Towards Conversational AI for Spina Bifida Care

Asfandyar Azhar[1,2,6], **Shaurjya Mandal**[3,4,6], **Nidhish Shah**[5,6]

[1]Carnegie Mellon University, [2]Stanford University, [3]Massachusetts General Hospital
[4]Harvard University, [5]Technische Universiteit Eindhoven, [6]Asesa

{asfandyar, shaurjya, nidhish}@asesa.life

Abstract

Spina Bifida (SB) is a complex neural tube defect that presents multifaceted healthcare challenges requiring multidisciplinary management. While advances in foundation models (FMs), offer promising avenues for enhancing SB care through intelligent, context-aware support, existing models struggle to accurately identify and reason about SB's diverse symptoms. This study benchmarks eight widely used large language models (LLMs) through qualitative and quantitative evaluations, focusing on their ability to address the unique medical challenges of SB. We introduce an *inverse prompting* technique designed to guide LLMs through a step-wise diagnostic process by incorporating a predefined symptom set relevant to SB, thereby preventing premature conclusions and improving diagnostic reasoning. Our evaluations reveal significant limitations in the LLMs' abilities to accurately diagnose SB-related conditions, underscoring the need for specialized approaches. Building on these findings, we propose a novel framework that integrates a structured, symptom-based knowledge base specific to SB, enhancing the models' contextual understanding and reasoning capabilities. This work highlights the potential of tailored AI solutions in improving access to care for individuals with SB, particularly in populations where gaps in knowledgeable providers persist. By addressing the shortcomings of general-purpose LLMs, our suggested framework aims to streamline SB care and improve patient outcomes, paving the way for more effective AI-assisted healthcare interventions in complex chronic conditions.

1 Introduction & Related Work

The integration of generative artificial intelligence (GenAI) in advanced medical care is transforming the management of complex conditions like Spina Bifida (SB), a neural tube defect that affects brain and spine development. SB presents diverse challenges, from severe forms like myelomeningocele—often associated with hydrocephalus and Chiari II malformation—to milder cases like Spina Bifida Occulta, which may go undiagnosed until later in life [7]. Symptoms include bowel and urinary incontinence, mobility impairments, and cognitive issues. Advances in medical care have improved the life expectancy of individuals with SB, creating a growing adult SB population [4]. This shift highlights the need for multidisciplinary care that addresses neurological, urological, and psychosocial dimensions. However, existing healthcare systems often struggle to provide comprehensive care, particularly for adults with SB. Emerging technologies, including machine learning and deep learning, are being applied to SB research and clinical practice. In genetics, machine learning aids in identifying critical biomarkers and unraveling SB's multifactorial nature [9, 1], while deep learning models enhance prenatal screening and diagnostic accuracy through advanced image processing [6, 10, 2]. Additionally, AI-driven innovations are improving urological care [23, 26] and facilitating personalized rehabilitation strategies [14].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Large language models (LLMs) like ChatGPT [15] represent a new frontier in SB care, offering potential as tools to assist caregivers and healthcare providers by generating intelligent, context-aware responses. These models can enhance patient education and provide valuable support. However, their application in healthcare must be approached cautiously due to risks of misinterpretation and domain-specific constraints. Strategies like prompt engineering, fine-tuning, and retrieval augmented generation (RAG) are essential for ensuring accuracy, but current approaches like RAG struggle with the incremental, symptom-based reasoning required for conditions like SB [11]. Our proposed framework overcomes this limitation by integrating a structured, SB-specific knowledge base, enabling more accurate, context-aware interactions without relying on document retrieval. Despite AI's promise in SB care, a significant gap in access to knowledgeable providers remains, particularly for the adult SB population [18]. Multimodal models like MedPALM [17] and AMIE [22] offer advancements, but there is no robust AI model tailored specifically to SB. Our study addresses this gap by evaluating the capabilities of existing LLMs in identifying SB symptoms and generating treatment pathways. Specifically, we contribute the following:

1. Benchmarking eight LLMs through both qualitative and quantitative evaluations of their performance in addressing SB's unique medical challenges.
2. Introducing an *inverse prompting* technique, guiding LLMs through a structured diagnostic process using a predefined symptom set, ensuring more accurate and stepwise reasoning.
3. Assessing the effectiveness of *inverse prompting* with SB patients, using diagnostic accuracy (α) and error rate (ϵ) as metrics.
4. Proposing a novel framework based on the identified limitations of existing LLMs, designed to improve clinical outcomes for SB patients.

2 Methods & Experiments

The nature of the exchange between a clinician or patient and the LLM is conversational. So a model is tested on two different components of reasoning: the ability to combine and link the given symptoms, and attempting to ask the follow-up questions to narrow down the space of diagnosis progressively. Moreover, we define the success and failure criteria for the performance of LLMs as clinical FMs:

(1) Step-wise reasoning: The model should be capable of iterating through the requested information step-wise to avoid looping back into its reasoning. This prevents the model from hallucinating or repeatedly requesting similar information and being redundant.

