Towards Conversational AI for Spina Bifida Care

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

1. Carnegie Mellon University, **2.** Stanford University, **3.** Harvard University, **4.** Massachusetts General Hospital, **5.** Asesa Labs

ABSTRACT

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

OBJECTIVES

1. Benchmark eight LLMs through both qualitative and quantitative evaluations of their performance in addressing SB's unique medical challenges.

2. Introduce an inverse prompting technique, guiding LLMs through a structured diagnostic process using a predefined symptom set, ensuring more accurate and stepwise reasoning.

3. Assess the effectiveness of inverse prompting with SB patients, using diagnostic accuracy (α) and error rate (ϵ) as metrics.

4. Propose a novel framework based on the identified limitations of existing LLMs, designed to improve clinical outcomes for SB patients.

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

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

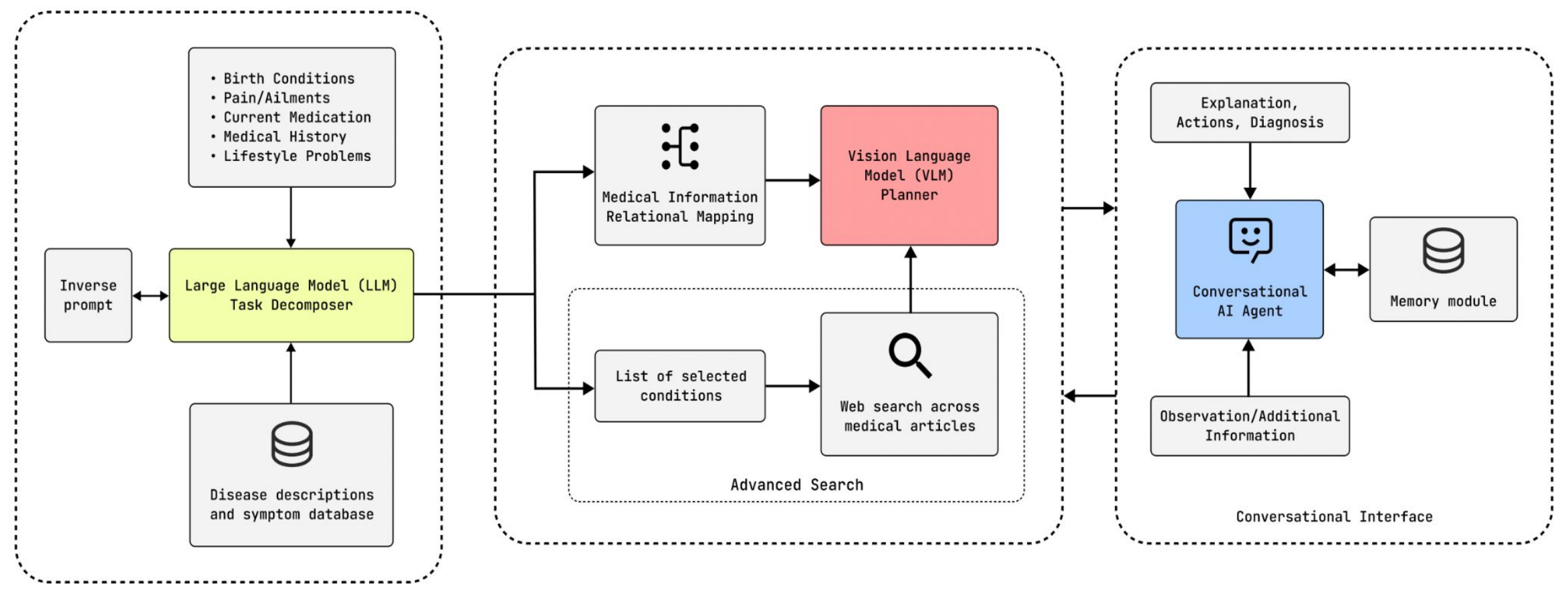
• Birth Conditions
• Pain/Ailments
• Current Medication
• Medical History
• Lifestyle Problems
Inverse prompt
Large Language Model (LLM) Task Decomposer
Disease descriptions and symptom database
Medical Information Relational Mapping
Vision Language Model (VLM) Planner
List of selected conditions
Web search across medical articles
Advanced Search
Explanation, Actions, Diagnosis
Conversational AI Agent
Memory module
Observation/Additional Information
Conversational Interface

Figure 1: Architectural overview of the proposed system

EXPERIMENTS

Evaluation Process:

1. 8 FMs were evaluated for diagnosing complications related to SB using reasoning and prompting methods.

