

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

Asfandyar Azhar

Dept. of Biomedical Engineering, Carnegie Mellon University, Pennsylvania, USA

Dept. of Radiology, Stanford University School of Medicine, California, USA

AAZHAR@ANDREW.CMU.EDU

Shaurjya Mandal

Mass General Hospital, Harvard Medical School, Massachusetts, USA

SMANDAL2@MGH.HARVARD.EDU

Zaid Khan

School of Medicine, Eastern Virginia Medical School, Virginia, USA

KHANZS@ODU.EDU

Nidhish Shah

Dept. of Computer Science, Eindhoven University of Technology, Eindhoven, NL

NIDHISHSHAHS@GMAIL.COM

Curtis Langlotz

Dept. of Radiology, Stanford University School of Medicine, California, USA

LANGLOTZ@STANFORD.EDU

Brad Dicianno

Dept. of Physical Medicine & Rehabilitation, University of Pittsburgh School of Medicine, Pennsylvania, USA

DICIANNO@PITT.EDU

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. This study presents 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 formulated in this study. 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, this study proposes 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 po-

tential 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.

Data and Code Availability This study provides methodological details sufficient for replication. Prompt templates (including inverse prompts and DR-CoT instructions), the definitions of α , ϵ , and β , and aggregation procedures are specified in Section 3 -to- 3.1. Tables report the counts used to compute metrics. Where raw transcripts cannot be shared due to model licensing, redacted exemplars have been provided that preserve interaction structure while omitting proprietary content (available via Github repository https://github.com/asfandyarazhar13/conversational_SBAI/tree/main).

Institutional Review Board (IRB) 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.

1. Introduction & Related Work

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 (Aguilar-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, deploying FMs in healthcare, particularly for SB, requires cautious navigation. 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., 2020). 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).

While prior work has utilized Retrieval-Augmented Generation (RAG) and reasoning frameworks like Chain-of-Thought (CoT) and Tree-of-Thoughts (ToT) to improve medical question answering, these methods are primarily evaluated on static datasets. In contrast, diagnosing SB complications requires sequential, conversational reasoning where an AI iteratively gathers information to narrow a hypothesis space. Furthermore, our goal is not to perform definitive static classification or outperform

clinicians. Rather, we propose this system as an accessible conversational decision-support and triage aid, helping an underserved adult SB populace surface relevant symptoms and seek timely, appropriate care. 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.

This study highlights the pressing need for a new conversational AI agent to assist in diagnosing and generating treatment pathways for SB. As noted, 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 defined metrics, namely *diagnostic accuracy* (α), *diagnostic error rate* (ϵ), and *model course correction rate* (β).

2. Methods

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,

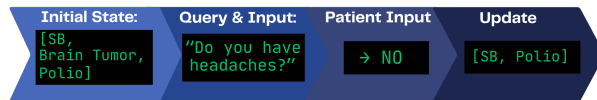


Figure 1: Visual depiction of PIE framework.

the success and failure criteria for the performance of LLMs as clinical FMs was defined as:

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 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 2 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, a pipeline of operations will be required (see Section 3.2) 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

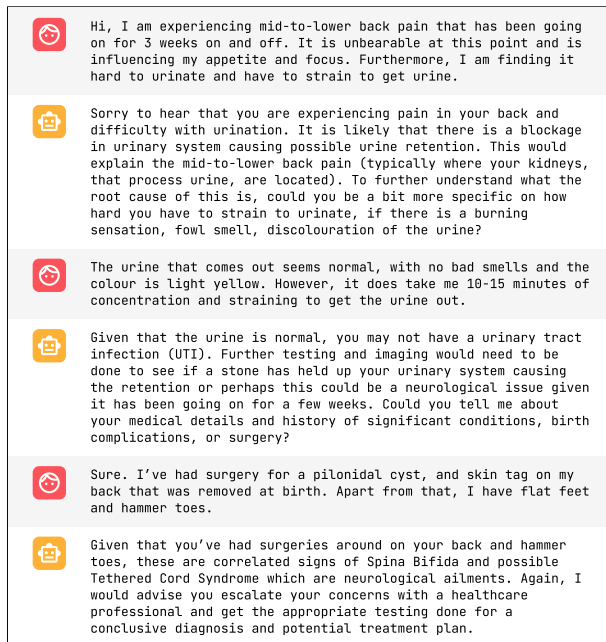


Figure 2: Example of an ideal diagnostic chat for SB.

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 3), low error rate (ϵ , defined in Section 3.1), and reduced need for human intervention (β , also defined in SubSection 3.1). Hence, the following formulation can help automate the diagnostic process in clinical settings, saving time and improving decision-making efficiency.

Our approach exclusively employs a stepwise reasoning process, progressively narrowing the potential disease space and converging on a singular condition or producing. To ensure the reasoning process is both transparent and clinically valid, clinical experts were consulted to design and validate the model’s outputs. Specifically, crafted prompts were created (see Appendix A) requiring the model to articulate its rationale or “thought process.” Furthermore, the inclusion of a reasoning or “thought” field in the JSON

output ensures that the model’s decision-making process is not a black-box. This transparency allows patients using a deployed version, such as within a mobile health application (Dicianno et al., 2015; Yu et al., 2019) to share the conversation and the model’s reasoning with their primary care physician. This enables physicians to review and contextualize the AI’s recommendations within the broader clinical picture, fostering trust and ensuring alignment with real-world clinical workflows.

Symptom Set for Each Condition. Given a set of conditions $C = \{c_1, c_2, \dots, c_m\}$, each condition c_i is associated with a set of symptoms $S_i = \{s_{i1}, s_{i2}, \dots, s_{in}\}$, where s_{ij} represents a specific symptom related to condition c_i . Let $U = \{s_1, s_2, \dots, s_n\}$ 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 $S_i \subseteq U$. It was assumed 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:

$$R(Q \cup \{s_k\}) =$$

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

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, Ω , 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