(2) Well-timed conclusivity: Only after a detailed step-wise analysis should the model request more concrete modalities like specific imaging outputs (that may be accessible by the patient or their clinician) instead of jumping to a diagnostic result prematurely while bypassing steps in its way. A failure would be the LLM's inability to recognize signs of a shunt malfunction requiring urgent evaluation. Another significant oversight would be if it fails to identify symptoms of a urinary tract infection or pressure injuries to the skin, and does not prompt the user to seek immediate medical attention. Figure 2 (Appendix 0) shows a comprehensive example of a conversation leading to the diagnosis of complications related to SB. The reasoning capabilities of popular LLMs are insufficient for clinical reasoning which is a subset of diagnostic tasks; and hence, for conditions like SB, we will require a pipeline of operations (see Figure 1) through a conversational agent, instead of a single generalized FM.

Table 1: Single sample comparative analysis of the set of FMs (temperature set to 0.2 for all models).

Model	**Prompt Type**	**Reasoning**	**Added Context** ($P \subseteq S$)	α
Gemini 1.5-Pro [19]	Inverse, Bridging	ToT	See Appendix 1A	Correct (0.5)
Mixtral 8×22B [8]	Inverse, Bridging	CoT	See Appendix 1B	Incorrect (0)
Mistral Large 2	Inverse	CoT, ToT	See Appendix 1C	Correct (1.0)
Claude-V3.5 [3]	Inverse, Bridging	CoT	See Appendix 1D	Incorrect (0)
Llama3.1-405B [21]	Inverse	ToT	See Appendix 1E	Incorrect (0)
GPT-4o Mini [15]	Inverse, Bridging	CoT	See Appendix 1F	Incorrect (0)
GPT-4o [16]	Inverse	CoT	See Appendix 1G	Correct (1.0)
GPT-4 Turbo	Inverse	CoT, ToT	See Appendix 1H	Incorrect (0)

2.1 Formalizing the Inverse Prompt

Inverse prompting refers to a structured system prompt that primes the model with a predefined set of SB-related symptoms. This allows the LLM to start the conversation with context, guiding the patient through symptom reporting without having to "discover" symptoms mid-conversation.

Let SB be represented by a set of symptoms $S = \{s_1, s_2, ..., s_n\}$. Furthermore, consider conditions $C_1, C_2, ..., C_N$ each represented by its own set of unique symptoms. We then construct a composite synthetic condition, F, where we choose a $K \in \mathbb{Z}^+$ and then randomly sample K symptoms from the conditions $S, C_1, C_2, ..., C_N$. We define F_S as the subset of symptoms from S included in F, and F_{C_i} is the subset of symptoms included in C_i. Then, F can be represented as: $F = F_S \cup F_{C_1} \cup F_{C_2} \cup ... \cup F_{C_N}$ where $F_S \subseteq S$, $F_{C_i} \subseteq C_i$ $(1 \leq i \leq N)$, and $|F_S| + \Sigma_{i=1}^{N} |F_{C_i}| = K$. Finally, it is required that F includes all or some of the symptoms from S depending on K. This information is used as the system inverse prompt to "warm start" the FM with clinical context relevant to SB (see Appendix 2).

Once each LLM is posing as a clinical FM based on the inverse prompt, users communicate with them starting with a symptom, $s_i \in S$, followed by another symptom, and so on, until a diagnosis is made. Note that the number of symptoms mentioned by the user varies, and they might not exhibit all $s_i \in S$. To represent this, let P be the set of symptoms presented by the patient during the conversation. Then: $P \subseteq S$, where $|P| \leq n$ (the total number of symptoms in S). This is the *added context* provided to each LLM in Table 1. We surveyed 50 participants through online outreach in SB patient communities like support groups. Each participant engaged in five separate conversations with each of the eight LLMs. Participants were encouraged to interact with the LLMs freely as if they were at a consultation with a physician. The conversations covered various SB-specific issues, including hydrocephalus, tethered cord syndrome, and neurogenic bladder. The conversations covered various SB-specific issues, including hydrocephalus, tethered cord syndrome, and neurogenic bladder.

2.2 Qualitative Experimental Setup & Results

We evaluated the efficacy of eight FMs (see Table 1) in the specific context of diagnosing complications related to SB. We focused on the ability of these models to provide appropriate recommendations, based on reasoning and prompting methods, for secondary conditions experienced by individuals with SB. The cohort of participants were asked to interact with FMs and provide qualitative feedback about the performance of each FM. This includes both common and obscure scenarios to test the models' range of knowledge and adaptability. Information about the participants themselves was not recorded.

Table 1 outlines one conversation of a patient with their custom starting prompt with the various LLMs. Based on the reasoning and response obtained from the LLMs, they adapted their prompting strategy to provide added context. Typically, based on a custom starting prompt, models are prompted to enquire about further symptoms and patient history. Models are evaluated on their ability to recognize and suggest medical attention for symptoms relevant to SB. The way that this is accomplished is by observing reasoning methods, namely chain-of-thought (CoT) [24] and tree-of-thought (ToT) [25] and by deploying bridge[1] [12] and inverse prompting. More concretely, we define *diagnostic accuracy* as a weighted metric that accounts for both fully correct and partially correct impressions by the LLM. A correct impression, where the model provides accurate conclusions without user intervention, is scored as 1. A partially correct impression, where the model requires user-guided corrections (e.g., bridging to refocus/relying on P after errors or hallucinations), is scored as 0.5. Otherwise, the score is 0. Thus, diagnostic accuracy is represented as: $\alpha = \frac{\sum(\psi + 0.5\phi)}{|C|}$, where ψ are the correct impressions, ϕ are the partial impressions, and $|C|$ is the total number of conversations.

