Multi-Agent AI Frameworks for Clinical Diagnosis Support: Benchmarking LLM Reasoning with Sleep Disorders as a Testbed

Aarushi Jaitly, MS
Carnegie Mellon University
Pittsburgh, PA, USA
ajaitly@cmu.edu

Helom Berhane, MS
Carnegie Mellon University
Pittsburgh, PA, USA
helom@cmu.edu

Deepa Burman, MD
UPMC Children's Hospital of Pittsburgh
Pittsburgh, PA, USA
deepa.burman@chp.edu

Anand Rao, PhD
Carnegie Mellon University
Pittsburgh, PA, USA
anandr2@andrew.cmu.edu

Ramayya Krishnan, PhD
Carnegie Mellon University
Pittsburgh, PA, USA
rk2x@andrew.cmu.edu

Rema Padman, PhD
Carnegie Mellon University
Pittsburgh, PA, USA
rpadman@andrew.cmu.edu

***Abstract*—Sleep is an important dimension of Quality of Life (QoL), affecting approximately 30–70% of older adults worldwide and contributing to a broad range of physical, cognitive, and psychological impairments [3]. Yet access to specialist sleep expertise remains severely limited, leaving caregivers and patients without timely, reliable guidance. Generative AI offers a potential pathway to scalable clinical support, but off-the-shelf large language models (LLMs) present well-documented challenges for safe clinical use: responses are often ungrounded in verified medical sources and lack the structured differential reasoning essential to clinical assessment.**

This paper presents a multi-agent AI framework for clinical decision support using sleep disorder management as a controlled testbed. The framework introduces three components: (1) a combinatorial synthetic patient persona corpus of 224 profiles ($8 \times 7 \times 4$) spanning clinically realistic comorbidity and etiology combinations; (2) 24-month longitudinal care pathway modeling across six clinically spaced episodes; and (3) a knowledge retrieval pipeline restricted to nine pre-approved medical sources, supporting a Testing, Evaluation, Verification, and Validation (TEVV) methodology. This framework is the subject of active research and development. We provide an overview of the current state of development, the evaluation infrastructure, and a pilot evaluation, and describe future work based on the preliminary findings reported here. For a single representative persona, all five benchmarked LLMs converged on plausible primary diagnoses, yet *none replicated the differential diagnostic reasoning process used by the physician benchmark*. These findings are preliminary and specific to one representative case; broader validation across the full 224-persona corpus is in preparation.

***Index Terms*—multi-agent AI, sleep disorder management, quality of life NLP, LLM benchmarking, retrieval-augmented generation, longitudinal care, differential diagnosis, TEVV pipeline, clinical decision support**

I. Introduction

Sleep disorders affect a substantial portion of the population [3], with consequences extending well beyond fatigue: inadequate or disrupted sleep is associated with elevated risks of cardiovascular disease, cognitive decline, depression, and reduced functional independence in older adults. Sleep disturbances affect up to 86% of patients with dementia and other neurodegenerative conditions [4]. Sleep disorders constitute an ideal testbed for evaluating clinical AI reasoning: their overlapping phenotypes — for example, Irregular Sleep-Wake Rhythm Disorder (ISWRD), REM Sleep Behavior Disorder (RBD), and circadian rhythm dysregulation share symptom profiles that require careful differential evaluation — their reliance on polysomnographic evidence for differential exclusion, and frequent off-guideline prescribing decisions in elderly populations demand precisely the structured diagnostic reasoning that off-the-shelf LLMs are least likely to demonstrate spontaneously. Specialist sleep expertise is geographically and temporally constrained, leaving caregivers without reliable guidance during out-of-hours periods.

LLMs encode substantial clinical knowledge [7] and can match or exceed physician quality on empathy and comprehensiveness [4], presenting a transformative opportunity for scalable clinical support. However, deploying off-the-shelf LLMs in clinical contexts introduces three limitations: ungrounded responses that may contradict guidelines or recommend contraindicated treatments; inability to reason longitudinally across a patient's evolving disease trajectory; and failure to replicate structured differential reasoning [8].