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 3 (Gemini Team, 2023)	Inverse, Bridging	ToT	See Appendix B1	\checkmark (0.5)
Mixtral 8 \times 22B (Jiang et al., 2024)	Inverse, Bridging	CoT	See Appendix B2	\times (0.0)
Mistral Large 3 (Jiang et al., 2023)	Inverse	CoT, ToT	See Appendix B3	\checkmark (1.0)
Claude-4.5 Sonnet (Anthropic, 2023)	Inverse, Bridging	CoT	See Appendix B4	\times (0.0)
Llama 4 (Touvron et al., 2023)	Inverse	ToT	See Appendix B5	\times (0.0)
GPT-4o (OpenAI, 2023)	Inverse, Bridging	CoT	See Appendix B6	\times (0.0)
GPT-5.1 (OpenAI, 2023)	Inverse	CoT	See Appendix B7	\checkmark (1.0)
GPT-5.2 (OpenAI, 2023)	Inverse	CoT, ToT	See Appendix B8	\times (0.0)

of conditions. The query selection rule is:

$$s_k^* = \arg \max_{s \in U \setminus Q} \left(\gamma \cdot \Omega(s, R(Q)) + (1 - \gamma) \cdot \min(|\{c_i \in R(Q) \mid s \in S_i\}|, |\{c_i \in R(Q) \mid s \notin S_i\}|) \right),$$

i.e.,

$$s_k^* = \arg \max_{s \in U \setminus Q} \left(\gamma \cdot \chi_{\text{info-gain}} + (1 - \gamma) \cdot \chi_{\text{set-balance}} \right),$$

where:

- $\chi_{\text{info-gain}}$ is the information gain from querying symptom s_k , calculated as

$$\phi(s_k) = H(C) - H(C \mid s_k)$$

with

$$\begin{aligned} H(C) &= - \sum P(c_i) \log P(c_i), \\ H(C \mid s_k) &= P(s_k = 1)H(C \mid s_k = 1) \\ &\quad + P(s_k = 0)H(C \mid s_k = 0). \end{aligned}$$

- $\chi_{\text{set-balance}}$ ensures the remaining conditions are as evenly split as possible.
- $\gamma \in [0, 1]$ balances the two terms (default 0.5).

Utility Function. The utility of each symptom query is further adjusted to account for α , ϵ , and β :

$$\begin{aligned} U(s_k) &= \gamma(\chi_{\text{info-gain}} + \lambda_\alpha P_\alpha - \lambda_\epsilon P_\epsilon - \lambda_\beta P_\beta) \\ &\quad + (1 - \gamma)\chi_{\text{set-balance}}, \end{aligned}$$

where $P_{\text{metric}} = P(\text{metric} \mid s_k)$ and λ_{metric} weight each term (default 0.33).

Conceptual Role of PIE & Assumptions: It is important to note that the PIE formulation and the associated utility functions serve as a conceptual framework modeling ideal diagnostic reasoning, rather than an explicit algorithm computed directly by the LLMs. The inverse prompt functionally constrains the model to emulate this active exclusion policy

Convergence Requirement. The process continues until $|R(Q)| = 1$, i.e.,

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

While real-world clinical presentations of SB frequently involve comorbidities (multiple coexisting conditions), this strict single-condition termination rule is employed exclusively as a simplifying abstraction for benchmarking consistency. In real-world clinical deployments, this constraint would be relaxed to output a ranked probability vector of conditions

Worked Conversational Example: To map the PIE framework operationally to the model’s behavior:

1. **Initial State:** The prompt primes the LLM with SB, Brain Tumor, and Polio ($|R(Q)| = 3$)
2. **Input:** Patient reports back pain. The LLM recognizes this exists across multiple conditions.
3. **Query Strategy:** To maximize information gain and split the remaining conditions, the LLM actively queries: ‘Do you have headaches?’ (indexing Brain Tumor).
4. **Update:** The patient denies headaches $s_k = 0$. The LLM excludes Brain Tumor, shrinking the diagnostic space to SB, Polio, and proceeds to the next query.

2.2. Constructing the Inverse Prompt

Rather than a fundamentally novel algorithmic architecture, *Inverse prompting* operates as a highly structured, constraint-based, knowledge-augmented system prompt. It differs from standard few-shot prompting, which relies on input-output exemplars, by explicitly encoding a symptom inclusion-exclusion constraint that restructures the LLM’s query strategy at inference time [Brown et al. \(2020\)](#). Furthermore, by embedding non-SB synthetic conditions (e.g., Brain Tumor, Polio) as explicit foils, IP forces the LLM into an active differential diagnosis task. Because the models must distinguish SB from these plausible alternatives, these foils act as internal negative controls, preventing the models from succeeding by simply defaulting to an SB confirmation bias.

Let SB be represented by a set of symptoms $S = \{s_1, s_2, \dots, s_n\}$. Furthermore, consider conditions C_1, C_2, \dots, C_N each represented by its own set of unique symptoms. A composite synthetic condition, F , is then constructed by choosing a $K \in \mathbb{Z}^+$ and randomly sampling K symptoms from the conditions S, C_1, C_2, \dots, C_N . Further, F_S is defined as the subset of symptoms from S included in F , and F_{C_i} is the subset of symptoms included in C_i . Then,

$$\begin{aligned} F &= F_S \cup F_{C_1} \cup F_{C_2} \cup \dots \cup F_{C_N}, \\ F_S &\subseteq S, \quad F_{C_i} \subseteq C_i, \\ |F_S| + \sum_{i=1}^N |F_{C_i}| &= K. \end{aligned}$$

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 A](#)).

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 the patient presents 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](#). Fifty participants were surveyed through

online outreach in SB patient communities like support groups (see [Appendix C](#)). 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.

Here is the completed table with the confidence intervals inserted into the appropriate cells while preserving the original structure and style.

3. Experimental Setup

The efficacy of eight FMs (see [Table 1](#)) was evaluated in the specific context of diagnosing complications related to SB. The ability of these models to provide appropriate recommendations was then scrutinized, 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 FM¹. This includes both common and obscure scenarios to test the models’ range of knowledge and adaptability. Information about the participants themselves was not saved.

To establish a rigorous ground truth, each of the 50 diagnostic scenarios was derived directly from the diagnoses and symptom interpretations documented in the actual clinical notes of the patients’ healthcare providers. These scenarios were validated by clinical experts to ensure authenticity, spanning the following distribution: urinary tract infections (9), neurogenic bladder (8), tethered cord syndrome (8), hydrocephalus (8), Chiari II malformation (7), pressure injuries (6), and CSF leaks (4). All 2,000 model interactions were independently scored by two clinical annotators blinded to the prompting method. Inter-rater reliability was strong (Cohen’s $\kappa = 0.81$). Disagreements were adjudicated by a third clinical reviewer, with final scores determined by majority vote (or 0.5 when all three reviewers differed). Additionally, because real patient conversations are frequently unstructured, we conducted a robustness analysis on ‘messy’ dialogues—characterized by non-sequential reporting and symptom noise (detailed in [Appendix F](#)). Under these conditions, IP demonstrated high resilience with only a marginal drop in accuracy.