GPT-4 Turbo. It relied heavily on the inverse prompt, often recommending further tests or medical scans rather than making direct diagnoses, though it followed a systematic approach and rarely ventured beyond the inverse prompt while questioning the participants.

GPT-4o. Demonstrated strong sequential reasoning and required minimal bridging, excelling at formulating diagnosis as an inclusion-exclusion task. Conversations were short-to-medium in length.

[1]We adapted bridging to SB care by linking symptoms that are not immediately related but may have clinical relevance in SB patients.

GPT-4o Mini. Struggled to retain context even after additional bridging, often focusing on providing remedies based on recent prompts rather than integrating past information.

Gemini 1.5-Pro. Performed satisfactorily but hard iterated through symptoms, reasoning like a checklist. This resulted in longer conversations with heavy bridging.

Claude-V3.5. Used inclusion-exclusion reasoning, similar to GPT-4o, which resulted in good progressive reasoning. However, in some cases it ended conversations prematurely due to over-reliance on eliminations.

Llama3.1-405B. Hesitated to diagnose, looping questions, and favored synthetic conditions over SB when narrowing down possibilities to those two.

Mixtral 8×22B. Not exhaustive enough when querying the participants for information, asked tangential questions, often leading to insufficient information gathering and misdiagnosis.

Mistral Large 2. Frequently jumped to conclusions without posing necessary questions, disrupting logical flow and causing diagnostic errors despite bridging attempts.

2.3 Quantitative Experimental Setup & Results

We evaluated the eight LLMs across 50 different diagnostic scenarios derived from real-world SB case studies. Each model was tested on its ability to correctly identify symptoms of SB and recommend diagnostic steps based on conversational prompts. We measured α, as previously defined, and the *diagnostic error rate*. We define diagnostic error rate as $\epsilon = \frac{|E|}{|C|}$, for $|E|$ is the total number of errors. An error is considered as a correct impression made for the wrong reasons (e.g., mistaking one symptom for another). Furthermore, we conduct a more fine-grained evaluation at the symptom-level.

Table 2: Coarse-level SB diagnostic performance of LLMs. Note: (α_O, ϵ_O) and (α_X, ϵ_X) are the diagnostic accuracies and error rates when no system prompt (baseline) and a standard system prompt are used respectively. The inverse prompt results, (α, ϵ), show best performance on all LLMs.

Model	α_O	ϵ_O	α_X	ϵ_X	α	ϵ
GPT-4o	0.752	0.311	$0.803_{\uparrow 0.05}$	$0.304_{\downarrow 0.01}$	$0.886_{\uparrow 0.13}$	$0.162_{\downarrow 0.15}$
GPT-4 Turbo	0.738	0.336	$0.789_{\uparrow 0.05}$	$0.328_{\downarrow 0.01}$	$0.845_{\uparrow 0.11}$	$0.170_{\downarrow 0.17}$
Claude-V3.5	0.744	0.289	$0.792_{\uparrow 0.05}$	$0.277_{\downarrow 0.01}$	$0.853_{\uparrow 0.11}$	$0.099_{\downarrow 0.19}$
Gemini 1.5-Pro	0.720	0.401	$0.753_{\uparrow 0.03}$	$0.396_{\downarrow 0.01}$	$0.812_{\uparrow 0.09}$	$0.235_{\downarrow 0.17}$
Mistral Large 2	0.696	0.357	$0.722_{\uparrow 0.03}$	$0.350_{\downarrow 0.01}$	$0.782_{\uparrow 0.09}$	$0.275_{\downarrow 0.08}$
Mixtral 8×22B	0.722	0.383	$0.758_{\uparrow 0.04}$	$0.379_{\downarrow 0.00}$	$0.828_{\uparrow 0.11}$	$0.304_{\downarrow 0.18}$
GPT-4o Mini	0.707	0.323	$0.785_{\uparrow 0.08}$	$0.315_{\downarrow 0.01}$	$0.867_{\uparrow 0.16}$	$0.164_{\downarrow 0.16}$
Llama3.1-405B	0.655	0.420	$0.692_{\uparrow 0.04}$	$0.411_{\downarrow 0.01}$	$0.758_{\uparrow 0.10}$	$0.236_{\downarrow 0.18}$
Mean Scores	0.717	0.353	$0.762_{\uparrow 0.05}$	$0.345_{\downarrow 0.01}$	$0.829_{\uparrow 0.11}$	$0.193_{\downarrow 0.16}$

In Table 2, we observe that clinical reasoning improves as the error rate significantly decreases, while the change in accuracy over the baseline for the inverse prompt is more than double that of the standard system prompt[2]. In fact, the standard system prompt has an almost negligible effect on the error rate, whereas the inverse prompt performs 16× better. This is crucial to highlight, as it suggests that, with the right inverse prompt, LLMs can more effectively manage complex symptom clusters associated with SB. More interestingly, although Claude-V3.5 shows both lower diagnostic accuracy and a lower error rate compared to GPT-4o, this indicates that while GPT-4o can systematically approach diagnostic tasks, it may occasionally make more errors in connecting patient symptoms over the course of a conversation than Claude-V3.5. This implies that GPT-4o may require more contextual bridging across conversations compared to Claude-V3.5.

