Toward Improving Diagnostic Reasoning for Spina Bifida Care: Benchmarking LLM-Patient Interactions

Anonymous ACL submission

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 present an *inverse prompting* technique aimed at guiding LLMs through a step-by-step diagnostic process. By incorporating a predefined set of symptoms relevant to SB, this approach prevents premature conclusions and enhances diagnostic reasoning, starting to address the Problem of Inclusion-Exclusion (PIE) as we have formulated it. 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 Works

The integration of advanced medical care and generative artificial intelligence (GenAI) is revolutionizing the management of complex chronic conditions like Spina Bifida (SB). SB, a neural tube defect affecting brain and spine development, presents multifaceted challenges in healthcare. The condition spans various forms, from myelomeningocele—the most severe, often coupled with hydrocephalus, Chiari II malformation, and Tethered Cord Syndrome (TCS)—to more benign occult forms. Symptoms range from bowel and urinary incontinence and mobility impairments to cognitive issues, with Spina Bifida Occulta cases sometimes remaining undetected until later in life (Dicianno et al., 2008).

Significant strides in medical science have increased the life expectancy of individuals with SB, leading to a predominantly adult population living with the condition (Briggs, 2022). This demographic shift underscores the need for improved, multidisciplinary care strategies that address neurological, urological, and psychosocial aspects. However, traditional healthcare often falls short, especially in adult SB care, resulting in gaps in access to comprehensive and coordinated management.

In response to these challenges, machine learning and deep learning techniques have emerged as pivotal tools in SB research and clinical practice. In genetics, machine learning has been instrumental in identifying critical biomarkers and elucidating the multifactorial nature of SB, as demonstrated by differential gene expression analysis (Karthik et al., 2022) and whole-genome sequence analysis (Aguiar-Pulido et al., 2021). Prenatal screening has also benefited, with deep learning models like Oct-U-Net (Chen et al., 2021), image processing (Konur et al., 2015) and segmentation algorithms (Ajitha and Punitha, 2022) enhancing the diagnostic accuracy of fetal ultrasound, thus improving early detection. In urology, DL models have refined the interpretation of complex clinical data, aiding patient management (Weaver et al., 2023), and innovations like AI-optimized catheter designs (Zhou et al., 2023) are addressing specific complications such as urinary tract infections. Moreover, AI tools

are contributing to the development of targeted rehabilitation strategies (McKernan et al., 2020) and aiding in the diagnosis of challenging cases like SBO from radiographs (Duran et al., 2023).

The emergence of large language models (LLMs) and foundation models (FMs) like ChatGPT (OpenAI, 2022), which have been trained on extensive and diverse datasets, offers a new frontier in SB care. These models hold promise for assisting caregivers and healthcare providers by offering intelligent, context-aware responses across various domains. By generating interactive agents and educational content, FMs can support individuals with SB and their caregivers in understanding SB's complexities and equip them with necessary care skills and education. However, the potential of FMs in healthcare, particularly for SB, must be navigated cautiously. Issues like the risk of misinterpretation and domain-specific constraints highlight the need for strategies like prompt engineering, fine-tuning, and retrieval augmented generation (RAG) to ensure accuracy and relevance. While RAG improves LLM performance by retrieving relevant documents during inference, it lacks the capability to adapt to incremental, symptom-based reasoning that conditions like SB require (Lewis et al., 2021). Our proposed framework differs from RAG in that it integrates a structured, symptom-based knowledge base specific to SB rather than retrieving documents during inference. The centralized database allows the model to maintain context throughout the conversation, linking symptoms progressively rather than relying on token generation via external retrieval.

Despite the potential benefits of AI in SB care, a critical gap remains in access to knowledgeable providers, particularly for the adult SB population. This gap represents the most pressing concern for individuals with SB and their caregivers (Struwe et al., 2022). To address this, there is an increasing focus on developing multimodal FMs that could supplement clinical outcomes. While models like MedPALM (Singhal et al., 2023) and AMIE (Tu et al., 2024) mark progress in this direction, no robust FM specifically designed for aiding those with SB, or similarly obscure conditions, currently exists. Such a model, informed by diverse approaches from related research, could significantly streamline access to care and improve patient outcomes.

Our study highlights the pressing need for a new conversational AI agent to assist in diagnosing and generating treatment pathways for SB, as we demonstrate that existing large language models struggle to accurately identify its relevant symptoms. Therefore, the specific contributions of this preliminary study include:

1. Benchmarking eight widely used LLMs through qualitative and quantitative evaluations, addressing the unique medical challenges presented by SB.

