

Ask WhAI: Probing Belief Formation in Role-Primed LLM Agents

Keith E. Moore*, Jun W. Kim, David Lyu, Jeffrey Heo, and Ehsan Adeli

Stanford University, Stanford, California, U.S.A.
{kem1, junwkim, dlyu, jeffheo, eadeli}@stanford.edu

Abstract

We introduce **Ask WhAI**, a breakpoint-based debugger for studying belief formation in role-primed large language model (LLM) agents. The system instruments specialist agents interacting across sequential encounters while reading from and writing to a shared electronic medical record (EMR). Each encounter serves as a breakpoint where belief state can be queried *out-of-band* without altering the dialogue, enabling inspection of priors and rationales.

Ask WhAI supports controlled experiments on multi-agent reasoning by recording and replaying encounters and by probing belief both in character and through character-agnostic inference (“Sherlock” mode). The system supports perturbations of a patient’s journey, including encounter reordering, document priming, probe reframing, and counterfactual evidence. Diagnostic evidence is governed by an oracle **LabAgent**, which reveals hidden results only when the corresponding tests are ordered.

We demonstrate the framework on a synthetic multi-specialty diagnostic scenario involving abrupt-onset neuropsychiatric symptoms. Across controlled perturbations, we observe strong role-conditioned priors, order effects in diagnostic reasoning, and substantial variation in expressed belief across probe designs.

By separating belief inspection from clinical dialogue and enabling replay under targeted perturbations, **Ask WhAI** provides a reproducible method for analyzing belief trajectories and epistemic dynamics in multi-agent LLM systems.

1 Introduction

Scientific progress depends on the accumulation of evidence and on the ability of experts to interpret evidence across disciplinary boundaries. In many areas of medicine, epistemic silos shaped by training, publication venues, and clinical norms inhibit productive consensus [7]. Large language models (LLMs) are increasingly used to emulate expert roles through prompts such as “Act like an experienced cardiologist.” Such role prompts can induce domain-specific priors, which may help performance but can also bias evidence interpretation, anchor decisions prematurely, and restrict consideration of alternatives.

*Corresponding author

Most evaluations of agentic systems emphasize final outputs (e.g., accuracy or task completion). Here we study *belief formation*: what a role-primed agent **appears to believe**, **how that belief changes** across interaction, and whether **changes are driven** by evidence, framing, or encounter order. Studying these dynamics requires tools that can inspect belief state during a multi-agent interaction without contaminating the dialogue itself.

To address this, we introduce **Ask WhAI**, a breakpoint-based debugger for probing belief formation in role-primed LLM agents (such as simulated medical encounters). The debugger instruments sequential encounters, supports out-of-band belief queries, and enables replay under controlled perturbations.

We evaluate the framework in a synthetic multi-specialty diagnostic workflow involving abrupt-onset neuropsychiatric symptoms. We chose this setting because epistemic silos between disciplines seem to manifest in the virtual agents (i.e., different perspectives from a psychiatrist and a neurologist agent when studying the same record). Our goal is not to settle the clinical controversy, but rather to use a patient’s journey as a controlled environment for measuring how role-based priors shape belief formation under changes in evidence exposure, probe design, and encounter order.

2 Related Work

Recent surveys describe rapid progress in multi-agent reasoning and coordination [11]. Agentic simulation frameworks have explored persistent personas, moderator roles, structured encounters, and sequential reasoning in clinical or social settings [10, 12, 1, 8]. Park et al. [10] model agents with distinct epistemic identities whose behavior reflects stable priors and social context. Swanson et al. [12] study collaborative workflows in which discipline-specific agents contribute specialized perspectives while a coordinating agent drives critique and integration. In the clinical domain, Agrawal et al. [1] use role-driven models to simulate structured patient vignettes and sequential reasoning.

Related work on reasoning analysis and debugging has examined replay, justification generation, output traceability, and message editing. Systems such as layered prompting [3], AGDebugger [2], and AgentRR [4] support inspection or modification of agent behavior, primarily in single-agent or post-hoc settings. More broadly, work in explainability and interpretability asks whether models can justify outputs or expose reasoning [15], but these typically focus on explaining final answers rather than measuring longitudinal belief formation. Inspired by these lines of work, we combine multi-agent, multi-encounter simulation with debugging-style instrumentation.

3 System Overview

3.1 Medical Case Simulator

The simulator models a diagnostic workflow as a sequence of encounters between large language model (LLM) agents assigned to distinct roles. Figure 1 illustrates the architecture of a single encounter. Additional implementation details are provided in Appendix A and the project repository [6].