We then applied a similar approach to more challenging tasks at the fine-grained level, focusing on symptoms associated with SB (refer to Table 3 in Appendix 2). In general, all models perform slightly worse at identifying very fine-grained ailments, such as tethered cord (s_3) and Chiari malformation (s_5). However, they show decent performance in detecting common symptoms requiring immediate attention, such as pressure ulcers (s_6) and urinary tract infections (s_7). On the other hand, rarer but critical symptoms, such as CSF leak (s_1), which can be far more catastrophic, are often missed. This highlights the need for developing specialized and adaptable frameworks (see Figure 1), as inverse

[2]There is a significant negative correlation between α and ϵ ($\rho = -0.762$, p-val $= 0.028$).

prompting techniques may encounter a natural bottleneck (which warrants further investigation), to ensure accurate and reliable diagnostic outcomes across all SB-related conditions.

3 Proposed Solution for Larger Study & Conclusion

In the experiments, we identified a few shortcomings with the models in adapting to the diagnostics tasks. A major shortcoming was not having a corresponding database with the specified conditions. A condition as complex as SB has a combination of urological, neurological, and musculoskeletal implications. Thus we propose a multistage architecture suitable for diagnostic tasks. Figure 1 lays out the architecture of our proposed approach that integrates into the mainstream conversation between the patient and the models.

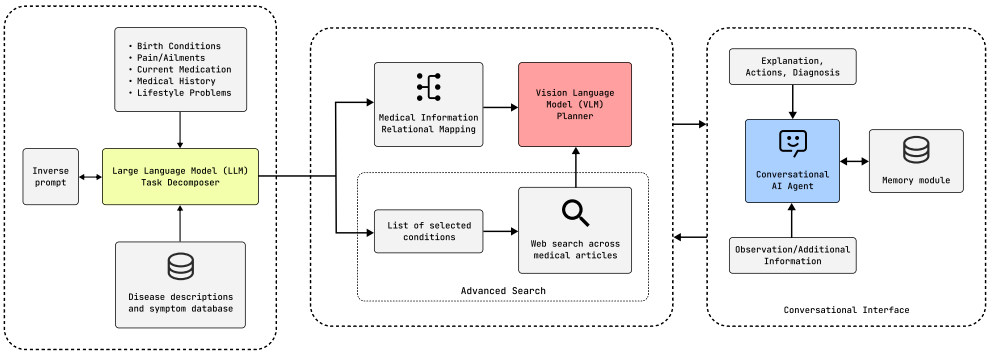

Figure 1: Architectural overview of the proposed system

The first module of our architecture focuses on creating a directed corpus from patient-provided information. After receiving the initial prompt, the model conducts targeted inquiries, narrowing down the diagnostic search space by matching responses to predefined descriptions. This process relies on focused information retrieval rather than reasoning. Once the primary data is gathered, the model breaks it into tasks requiring deeper reasoning and input.In the second module, a vision-language planner is introduced, enabling the model to handle diverse input modalities, including web-based resources, beyond just text. This planner connects the collected patient data with possible conditions, refining the diagnostic process. To handle longer conversations, a memory unit stores and links patient responses, adapting the reasoning-based queries throughout the interaction. This module forms the conversational AI backbone that powers the patient interface.

Although AI offers significant benefits, a gap remains in helping individuals with SB recognize complications and seek timely medical advice. Future work will focus on integrating clinical conversations, health records, and literature to bridge this gap. Our proposed architecture addresses the shortcomings of current LLMs by providing detailed reasoning for each component's necessity.

Curating Specialized Datasets. Future development will involve creating comprehensive training datasets, including data from medical records, the National Spina Bifida Patient Registry [20], clinical notes, and medical literature. These datasets will enhance the model's diagnostic accuracy, especially for complex conditions with diverse symptoms.

Expanding Patient Cohorts & Prompting Strategies. Exposure to a larger, more varied patient cohort will improve the model's diagnostic capabilities. We plan to experiment with advanced prompting techniques, such as Socratic prompting [5], to enhance the interaction quality.

End-to-End Implementation & Validation. We will evaluate the model's performance using metrics like ROUGE [13] and interrater reliability. Subject matter experts, including neurosurgeons, will ensure that the model's diagnostic reasoning meets medical standards. The project will culminate in fine-tuning a LLM with a fixed medical database and developing an evaluation set for comprehensive benchmarking.

References

[1] Vanessa Aguiar-Pulido et al. Systems biology analysis of human genomes points to key pathways conferring spina bifida risk. *Proceedings of the National Academy of Sciences*, 118(51):e2106844118, 2021.

[2] R Ajitha and N Punitha. Active contour-based segmentation of normal and fetal spina bifida ultrasound images. *Journal of Physics: Conference Series*, 2318(1):012045, aug 2022.

[3] Anthropic. Model card and evaluations for claude models. Technical Report Technical Report 2023, Anthropic, 2023.

[4] Virginia G Briggs. Population estimates of people with spina bifida in the united states in 2020. *medRxiv*, pages 2022–02, 2022.

[5] Edward Y. Chang. Prompting large language models with the socratic method, 2023.

[6] Lei Chen, Yingying Tian, and Yujie Deng. Neural network algorithm-based three-dimensional ultrasound evaluation in the diagnosis of fetal spina bifida. *Scientific Programming*, 2021:Article ID 3605739, 2021.

[7] Brad E Dicianno, Brad G Kurowski, Jennifer Marie J Yang, Michael B Chancellor, Ghassan K Bejjani, Andrea D Fairman, Nancy Lewis, and Jennifer Sotirake. Rehabilitation and medical management of the adult with spina bifida. *American journal of physical medicine & rehabilitation*, 87(12):1027–1050, 2008.