2. Introducing an *inverse prompting* technique, designed to guide LLMs through a step-by-step diagnostic process by incorporating a predefined set of symptoms relevant to SB and solving a formulated inclusion-exclusion task. This approach builds on, yet differs from, standard system prompts by enforcing constraints that prevent models from reaching premature conclusions.

3. Evaluating the interactions of individuals with SB using popular LLMs, leveraging the *inverse prompt* to assess its effectiveness through metrics we define, namely *diagnostic accuracy* (α), *diagnostic error rate* (ϵ), and *model course correction rate* (β).

4. Building on the identified limitations of the eight LLMs, particularly from a clinical FM perspective, we propose a novel framework designed to overcome these shortcomings.

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 progressively narrow down the space of diagnosis. 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 in a step-wise manner to avoid looping back into its own 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

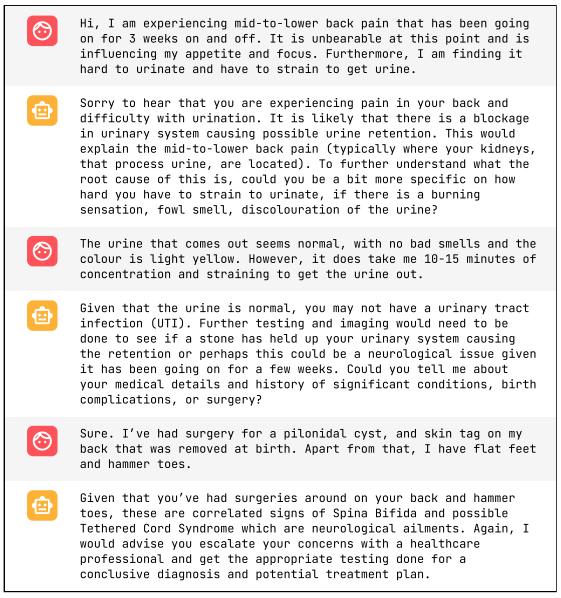

Figure 1: Example of an ideal diagnostic chat for SB.

the patient or their clinician) instead of jumping to a diagnostic result prematurely while bypassing steps in its way.

A major 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 1 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 2) through a conversational agent, instead of a single generalized foundation model.

2.1 Formulating The Problem of Inclusion-Exclusion (PIE)

The process of diagnosing a condition involves systematically gathering information to converge on the most likely diagnosis given a patient's symptoms. This can be framed as a Problem of Inclusion-Exclusion (PIE), where the LLM iteratively queries the user about symptoms, refining its understanding by identifying the most relevant symptoms for differential diagnosis and eliminating less relevant ones as the conversation transpires. The objective is to minimize the number of unnecessary queries while ensuring high diagnostic accuracy (α, defined in Section 2.4), low error rate (ϵ, defined in Section 2.6), and reduced need for human intervention (β, defined in Section 2.7). Hence, the following formulation can help automate the diagnostic process in clinical settings, saving time and improving decision-making efficiency.

Symptom Set for Each Condition. Given a set of conditions $C = \{c_1, c_2, \ldots, c_m\}$, each condition c_i is associated with a set of symptoms $S_i = \{s_{i1}, s_{i2}, \ldots, s_{in_i}\}$, where s_{ij} represents a specific symptom related to condition c_i. Let $U = \{s_1, s_2, \ldots, s_n\}$ be the universal set of all possible symptoms across all conditions. The objective is to identify the correct condition $c^* \in C$ by querying the user about the presence or absence of symptoms in U. To do this efficiently, the LLM uses an inclusion-exclusion approach, querying about a symptom s_i if it contributes to the differentiation between multiple conditions. Each condition c_i is characterized by a set of symptoms $S_i \subseteq U$. We assume that each condition has a unique combination of symptoms, though overlap in symptoms between conditions is common.

Query Process. Let $Q \subseteq U$ be the set of symptoms that have been queried so far, and let $R(Q) \subseteq C$ represent the remaining set of possible conditions based on the symptoms identified from the queries. Initially, $R(Q) = C$. The model uses an inclusion-exclusion update rule to dynamically refine $R(Q)$ based on the user's response:

$$R(Q \cup \{s_k\}) = \begin{cases} R(Q) \cap \{c_i \mid s_k \in S_i\} \\ \text{if user confirms } s_k \quad (sk = 1) \\ R(Q) \setminus \{c_i \mid s_k \in S_i\} \\ \text{if user denies } s_k \quad (sk = 0) \end{cases}$$