1. Find summarized qualitative results in [Section 3.2](#).

Table 2: Coarse-level diagnostic performance across different prompting strategies. For each model, the performance of a no-prompt baseline (O), a standard system prompt (X), Dr-CoT (DC), and our inverse prompt (I) was compared. Metrics shown are diagnostic accuracy (α), error rate (ϵ), bridging rate (β), and mean conversation length (\bar{L}). Metrics are reported as means \pm 95% bootstrap confidence intervals.

Model	Condition	α	ϵ	β	\bar{L}
GPT-5.1	Baseline (O)	0.752 [0.695, 0.801]	0.311 [0.257, 0.371]	0.351 [0.294, 0.412]	6.22 [6.02, 6.42]
	Standard (X)	0.803 \uparrow .05 [0.749, 0.848]	0.304 \downarrow .01 [0.250, 0.364]	0.316 \downarrow .04 [0.262, 0.376]	6.05 \downarrow .17 [5.87, 6.23]
	DR-CoT (DC)	0.848 \uparrow .10 [0.798, 0.887]	0.210 \downarrow .10 [0.164, 0.265]	0.277 \downarrow .07 [0.225, 0.336]	5.55 \downarrow .67 [5.39, 5.71]
	Inverse (I)	0.886 \uparrow .13 [0.841, 0.920]	0.162 \downarrow .15 [0.122, 0.213]	0.238 \downarrow .11 [0.189, 0.295]	5.15 \downarrow 1.07 [5.00, 5.30]
GPT-5.2	Baseline (O)	0.738 [0.680, 0.789]	0.336 [0.280, 0.397]	0.389 [0.331, 0.451]	6.94 [6.74, 7.14]
	Standard (X)	0.789 \uparrow .05 [0.734, 0.835]	0.328 \downarrow .01 [0.273, 0.388]	0.354 \downarrow .04 [0.297, 0.415]	6.82 \downarrow .12 [6.64, 7.00]
	DR-CoT (DC)	0.835 \uparrow .10 [0.784, 0.876]	0.195 \downarrow .14 [0.151, 0.249]	0.265 \downarrow .12 [0.214, 0.323]	5.75 \downarrow 1.19 [5.59, 5.91]
	Inverse (I)	0.845 \uparrow .11 [0.795, 0.885]	0.170 \downarrow .17 [0.129, 0.221]	0.226 \downarrow .16 [0.179, 0.282]	5.47 \downarrow 1.47 [5.32, 5.62]
Claude-4.5 Sonnet	Baseline (O)	0.744 [0.686, 0.794]	0.289 [0.236, 0.348]	0.363 [0.306, 0.424]	5.85 [5.65, 6.05]
	Standard (X)	0.792 \uparrow .05 [0.737, 0.838]	0.267 \downarrow .02 [0.216, 0.325]	0.355 \downarrow .01 [0.298, 0.416]	5.73 \downarrow .12 [5.55, 5.91]
	DR-CoT (DC)	0.824 \uparrow .08 [0.772, 0.866]	0.150 \downarrow .14 [0.111, 0.200]	0.230 \downarrow .13 [0.182, 0.286]	5.23 \downarrow .62 [5.07, 5.39]
	Inverse (I)	0.853 \uparrow .11 [0.804, 0.892]	0.099 \downarrow .19 [0.068, 0.142]	0.193 \downarrow .17 [0.149, 0.246]	4.98 \downarrow .87 [4.83, 5.13]
Gemini 3	Baseline (O)	0.720 [0.661, 0.772]	0.401 [0.342, 0.463]	0.447 [0.387, 0.509]	7.60 [7.40, 7.80]
	Standard (X)	0.753 \uparrow .03 [0.696, 0.802]	0.396 \downarrow .01 [0.337, 0.458]	0.439 \downarrow .01 [0.379, 0.501]	7.52 \downarrow .08 [7.34, 7.70]
	DR-CoT (DC)	0.790 \uparrow .07 [0.735, 0.836]	0.290 \downarrow .11 [0.237, 0.349]	0.340 \downarrow .11 [0.284, 0.401]	7.00 \downarrow .60 [6.84, 7.16]
	Inverse (I)	0.812 \uparrow .09 [0.759, 0.856]	0.235 \downarrow .17 [0.187, 0.291]	0.295 \downarrow .15 [0.242, 0.354]	6.86 \downarrow .74 [6.71, 7.01]
Mistral Large 3	Baseline (O)	0.696 [0.636, 0.750]	0.357 [0.300, 0.418]	0.409 [0.350, 0.471]	7.28 [7.08, 7.48]
	Standard (X)	0.722 \uparrow .03 [0.663, 0.774]	0.350 \downarrow .01 [0.294, 0.411]	0.387 \downarrow .02 [0.329, 0.449]	7.17 \downarrow .11 [6.99, 7.35]
	DR-CoT (DC)	0.760 \uparrow .06 [0.703, 0.809]	0.300 \downarrow .06 [0.247, 0.359]	0.350 \downarrow .06 [0.294, 0.411]	6.70 \downarrow .58 [6.54, 6.86]
	Inverse (I)	0.782 \uparrow .09 [0.727, 0.829]	0.275 \downarrow .08 [0.223, 0.333]	0.259 \downarrow .15 [0.209, 0.317]	6.58 \downarrow .70 [6.43, 6.73]
Mixtral 8 \times 22B	Baseline (O)	0.722 [0.663, 0.774]	0.383 [0.325, 0.445]	0.377 [0.319, 0.439]	6.82 [6.62, 7.02]
	Standard (X)	0.758 \uparrow .04 [0.701, 0.807]	0.379 \downarrow .00 [0.321, 0.441]	0.371 \downarrow .01 [0.313, 0.432]	6.79 \downarrow .03 [6.61, 6.97]
	DR-CoT (DC)	0.778 \uparrow .06 [0.722, 0.825]	0.310 \downarrow .07 [0.256, 0.370]	0.320 \downarrow .06 [0.265, 0.380]	6.55 \downarrow .27 [6.39, 6.71]
	Inverse (I)	0.828 \uparrow .11 [0.776, 0.870]	0.304 \downarrow .08 [0.250, 0.364]	0.274 \downarrow .10 [0.222, 0.332]	5.63 \downarrow 1.19 [5.48, 5.78]

Existing medical and biomedical FMs (Moor et al., 2023; Luo et al., 2022), while trained on specialized medical data, were not utilized due to their limited accessibility and lack of commercial availability. Instead, the focus of this study was to evaluate widely accessible and available LLMs, ensuring relevance to a broader patient population. Furthermore, ground-truth for these evaluations was established using the diagnoses and symptom interpretations provided in the clinical notes of users’ doctors, grounding the

study in a specific condition, Spina Bifida, and a targeted clinical task, diagnostic reasoning.