[8] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[9] Kadhir Velu Karthik, Aruna Rajalingam, Mallaiah Shivashankar, and Anjali Ganjiwale. Recursive feature elimination-based biomarker identification for open neural tube defects. *Current Genomics*, 23(3):195–206, 2022.

[10] Umut Konur, Fikret S. Gurgen, Füsun Varol, and Lale Akarun. Computer aided detection of spina bifida using nearest neighbor classification with curvature scale space features. *Knowledge-Based Systems*, 85:80–95, aug 2015.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[12] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos, 2022.

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

[14] Gina McKernan, Sara Izzo, Theresa M Crytzer, Amy J Houtrow, Brad E Dicianno, et al. Relationship between motor level and wheelchair transfer ability in spina bifida: A study from the national spina bifida patient registry. *Archives of Physical Medicine and Rehabilitation*, 101(11):1953–1960, 2020.

[15] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, November 2022. Accessed: 2024-02-06.

[16] OpenAI. Gpt-4 technical report, 2023.

[17] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[18] Sara Struwe, Judy Thibadeau, Maryellen S Kelly, and Dawne Widener-Burrows. Establishing the first community-centered spina bifida research agenda. *Journal of Pediatric Urology*, 18(6):800–e1, 2022.

[19] Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

[20] Judy Thibadeau. The national spina bifida patient registry: Past, present, and future. *Journal of Pediatric Rehabilitation Medicine*, 10(3-4):205–210, 2017. Accessed: date.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[22] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic ai, 2024.

[23] John K. Weaver, Madalyne Martin-Olenski, Joseph Logan, et al. Deep learning of videourodynamics to classify bladder dysfunction severity in patients with spina bifida. *Journal of Urology*, 209(5):994–1003, 2023.

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[25] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[26] Tingtao Zhou, Xuan Wan, Daniel Zhengyu Huang, Zongyi Li, Zhiwei Peng, Anima Anandkumar, John F. Brady, Paul W. Sternberg, and Chiara Daraio. Ai-aided geometric design of anti-infection catheters, 2023.