Query Selection. To minimize the number of queries, the LLM selects a symptom $s_k \in U \setminus Q$ that both maximizes an information gain function, ϕ, to reduce uncertainty and divides the remaining set of conditions $R(Q)$ as evenly as possible based on symptom presence or absence. This combined approach ensures that each query contributes to reducing diagnostic ambiguity while maintaining an even partition of conditions. The query selection rule is: $s_k^* = \arg\max_{s \in U \setminus Q}(\gamma \cdot \phi(s, R(Q)) + (1-\gamma) \cdot \min(|\{c_i \in R(Q) \mid s \in S_i\}|, |\{c_i \in R(Q) \mid s \notin S_i\}|)) = \arg\max_{s \in U \setminus Q}(\gamma \cdot \chi_{\text{info-gain}} + (1-\gamma) \cdot \chi_{\text{set-balance}})$, where:

- $\chi_{\text{info-gain}}$ is the information gain from querying symptom s_k, calculated as: $\phi(s_k) =$

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 (Team, 2023)	Inverse, Bridging	ToT	See Appendix A1	✓(0.5)
Mixtral 8×22B (Jiang et al., 2024)	Inverse, Bridging	CoT	See Appendix A2	✗(0.0)
Mistral Large 2	Inverse	CoT, ToT	See Appendix A3	✓(1.0)
Claude-3.5 Sonnet (Anthropic, 2023)	Inverse, Bridging	CoT	See Appendix A4	✗(0.0)
Llama-3.1-405B (Touvron et al., 2023)	Inverse	ToT	See Appendix A5	✗(0.0)
GPT-4o Mini (OpenAI, 2022)	Inverse, Bridging	CoT	See Appendix A6	✗(0.0)
GPT-4o (OpenAI, 2023)	Inverse	CoT	See Appendix A7	✓(1.0)
GPT-4 Turbo	Inverse	CoT, ToT	See Appendix A8	✗(0.0)

$H(C) - H(C|s_k)$, where:

- $H(C) = -\sum P(c_i) \cdot \log(P(c_i))$
- $H(C|s_k) = P(s_k = 1) \cdot H(C|s_k = 1) + P(s_k = 0) \cdot H(C|s_k = 0)$
- Here, $P(c_i)$ is the probability of condition c_i given the patient's responses up to a given point, $P(s_k = r \in \mathbb{B})$ are the probabilities of receiving a response that indicates the presence (or absence) of symptom s_k, and $H(C|s_k = r \in \mathbb{B})$ are the entropies of the condition set after receiving the respective response.

$H(C)$ is the entropy of the set of possible conditions before querying any new symptom and $H(C|s_k)$ is the conditional entropy after the query. In general, when the LLM queries a symptom, the response updates the model's belief about the likelihood of each condition, allowing it to efficiently narrow down the set of possible conditions and that is the overall function ϕ.

- $\chi_{\text{set-balance}}$ ensures the remaining conditions are as evenly split as possible between those where the symptom is present and those where it is absent.

- γ is a weighting factor that balances the trade-off between maximizing information gain and maintaining set balance (defaulted to 0.5).

With this setup, the importance of the set $R(Q)$ cannot be overstated, as both information gain and set balancing depend on the current state of $R(Q)$, not just the global set of conditions C. As the model queries symptoms and updates $R(Q)$, the relative value of querying any symptom $s_k \in U \setminus Q$ changes based on how effectively it helps differentiate between the remaining conditions in $R(Q)$.

Utility Function. To further optimize and measure the process, we adjust the utility of each symptom query to account for α, ϵ, and β. The full utility function becomes: $\mathcal{U}(s_k) = \gamma(\chi_{\text{info-gain}} + \lambda_\alpha P_\alpha - \lambda_\epsilon P_\epsilon - \lambda_\beta P_\beta) + (1-\gamma)\chi_{\text{set-balance}}$, where:

- $P_{\text{metric}} = P(\text{metric}|s_k)$ are the probabilities that querying s_k increases accuracy, decreases errors, or reduces the need for bridging, respectively.

- λ_{metric} are weights controlling the importance of these metrics in the query selection (defaulted to $0.\overline{33}$).

Convergence Requirement. The process continues until $|R(Q)| = 1$, at which point the LLM has identified the most likely condition c^*. The objective is to minimize the number of queries, that is, to find the smallest $|Q|$ such that:

$$\arg\min_{Q \subseteq U} |Q| \text{ subject to } |R(Q)| = 1$$

This formulation provides a framework for LLMs to efficiently converge on a diagnosis by minimizing the number of queries and using an inclusion-exclusion strategy to rule out possible conditions. This approach can be extended to handle uncertainty in user responses or symptoms that occur probabilistically across conditions. However, this is a preliminary study to integrate a logical querying strategy into LLMs for diagnostic tasks. The number of SB symptoms (that we surveyed) required for convergence is relatively small, so the performance gain from an adaptive formulation of PIE is limited in this case. As a result, our current design of the *inverse prompt* employs a brute-force method to traverse the symptoms. At a later stage, we plan to integrate an agent-based solver that progressively computes the entropy of querying outcomes, which should demonstrate superior performance compared to the brute-force method.