Each encounter begins when the moderator agent¹ and the specialist internalize any private

¹The architecture supports peer-to-peer messaging, but without a moderator, agents often talk past one another. Following [12], we use a moderator (e.g., parent or lead reviewer) to keep discussions focused through

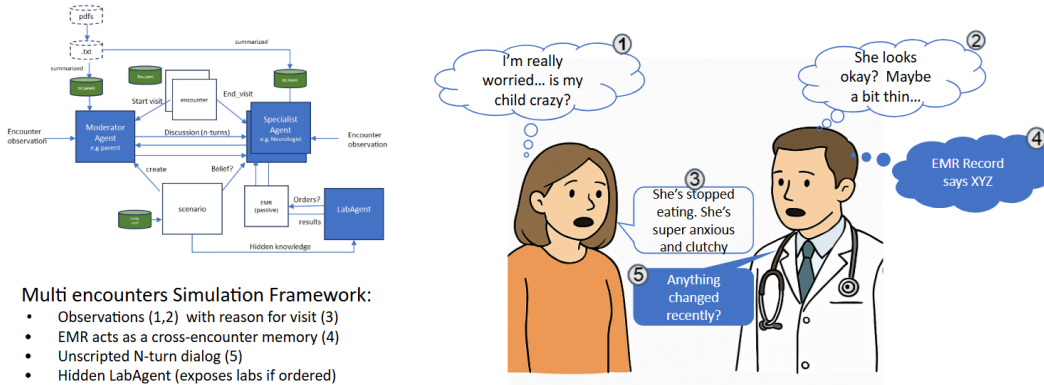


Figure 1: Conceptual architecture of an encounter with a specialist agent.

observations (steps 1 and 2). The specialist may also review the patient’s EMR (step 4), which serves as a timestamped, cross-agent memory. The moderator and specialist then engage in unscripted dialogue (step 5), triggered solely by the scripted `reason_for_visit` (step 3). All EMR updates and messages are logged.

The simulator supports:

- Generating role-specific interactions from scripted scenarios
- Maintaining a longitudinal EMR with timestamped entries and evolving lab results
- Enforcing access controls based on role and encounter timing

3.2 Ask WhAI Debugger

Ask WhAI wraps the simulator to enable structured inspection and targeted perturbation of belief state. It treats each encounter boundary as a *breakpoint* (i.e., a pause where agent beliefs can be queried, replayed, or modified). This enables us to trace belief evolution and test whether shifts are driven by evidence, framing, prompt design, or entrenched role-based priors.

Most experiments preserve the original sequence of encounters, varying only what information is shown, how it is framed, or how confidently it is presented. For example, we may inject counter-belief documents, vary agent tone, or display lab results that were never originally ordered.

Ask WhAI also supports reordering encounters (i.e., visiting a psychiatrist before a neurologist); however, since these encounters surface test results, we generate lab tests when they are ordered. This means that, in some situations, a specialist sees a test result because a prior specialist ordered it. This lets us probe how belief formation depends on content, attribution, and **timing**.

Ask WhAI exposes seven independent controls (Table 1) that can be applied independently or jointly to simulate variation in workflow.

4 Use Case: Abrupt-Onset Neuropsychiatric Scenario

To demonstrate Ask WhAI, we simulate a multi-specialty diagnostic process for a child presenting with handwriting changes, separation anxiety, and sudden-onset severe OCD following a recent challenges and rebuttals.

Control	Description
Priming	Provide agents with documents or evidence before an encounter.
Exposing	Reveal EMR state to agents before they engage.
Probing	Modify the belief query or the format of the response.
Encounter order	Reorder the sequence of specialist interactions.
Lab results	Insert or modify lab findings shown to the agent.
Persona / voice	Adjust the style or assertiveness of agent speech.
Forced reflection	Alter EMR prompts to elicit more reflective reasoning.

Table 1: Key controls available to influence specialist behavior and belief formation.

infection. We use this scenario as a stress test because the surrounding literature and clinical interpretations are often specialty-dependent, making it a natural setting to examine how role-based priors, evidence exposure, and encounter order shape belief trajectories. This symptom constellation is consistent with a class of **abrupt-onset neuropsychiatric syndromes** known as **PANS/PANDAS** [13, 14] (Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcus). The simulated case is a useful stress test because it centers on a long-standing causality controversy.²

Each specialist agent is initialized with a role-specific persona (e.g., “you are a pediatric neurologist”) and receives identical access to the patient’s current EMR and lab results unless scenario controls specify otherwise. These role prompts encode strong domain-specific assumptions learned during training, which can anchor interpretation and bias evidence integration. The simulated case of abrupt onset symptoms offers an environment where such entrenched priors are likely to surface, allowing us to observe cases in which agents:

- Ignore or discount relevant findings inconsistent with their priors
- Make logical errors even on evidence they accept
- Reach divergent conclusions despite having access to the same information

Table 2 summarizes the baseline diagnostic stance of each persona after the *first* encounter, where all agents saw the same initial record and were asked whether an infection-triggered etiology was plausible.

Agent Persona	Baseline perspective on PANS/PANDAS
Neurologist	Views strep as too common to indicate causality. Often cites the Kaplan & Kurlan study [5]. Rationale: Correlation is not causation.
Immunologist	Views PANDAS as similar to other post-infectious autoimmune syndromes (e.g., Sydenham chorea, Lyme disease, NMDAR encephalitis). Rationale: Relies on anti-neuronal biomarkers.
Psychiatrist	Sees PANDAS as severe idiopathic pediatric OCD. Rationale: Without more evidence, treat as severe OCD.
Pediatrician	Refers to a specialist (neurology or psychiatry). Rationale: Follows AAP guidance and standard of care.