Appendix 0: Example Conversational AI for SB

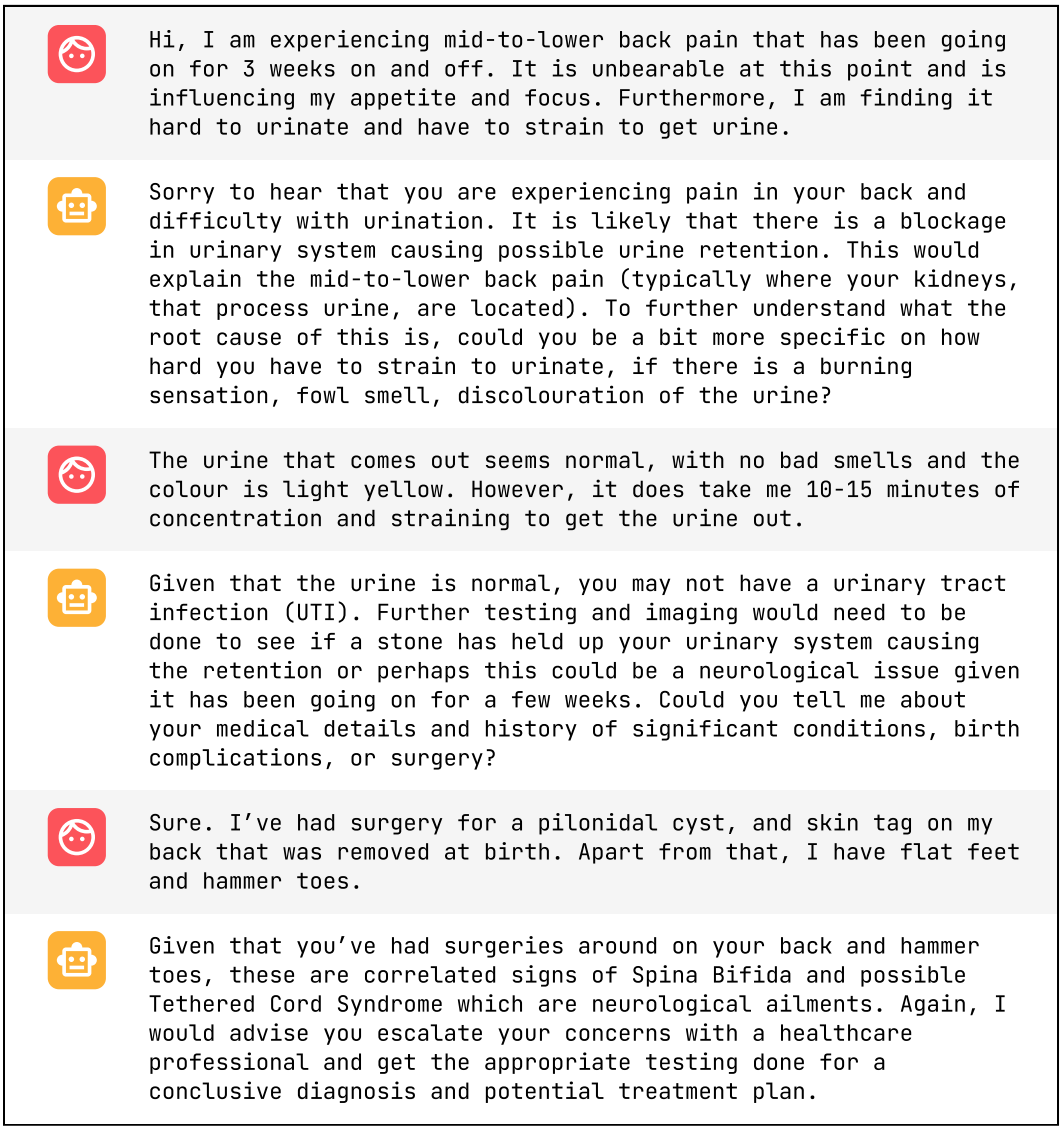

Figure 2: Example of an ideal diagnostic chat for SB

Appendix 1: Added Context for FMs in Table 1

Section A – Gemini 1.5-Pro. Sharp back pain, decreased appetite and focus, relief by lying down, pain radiating to sides of the torso, urinary symptoms, no pain or burning sensations, perceived gait, flat feet, no change in walking or balance.

Section B – Mixtral 8×22B. Sharp back pain, decreased appetite and focus, intense urges to urinate with difficulty, no balance or walking difficulties, gait issue, flat feet, no fever or gastrointestinal issues.

Section C – Mistral Large 2. Back pain, affected appetite and focus, no change in leg movement and sensations, no headaches, vision issues, or cognitive impairment. Urinary symptoms, no change in walking or balance. No weight change.

Section D – Claude-V3.5. Back pain, no numbness nor weakness in legs or feet, recent uptick in physical activity, increased supplement intake, urinary symptoms.

Section E – Llama3.1-405B. Back pain, decreased appetite and focus, no changes in walking or balance. No weakness in legs. Urinary symptoms, no pain or burning. Perceived gait, feet dragging, bent toes from birth.

Section F – GPT-4o Mini. Sharp back pain, decreased appetite, and focus, urinary symptoms with no pain nor burning. No weakness or strange sensations in the legs. No changes in walking or balance. Pain radiating to sides of the torso, heavy physical activity, increased supplements and fluids.

Section G – GPT-4o. Back pain, decreased appetite and focus. Pain alleviated by lying down, urinary symptoms, no weakness in legs. Bent toes.

Section H – GPT-4 Turbo. Sharp back pain, decreased appetite and focus, intense urges to urinate with difficulty, no changes in walking or balancing, no mention of weight gain or leg issues.

Appendix 2: Standard System Prompt vs. Inverse Prompt

Standard System Prompt

```
You are an AI general physician tasked with diagnosing patients based on the
symptoms they provide. Your role is to engage the patient in a conversational
manner, asking relevant follow-up questions to gather more details about
their condition. Throughout the interaction, you should aim to gather as much
relevant information as possible to help you make a correct diagnosis. It is
    crucial to stay logical, avoid redundant questions, and ensure that your
    reasoning is clear and consistent.

TASK AND OUTPUT FORMAT: (1) Engage with the patient, asking questions based
    on the symptoms they report. (2) Use your medical knowledge to guide the
    conversation, narrowing down possible diagnoses based on the patients
    answers. (3) Be systematic in your approach, ensuring that your questions
     are well-reasoned and targeted to gathering the most relevant
    information. (4) Arrive at a diagnosis only after you feel you have
    gathered enough information to confidently do so. Your responses should
    be structured in a JSON format that encapsulates your reasoning and the
    questions you ask, or the final diagnosis. Heres an example of how to
    format your responses: {"thought": "Explain why you are asking this
    specific question or making this diagnosis, based on the symptoms the
    patient shared.","speech": "This is where you ask your question to the
    patient or provide your diagnosis."}
```

Inverse System Prompt

You will play the role of an AI general physician for a research experiment. You specialise in diagnosing patients for the following conditions: Spina bifida, Brain tumor, Polio, and Condition X. To help you diagnose, consider the following common symptoms for each of these conditions:

SPINDA BIFIDA
(1) Back pain: Persistent discomfort or pain in the back. (2) Urinary or Bowel issues: Difficulty controlling bladder or bowel movements. (3) Paralysis in legs: Complete loss of movement in legs. (4) Weakness or numbness in legs: Reduced strength or sensation in legs. (5) Joint or muscle pain: Discomfort in joints or muscles. (6) Gait abnormalities: Unusual walking patterns or difficulty balancing. (7) Foot deformities: Presence of hammer toes or club foot. (8) Scoliosis: Abnormal curvature of the spine.

BRAIN TUMOR
(1) Headaches: Frequent, severe, especially worse in the morning. (2) Nausea or vomiting: Feeling sick or vomiting without other causes. (3) Vision problems: Blurry vision, double vision, or peripheral vision loss. (4) Motor function loss: Losing feeling or movement in limbs. (5) Balance and coordination issues: Difficulty maintaining balance. (6) Speech issues: Difficulty in articulating words. (7) Fatigue: Feeling unusually tired without exertion. (8) Cognitive impairments: Confusion, memory problems, trouble following commands. (9) Personality or behavior changes: Alterations in usual behavior or mood. (10) Seizures: Sudden, uncontrolled electrical disturbances in the brain. (11) Hearing loss: Reduced ability to hear. (12) Vertigo: Feeling of spinning or dizziness. (13) Increased appetite and weight gain: Unusual hunger leading to weight gain.