2.2 Constructing the Inverse Prompt

Inverse prompting refers to a structured system prompt that primes the model with a predefined set

of SB-related symptoms, enabling the LLM to initiate the conversation with context relevant to SB. This approach helps guide patients through symptom reporting, thereby avoiding the need for the model to "discover" symptoms mid-conversation. In other words, we seek to minimize the risk of the LLM inadvertently generating new symptoms based on assumption or stochastic hallucination. Instead, the inverse prompt steers the model towards actively querying the patient for relevant symptoms, following a well-defined pathway. By framing symptom discovery as a Problem of Inclusion-Exclusion (PIE), inverse prompting mitigates the model's tendency to hallucinate irrelevant symptoms.

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 B).

2.3 User Interaction Logic

Now that 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, TCS, and neurogenic bladder.

2.4 Qualitative Experimental Setup

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. 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) (Wei et al., 2023) and tree-of-thought (ToT) (Yao et al., 2023) and by deploying bridge[1] (Li et al., 2022) 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: $\alpha = \frac{\sum(\psi + 0.5\phi)}{|L|}$, where ψ are the correct impressions, ϕ are the partial impressions, and $|L|$ is the total number of conversations.

2.5 Qualitative Experimental Results

GPT-4 Turbo. While bridging was not required, it was surprisingly overly reliant on the details given by the inverse prompt. It tended to be cautious, often recommending additional tests or medical scans rather than making a direct diagnosis. Rarely did

[1] We adapted bridging to SB care by linking symptoms that are not immediately related but may have clinical relevance in SB patients. For example, urinary incontinence may signal both neurogenic bladder and TCS, which are not typically linked in general diagnostics but are critical in SB.

it venture beyond the information from the inverse prompt, yet it was competent in proceeding with a systematic approach to questioning a patient.

GPT-4o. It developed a very structured understanding of the inverse prompt. It could reason progressively, thus building upon each of the responses from the patient. Thus it needed the least bridging or additional context. The nature of the conversations (typically short to medium in length) indicated its ability to formulate the diagnosis task as a PIE problem and try to converge instead of adopting an adjacent sequence of reasoning. In other words, GPT-4o had the best sequential inference ability of the eight FMs.

GPT-4o Mini. It failed to retain relevant context even after additional information via bridging. In practice, it focused on providing remedies based on the most recent prompts rather than forming connections with the previous information into a collective result.

Gemini 1.5-Pro. It did a satisfactory job of aligning with the context of the inverse prompt but chose to hard iterate through the symptoms instead of using a flow of linkage or just medical reasoning. Therefore a lot of information had to be bridged to drive it to a correct diagnosis thereby leading to lengthier conversations.

Claude-3.5 Sonnet. Similar to GPT-4o, it adopts the PIE formulation resulting in progressive reasoning with usually a better diagnostic accuracy. Although in some cases it tends to be more limited in its reasoning due over-reliance on processing eliminations. This leads to abrupt ends in conversations and arriving to conclusions prematurely.

Llama-3.1-405B. It was often too hesitant to make a diagnosis, leading to the asking of multiple questions and looping on the information (i.e., asking the same questions repeatedly in the same conversation). In the cases where it did make a diagnosis, particularly when its reasoning narrowed down to either SB or the composite synthetic condition, it invariably resorted to choosing the composite synthetic condition.

Mixtral 8×22B. It was not comprehensive about asking for information, and the obtained reasoning was usually tangential to the expected medical reasoning. This prevented it from asking for enough relevant information, which often resulted in a case of misdiagnosis after a few exchanges with the user.

Mistral Large 2. It had a tendency to leap to unjustified conclusions without first posing the necessary questions. This habit could disrupt its logical sequence of thought, resulting in an inaccurate diagnosis despite attempts at forced bridging.

2.6 Quantitative Experimental Setup

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 $\epsilon = \frac{|E|}{|L|}$, where $|E|$ 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.

2.7 Quantitative Experimental Results

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 $16\times$ better. The standard system prompt is likely making the correct diagnoses for the wrong reasons at the same rate as when there is no system prompt present – validating our hypothesis that inverse prompting effectively handles the PIE inherent in complex symptom reporting. Additionally, the set of LLMs demonstrates increased precision, with fewer conversations requiring bridging or forced user interventions for course correction ($\beta = \frac{|B|}{|L|}$, where $|B|$ is the total times bridging is needed), while simultaneously shortening the overall length of model-patient interactions ($\bar{L}$).