Table 2: Divergent specialist interpretations of PANS/PANDAS causality

These perspectives mirror real-world clinical encounters, where entrenched role-specific priors can lead to critical omissions. For example, if a child presents with contamination fears but

²The association between infections and psychiatric symptoms has been noted since Osler’s study of obsessional behaviors following Sydenham’s chorea [9]. Debate continues over whether such associations are causal or coincidental.

no reported sore throat, the pediatrician may not consider strep as a possible cause and therefore not conduct a strep test. Similarly, when we asked the psychiatrist agent why a strep test had not been ordered, the reply was: “Because we don’t do strep tests. That is done by a pediatrician.” This highlights how divided responsibility, combined with implicit role boundaries, can prevent simple diagnostic actions, even when clinically relevant.³

5 Experiments and Results

5.1 Reproducibility

Simulations are reproducible under fixed conditions by caching agent responses based on prompt content, model version (e.g., `gpt-4o-2024-08-06`), model temperature, and message order. Timestamps and other non-behavioral elements are excluded from the cache key. Cached responses are reused unless bypassed explicitly (e.g., for stochastic probing). All cached outputs, scenario configurations, and simulator code are versioned and checked into our repository [6].

5.2 Experimental Setup

Starting from the baseline beliefs in Table 2, we ran a series of experiments, each varying a single debugger control (Table 1). Each experiment begins with an analytic question (“What do we want to vary?”), identifies a simulation *hook*, modifies the simulator to expose it, and configures the debugger to isolate and measure its effect.

Belief is probed after each encounter using a configurable prompt.⁴ In the first scenario, this was a simple categorical query (“Is this an infection-triggered etiology?”) with responses mapped to {skeptical, neutral, believes}. Prompts are configurable to return categorical or continuous variable responses (such as confidence in a diagnosis).

In the following subsection, we present three of the experiments (details of which are in the appendices).

5.3 Experiment: Priming with Facts

Goal: Pre-encounter document ingestion

Method: Before an encounter, agents can be **primed** with documents (articles, papers, bulletins) that persist in agent memory. Because each agent reads documents “in character,” the same document can produce different summaries. Although any number of documents can be added, a system document, `0.txt`, is provided to all agents to establish common priming or facts.

Intervention: Added to `0.txt` a recent bulletin stating that the American Academy of Pediatrics (AAP) recognized infection-triggered abrupt-onset neuropsychiatric syndrome as a research focus.

Outcome: Under priming (see Table 3, column 3b), the pediatrician shifted from skeptical (Encounter 1) to neutral (Encounter 3) and ultimately to belief in PANDAS by Encounter 15. The same experiment was not able to move the neurologist’s perspective (not shown).

Observation: Priming softened skepticism in most agents. Because the hook isolates document exposure, changes can be attributed to this intervention rather than uncontrolled history drift.

³Ethical note: This simulation is synthetic and illustrative. It is intended solely to test whether multi-agent systems reproduce characteristic reasoning dynamics observed in interdisciplinary settings and is not diagnostic.

⁴Prompts are configured in a scenario `.yaml` file.

Encounter	3a (Baseline)	3b Priming	3c (Sherlock Mode)	3d (Entangled)
1	Skeptical	Skeptical	Viral pharyngitis	Skeptical
2	Skeptical	Skeptical	PANS	Skeptical
3	Skeptical	Neutral	PANS	Skeptical
10	Skeptical	Neutral	PANDAS	Skeptical
15	Neutral	Believes	PANDAS	Believes

Table 3: Pediatrician belief evolution across identical clinical scenarios.

5.4 Experiment: Information Exposure and Belief Assessment

Goal: Out-of-band belief probing

Method: Using the same scenario as Section 5.3, we compared three probe styles:

1. **Multiple Choice (a,b):** Ask if the case is infection triggered [skeptical, neutral, believes] and request justification.
2. **List-first (c):** Request top 10 diagnoses before any stance query.
3. **Entangled (d):** Ask for top diagnosis, then stance and justification for that diagnosis.

Intervention: Changed only the probe format via the hook, leaving all other simulation elements fixed.

Outcome: In list-first probing, infection triggered surfaced earlier and stance shifted toward belief sooner. Entangled format produced early top-diagnosis commitment with explicit belief.

Observation: A single-word change (“diagnoses”) altered role behavior, overriding the cautious pediatrician persona. We referred to this mode as “Sherlock mode,” shown in Table 3, column 3c. This style of probe allowed us to measure the internal difference between what the model had diagnosed and what the model would report “in character.” In prompt 3d (described in the appendix), we blend the two prompts to show both the “out-of-character” internal diagnostic state and the “in-character” post hoc belief.

5.5 Experiment: Order Effects on Belief Dynamics

Goal: Measure how belief formation in the pediatrician agent is shaped by the sequence and identity of prior specialist encounters.

Method: We ran six permutations of a fixed 16-encounter diagnostic scenario, varying the order of neurology, psychiatry, and rheumatology specialists. Pediatrician agents were inserted between specialist blocks and at the end to assess belief evolution. Belief scores (0–10) were derived from stance labels and averaged over three simulation runs at constant temperature.

Intervention: Only the specialist encounter order was varied; all other simulation parameters were held constant.

Outcome: Pediatrician belief scores varied significantly by encounter sequence (Figure 2):

- (a) Belief increased cumulatively with successive encounters (ANOVA $p < 0.0001$).
- (b) Rheumatologists had the strongest influence when seen early; neurologists had the weakest (ANOVA $p < 0.05$).
- (c) Omitting rheumatology suppressed belief formation (ANOVA $p < 0.05$).
- (d) Rheumatologist beliefs were higher when preceded by neurologists rather than psychiatrists (ANOVA $p < 0.0001$).