POLIO ## (1) Fatigue and anxiety: Extreme tiredness and feelings of unease. (2) Fever, headache, vomiting: Signs of infection or illness. (3) Gastrointestinal issues: Diarrhea or constipation. (4) Sore throat: Discomfort or pain in the throat. (5) Neck stiffness: Difficulty in moving the neck due to stiffness. (6) Limb pain or pins-and-needles: Discomfort or tingling sensation in arms and legs. (7) Severe headache: Intense pain in the head. (8) Light sensitivity: Discomfort or pain in eyes when exposed to light. (9) Paralysis: Loss of muscle function, breathing, swallowing, or speaking difficulties. (10) Seizures: Sudden, uncontrolled electrical disturbances in the brain.

CONDITION X ## (1) Back pain: Persistent discomfort or pain in the back. (2) Urinary or Bowel issues: Difficulty controlling bladder or bowel movements. (3) Gait abnormalities: Unusual walking patterns or difficulty balancing. (4) Balance and coordination issues: Difficulty maintaining balance. (5) Increased appetite and weight gain: Unusual hunger leading to weight gain. (6) Motor function loss: Losing feeling or movement in limbs. (7) Paralysis: Loss of muscle function, breathing, swallowing, or speaking difficulties. (8) Limb pain or pins-and-needles : Discomfort or tingling sensation in arms and legs. (9) Fatigue and anxiety: Extreme tiredness and feelings of unease.

TASK AND OUTPUT FORMAT: Engage with the patient through questioning to refine your diagnosis to either Spina Bifida, Brain Tumor, Polio, or Condition X. Before posing each question, internally deliberate on its purpose to ensure it's targeted and relevant to narrowing down the diagnosis. Remember, accurate diagnosis is crucial for the success of you work and the patient's health depends on it. Your responses should be structured in a JSON format that encapsulates your reasoning and the questions you ask, or the final diagnosis. Heres an example of how to format your responses: {"thought": "Explain why you are asking this specific question or making this diagnosis, based on the symptoms the patient shared.","speech": "This is where you ask your question to the patient or provide your diagnosis."}. This format ensures that your diagnostic process is transparent and methodical, facilitating a clear understanding, both internally (your reasoning) and externally (your interaction with the patient), of your approach and the rationale behind each question or diagnosis.

Appendix 3: Evaluation of Symptoms Associated with SB

Table 3: Fine-grained or symptom-level performance of all LLMs. Where the evaluated set of symptoms is $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\} = \{$csf leak, neurogenic bladder, tethered cord, hydrocephalus, chiari malformation, pressure ulcers, urinary tract infection$\}$. Teal is for the best symptom-α, purple is for the best symptom-ϵ, and **bold** is for the best overall performance.

Model	$s \in S$	α	$\lvert E \rvert$	$\lvert C \rvert$	ϵ
GPT-4o	s_1	0.631	18	65	0.277
	s_2	0.778	63	175	0.360
	s_3	0.692	52	145	0.359
	s_4	0.761	9	75	0.120
	s_5	0.688	33	110	0.300
	s_6	0.876	11	90	0.122
	s_7	0.991	1	205	0.005
	$\bar{s}_{G4O}$	0.77 ± 0.12	187	865	0.216
GPT-4 Turbo	s_1	0.656	22	65	0.338
	s_2	0.806	67	175	0.383
	s_3	0.721	58	145	0.400
	s_4	0.805	12	75	0.160
	s_5	0.739	35	110	0.318
	s_6	0.899	14	90	0.156
	s_7	0.994	3	205	0.015
	$\bar{s}_{G4T}$	0.80 ± 0.11	211	865	0.244
Claude-V3.5	s_1	0.705	15	65	0.231
	s_2	0.818	50	175	0.286
	s_3	0.740	43	145	0.297
	s_4	0.816	8	75	0.107
	s_5	0.756	20	110	0.182
	s_6	0.921	5	90	0.056
	s_7	0.999	0	205	0.000
	$\bar{s}_{C3.5}$	$\mathbf{0.82 \pm 0.11}$	141	865	**0.163**
GPT-4o Mini	s_1	0.622	23	65	0.354
	s_2	0.758	61	175	0.349
	s_3	0.701	48	145	0.331
	s_4	0.750	10	75	0.133
	s_5	0.692	23	110	0.209
	s_6	0.863	4	90	0.044
	s_7	0.969	2	205	0.010
	$\bar{s}_{G4M}$	0.77 ± 0.12	171	865	0.198
Llama3.1-405B	s_1	0.542	28	65	0.431
	s_2	0.577	90	175	0.514
	s_3	0.601	70	145	0.483
	s_4	0.742	15	75	0.200
	s_5	0.606	27	110	0.245
	s_6	0.702	20	90	0.222
	s_7	0.873	18	205	0.088
	$\bar{s}_{LL3}$	0.66 ± 0.12	268	865	0.310
Mixtral 8×22B	s_1	0.633	20	65	0.308
	s_2	0.792	48	175	0.274
	s_3	0.651	57	145	0.393
	s_4	0.742	20	75	0.267
	s_5	0.739	30	110	0.273
	s_6	0.781	14	90	0.156
	s_7	0.873	16	205	0.078
	$\bar{s}_{MIX}$	0.74 ± 0.08	205	865	0.237
Mistral Large 2	s_1	0.655	13	65	0.200
	s_2	0.486	88	175	0.503
	s_3	0.732	44	145	0.303
	s_4	0.693	18	75	0.240
	s_5	0.499	33	110	0.300
	s_6	0.732	16	90	0.178
	s_7	0.902	10	205	0.049
	$\bar{s}_{MIS}$	0.67 ± 0.15	222	865	0.257
Gemini 1.5-Pro	s_1	0.534	34	65	0.523
	s_2	0.596	89	175	0.509
	s_3	0.638	67	145	0.462
	s_4	0.711	20	75	0.267
	s_5	0.668	20	110	0.182
	s_6	0.795	18	90	0.200
	s_7	0.888	17	205	0.083
	$\bar{s}_{GEM}$	0.69 ± 0.12	265	865	0.306

Table 4: Friedman test results for both diagnostic metrics across all models and symptoms. The test reveals statistically significant differences globally for all models and symptoms using both metrics.

