOAP Notes

In this work, we build upon the traditional SOAP framework (Podder, Lew, and Ghassemzadeh 2023), long used to structure clinical encounters, by embedding two novel modules that capture both longitudinal patient context and discharge-specific information. First, the Comprehensive Patient Profile (CPP) aggregates key administrative and identification details (e.g., patient name, date of birth, gender, medical record number, and emergency contacts) and records the timestamp of the last CPP update. Second, the Discharge Conditions (DC) section compiles the patient's physical exam findings at discharge, overall discharge condition, tailored discharge instructions, final disposition, and a concise hospital-course summary. Together, these enhancements ensure that every note not only documents the clinical encounter itself but also provides a readily accessible snapshot of the patient's identity, care trajectory, and post-hospitalization plan.

Evaluation

Evaluation was limited to 11 unstructured discharge summaries from the MIMIC-IV dataset (Johnson et al. 2023), chosen as an initial test case to demonstrate system feasibility. Each clinical note was reformulated into a prompt and processed by the framework to generate both structured clinical reports and corresponding knowledge graphs. We focused on a structural coverage metric to verify formatting correctness, leaving clinical correctness, factual accuracy, and expert validation for future work. This design choice enabled us to test pipeline feasibility while deferring full-scale validation to later studies. The reward function quantifies the extent to which the expected clinical headings, across four hierarchical levels of granularity, are accurately represented in the structured output. This metric provides a systematic but narrow evaluation of response accuracy. For each sample, we computed the reward score and reported the mean and standard deviation across all examples.

Results

Using the DeepSeek-R1 (Guo et al. 2025) 70B model, Hermes successfully generated structured reports from the 11 unstructured discharge summaries sampled from the MIMIC-IV dataset (Johnson et al. 2023). The average reward across all generated reports was 1.0 with a standard deviation of 0, indicating that Hermes consistently produced outputs that fully adhered to the desired structure in this initial proof-of-concept evaluation. A light qualitative inspection of the generated reports further confirms the framework's effectiveness in extracting and organizing clinically relevant information. The structured outputs not only met the required format but also preserved the semantic integrity of the original notes, presenting the data in a more accessible and interpretable form. For example, while using DeepSeek-R1 32B as the backbone model for the first iteration of the Hermes-R agent, for a chosen representative example, the output omitted 12 required section headings. However, in the second iteration, the agent successfully incorporated all necessary headings, demonstrating the framework's ability to iteratively refine and correct its outputs. The final structured report is accessible on GitHub (Data/1 folder), and its knowledge graph is presented in Figure 2.

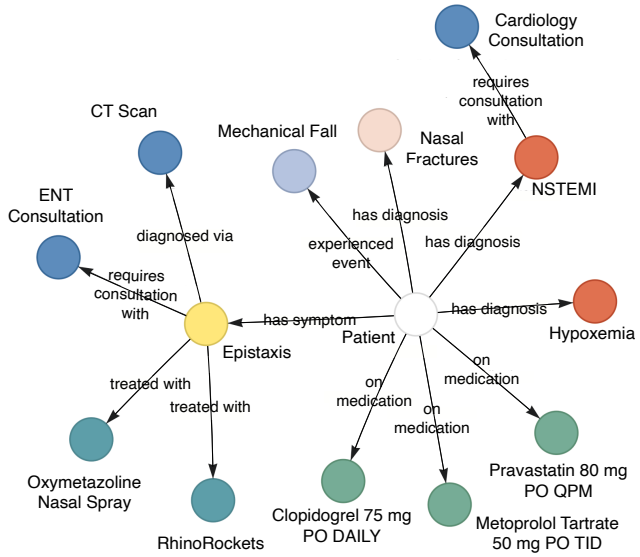

Figure 2: Knowledge graph generated by Hermes-G. Medical entities are color-coded for intuitive visualization: patient = white; confirmed diagnoses = red; possible diagnoses = light red; lab results = orange; symptoms = yellow; on medications = green; suggested tests = blue

These results should be interpreted cautiously given the small evaluation set, the lack of quantitative comparisons to existing clinical NLP and knowledge graph methods, and the focus on structural rather than clinical correctness metrics. Future work will include systematic benchmarking against state-of-the-art pipelines, evaluation across diverse institutions and note types, and expert clinical review. Preliminary findings suggest Hermes could be a reliable and effective tool for transforming unstructured clinical narratives into structured documentation, with strong potential for integration into clinical workflows and decision support systems.

Conclusion

In this paper, we introduced Hermes, a novel framework for transforming unstructured clinical notes into structured documentation by combining information extraction, knowledge representation, and validation within a single system. Built on top of large language models, Hermes generates structured clinical reports that preserve the semantic integrity and contextual relevance of the original notes. Our evaluation on MIMIC-IV discharge summaries showed that Hermes consistently produces outputs in the correct format and iteratively improves report completeness through reinforcement. While the current results are promising, several directions remain for future work.

Future Work

A key next step is rigorous evaluation, including (1) expanding datasets to include multiple clinical note types and institutions, (2) incorporating qualitative expert review to validate factual and clinical accuracy, and (3) benchmarking Hermes against existing clinical NLP and knowledge graph construction systems. These steps will clarify the framework's real-world applicability and guide further refinement. Additionally, we plan to integrate additional large language models, such as LLaMA (Dubey et al. 2024) and Qwen (Team 2024), into the Hermes framework and conduct comparative evaluations to assess their performance. We also aim to fine-tune smaller LLMs for specific subtasks such as structured report generation and knowledge graph construction, improving the modularity and computational efficiency of the overall system. Incorporating formal verification techniques could further validate the correctness and consistency of both the structured outputs and the generated knowledge graphs, thereby increasing the reliability of Hermes in clinical applications. Finally, extending the framework to accommodate multimodal data sources, such as laboratory results and medical imaging, offers a valuable opportunity to enrich the semantic depth of the outputs and improve downstream decision support.