Observation: The presence and order of specialist voices shaped belief formation in both pediatricians and rheumatologists. Rheumatology emerged as a prerequisite for belief change, while prior neurologic framing amplified its effect. More details on the longitudinal belief trajectories are provided in Appendix D, Figure 8.

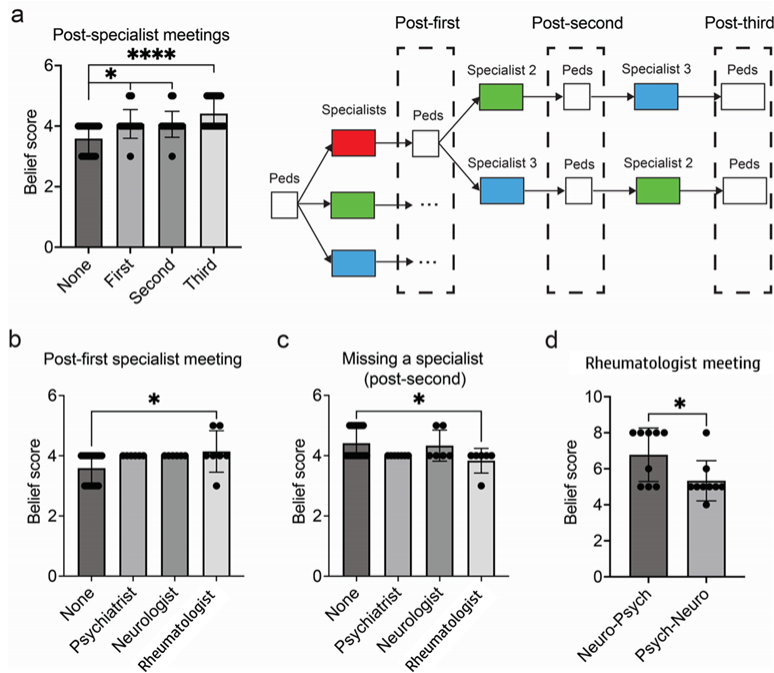


Figure 2: Order effects on pediatrician and specialist belief scores (a–d). Belief increased cumulatively across encounters and varied by specialist. Panel (d) shows a scale shift in y-axis to highlight neurologist influence on rheumatologist belief.

6 Discussion and Insights

6.1 Insights About Multi-Agent Simulation

Several patterns emerged from the debugger-based simulations that may generalize beyond the simulated use case:

1. **Reflection changes minds** (Moderator, Reflection Prompt). Belief shifts occurred most reliably when contradictions were explicitly surfaced (by the moderator, or a reflection prompt/EMR record requiring the agent to address inconsistencies).
2. **Early notes anchor thinking** (Evidence Injection, Selective Visibility). Once a record is added to the EMR, it strongly influenced downstream reasoning, even when later evidence contradicted it. This reflects the real-world risk of “narrative lock-in,” where an ICD-10 code can outweigh contradictory findings. Hiding early anchoring notes reduced this effect.
3. **Inference errors persist** (Priors, Persona, Evidence Injection). Agents often misinterpreted failure to reject the null hypothesis as proving the null. This happened often enough to suggest the need for targeted interventions against common reasoning errors. This error could be reduced with targeted reminders in the `0.txt` file.
4. **Order effects shape trajectories** (Encounter Reordering). The stance of early encounters (for example, skeptical neurologists versus open rheumatologists) significantly influenced how downstream agents interpreted the same data.

6.2 Insights from the Abrupt-Onset Scenario

While the scenario was based on a generic post-infectious hypothesis, it illustrates how disciplinary priors shape interpretation of ambiguous evidence. Even as a synthetic testbed, it surfaced reasoning patterns consistent with real epistemic divides across specialties.

- The neurologist agent persistently cited [5] as refuting an infection-triggered hypothesis, despite the study not supporting that claim.
- All agents assumed that a *current* strep infection was required to consider an infection-triggered diagnosis. This was a surprising constraint, given that other post-infectious conditions (such as Sydenham chorea or Long COVID) do not require an active infection.

Attempts to override these assumptions using document priming, counterfactuals, or prompting strategies had limited effect. Shifts only occurred when we inserted a *simulated* AAP article asserting that an infection can trigger abrupt-onset neuropsychiatric symptoms, suggesting that agents attend not only to content but to attribution and perceived authority (revealing how disciplinary priors interact with source credibility).

7 Future Work

While the framework provides a powerful testbed for epistemic modeling, several limitations remain:

- **Real-case validation:** Scenarios were realistic but still synthetic. Applying the debugger to real EMR cases could reveal how lab orders, ICD codes, and notes evolve over time.
- **Multiple test ordering:** The `LabAgent` does not yet support re-ordering or time-dependent changes. Enabling this would better reflect real-world longitudinal care.
- **Maintaining character:** In agent-to-agent discussions, occasional “Sherlock mode” breaks occurred when models stepped out of persona. More work is needed on prompt design to preserve in-character reasoning.
- **Broader scenarios:** Early experiments outside clinical diagnosis (such as multi-disciplinary paper review and cross-specialty negotiation of clinical trial design) showed that the framework can surface similar order effects, anchoring, and reflection gaps. Systematically extending to these contexts could test its generality.