2. 50 participants interacted with the models, providing qualitative feedback on their performance across common and obscure scenarios.

3. The model set's performance is tested in reasoning through combining symptoms and asking follow-up questions to narrow down diagnoses.

Success and Failure Definitions:

Moreover, we define the success and failure criteria for the performance of LLMs as clinical FMs:

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

2. Well-timed conclusivity: Only after a detailed step-wise analysis should the model request more concrete modalities like specific imaging outputs (that may be accessible by the patient or their clinician) instead of jumping to a diagnostic result prematurely while bypassing steps in its way.

Formalizing the Inverse Prompt 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 $K \in Z^+$ 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.

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.
```

RESULTS

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

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

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

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

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

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

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

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

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

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

Metrics

Diagnostic accuracy measures the proportion of correct impressions (fully or partially correct). It is defined as: $\alpha = \frac{\sum(\psi + 0.5\phi)}{|\mathcal{N}|}$, where ψ is the number of correct diagnoses, is the number of partially correct diagnoses (e.g., identified some symptoms but led to the wrong conclusion), and $|\mathcal{N}|$ is the total number of conversations with the LLM. Errors (e.g., making the right diagnosis for the wrong reasons) arise when a diagnosis is based on incorrect reasoning. The *error rate* is defined as: $\epsilon = \frac{|E|}{|\mathcal{N}|}$, where $|E|$ is the number of errors in diagnoses and $|\mathcal{N}|$ is the total number of conversations. The *user intervention rate* measures frequency of user-guided interventions to refocus the model. It is defined as: $\beta = \frac{\text{Number of user interventions}}{|\mathcal{N}|}$. The criteria for bridging includes: failing to link symptoms that are clinically relevant, introducing unrelated impressions or findings, repeats queries or fails to progress reasoning, does not retain key information from earlier interaction.

PROPOSED METHOD, FUTURE WORK, & SYMPTOM-LEVEL FINDINGS

A multistage architecture is proposed to better handle diagnostic tasks through integrated patient-model conversation (Figure 1).

Module 1:

- ***Directed corpus formation:*** Targets patient-specific information to narrow down diagnosis search space.
- ***Information retrieval:*** Focuses on retrieving relevant data rather than reasoning, decomposing it into viable tasks needing further inputs.

Module 2:

- ***Planner module:*** Utilizes a vision-language planner for diverse input requests and better interpretation of web-based corpus related to conditions.
- ***Relational mapping:*** Maps top-level patient information to selected conditions.
- ***Conversation AI backend:*** Connects to the interface, facilitating information linkage with a memory unit across longer conversations.

Curating Specialized Datasets: Should incorporate diverse medical records (with emphasis on comprehensive data from the National Spina Bifida Patient Registry and other medical sources), clinical notes, and literature to enrich the knowledge base, increasing diagnostic reliability for complex conditions.

Larger Participant Cohorts: Enhances model effectiveness through varied patient interactions.

Prompting Strategy Experimentation: Testing strategies like Socratic prompting for improved diagnostic interactions.

End-to-End Implementation & Validation:

- ***Quantifiable Metrics:*** Evaluating model effectiveness using metrics like ROUGE and interrater reliability.
- ***Expert Involvement:*** Neurosurgeons and other experts to ensure alignment with medical standards.
- ***LLM Finetuning:*** Using a fixed medical database for comprehensive evaluation and benchmarking.

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

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