This paper presents a multi-agent AI framework addressing these limitations using sleep disorder management as a controlled testbed. The framework grounds model responses in a knowledge retrieval pipeline restricted to nine pre-approved medical sources, supports longitudinal reasoning across 24-month care pathways, and incorporates a clinically derived classification template to scaffold differential diagnostic reasoning [1], [2]. The pilot evaluation reported here assesses LLM diagnostic reasoning as a foundation for the Doctor Agent component currently under construction.

Contributions include: (1) an in-development multi-agent clinical AI framework with a processing-layer disorder clas-

This research was funded in part by the National Institute of Standards and Technology (NIST) (ror.org/05xpvk416) and the Carnegie Mellon University AI Measurement Science and Engineering Center (AIMSEC) (https://ror.org/05x2bcf33). Aarushi Jaitly, Helom Berhane, Ramayya Krishnan, Anand Rao, and Rema Padman were supported in part by NIST through Federal Award ID Number 60NANB24D231.

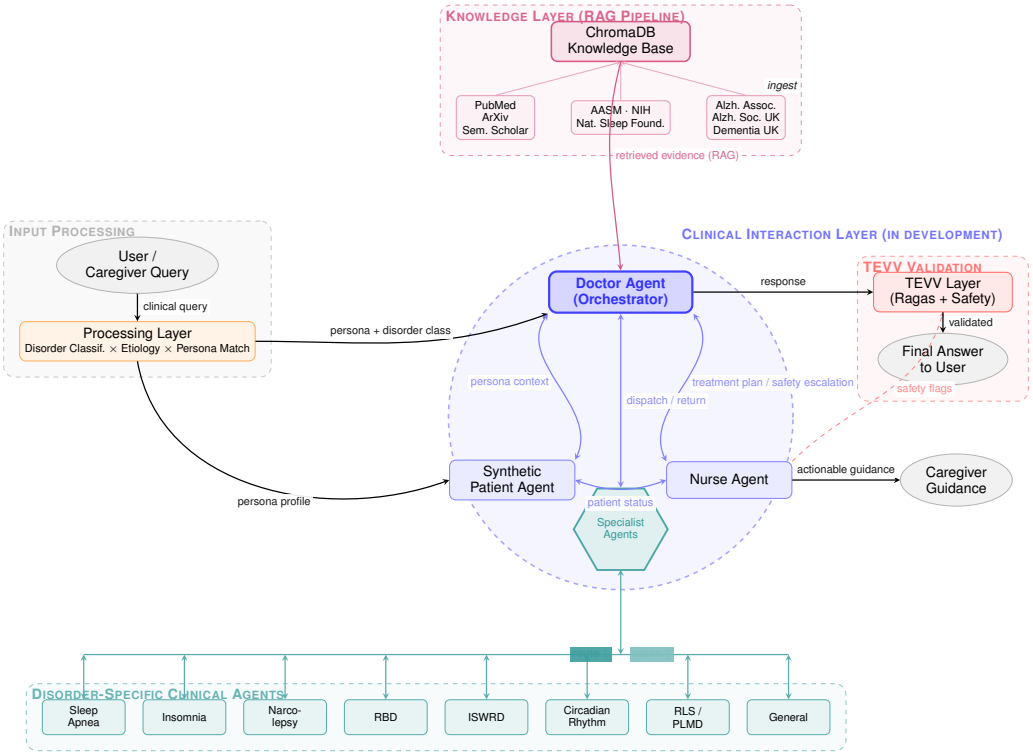

Fig. 1. Multi-Agent Clinical AI Architecture. **(1) Input Processing:** the Processing Layer classifies disorder type, etiology, and matches a synthetic persona from the 224-profile corpus, then routes to the Doctor Agent. **(2) Clinical Interaction Layer (in development):** the Doctor Agent (Orchestrator), Nurse Agent, Synthetic Patient Agent, and Specialist Agents form a circular topology; the Doctor Agent dispatches to domain-scoped Specialist Agents and integrates reasoning; the Nurse Agent manages caregiver communication and safety escalation; the Synthetic Patient Agent supplies persona context. **(3) Disorder-Specific Clinical Agents:** eight specialists receive routed queries and return domain-scoped reasoning. **(4) Knowledge Layer:** a closed-world RAG pipeline ingests documents from nine pre-approved sources into ChromaDB, supplying retrieved evidence to the Doctor Agent at query time. All responses pass through the TEVV Layer before delivery.

sifier, domain-scoped specialist agents, and synthetic longitudinal persona modeling; (2) a structured dual-prompt pilot LLM evaluation using a clinician-validated persona; (3) a clinical scoring methodology grounded in AASM standards; and (4) preliminary documentation of a differential diagnosis gap in one representative case, with direct implications for clinical AI deployment. Clinical guidelines follow American Academy of Sleep Medicine (AASM) standards throughout.

II. Related Work

QoL and sleep in clinical NLP (Natural Language Processing). NLP research addressing QoL has primarily focused on entity extraction from clinical notes [5], capturing structured signals in single-encounter settings but not the longitudinal reasoning that characterizes real-world sleep QoL assessment. This framework addresses that gap through longitudinal persona modeling and multi-episode care pathway generation.