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 inquire 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., 2022) and tree-of-thought (ToT) (Yao et al., 2023) and by deploying bridge² (Li et al., 2022) and inverse prompting. More concretely, *diagnostic accuracy* is defined as a weighted metric that accounts for fully and partially correct impressions (defined in Appendix D) 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.

3.1. Quantitative Experimental Results

In Table 2, it is observed 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 prompt³. 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 the 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}).

2. Bridging to SB care was adapted 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.

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

The introduction of Diagnostic-Reasoning Chain of Thought (DR-CoT) (Wu et al., 2023) adds another layer of performance comparison. It improves diagnostic accuracy by explicitly prompting LLMs to summarize evidence, formulate ranked differential diagnoses, and iteratively refine inquiries based on the evidence gathered. While DR-CoT performs well and enhances diagnostic reasoning, the inverse prompt still achieves the best overall performance across models.

The fine-grained performance of the models is detailed in Table 3. The evaluated set of symptoms, S , corresponds to: s_1 : CSF leak, s_2 : neurogenic bladder, s_3 : TCS, s_4 : hydrocephalus, s_5 : Chiari II malformation, s_6 : pressure injuries, and s_7 : urinary tract infection.

These findings collectively reinforce the effectiveness of inverse prompting in clinical applications and its ability to surpass both traditional internet searches and other reasoning methodologies for LLMs in delivering accurate, patient-centered outcomes (Figure 3).

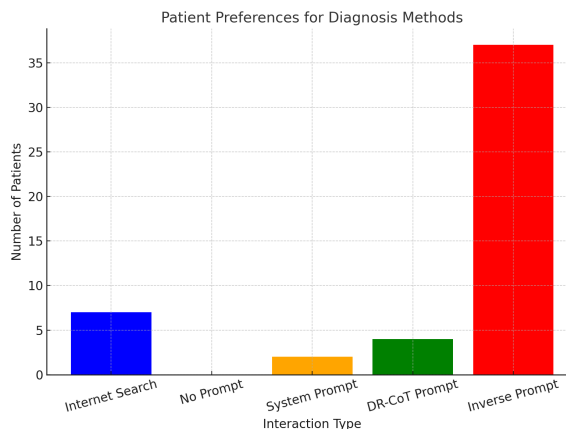


Figure 3: Surveyed patient preferences on diagnostic methods, $n = 50$.

3.2. Qualitative Patient and Model Feedback

GPT-5.2. 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 it venture beyond the information from the inverse

Table 3: Fine-grained performance. Best performance for each metric is shown in *italics*. Models are columns: **G51** (GPT-5.1), **G52** (GPT-5.2), **C4S** (Claude-4.5 Sonnet), **G4o** (GPT-4o), **L4** (Llama 4), **M8x** (Mixtral 8x22B), **ML3** (Mistral Large 3), **G3** (Gemini 3). Metrics are reported as means \pm 95% bootstrap confidence intervals.

Symptom	Metric	G51	G52	C4S	G4o	L4	M8x	ML3	G3
CSF Leak (s_1)	α	.631 [0.415, 0.805]	.656 [0.439, 0.823]	.705 [0.486, 0.858]	.622 [0.407, 0.798]	.542 [0.335, 0.735]	.633 [0.417, 0.806]	.655 [0.438, 0.822]	.534 [0.328, 0.729]
	ϵ	.277 [0.130, 0.496]	.338 [0.172, 0.556]	.231 [0.100, 0.449]	.354 [0.184, 0.571]	.431 [0.243, 0.641]	.308 [0.151, 0.527]	.200 [0.081, 0.416]	.523 [0.319, 0.720]
		.778 [0.628, 0.879]	.806 [0.659, 0.899]	.818 [0.673, 0.908]	.758 [0.607, 0.864]	.577 [0.424, 0.717]	.792 [0.644, 0.889]	.486 [0.339, 0.635]	.596 [0.442, 0.733]
Neuro. Bladder (s_2)	α	.360 [0.230, 0.515]	.383 [0.249, 0.538]	.286 [0.170, 0.440]	.349 [0.221, 0.504]	.514 [0.365, 0.661]	.274 [0.160, 0.427]	.503 [0.355, 0.651]	.509 [0.360, 0.656]
	ϵ	.692 [0.537, 0.813]	.721 [0.567, 0.836]	.740 [0.587, 0.851]	.701 [0.547, 0.820]	.601 [0.447, 0.737]	.651 [0.496, 0.779]	.732 [0.579, 0.844]	.638 [0.483, 0.769]
		.359 [0.229, 0.514]	.400 [0.263, 0.554]	.297 [0.178, 0.451]	.331 [0.206, 0.486]	.483 [0.337, 0.632]	.393 [0.257, 0.547]	.303 [0.183, 0.457]	.462 [0.318, 0.613]
Hydrocephalus (s_4)	α	.761 [0.610, 0.866]	.805 [0.658, 0.899]	.816 [0.670, 0.906]	.750 [0.598, 0.858]	.742 [0.590, 0.852]	.742 [0.590, 0.852]	.693 [0.539, 0.814]	.711 [0.557, 0.828]
	ϵ	.120 [0.052, 0.255]	.160 [0.077, 0.302]	.107 [0.044, 0.239]	.133 [0.060, 0.271]	.200 [0.105, 0.348]	.267 [0.155, 0.420]	.240 [0.134, 0.391]	.267 [0.155, 0.420]
		.688 [0.523, 0.816]	.739 [0.575, 0.855]	.756 [0.593, 0.868]	.692 [0.527, 0.819]	.606 [0.441, 0.750]	.739 [0.575, 0.855]	.499 [0.342, 0.656]	.668 [0.502, 0.800]
Chiari II (s_5)	α	.300 [0.174, 0.465]	.318 [0.188, 0.484]	.182 [0.088, 0.339]	.209 [0.107, 0.369]	.245 [0.133, 0.408]	.273 [0.154, 0.437]	.300 [0.174, 0.465]	.182 [0.088, 0.339]
	ϵ	.876 [0.714, 0.952]	.899 [0.743, 0.965]	.921 [0.771, 0.976]	.863 [0.699, 0.945]	.702 [0.523, 0.835]	.781 [0.606, 0.892]	.732 [0.554, 0.857]	.795 [0.621, 0.902]
		.122 [0.047, 0.283]	.156 [0.067, 0.323]	.056 [0.014, 0.199]	.044 [0.009, 0.182]	.222 [0.110, 0.397]	.156 [0.067, 0.323]	.178 [0.081, 0.349]	.200 [0.095, 0.373]
UTI (s_7)	α	.991 [0.906, 0.999]	.994 [0.911, 1.000]	.989 [0.920, 1.000]	.969 [0.871, 0.993]	.873 [0.746, 0.942]	.873 [0.746, 0.942]	.902 [0.781, 0.960]	.888 [0.764, 0.951]
	ϵ	.005 [0.000, 0.088]	.015 [0.002, 0.104]	.000 [0.000, 0.079]	.010 [0.001, 0.096]	.088 [0.035, 0.206]	.078 [0.029, 0.193]	.049 [0.014, 0.155]	.083 [0.032, 0.200]

prompt, yet it was competent in proceeding with a systematic approach to questioning a patient.