8 Conclusion

This project demonstrated per-discipline biases in diagnostic reasoning. In the abrupt-onset neuropsychiatric scenario, the debugger showed how early documentation and encounter sequencing shape downstream reasoning, how structured reflection enables belief change, and how reasoning flaws common in practice (such as treating absence of evidence as evidence of absence) persist without targeted interventions. Belief change required more than data. It required forced reflection. Prompts must explicitly elicit this action to resolve inconsistencies; moderation helped, though breakdowns still occurred.

The framework generalizes without code changes to other conditions and decision-making contexts, from simulating patient journeys to joint review of research. As medicine faces syndromes that outpace consensus, such simulations may help illuminate not only what experts believe but also why they hold those beliefs, and what, if anything, might change their minds.

References

- [1] Moni Agrawal, Zijian Wang, Shujie Xie, Patrick A. Keane, Anirudh Kumar, Mo Yu, and Marzyeh Ghassemi. Do large language models understand clinical guidelines? a case study in pediatric asthma. *arXiv preprint arXiv:2412.11716*, 2023.
- [2] William Epperson, Gaurav Bansal, Valentine Dibia, Adam Fourney, Jeroen Gerrits, Ed Zhu, and Saleema Amershi. Interactive debugging and steering of multi-agent ai systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. arXiv:2503.02068.
- [3] Bashir Faisal and David Tunkel. Explainable ai in multi-agent systems: Advancing transparency with layered prompting. *arXiv preprint arXiv:2502.11455*, February 2025.
- [4] Erhu Feng, Wenbo Zhou, Zibin Liu, Le Chen, Yunpeng Dong, Cheng Zhang, Yisheng Zhao, Dong Du, Zhichao Hua, Yubin Xia, and Haibo Chen. Get experience from practice: Llm agents with record & replay. *arXiv preprint arXiv:2505.17716v1*, May 2025.
- [5] Roger Kurlan, David Johnson, and Edward L. Kaplan. Streptococcal infection and exacerbations of childhood tics and obsessive-compulsive symptoms: a prospective blinded cohort study. *Pediatrics*, 121(6):1188–1197, 2008.
- [6] Keith Moore, Jun W. Kim, David Lyu, Jeffrey Heo, and Ehsan Adeli. Simulation code for agentic modeling of medical controversy. https://github.com/buster4947/llm_pandas, 2025.
- [7] Robin P. Newhouse and Bonnie Spring. Interdisciplinary evidence-based practice: moving from silos to synergy. *Nursing Outlook*, 58(6):309–317, Nov–Dec 2010.
- [8] Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025.
- [9] William Osler. *On Chorea and Choreiform Affections*. H.K. Lewis, London, 1894.
- [10] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.
- [11] Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models: A survey. *arXiv preprint arXiv:2503.23037*, 2025.
- [12] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, 2024.
- [13] S. E. Swedo, J. Seidlitz, M. Kovacevic, et al. Clinical presentation of pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections in research and community settings. *Journal of Child and Adolescent Psychopharmacology*, 25(1):26–30, 2015.

- [14] Susan E. Swedo, Henrietta L. Leonard, Margo A. Garvey, Beth Mittleman, Andrew J. Allen, S. Perlmutter, L. Lougee, S. Dow, J. Zamkoff, and B. K. Dubbert. Pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections: clinical description of the first 50 cases. *American Journal of Psychiatry*, 155(2):264–271, February 1998. Erratum in: *Am J Psychiatry* 1998 Apr;155(4):578.
- [15] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):20, 2024.

A Agentic System Architecture and Roles

A.1 System Actors and Interactions

The simulation involves six primary components:

- **Specialist:** Each clinical agent (e.g., psychiatrist, neurologist, rheumatologist, pediatrician) is instantiated with a fixed persona and voice. The *persona* captures their epistemic stance and disciplinary priors, while the *voice* governs how they express uncertainty, cite evidence, and engage in dialogue. Agents operate independently and retain long-term beliefs across encounters.
- **Moderator:** A neutral agent tasked with surfacing contradictions, probing justifications, and prompting reflection. The moderator is especially active during *moderated review* and *sequential encounters*, acting as a reflective foil to elicit deeper reasoning. The moderator can also be a *patient* or a parent advocating for a patient.
- **Lab (Oracle):** A hidden agent representing ground truth or unavailable diagnostic data. The LabAgent releases lab results (consistent with hidden diagnosis) if ordered during an encounter.
- **EMR (Electronic Medical Record):** A structured memory artifact that aggregates case observations, lab results, and planned follow-up. This enables continuity across encounters and serves as a shared substrate for agent reasoning.
- **Private reflection:** We also support querying agents for their "private thoughts" (i.e., why they hold a belief). This is particularly useful for seeing what they are paying attention to.
- **Encounter:** An office visit (time passes between encounters)