References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Arsenyan, V.; Bughdaryan, S.; Shaya, F.; Small, K. W.; and Shahnazaryan, D. 2024. Large Language Models for Biomedical Knowledge Graph Construction: Information extraction from EMR notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 295–317. Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chen, X.; Yi, H.; You, M.; et al. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8: 159.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Grothey, B.; Odenkirchen, J.; Brkic, A.; Schömig-Markiefka, B.; Quaas, A.; Büttner, R.; and Tolkach, Y. 2025. Comprehensive testing of large language models for extraction of structured data in pathology. *Communications Medicine*, 5(96). Published: 31 March 2025.

Gunter, D.; Puac Polanco, P.; Miguel, O.; Thornhill, R.; Yu, A.; Liu, Z.; Mamdani, M.; Pou-Prom, C.; and Aviv, R. 2022. Rule-based natural language processing for automation of stroke data extraction: a validation study. *Neuroradiology*, 64.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han, X.; Ma, Y.; Liu, Y.; Fan, Y.; Liu, H.; and Wang, Y. 2023. MedAlpaca: A Biomedical Conversational Model Based on LLaMA and Alpaca. *arXiv preprint arXiv:2304.09817*.

He, J.; Treude, C.; and Lo, D. 2025. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision and the Road Ahead. *ACM Transactions on Software Engineering and Methodology*.

Huang, J.; Yang, D. M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L. J.; Xiao, G.; Peterson, E. D.; Zhan, X.; and Xie, Y. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, 7: 106.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Leong, H. Y.; Gao, Y. F.; Shuai, J.; Zhang, Y.; and Pamuksuz, U. 2024. Efficient fine-tuning of large language models for automated medical documentation. *arXiv preprint arXiv:2409.09324*.

Lyu, K.; Tian, Y.; Shang, Y.; Zhou, T.; Yang, Z.; Liu, Q.; Yao, X.; Zhang, P.; Chen, J.; and Li, J. 2023. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. *Journal of Biomedical Informatics*, 139: 104298.

Meystre, S.; Savova, G.; Kipper-Schuler, K.; and Hurdle, J. 2007. Extracting Information From Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearb Med Inform*, 128–144.

Mykowiecka, A.; Marciniak, M.; and Kupć, A. 2009. Rule-based information extraction from patients' clinical data. *J. of Biomedical Informatics*, 42(5): 923–936.

Patrick, J.; and Li, M. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17: 524–7.

Podder, V.; Lew, V.; and Ghassemzadeh, S. 2023. SOAP Notes. In *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing. [Updated 2023 Aug 28].

Raghupathi, W.; and Raghupathi, V. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1): 3.

Rajaganapathy, S.; Chowdhury, S.; Li, X.; Buchner, V.; He, Z.; Zhang, R.; Jiang, X.; Yang, P.; Cerhan, J. R.; and Zong, N. 2025. Synoptic reporting by summarizing cancer pathology reports using large language models. *npj Health Systems*, 2(11).

Ramachandran, A. 2024. A Survey of Agentic AI, Multi-Agent Systems, and Multimodal Frameworks: Architectures, Applications, and Future Directions. *ResearchGate*. Available at https://www.researchgate.net/publication/387577302.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Babiker, A.; Schärli, N.; Chowdhery, A.; Mansfield, P.; Demner-Fushman, D.; Agüera y Arcas, B.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Sinsky, C.; Colligan, L.; Li, L.; Prgomet, M.; Reynolds, S.; Goeders, L.; Westbrook, J.; Tutty, M.; and Blike, G. 2016. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165(11): 753–760. PMID: 27595430.

Sohn, S.; Clark, C.; Halgrim, S.; Murphy, S.; and Liu, H. 2014. MedXN: an Open Source Medication Extraction and Normalization Tool for Clinical Text. *Journal of the American Medical Informatics Association : JAMIA*, 21.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tran, K.-T.; Dao, D.; Nguyen, M.-D.; Pham, Q.-V.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Uzuner, O.; Bodnari, A.; Shen, S.; Forbush, T.; Pestian, J.; and South, B. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, 19: 786–91.

Veen, D. V.; Uden, C. V.; Blankemeier, L.; Delbrouck, J.-B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E. P.; Seehofnerová, A.; Rohatgi, N.; Hosamani, P.; Collins, W.; Ahuja, N.; Langlotz, C. P.; Hom, J.; Gatidis, S.; Pauly, J.; and Chaudhari, A. S. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30: 1134–1142. Published: 27 February 2024.

Wang, W.; Ma, Z.; Wang, Z.; Wu, C.; Ji, J.; Chen, W.; Li, X.; and Yuan, Y. 2025. A Survey of LLM-based Agents in Medicine: How far are we from Baymax? *arXiv preprint arXiv:2502.11211*.

Xu, T.; Gu, Y.; Xue, M.; Gu, R.; Li, B.; and Gu, X. 2024. Knowledge graph construction for heart failure using large language models with prompt engineering. *Frontiers in Computational Neuroscience*, 18.