LLMs in clinical settings. GPT-4 achieves near-expert accuracy on the USMLE (United States Medical Licensing Examination) [6]; Singhal et al. showed LLMs encode substantial clinical knowledge [7]; Ayers et al. found LLM responses often exceeded physician quality on empathy and comprehensiveness [4]. However, clinical LLM use faces challenges beyond hallucination [8]: responses lack longitudinal memory, are ungrounded in verified clinical sources, and do not perform structured differential reasoning essential to safe diagnosis, motivating the approach of grounding responses in a curated sleep medicine literature base with synthetic personas derived from peer-reviewed sources.

Multi-agent medical AI. MDAgents [1], Agent Hospital [2], MedAgent-Zero [9], and AgentMental [10] collectively demonstrate that specialist agent recruitment, adaptive questioning, and structured memory improve diagnostic accuracy over static prompting. These frameworks, however, do not generally combine closed-world source restriction with longitudinal patient memory, which can leave room for ungrounded generation and limit reasoning across extended care trajectories. The framework described here addresses this through closed-world source restriction, longitudinal persona memory, and a physician-derived classification template to scaffold differential reasoning [13].

III. The Multi-Agent AI Framework

A. Agent Architecture

The framework is built around four classes of agents interacting across structured clinical episodes, illustrated in Fig. 1.

The **Research Agent** layer comprises nine source-specific crawlers that ingest documents from pre-approved medical sources into a shared ChromaDB vector database, forming the closed-world knowledge base used at query time. A **Processing Layer** receives each incoming query and performs disorder classification, etiology analysis, and persona matching before routing to the Doctor Agent. The **Doctor Agent** (Orchestrator) coordinates the diagnostic workflow through a classification-then-dispatch pattern: the processing layer routes the classified query to the appropriate **Disorder-Specific Clinical Agent**, each defined with a distinct system prompt and bounded knowledge scope; the Doctor Agent integrates the specialist output with TEVV-enforced guideline constraints to produce a differential assessment and treatment recommendation. Eight specialist agents cover Sleep Apnea, Insomnia, Narcolepsy, REM Sleep Behavior Disorder (RBD), Irregular Sleep-Wake Rhythm Disorder (ISWRD), Circadian Rhythm, Restless Legs Syndrome and Periodic Limb Movement Disorder (RLS/PLMD), and General presentations, adapting the Agent Hospital model [2] to domain-scoped clinical reasoning.