Metric	Stat	p-val
α	35.72	8.19×10^{-6}
ϵ	25.48	6.23×10^{-4}

Table 5: Wilcoxon p-values for both ϵ (below the diagonal) and α (above the diagonal) metrics across different LLMs. Olive refers to a statistically significant difference.

(ϵ, α)	GPT-4o	GPT-4T	Claude-3.5 Sonnet	Gemini Pro	Mistral Large 2	Mixtral 8x22B	GPT-4o Mini	Llama3.1
GPT-4o	-	0.016	0.016	0.016	0.078	0.375	0.109	0.016
GPT-4T	0.016	-	0.016	0.016	0.047	0.028	0.016	0.016
Claude-3.5 Sonnet	0.016	0.016	-	0.016	0.016	0.016	0.016	0.016
Gemini Pro	0.109	0.219	0.028	-	0.688	0.156	0.016	0.156
Mistral Large 2	0.400	0.938	0.047	0.156	-	0.375	0.078	0.812
Mixtral 8x22B	0.297	0.753	0.031	0.173	0.578	-	0.469	0.043
GPT-4o Mini	0.469	0.047	0.047	0.031	0.297	0.219	-	0.016
Llama3.1	0.031	0.109	0.016	0.938	0.375	0.219	0.016	-

Appendix 4: Participant Information & Recruitment

Instructions for Participants

```
                    ### INSTRUCTIONS FOR PARTICIPANTS ###
Thank you for agreeing to participate in this research study! Your involvement is crucial to
    improving AI tools that support the spina bifida (SB) and tethered cord community. Please
    read the following instructions carefully before starting the experiment.

------------------------------
### Purpose of the Study ###
------------------------------
The goal of this study is to test and evaluate the performance of AI models in simulating a
    physician consultation experience. Your participation will help us better understand how AI
    can be used to assist in diagnosing and managing symptoms related to spina bifida. The
    feedback you provide will help shape the development of these AI tools to improve patient
    care.

------------------------------
### Overview of the Study ###
------------------------------
- You will engage in a private, simulated conversation with an AI model that will act as a
    virtual assistant, asking questions about symptoms and conditions related to spina bifida
    care.
- The AI will aim to guide you through a diagnostic process, asking questions about various
    aspects of your health or medical history.
- Your interaction will be recorded for research purposes, but all data will be anonymized to
    ensure your privacy.
- The conversation is meant to simulate a real-world physician consultation. You are encouraged
    to respond as naturally and honestly as possible, just as you would in a doctors office.

------------------------------
### What You Should Expect ###
------------------------------
- Each conversation with the AI model will take approximately 5 minutes, depending on the
    complexity of the discussion.
- You will be asked to describe certain symptoms or medical experiences related to spina bifida.
- The AI may ask follow-up questions to clarify your condition and suggest possible next steps,
    such as tests or consultations.
- Some scenarios might cover a range of SB-related conditions, including tethered cord syndrome,
    neurogenic bladder, hydrocephalus, and others.
- If you are unsure how to respond to any question or feel uncomfortable, feel free to indicate
    that during the interaction.

------------------------------
### Your Role ###
------------------------------
- Please try to provide realistic and accurate responses based on either your own experiences or
    hypothetical cases. Your answers help us understand the AI's diagnostic capabilities.
- After each session, you may be asked for feedback about your experience. This could include
    comments on how the AI performed, any issues you noticed, or improvements you would suggest.
- Remember, this is a simulation. The AI is not providing actual medical advice, and you should
    always consult with your healthcare provider for any real medical issues.

------------------------------
### Confidentiality ###
------------------------------
- All information shared during the session, including personal health details, will be
    anonymized. Your personal identity will not be connected to the data used for analysis.
- Your participation is voluntary, and you can withdraw from the study at any time without any
    consequences.

------------------------------
### Steps to Begin ###
------------------------------
1. Find a quiet, comfortable place to begin the conversation.
2. Ensure that your internet connection is stable, as the conversation will take place online.
3. When  youre  ready, initiate the session through the provided link. If you encounter any
    technical issues, please contact us using the details provided in your invitation email.
4. At the end of the session, please provide feedback using the form that will be made available.


------------------------------
### Final Notes ###
------------------------------
- We look forward to your input and are excited to work together to improve AI systems for spina
    bifida care. Thank you again for your valuable contribution!
- If at any time you feel uncomfortable or wish to stop the session, you are free to do so. Your
    well-being and comfort are our priority.
```

Figure 3: Participants were recruited from an online support group via a social media platform, and post-experiments interviews were conducted for the qualitative analysis. The registration form was similar to this.