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. Interestingly, although Claude-3.5 Sonnet 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-3.5 Sonnet. This implies that GPT-4o may require more contextual bridging across conversations compared

[2]There is a significant negative correlation between α and ϵ ($\rho = -0.762$, p-val $= 0.028$). Find additional statistical tests in Appendix C.

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_I, ϵ_I), show best performance on the entire LLM set. β is the portion of conversations that required bridging and $\bar{L}$ is the mean conversation length by prompts required to make a diagnostic conclusion.

Model	α_O	ϵ_O	β_O	$\bar{L}_O$	α_X	ϵ_X	β_X	$\bar{L}_X$	α_I	ϵ_I	β_I	$\bar{L}_I$
GPT-4o	0.752	0.311	0.351	6.22	$0.803_{\uparrow 0.05}$	$0.304_{\downarrow 0.01}$	$0.316_{\downarrow 0.04}$	$6.05_{\downarrow 0.17}$	$0.886_{\uparrow 0.13}$	$0.162_{\downarrow 0.15}$	$0.238_{\downarrow 0.11}$	$5.15_{\downarrow 1.07}$
GPT-4 Turbo	0.738	0.336	0.389	6.94	$0.789_{\uparrow 0.05}$	$0.328_{\downarrow 0.01}$	$0.354_{\downarrow 0.04}$	$6.82_{\downarrow 0.12}$	$0.845_{\uparrow 0.11}$	$0.170_{\downarrow 0.17}$	$0.226_{\downarrow 0.16}$	$5.47_{\downarrow 1.47}$
Claude-3.5 Sonnet	0.744	0.289	0.363	5.85	$0.792_{\uparrow 0.05}$	$0.267_{\downarrow 0.02}$	$0.355_{\downarrow 0.01}$	$5.73_{\downarrow 0.12}$	$0.853_{\uparrow 0.11}$	$0.099_{\downarrow 0.19}$	$0.193_{\downarrow 0.17}$	$4.98_{\downarrow 0.87}$
Gemini 1.5-Pro	0.720	0.401	0.447	7.60	$0.753_{\uparrow 0.03}$	$0.396_{\downarrow 0.01}$	$0.439_{\downarrow 0.01}$	$7.52_{\downarrow 0.08}$	$0.812_{\uparrow 0.09}$	$0.235_{\downarrow 0.17}$	$0.295_{\downarrow 0.15}$	$6.86_{\downarrow 0.74}$
Mistral Large 2	0.696	0.357	0.409	7.28	$0.722_{\uparrow 0.03}$	$0.350_{\downarrow 0.01}$	$0.387_{\downarrow 0.02}$	$7.17_{\downarrow 0.11}$	$0.782_{\uparrow 0.09}$	$0.275_{\downarrow 0.08}$	$0.259_{\downarrow 0.15}$	$6.58_{\downarrow 0.70}$
Mixtral 8×22B	0.722	0.383	0.377	6.82	$0.758_{\uparrow 0.04}$	$0.379_{\downarrow 0.00}$	$0.371_{\downarrow 0.01}$	$6.79_{\downarrow 0.03}$	$0.828_{\uparrow 0.11}$	$0.304_{\downarrow 0.18}$	$0.274_{\downarrow 0.10}$	$5.63_{\downarrow 1.19}$
GPT-4o Mini	0.707	0.323	0.438	6.95	$0.785_{\uparrow 0.08}$	$0.315_{\downarrow 0.01}$	$0.429_{\downarrow 0.01}$	$6.92_{\downarrow 0.03}$	$0.867_{\uparrow 0.16}$	$0.164_{\downarrow 0.16}$	$0.230_{\downarrow 0.21}$	$5.19_{\downarrow 1.76}$
Llama-3.1-405B	0.655	0.420	0.451	9.51	$0.692_{\uparrow 0.04}$	$0.411_{\downarrow 0.01}$	$0.446_{\downarrow 0.01}$	$9.29_{\downarrow 0.22}$	$0.758_{\uparrow 0.10}$	$0.236_{\downarrow 0.18}$	$0.326_{\downarrow 0.13}$	$7.95_{\downarrow 1.56}$
Mean Scores	0.717	0.353	0.403	7.15	$0.762_{\uparrow 0.05}$	$0.345_{\downarrow 0.01}$	$0.387_{\downarrow 0.02}$	$7.03_{\downarrow 0.12}$	$0.829_{\uparrow 0.11}$	$0.193_{\downarrow 0.16}$	$0.255_{\downarrow 0.15}$	$5.98_{\downarrow 1.17}$