The interactions between these components is shown in Figure 3

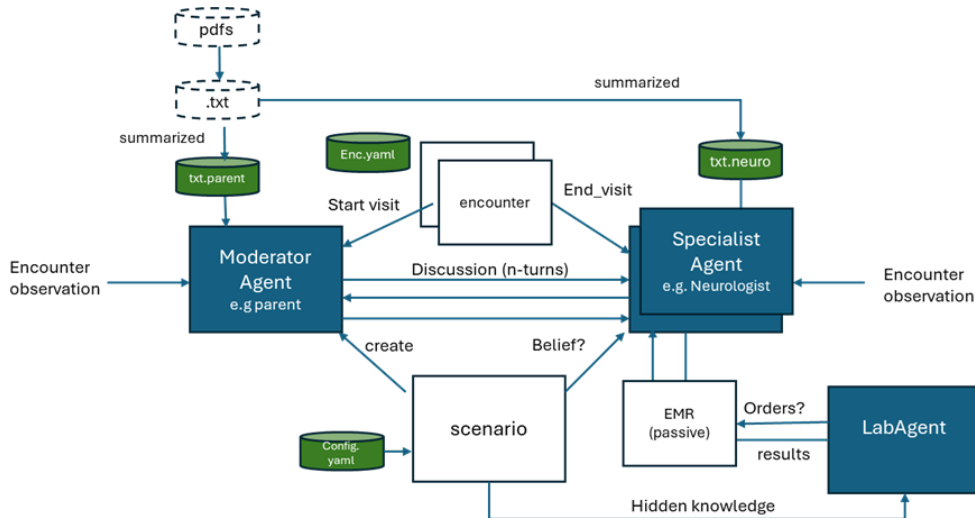


Figure 3: Overall System Architecture

The 4.0 version of OpenAI that we used did not support direct uploading of `.pdf` files. Therefore, we convert each `.pdf` to `.txt` and then have the agent convert the document, based on its persona, into a version it considers a summary. This both reduced the tokens used in prompts and allowed comparison of how two different agents might read the same document, including what information was retained and what was dropped.

A.2 Encounters and Moderated Multi-Turn Interactions

The heart of the simulation is the `encounters.yaml` file for each scenario. An encounter is a clinical interaction between one agent and the moderator. Encounters include:

- a **reason for visit** (e.g., symptom presentation or parent concern),
- optional **labs, documents, or private observations** (e.g., things observed in the encounter)

The conversation is intentionally **not** scripted and emerges from the multi-turn exchange between the moderator and the specialist. The personas and voices of the participants determine the conversation. The reason for visit can trigger escalation (e.g., referral to another specialty). Official encounter summaries are placed in the electronic medical record. Encounters are coordinated by a scenario that also queries private beliefs for analysis.

A.3 Configuring Scenarios

Scenarios are established in the `config/config.yaml` file. This file configures the agents and the labs and controls when and how beliefs are queried. This architecture separates concerns to allow targeted epistemic probing, flexible experiment design, and longitudinal belief tracking.

`config/config.yaml`

```

default: 3
summaries: summaries
scenarios:

* id: 1
  model: gpt-4o
  max_tokens: 6000
  moderator: parent
  doctor_prefix: |
    You are evaluating a child who presents with the following
    clinical observations:
  encounters: config/scenario1/encounters.yaml
  persona_prompt_dir: prompts/personas
  voice_prompt_dir: prompts/voices

```

This file specifies which agent model to use, where to find the `persona` and `voice` prompt templates, and which encounter script to run. It also supports optional reflection and EMR generation prompts, enabling structured memory and belief parsing.

A.3.1 Configuring Encounters

encounters.yaml

```
* doctor: pediatrician
  id: 1
  doctor_preread: [0]
  lab_results: # in visit
  - test: Rapid Antigen Test
    result: Negative for Group A Streptococcus
  doctor_context: 7 year old female patient who appears to have ...
  moderator_context: "Very worried about daughter.
    She's been acting strangely..."
  reason_for_visit: She won't eat. Keeps saying she's
    'not allowed' to eat...
```

Each encounter provides the opportunity to pre_read a paper, make an observation, establish the reason for the visit and force an in-office lab result. Encounters can trigger lab orders, update the shared EMR, or pass information between agents. They also support narrative branching and escalation to other specialties.

A.3.2 EMR Longitudinal Memory

Each encounter results in a public EMR summary. These summaries accumulate longitudinally and are passed to downstream agents, simulating anchoring and continuity across encounters. The agentic system passes the history of the encounter, which includes prior EMR records, observations, and conversations, for summarization and returns a structured EMR record. Given a structured prompt, an example EMR record is shown in Figure 4:

```
PEDIATRICIAN added EMR note:
[EMR Update]

**ICD-10 Codes:**
- F42: Obsessive-Compulsive Disorder
- F93.0: Separation Anxiety Disorder of Childhood

**Clinical Observations:**
- Patient: 7-year-old girl
- Presenting Symptoms: Sudden-onset obsessive behaviors, food refusal, anxiety, ritualistic behaviors, new-onset fears (e
  specially around contamination), urinary urgency, separation anxiety
- Vital Signs / Physical Findings: Physical exam unremarkable, no fever
- Labs / Tests: Rapid strep test negative
- Relevant History: Recent sore throat, acute behavioral changes over the past 1-2 weeks

**Assessment:**
- The patient presents with acute onset of obsessive-compulsive behaviors and anxiety, which could be indicative of Pedia
  tric Acute-onset Neuropsychiatric Syndrome (PANS) or another psychiatric condition. The negative rapid strep test reduces
  the likelihood of a current strep infection as a trigger. The symptoms may be related to a psychiatric or neurological c
  ondition, necessitating further evaluation.

**Plan:**
1. Refer to a child psychiatrist or psychologist for a comprehensive evaluation of anxiety and obsessive behaviors.
2. Consider a neurology consultation if symptoms persist or worsen, to rule out any neurological causes.
3. Monitor symptoms closely and provide support at home, focusing on stress reduction and maintaining routines.

**Patient advice:**
- Encourage the child to engage in calming activities, such as drawing, which she finds helpful.
```