GPT-5.1. 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-5.1 had the best sequential inference ability of the eight FMs.

GPT-4o. 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 3. 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-4.5 Sonnet. Similar to GPT-5.1, it adopts the PIE formulation resulting in progressive reasoning with usually a better diagnostic accuracy. However, in some cases, it tends to be more limited in its reasoning due to over-reliance on processing eliminations. This leads to abrupt ends in conversations and arriving to conclusions prematurely.

Llama 4. 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 8x22B. 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 3. It tended 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.

In the experiments, several shortcomings with the models were identified 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 as complex as SB has a combination of urological, neurological, and musculoskeletal implications. Thus this study proposes a multistage architecture suitable for diagnostic tasks. Figure 5 in Appendix G lays out the architecture of our proposed approach that integrates into the mainstream conversation between the patient and the models.

3.3. Conclusion & Future Work

Eight LLMs were holistically evaluated to assess their ability to diagnose SB, introducing inverse prompting to guide step-wise reasoning. The 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. The approach and proposed framework, based on experimental insights, addresses current LLM shortcomings and aims to improve healthcare for SB patients through:

Curating Specialized Datasets. Refining training data by integrating diverse samples, including records from the National Spina Bifida Patient Registry (Thibadeau, 2017), clinical notes, and preprocessed literature, to enhance diagnostic reliability.

Patient Cohorts & Solving Strategies. Involving larger patient cohorts and using strategies like discrete querying agents and Socratic prompts (Chang, 2023) to improve model performance.

End-to-End Implementation & Validation. Future work will assess the model using metrics like ROUGE (Lin, 2004) and inter-rater 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. To address the limitations identified in this benchmark, we propose a conceptual multi-stage architecture utilizing a vision-language planner and structured symptom database (detailed in Appendix G and Figure 5). Notably, while Figure 5 represents a conceptual framework, we present it as a sub-contribution, to address the limitations patients encountered with ex-

isting LLMs during our experiments. The implementation and validation of this will be pursued as part of future work, with the goal of establishing a more robust/diagnostically reliable system that overcomes the constraints of current language models. Furthermore, future large-scale evaluations will incorporate true non-SB negative control cases to rigorously validate the models’ discriminative capabilities.

Limitations. Our study’s limitations include its relatively borderline sample size (50 participants) and focus on specific SB-related symptoms, leaving room to explore our framework’s utility for other conditions (which would potentially yield similar results) at a larger scale. 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.

LLM Usage

No generative AI systems were used to write or edit this manuscript. Large language models were research subjects only; all analyses and writing were performed by the authors.

Ethical Statement

Misdiagnosis and Privacy. 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.

Applied Deployment. Our proposed system could also complement human clinicians, serving as a decision-support tool rather than a standalone diagnostic system. It integrates a reasoning/thought field in its output, ensuring transparency by documenting the rationale behind its responses. This enables patients using a deployed version (e.g., in a mobile health app) to share conversations with their primary care physician, who can review the reasoning of the AI. Deployment challenges include aligning with clinical workflows, maintaining up-to-date medical knowledge, and addressing the complexities of interdisciplinary care. By prioritizing transparency, safety, and ethical standards, the system aims to enhance, not replace, clinical expertise.

Mental Health. In addition to its diagnostic potential, the system can offer significant mental health benefits. Many patients feel hesitant or uncomfortable discussing sensitive medical concerns with healthcare providers, which can lead to delays in seeking care or incomplete symptom reporting. By providing a safe, non-judgmental environment for patients to articulate their symptoms and concerns, the system can act as a “training ground.” This enables patients to rehearse difficult conversations, gain clarity about their symptoms, and build confidence before consulting a healthcare professional. Such interactions can foster more informed and productive patient-clinician dialogues, potentially alleviating anxiety associated with medical consultations.

Acknowledgments

The authors have no individuals or organizations to acknowledge for this work. The authors received no financial support for the research, authorship, or publication of this study. The authors declare no conflicts of interest or relevant financial disclosures.

Author Contributions

AA led the project and contributed to conceptualization/study design, methodology, software/model development, formal analysis, data interpretation, and manuscript drafting/revision. **SM** contributed to study design, methodology, data interpretation, and manuscript drafting/revision. **ZK** contributed to study design, literature review, clinical interpretation, data interpretation, and manuscript drafting/revision. **NS** contributed to software/model development, data processing, analysis/visualizations,

and manuscript revision. **CL** and **BD** supervised the project and contributed to methodological ideation, critical revision of the manuscript, general oversight, and clinical interpretation.

References

- 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. doi: 10.1073/pnas.2106844118.
- R Ajitha and N Punitha. Active contour-based segmentation of normal and fetal spina bifida ultrasound images. In *Journal of Physics: Conference Series*, volume 2318, page 012045. IOP Publishing, aug 2022. doi: 10.1088/1742-6596/2318/1/012045. URL <https://dx.doi.org/10.1088/1742-6596/2318/1/012045>.
- Anthropic. Model card and evaluations for Claude models. Technical report, Anthropic, 2023.
- Virginia G. Briggs. Population estimates of people with spina bifida in the united states in 2020. *medRxiv*, 2022. doi: 10.1101/2022.02.03.22268639. URL <https://www.medrxiv.org/content/10.1101/2022.02.03.22268639v1>. preprint.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Edward Y. Chang. Prompting large language models with the Socratic method, 2023. URL <https://arxiv.org/abs/2303.08769>.
- 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. doi: 10.1155/2021/3605739.