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

Model	s	α	$\lvert E \rvert$	$\lvert L \rvert$	ϵ	**Model**	s	α	$\lvert E \rvert$	$\lvert L \rvert$	ϵ
	s_1	0.631	18	65	0.277		s_1	0.656	22	65	0.338
	s_2	0.778	63	175	0.360		s_2	0.806	67	175	0.383
	s_3	0.692	52	145	0.359		s_3	0.721	58	145	0.400
GPT-4o	s_4	0.761	9	75	0.120	GPT-4 Turbo	s_4	0.805	12	75	0.160
	s_5	0.688	33	110	0.300		s_5	0.739	35	110	0.318
	s_6	0.876	11	90	0.122		s_6	0.899	14	90	0.156
	s_7	0.991	1	205	0.005		s_7	0.994	3	205	0.015
	$\bar{s}_{G4O}$	0.77 ± 0.12	187	865	0.216		$\bar{s}_{G4T}$	0.80 ± 0.11	211	865	0.244
	s_1	0.705	15	65	0.231		s_1	0.622	23	65	0.354
	s_2	0.818	50	175	0.286		s_2	0.758	61	175	0.349
	s_3	0.740	43	145	0.297		s_3	0.701	48	145	0.331
Claude-3.5 Sonnet	s_4	0.816	8	75	0.107	GPT-4o Mini	s_4	0.750	10	75	0.133
	s_5	0.756	20	110	0.182		s_5	0.692	23	110	0.209
	s_6	0.921	5	90	0.056		s_6	0.863	4	90	0.044
	s_7	0.999	0	205	0.000		s_7	0.969	2	205	0.010
	$\bar{s}_{C3.5}$	$\mathbf{0.82 \pm 0.11}$	141	865	**0.163**		$\bar{s}_{G4M}$	0.77 ± 0.12	171	865	0.198
	s_1	0.542	28	65	0.431		s_1	0.633	20	65	0.308
	s_2	0.577	90	175	0.514		s_2	0.792	48	175	0.274
	s_3	0.601	70	145	0.483		s_3	0.651	57	145	0.393
Llama-3.1-405B	s_4	0.742	15	75	0.200	Mixtral 8×22B	s_4	0.742	20	75	0.267
	s_5	0.606	27	110	0.245		s_5	0.739	30	110	0.273
	s_6	0.702	20	90	0.222		s_6	0.781	14	90	0.156
	s_7	0.873	18	205	0.088		s_7	0.873	16	205	0.078
	$\bar{s}_{LL3}$	0.66 ± 0.12	268	865	0.310		$\bar{s}_{MIX}$	0.74 ± 0.08	205	865	0.237
	s_1	0.655	13	65	0.200		s_1	0.534	34	65	0.523
	s_2	0.486	88	175	0.503		s_2	0.596	89	175	0.509
	s_3	0.732	44	145	0.303		s_3	0.638	67	145	0.462
Mistral Large 2	s_4	0.693	18	75	0.240	Gemini 1.5-Pro	s_4	0.711	20	75	0.267
	s_5	0.499	33	110	0.300		s_5	0.668	20	110	0.182
	s_6	0.732	16	90	0.178		s_6	0.795	18	90	0.200
	s_7	0.902	10	205	0.049		s_7	0.888	17	205	0.083
	$\bar{s}_{MIS}$	0.67 ± 0.15	222	865	0.257		$\bar{s}_{GEM}$	0.69 ± 0.12	265	865	0.306

to Claude-3.5 Sonnet – which is evidenced by their respective β values.

We then applied a similar approach to more challenging tasks at the fine-grained level, focusing on symptoms associated with SB (Table 3). In general, all models perform slightly worse at identifying very fine-grained ailments, such as TCS (s_3) and Chiari II malformation (s_5). However, they show decent performance in detecting common symptoms requiring immediate attention, such as pressure injuries (s_6) and urinary tract infections (s_7). On the other hand, rarer but critical symptoms, such as a cerebrospinal fluid (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 2), as inverse prompting techniques may encounter a natural bottleneck (which warrants further investigation as discussed in Section 2.1), to ensure accurate and reliable outcomes across all SB-related conditions.