Figure 4: EMR example written by PEDIATRICIAN agent

A.3.3 Lab/Oracle Interaction

At the end of each encounter, the **LabAgent** adds lab results for *ordered* tests. The agent is also an LLM agent and attempts to determine whether an order could conceivably yield the result. Like all agents, the **LabAgent** has a persona but is typically configured to be clinical and conservative.

The **LabAgent** primarily operates on the electronic medical record. It knows a priori which hidden tests are configured in a `.yaml` file for each scenario. The **LabAgent**'s role is to keep these tests hidden until explicitly called for, while also recognizing when an order could conceivably include the test (e.g., a lumbar puncture also resulting in a finding of protein in the CSF).

The contents of the `hidden_labs.yaml` were:

```
mri: |
  MRI Brain w/ contrast reveals a 2.5 cm mass in the posterior fossa
  compressing the fourth ventricle. Suggestive of obstructive
  hydrocephalus.

lp: |
  Lumbar puncture shows elevated opening pressure. No signs
  of infection. Protein and glucose levels are within normal
  limits.

cbc: |
  mild elevation of eosinophils
```

The orders by the doctor were:

```
**Plan:**
1. Order MRI brain to assess for any structural...
...
```

This resulted in the **LabAgent** releasing the MRI result into the electronic medical record. The lumbar puncture and blood tests were not released because the neurologist did not order them. In this case, the MRI was sufficient for diagnosing the condition, but it was interesting that rule-out tests were not part of orders.

A.3.4 Private "belief" prompts

A sample belief prompt is shown in Figure 5:

```
- id: 12 # using for kem1
model: gpt-4o
max_tokens: 6000
moderator: patient_2
labs_hidden: config/scenario_kem2/labs_hidden.yaml
doctor_prefix: A 7-year-old girl presents with sudden-onset obsessive behaviors, food refusal, and anxiety. Her physical exam is unremarkable.
encounters: config/scenario_kem2/encounters.yaml
persona_prompt_dir: prompts/personas
voice_prompt_dir: prompts/voices
n_turns: 4
belief_parse_expr: "Belief:\\s*(.+?)\\s*(\\n|$)"
reflection: |
  You are now privately reflecting on your clinical impression of this case, and what is most likely.
  What do you think might be the top 10 diagnoses and which one do you think it might be?
  Please respond in this format:
  Top diagnoses: [Your answer]
  Belief: [Your top belief]
  Justification: [clinical reasoning, as if writing a candid note to a trusted colleague]
  Be sure to address any conflicting evidence or perspectives in the EMR
```

Figure 5: Sample Belief Prompt

Notice that `belief_parse_expr` is a regular expression used to extract the belief from the prompted reply. This provides considerable flexibility in obtaining categorical or numeric responses.

This prompt produces reflections like the following:

```
-- psychiatrist Initial Belief --
[BELIEF CHECKPOINT] Belief: Pediatric Acute-onset Neuropsychiatric Syndrome (PANS)
Justification: Top diagnoses:
1. Pediatric Acute-onset Neuropsychiatric Syndrome (PANS)
2. Obsessive-Compulsive Disorder (OCD)
3. Anxiety Disorder
4. Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections (PANDAS)
5. Acute Stress Reaction
6. Autoimmune Encephalitis
7. Adjustment Disorder
8. Somatic Symptom Disorder
9. Functional Neurological Disorder
10. Viral Encephalitis

Belief: Pediatric Acute-onset Neuropsychiatric Syndrome (PANS)

Justification: Given the sudden onset of obsessive-compulsive symptoms, food refusal, and anxiety in the context of a recent minor illness, PANS emerges as the most compelling diagnosis. The rapid onset and symptom constellation, including separation anxiety and frequent urination, align closely with PANS criteria. Although the negative rapid strep test reduces the likelihood of a current strep infection, it doesn't rule out PANS, which can be triggered by various infections or stressors.

While classic OCD typically has a gradual onset and wouldn't usually present with such acute changes, the abrupt appearance of symptoms here suggests the possibility of an underlying medical trigger. Anxiety disorders and adjustment disorders could explain some of the symptoms, but the overall presentation and the timing with the recent illness strengthen the case for PANS.
```

Figure 6: Belief response

B Cost and Performance

The system relies on many agent interactions, including dialogue between agents (typically three or four turns), interpretation of the EMR, writing to the EMR, belief queries, and possible pruning of history. This results in approximately 180–200 interactions with the LLM for a 15-encounter run. While individual exchanges are modest, the calls accumulate (see Figure 7 and Table 4).

Caching: To address this cost, and to allow repetition of experiments, we provide a cache that returns prior responses to identical prompts and histories. This means that changes in the order of exchanges can cause a cache miss for a given exchange. The cache can be disabled either by changing the `encounter.yaml` file or by using the `--nocache` command-line option.