- Brad E Dicianno, Brad G Kurowski, Jennifer Marie J Yang, Michael B Chancellor, Ghassan K Bejjani, Andrea D Fairman, Nancy Lewis, and Jennifer Sotirake. Rehabilitation and medical management of the adult with spina bifida. *American Journal of Physical Medicine & Rehabilitation*, 87(12):1027–1050, 2008.
- Brad E Dicianno, Bambang Parmanto, Andrea D Fairman, Theresa M Crytzer, Daihua X Yu, Gede Pramana, Derek Coughenour, and Alan A Petrazzi. Perspectives on the evolution of mobile (mhealth) technologies and application to rehabilitation. *Physical Therapy*, 95(3):397–405, 2015.
- Semra Duran, Kemal Üreten, Yüksel Maraş, Hadi Maraş, Kevser Gök, Ebru Atalar, and Velihan Çayhan. Automatic detection of spina bifida occulta with deep learning methods from plain pelvic radiographs. *Research on Biomedical Engineering*, 39:655–661, July 2023. doi: 10.1007/s42600-023-00296-6.
- Gemini Team. Gemini: A family of highly capable multimodal models, 2023. URL <https://arxiv.org/abs/2312.11805>.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- 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, Léo 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. URL <https://arxiv.org/abs/2401.04088>.
- Kadhir Velu Karthik, Aruna Rajalingam, Malliah Shivashankar, and Anjali Ganjiwale. Recursive feature elimination-based biomarker identification for open neural tube defects. *Current Genomics*, 23(3):195–206, 2022. doi: 10.2174/1389202923666220511162038.
- Umut Konur, Fikret S. Gürgen, Füsün 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, 2015. doi: 10.1016/j.knosys.2015.04.021.
- 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, 2020. URL <https://arxiv.org/abs/2005.11401>.
- 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. URL <https://arxiv.org/abs/2203.14104>.
- 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. URL <https://aclanthology.org/W04-1013>.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022. doi: 10.1093/bib/bbac409.
- 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. doi: 10.1016/j.apmr.2020.06.016.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkas, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium (ML4H)*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 2023. URL <https://proceedings.mlr.press/v225/moor23a.html>.

- OpenAI. Introducing ChatGPT. <https://openai.com/blog/chatgpt>, nov 2022. Accessed: 2025-08-07.
- OpenAI. GPT-4 technical report, 2023. URL <https://arxiv.org/abs/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. Towards expert-level medical question answering with large language models, 2023. URL <https://arxiv.org/abs/2305.09617>.
- 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.
- Judy Thibadeau. The national spina bifida patient registry: Past, present, and future. *Journal of Pediatric Rehabilitation Medicine*, 10(3-4):205–210, 2017. doi: 10.3233/PRM-170463.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutvi 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. URL <https://arxiv.org/abs/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. Towards conversational diagnostic AI, 2024. URL <https://arxiv.org/abs/2401.05654>.
- John K. Weaver, Madalyne Martin-Olenski, Joseph Logan, et al. Deep learning of videourodynamics to classify bladder dysfunction severity in patients with spina bifida. *The Journal of Urology*, 209(5):994–1003, 2023. doi: 10.1097/JU.0000000000003267.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022. URL <https://arxiv.org/abs/2201.11903>.
- Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. Large language models perform diagnostic reasoning, 2023. URL <https://arxiv.org/abs/2307.08922>.
- 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. URL <https://arxiv.org/abs/2305.10601>.
- Daihua Yu, Bambang Parmanto, and Brad Dicianno. An mhealth app for users with dexterity impairments: Accessibility study. *JMIR mHealth and uHealth*, 7(1):e202, 2019. doi: 10.2196/9931.
- 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. URL <https://arxiv.org/abs/2304.14554>.

Appendix A. Standard System Prompt vs. Inverse Prompt

The following detail both prompting methods.

A.1. 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.”

A.2. Inverse System Prompt

You will play the role of an AI general physician for a research experiment. You specialize 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.

SPINA 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 impairment: 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 your 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 B. Added Context for FMs in Table 1

1. **Gemini 3** 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. **Mistral 8x22B** 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 3** 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-4.5 Sonnet** Back pain, no numbness nor weakness in legs or feet, recent uptick in physical activity, increased supplement intake, urinary symptoms.
5. **Llama 4** 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** 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-5.1** Back pain, decreased appetite and focus. Pain alleviated by lying down, urinary symptoms, no weakness in legs. Bent toes.
8. **GPT-5.2** 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 C. Participant Recruitment Instructions and Demographic Information

Patients in Management Stage. Participants were recruited from a social media support group focused on the condition, resulting in a cohort that was relatively tech-savvy, informed, and articulate. It is acknowledged that this group may not be fully representative of a broader patient population, as many participants were in the management stage of their condition and less likely to digress from the main context of the conversation. Consequently, the "messy" nature of real-world patient interactions was limited, allowing us to gain valuable insights and benchmark existing models in a more controlled setting. As noted in the main text, the inverse prompt embeds non-SB conditions as foils, serving as an internal negative control.

Patient Symptom Diversity. Our cohort includes a diverse range of Spina Bifida cases covering various types of meningocele. Severity levels range from mild mobility impairments and isolated neurogenic bladder to complex presentations involving Chiari II malformation, tethered cord syndrome, and hydrocephalus. This breadth allowed us to evaluate the models across both common and rare SB-related

complications, as demonstrated in the symptom-level experiments in Table 3.

Patient Background. The dialogues were collected from adult patients representing diverse sub-demographics, including varying socioeconomic backgrounds, racial groups, and access to care. While the participants were not currently active or hospitalized patients (this was not the focus of the study), the scenarios were validated by clinical experts to ensure authenticity and representativeness.

Circumventing Negative Controls. Non-SB cases were not included, as the primary objective of this user study was to evaluate whether LLMs could perform accurate differential diagnoses within the SB domain. To address this, the inverse prompting method itself incorporated synthetic conditions and non-SB differential diagnoses as part of the prompts, serving as internal controls for evaluation.

Accessibility. Our recruitment did not include elderly individuals or those with significant cognitive impairments that can sometimes accompany SB. It is acknowledged that working with these populations presents unique challenges in gathering accurate symptom information. Future work could explore adaptations to better accommodate such populations, such as integrating multi-modal input systems or modules (e.g., voice/visual cues) or by designing more accessible and patient-friendly interfaces.

Participant Instructions. 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 <https://forms.gle/fi58JU9f2J2V22cz8>. Exact instructions visualized below in Figure 4

Appendix D. Definitions

Impression. In the clinical context, an impression refers to a concise interpretation or summary of observed findings, often provided by a healthcare professional or diagnostic tool. It synthesizes available data, such as imaging results, patient history, or physical examination findings, to propose a potential diagnosis, conclusion, or recommendation for further management. For instance, in radiology, an

Table 4: Nemenyi test results for β and \bar{L} values. **Bold** refers to a statistically significant difference.