The **Nurse Agent** handles caregiver-facing communication, translating clinical outputs into actionable guidance and monitoring for safety flags, escalating threshold concerns to the Doctor Agent. The **Synthetic Patient Agent** encodes a persona from the 224-profile corpus and is designed to maintain structured episode history across the 24-month care pathway to support longitudinal reasoning. Inter-agent communication protocols and longitudinal memory management for these two agents are currently under development and will be reported in future work. The pilot evaluation reported in Section IV focuses on the Doctor Agent role, benchmarking five frontier LLMs as candidate reasoning engines for that component.

B. System Architecture

The system operates across two functionally distinct layers illustrated in Fig. 1.

Knowledge Layer. Nine source-specific crawlers ingest documents from pre-approved medical sources into a shared ChromaDB vector database. Sources span peer-reviewed journals (PubMed, ArXiv, Semantic Scholar, Cochrane), clinical guideline bodies (AASM, NIH, National Sleep Foundation), and patient-facing organizations (Alzheimer's Association, Alzheimer's Society UK, Dementia UK). At query time, a Researcher Agent Orchestrator executes parallel vector searches and passes ranked, source-attributed chunks to the Doctor Agent via in-memory Retrieval-Augmented Generation (RAG) pass-through, grounding responses in verified source documents.

Clinical Interaction Layer. Four agent classes constitute the agentic core, currently under development. The Doctor Agent coordinates diagnostic workflow, integrating specialist outputs and applying AASM guideline constraints via the TEVV layer. The Nurse Agent handles caregiver-facing communication and monitors safety flags, escalating concerns to the Doctor Agent. The eight Disorder-Specific Clinical Agents receive routed queries and return domain-scoped reasoning. The Synthetic Patient Agent encodes a persona from the 224-profile corpus and will maintain structured episode history for longitudinal reasoning.

The pilot evaluation benchmarks five frontier LLMs in the Doctor Agent role. All responses pass through the TEVV layer, which combines Ragas-based faithfulness scoring [12] with rule-based safety checks against AASM contraindication criteria.

C. Synthetic Persona Generator (224 Profiles)

The framework employs a combinatorial persona structure: 8 disorder types $\times$ 7 etiological causes (age-related, medication-induced, comorbidity, environmental, neurological, behavioral, psychological) $\times$ 4 treatment modalities (CBT-I: Cognitive Behavioral Therapy for Insomnia; pharmacological; light therapy; PAP: Positive Airway Pressure) = 224 unique profiles, enabling evaluation against the full cross-product of realistic clinical presentations without requiring real patient data. The taxonomy was derived through LLM-assisted extraction across peer-reviewed sleep disorder literature and is subject to ongoing refinement. Each persona encodes sociodemographic characteristics, comorbid conditions, medication history, and psychosocial context. Dr. Burman (UPMC Children's Hospital of Pittsburgh), a physician author of this paper, is conducting clinical review to validate comorbidity combinations and treatment-etiology pairings.

D. 24-Month Longitudinal Care Pathways

The system is designed to generate 24-month care pathway episodes across six clinically spaced interactions (months 0, 1.5, 3, 6, 12, 24) [3] encoding four phases: initial presentation and safety establishment; early intervention and titration; maintenance and setback management; and long-term adaptation. Two 24-month case studies reviewed by Dr. Burman serve as longitudinal pathway prototypes. Evaluation of longitudinal reasoning capabilities is currently underway.

E. TEVV Assessment Pipeline

The system incorporates a TEVV methodology as a structured research-stage framework for assessing response quality and safety. *Testing* exercises the system across the 224-persona corpus to identify failure modes; *Evaluation* applies the Ragas framework [12] to measure faithfulness and guideline adherence against AASM standards; *Verification* checks outputs against structured ground truth templates; and *Validation*, in the sense of independent clinical confirmation of system appropriateness, will be conducted by Dr. Burman and domain specialists as described in the Validation Roadmap (Section VI-D).