3 Proposed Solution for Larger Study

In the experiments, we identified several shortcomings with the models in adapting to diagnostic tasks. The primary issue was the lack of a corresponding database for the specified conditions. The use of inverse prompting served as an initial inductive step (or proof of concept) to empirically demonstrate this limitation within our experiments. A condition

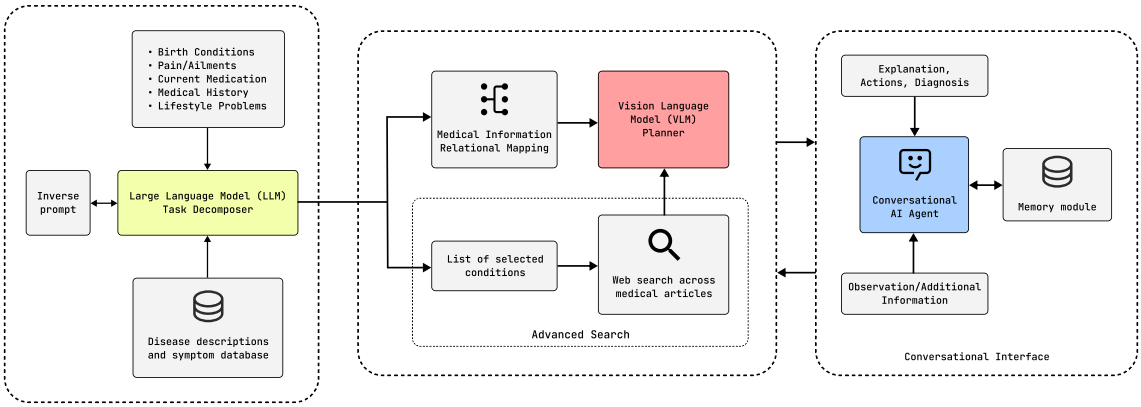

Figure 2: Architectural overview of the proposed system

as complex as SB has a combination of urological, neurological, and musculoskeletal implications. Thus we propose a multistage architecture suitable for diagnostic tasks. Figure 2 lays out the architecture of our proposed approach that integrates into the mainstream conversation between the patient and the models.

3.1 Model Architecture

The first module is responsible for forming its own directed corpus based on targeted information from the patient. The model would carry out these sectional inquiries after the conversation starts and the model obtains its primary prompt. The model repeatedly matches the information to the predefined descriptions to formally narrow down the search space for diagnosis. The nature of the expected responses involves a more focused information retrieval setup rather than a reasoning setup. Once the top-level information has been obtained, the LLM decomposes it into several viable tasks that require more reasoning and fundamental inputs.

In the second module, the planner module starts with two main setups; a relational mapping between the top-level information collected from the patient, and the list of selected conditions obtained from the matching in the predecessor module. We propose deploying a vision-language planner instead of an LLM planner because it extends the model's scope to request more diverse modalities as input in this stage. Also, it would allow the model to interpret better the web-based corpus related to the specific conditions, since the information available may not be restricted to text. This second module essentially loops to form the conversation AI backend that connects to the interface. To adapt the type of reasoning-based queries obtained from patient responses, we predefine the nature of inputs and responses extracted from the text along with a memory unit to facilitate information linkage across longer conversations. Lastly, adopting discrete model blocks enables more efficient agents that follow PIE heuristics to converge on a diagnosis.

3.2 Conclusion & Future Work

We evaluated eight LLMs using qualitative and quantitative methods to assess their ability to diagnose SB, introducing inverse prompting to guide step-wise reasoning. Our results revealed strengths and weaknesses in handling SB cases and highlighted a gap in AI's ability to help patients recognize complications and seek timely medical advice. This underscores the need for integrating clinical conversations, health records, and literature evidence. Our approach and proposed framework, based on experimental insights, addresses current LLM shortcomings and aims to improve healthcare for SB patients through:

Curating Specialized Datasets. We plan to refine training datasets by integrating diverse medical data, including records, the National Spina Bifida Patient Registry (Thibadeau, 2017), clinical notes, and preprocessed literature, to enhance diagnostic reliability.

Patient Cohorts & Solving Strategies. Future work will involve larger patient cohorts and experimenting with strategies like discrete querying agents and Socratic prompts (Chang, 2023) to improve model performance.

End-to-End Implementation & Validation. We will assess the model using metrics like ROUGE (Lin, 2004) and interrater reliability, with expert validation to ensure diagnostic accuracy. A fixed medical database will be used to fine-tune the LLM and develop a benchmark evaluation set.

4 Limitations & Ethical Implications

The use of conversational AI in healthcare, especially for complex conditions like SB, raises ethical concerns. A key issue is the risk of misdiagnosis as LLMs lack the deep medical understanding of healthcare professionals. Incorrect advice could delay treatment or lead to inappropriate care for individuals with SB. Our proposed system aims to mitigate this risk through a structured, symptom-based knowledge repository and inverse prompting, but it could face limitations such as data gaps, biased training, and language ambiguities if not developed diligently. Privacy is another concern as data used to train models must follow strict confidentiality rules. Although our study does not collect patient data, future use of specialized datasets must comply with frameworks like HIPAA, ensuring anonymity and informed consent.