		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

impression might read: “Spinal dysraphism noted at the L4-L5 level with a myelomeningocele sac; clinical symptoms are consistent with tethered cord syndrome. A surgical consultation is advised to evaluate treatment options.” This serves as a guiding statement for clinical decision-making and patient care.

Appendix E. Statistical Testing for Results in Table 2

Nemenyi test results from Table 4 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.

Bonferroni-Dunn test results in Table 5 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.

Concerning the Wilcoxon p -values in Table 6, our results vary dramatically depending on the performance metric used, highlighting that a model’s relative performance is highly task-dependent. For α , GPT-5.2 is the primary standout, showing a statis-

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 you're 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 4: Instructions shown to participants.

Table 5: Bonferroni-Dunn test results for β and \bar{L} values. **Bold** refers to a statistically significant difference.

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

tically significant performance difference when compared to nearly every other LLM. Conversely, ϵ reveals a greater number of significant differences between the various models; however, Mistral Large 3 shows no statistically significant performance difference against any competitor, suggesting its capabilities on this specific task are statistically indistinguishable from the rest of the field.

Examining the Friedman test results in Table 7 reveal statistically significant differences globally for all models and symptoms using all metrics.

Appendix F. Performance Change for Messy Dialogues

The following criteria are proposed identifying messy dialogues within our set of conversations.

1. Temporal and Clinical Coherence:

- Non-sequential reporting mixing current symptoms with historical medical events (e.g., discussing current back pain, then mentioning a surgery from years ago, then returning to current symptoms)
- Discrepancies in symptom descriptions across the conversation (e.g., describing pain as “constant” in one message and “intermittent” in another)
- Interleaving of medical symptoms with contextual life details (e.g., describing bladder symptoms alongside details about work schedule or daily routine)
- Usage of non-standardized terminology for medical conditions (e.g., describing neurological symptoms as “feeling off” or “not quite right”)

2. Conversation Length and Complexity:

- Dialogues exceeding the mean conversation length or involving numerous prompt-response cycles
- Presence of looping or repetitive patterns without resolution

3. Content Alignment and Response Quality:

- Misinterpretation of patient inputs or provision of irrelevant/incorrect information
- Abrupt topic shifts or responses misaligned with patient queries
- Frequent requests for clarification or revision of the LLM’s responses

4. Diagnostic Reasoning Quality:

- Ambiguous or overly verbose reasoning in the LLM’s thought process (from JSON outputs)
- Patient expressions of confusion or dissatisfaction with diagnostic reasoning (e.g., “*This isn’t helpful*”)

5. Error Metrics:

- Conversations with significantly higher proportions of bridging prompts compared to the cohort average
- Multiple corrections or clarifications needed within the same diagnostic thread

Based on the criteria, the messy prompts are segregated from the general conversations and test the performance of the LLMs for the different settings (no prompt, system prompt, DR-CoT prompt, and inverse prompt). As shown in Table 8, it should be noted that even for more messy and noisier conversation/user-prompts, there is only a slight performance drop, validating the robust nature of our inverse prompt method when comparing values with Table 2.

Appendix G. Proposed Multi-Stage 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

Table 6: Wilcoxon p -values for both ϵ (below the diagonal) and α (above the diagonal) metrics across different LLMs. **Bold** refers to a statistically significant difference.

(ϵ, α)	GPT-5.1	GPT-5.2	Claude-4.5 Sonnet	Gemini 3
GPT-5.1	-	0.016	0.016	0.016
GPT-5.2	0.016	-	0.016	0.016
Claude-4.5 Sonnet	0.016	0.016	-	0.016
Gemini 3	0.109	0.219	0.028	-
Mistral Large 3	0.400	0.938	0.047	0.156
Mixtral 8x22B	0.297	0.753	0.031	0.173
GPT-4o	0.469	0.047	0.047	0.031
Llama 4	0.031	0.109	0.016	0.938
(ϵ, α)	Mistral Large 3	Mixtral 8x22B	GPT-4o	Llama 4
GPT-5.1	0.078	0.375	0.109	0.016
GPT-5.2	0.047	0.028	0.016	0.016
Claude-4.5 Sonnet	0.016	0.016	0.016	0.016
Gemini 3	0.688	0.156	0.016	0.156
Mistral Large 3	-	0.375	0.078	0.812
Mixtral 8x22B	0.578	-	0.469	0.043
GPT-4o	0.297	0.219	-	0.016
Llama 4	0.375	0.219	0.016	-

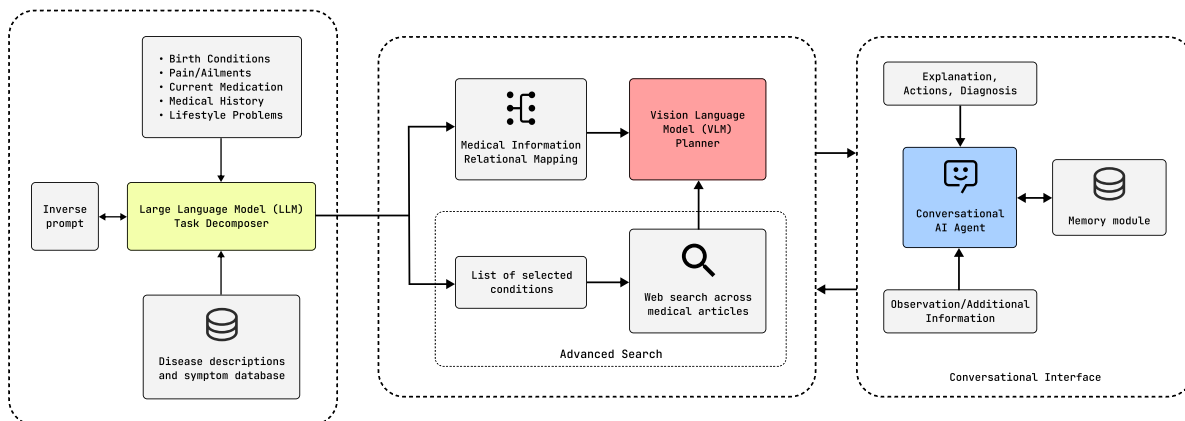


Figure 5: Architectural overview of the proposed system

Table 7: Friedman test results for both diagnostic metrics, α and ϵ , and both conversational metrics, β and \bar{L} , across all models and symptoms.