Experiment Type	Encounters	Total Tokens (k)	Total Cost (USD)
Moderated Reflection	4	234	\$0.42
Document Exposure	4	916	\$1.60
Case Study	15	381	\$0.69
Case - Sherlock Mode	15	570	\$1.03

Table 4: Token and cost estimates by simulation type

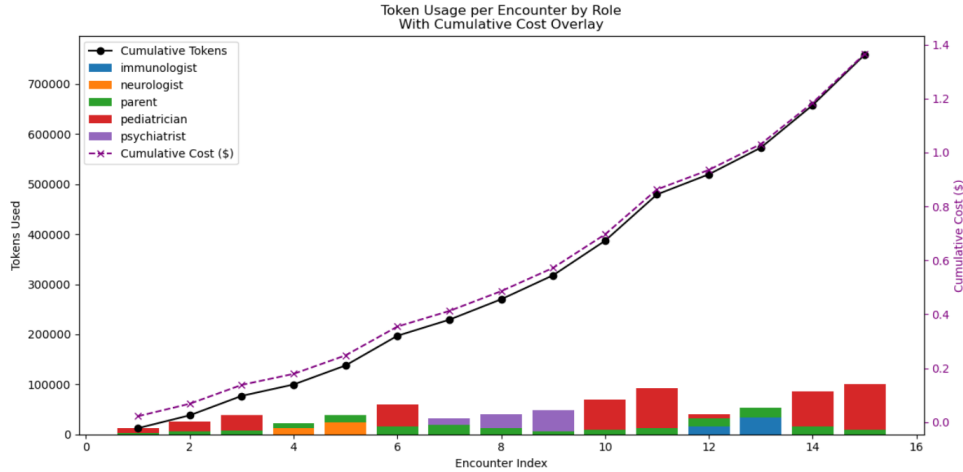


Figure 7: Token use per encounter

C Internal State vs Character Belief

In this section, we expand on Section 5.4, where we noticed a difference between asking for a diagnosis and asking for an "in-character" belief.

For the "in-character" setting in Table 5, Scenario 3a (replicated below for convenience), we used a prompt like:

Encounter	3a (Baseline)	3b Priming	3c (Sherlock Mode)	3d (Entangled)
1	Skeptical	Skeptical	Viral pharyngitis	Skeptical
2	Skeptical	Skeptical	PANS	Skeptical
3	Skeptical	Neutral	PANS	Skeptical
10	Skeptical	Neutral	PANDAS	Skeptical
15	Neutral	Believes	PANDAS	Believes

Table 5: Duplicate of Table 3: belief evolution across identical clinical scenarios.

Belief: [rejects | skeptical | neutral | believes]
 Justification: [clinical reasoning as if writing to a trusted colleague]

This format encourages thresholded, persona-consistent reasoning. The pediatrician begins skeptical, expresses neutrality only after multiple confirmatory signals, and ultimately endorses an infection-triggered diagnosis only after specialist annotation and longitudinal observation.

By contrast, Table 5, Scenario 3c ("Sherlock mode"), returned something beyond the character. Here we used a more open-ended prompt:

Belief: [most likely]

Top diagnoses: [freeform list]
 Justification: [clinical reasoning]

This format unexpectedly loosened the constraints of persona and allowed the model to reason as an independent diagnostic synthesizer (we refer to "Sherlock mode"). The same pediatrician agent, when prompted this way, surfaces **infection-triggered** as the most likely diagnosis by Encounter 2 and acts accordingly, despite having no more evidence than in Scenario 3a. Finally, Scenario 3d combines the two prompts to show that asking for the diagnosis does not actually affect the in-character belief:

Top diagnoses: [ranked differential]
 Belief: [rejects | skeptical | neutral | believes]
 Justification: [reasoning that reconciles diagnostic list with belief]

D Order Effects on Diagnostic Belief States

Expanding on Figure 2 from Section 5.5, we monitored the agent's beliefs across six different encounter series. The series permuted the order of the patient's visit with specialists. At the end of each patient encounter, a belief score was recorded on a scale from 0 (no belief) to 10 (strong belief that the case was infection-triggered).

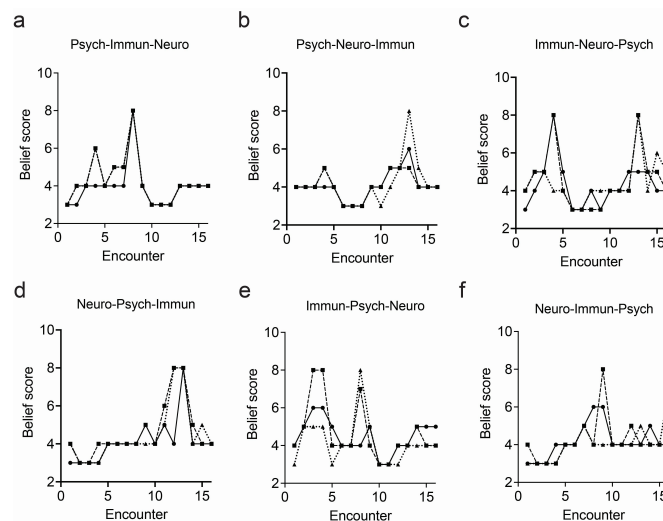


Figure 8: Tracking of infection etiology belief scores across encounters.