Our study's limitations include its relatively small sample size (50 participants) and focus on specific SB-related conditions, limiting generalizability. While inverse prompting improved diagnostics, it relies on potentially outdated knowledge, underscoring the need for regular updates and adaptation to different languages and cultures. Finally, AI-assisted care should complement, not replace, medical professionals. It can enhance access to information, especially in underserved areas, but lacks the nuanced judgment of a clinician. A collaborative approach between AI and healthcare providers is essential to ensure safe and effective patient care.

This study was conducted following approval from the Institutional Review Board (IRB) at Stanford University (Protocol 76331, IRB8, Registration 6208), which determined the research to involve minimal risk to participants. Recruitment and data collection adhered to the IRB's guidelines, including the use of informed consent procedures. The study complies with federal regulations for the protection of human subjects, including 45CFR46, and retains participant confidentiality in accordance with Stanford's Research Policy Handbook.

References

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

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

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

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

Edward Y. Chang. 2023. Prompting large language models with the socratic method. *Preprint*, arXiv:2303.08769.

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

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

Semra Duran, Kemal Ureten, Yüksel Maraş, Hadi Maraş, Kevser Gök, Ebru Atalar, and Velihan Çayhan. 2023. Automatic detection of spina bifida occulta with deep learning methods from plain pelvic radiographs. *Research on Biomedical Engineering*, 39:655–661.

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. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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

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

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. 2022. Bridge-prompt: Towards ordinal action understanding in instructional videos. *Preprint*, arXiv:2203.14104.

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

Gina McKernan, Sara Izzo, Theresa M Crytzer, Amy J Houtrow, Brad E Dicianno, et al. 2020. 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.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2024-02-06.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

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

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

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. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Tingtao Zhou, Xuan Wan, Daniel Zhengyu Huang, Zongyi Li, Zhiwei Peng, Anima Anandkumar, John F. Brady, Paul W. Sternberg, and Chiara Daraio. 2023. Ai-aided geometric design of anti-infection catheters. *Preprint*, arXiv:2304.14554.

A Added Context for FMs in Table 1

Section	Description
1. 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.
2.. 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.
3. 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.
4. Claude-3.5 Sonnet	Back pain, no numbness nor weakness in legs or feet, recent uptick in physical activity, increased supplement intake, urinary symptoms.
5. 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.
6. 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.
7. GPT-4o	Back pain, decreased appetite and focus. Pain alleviated by lying down, urinary symptoms, no weakness in legs. Bent toes.
8. 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.

B 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.

C Statistical Testing for Results in Table 2

Table 2.1: Friedman test results for both diagnostic metrics, α and ϵ, and both conversational metrics, β and $\bar{L}$, across all models and symptoms. The tests reveal statistically significant differences globally for all models and symptoms using all metrics.

Metric	Stat	p-val
α	35.72	8.19×10^{-6}
ϵ	25.48	6.23×10^{-4}
β	16.00	3.35×10^{-4}
$\bar{L}$	15.55	4.20×10^{-4}

Table 2.2: 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	-

Table 2.3: Nemenyi test results for β and $\bar{L}$ values. Olive refers to a statistically significant difference. Results indicate that there are statistically significant differences in the comparisons involving β_O vs. β_I and $\bar{L}_O$ vs. $\bar{L}_I$, with p-values of 0.001, which are well below conventional significance thresholds. This implies that the inverse prompt (β_I, $\bar{L}_I$) leads to significant improvements compared to the baseline (β_O, $\bar{L}_O$). On the other hand, no significant differences were found between other comparisons, suggesting comparable performance in those pairs.

		O	**X**	**I**
β	β_O	1.000	0.112	0.001
	β_X	0.112	1.000	0.112
	β_I	0.001	0.112	1.000
$\bar{L}$	$\bar{L}_O$	1.000	0.187	0.001
	$\bar{L}_X$	0.187	1.000	0.085
	$\bar{L}_I$	0.001	0.085	1.000

Table 2.4: Bonferroni-Dunn test results for β and $\bar{L}$ values. Olive refers to a statistically significant difference. Results confirm significant differences for β_O vs. β_I with a corrected p-value of 0.000, suggesting a strong improvement with the inverse prompt over the baseline. Similarly, for $\bar{L}$, there was a near-significant result in the comparison between $\bar{L}_O$ and $\bar{L}_I$ (p-val = 0.099), indicating that while the inverse prompt seems to reduce conversation length effectively, the evidence is not as strong as for β. The remaining comparisons do not show statistically significant differences, implying consistent results across these pairs.

Comparison	β p-val (corrected)	$\bar{L}$ p-val (corrected)
O vs. X	0.941	1.725
O vs. I	0.000	0.099
X vs. I	0.112	0.084

D Participant Recruitment and Instructions

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.