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}

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.

Table 8: “Messy” dialogue SB diagnostic performance of LLMs. Table interpretation details follow Table 2.

Model	α_O	ϵ_O	β_O	L_O	α_X	ϵ_X	β_X	L_X
GPT-5.1	0.733 [0.675, 0.784]	0.304 [0.250, 0.364]	0.344 [0.288, 0.405]	6.08 [5.86, 6.30]	0.779 \uparrow 0.05 [0.724, 0.826]	0.296 \downarrow 0.01 [0.243, 0.355]	0.307 \downarrow 0.04 [0.253, 0.367]	5.89 \downarrow 0.19 [5.69, 6.09]
GPT-5.2	0.715 [0.656, 0.767]	0.326 [0.271, 0.386]	0.373 [0.315, 0.434]	6.61 [6.39, 6.83]	0.769 \uparrow 0.05 [0.713, 0.817]	0.312 \downarrow 0.01 [0.258, 0.372]	0.339 \downarrow 0.03 [0.283, 0.400]	6.55 \downarrow 0.06 [6.35, 6.75]
Claude-4.5 Sonnet	0.723 [0.664, 0.775]	0.283 [0.231, 0.342]	0.348 [0.292, 0.409]	5.63 [5.41, 5.85]	0.769 \uparrow 0.05 [0.713, 0.817]	0.256 \downarrow 0.03 [0.206, 0.314]	0.342 \downarrow 0.01 [0.286, 0.403]	5.47 \downarrow 0.16 [5.27, 5.67]
Gemini 3	0.690 [0.630, 0.744]	0.393 [0.335, 0.455]	0.428 [0.368, 0.490]	7.40 [7.18, 7.62]	0.722 \uparrow 0.03 [0.663, 0.774]	0.378 \downarrow 0.02 [0.320, 0.440]	0.426 \downarrow 0.00 [0.366, 0.488]	7.26 \downarrow 0.14 [7.06, 7.46]
Mistral Large 3	0.680 [0.620, 0.735]	0.339 [0.283, 0.400]	0.393 [0.335, 0.455]	7.01 [6.79, 7.23]	0.706 \uparrow 0.03 [0.647, 0.759]	0.335 \downarrow 0.00 [0.279, 0.396]	0.377 \downarrow 0.02 [0.319, 0.439]	6.90 \downarrow 0.11 [6.70, 7.10]
Mixtral 8 \times 22B	0.691 [0.631, 0.745]	0.365 [0.308, 0.426]	0.367 [0.310, 0.428]	6.64 [6.42, 6.86]	0.736 \uparrow 0.05 [0.678, 0.787]	0.366 \uparrow 0.00 [0.309, 0.427]	0.353 \downarrow 0.01 [0.296, 0.414]	6.52 \downarrow 0.12 [6.32, 6.72]
Mean	0.705	0.337	0.371	6.50	0.737 \uparrow 0.03	0.324 \downarrow 0.01	0.354 \downarrow 0.02	6.48 \downarrow 0.02
Model	α_{DC}	ϵ_{DC}	β_{DC}	L_{DC}	α_I	ϵ_I	β_I	L_I
GPT-5.1	0.821 \uparrow 0.09 [0.769, 0.864]	0.201 \downarrow 0.10 [0.156, 0.255]	0.266 \downarrow 0.08 [0.215, 0.324]	5.31 \downarrow 0.77 [5.12, 5.50]	0.861 \uparrow 0.13 [0.813, 0.898]	0.154 \downarrow 0.15 [0.115, 0.204]	0.231 \downarrow 0.11 [0.183, 0.287]	5.01 \downarrow 1.07 [4.83, 5.19]
GPT-5.2	0.801 \uparrow 0.09 [0.747, 0.846]	0.191 \downarrow 0.14 [0.147, 0.244]	0.252 \downarrow 0.12 [0.202, 0.309]	5.47 \downarrow 1.14 [5.28, 5.66]	0.808 \uparrow 0.09 [0.755, 0.852]	0.166 \downarrow 0.16 [0.125, 0.217]	0.216 \downarrow 0.16 [0.169, 0.271]	5.34 \downarrow 1.27 [5.16, 5.52]
Claude-4.5 Sonnet	0.796 \uparrow 0.07 [0.742, 0.841]	0.145 \downarrow 0.14 [0.107, 0.194]	0.225 \downarrow 0.12 [0.178, 0.281]	5.05 \downarrow 0.58 [4.86, 5.24]	0.835 \uparrow 0.11 [0.784, 0.876]	0.097 \downarrow 0.19 [0.066, 0.140]	0.185 \downarrow 0.16 [0.142, 0.238]	4.82 \downarrow 0.81 [4.64, 5.00]
Gemini 3	0.752 \uparrow 0.06 [0.695, 0.801]	0.279 \downarrow 0.11 [0.227, 0.338]	0.325 \downarrow 0.10 [0.270, 0.385]	6.76 \downarrow 0.64 [6.57, 6.95]	0.792 \uparrow 0.10 [0.737, 0.838]	0.223 \downarrow 0.17 [0.176, 0.279]	0.281 \downarrow 0.15 [0.229, 0.340]	6.63 \downarrow 0.77 [6.45, 6.81]
Mistral Large 3	0.730 \uparrow 0.05 [0.672, 0.781]	0.288 \downarrow 0.05 [0.235, 0.347]	0.334 \downarrow 0.06 [0.278, 0.395]	6.44 \downarrow 0.57 [6.25, 6.63]	0.764 \uparrow 0.08 [0.708, 0.812]	0.264 \downarrow 0.08 [0.213, 0.322]	0.251 \downarrow 0.14 [0.201, 0.308]	6.25 \downarrow 0.76 [6.07, 6.43]
Mixtral 8 \times 22B	0.749 \uparrow 0.06 [0.692, 0.799]	0.302 \downarrow 0.06 [0.248, 0.362]	0.305 \downarrow 0.06 [0.251, 0.365]	6.39 \downarrow 0.25 [6.20, 6.58]	0.802 \uparrow 0.11 [0.748, 0.847]	0.289 \downarrow 0.08 [0.236, 0.348]	0.264 \downarrow 0.10 [0.213, 0.322]	5.35 \downarrow 1.29 [5.17, 5.53]
Mean	0.776 \uparrow 0.07	0.235 \downarrow 0.10	0.284 \downarrow 0.09	6.25 \downarrow 0.25	0.794 \uparrow 0.09	0.203 \downarrow 0.13	0.237 \downarrow 0.13	5.41 \downarrow 1.09

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 categories and structures 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.