Table I presents pipeline metrics from initial testing. These figures are drawn from the single representative case in this paper, scored across two prompt conditions and five models, yielding ten responses in total. Factual Consistency and Retrieval Quality are computed via the Ragas framework; Guideline Adherence reflects rule-based alignment with AASM clinical practice guidelines; Communication Quality

reflects structured review of empathy, readability, and utility. The Safety Score of 100% indicates that none of the ten responses triggered AASM contraindication flags; this reflects the constrained scope of this pilot and should not be interpreted as a general safety guarantee. All figures will be refined as evaluation is extended to the full persona corpus.

IV. Evaluation Design

This section reports a pilot evaluation using one representative case from the 224-persona corpus, assessing LLM diagnostic reasoning as a foundation for the Doctor Agent component. Findings are preliminary and specific to this single persona.

A. Patient Persona: Mrs. C

Mrs. C is an 82-year-old woman with a three-year dementia history (MMSE[1] 25/30) and managed hypothyroidism. Sleep is fragmented across the full 24-hour cycle: she retires at 10 PM, wakes 2–4 times nightly for nocturia (nighttime urination), and rises at 8–8:30 AM. Nighttime awakenings involve confusional arousals (episodes of incomplete awakening with disoriented, automatic behavior) and visual hallucinations at sleep-wake transitions. PSG[2] excluded clinically significant Obstructive Sleep Apnea (OSA; Apnea-Hypopnea Index AHI <5, where AHI measures breathing interruptions per hour of sleep) but confirmed fragmentation (arousal index 22.1/hour; efficiency 71%).

Clinical ground truth was established from Dr. K. Sharkey's documented management of an equivalent ISWRD case in the AASM 2016 case study series [13]: primary diagnosis ISWRD; sedative-hypnotics contraindicated per AASM 2015; melatonin prescribed off-guideline at 3 mg immediate-release, self-titrated to 1.5 mg with monitoring for depressive symptoms.

B. Prompt Conditions

Prompt 1 (P1): Mrs. C's case presented as an unstructured clinical narrative (180–250 words). Models required to state primary diagnosis, differential considerations, diagnostic reasoning, suggested treatment, and confidence level.

Prompt 2 (P2): Same case organized within a structured 15-domain sleep intake classification template developed by the physician author of this paper (Dr. Burman), instructing models to use the template as their primary classification framework (220–300 words). Conditions differ in organization, not information content, enabling a controlled assessment of structure's effect on reasoning quality [11]. The per-criterion scoring approach follows fine-grained evaluation principles established by MT-Bench-101 [14].

[1]Mini-Mental State Examination (MMSE): a 30-point cognitive screening instrument; scores below 24 indicate cognitive impairment.

[2]Polysomnography (PSG): an overnight sleep study recording brain activity, eye movements, heart rate, and muscle activity, used to diagnose sleep disorders.

C. Evaluation Rubric and Model Specifications

Rubric A (clinical, max 30 pts): six criteria at 1–5 each, with each score level explicitly defined and requiring a direct response excerpt as justification, grounded in AASM clinical practice standards [13]. Table II summarises the criteria. A second general-quality rubric is under development and will be reported in future work.

Table III lists the five models evaluated. All models were queried via API.

V. Results

A. Prompt 1: Unstructured Vignette

All five models converged on ISWRD or dementia-related circadian rhythm sleep-wake disorder as the primary diagnosis for this representative persona. Guideline fidelity scored 5/5 uniformly, reflecting adequate baseline clinical knowledge for a canonical presentation. DeepSeek-V2 and Claude 3.5 Sonnet produced the most mechanistically precise responses (27/30); GPT-4o and Gemini 1.5 Pro scored 26/30; Llama 3 70B scored 24/30. Table IV presents criterion-level scores.

B. Prompt 2: Structured Clinical Template

Adding the physician-derived template produced consistent improvements. Largest gains appeared in personalization and safety flagging. GPT-4o, DeepSeek-V2, and Claude 3.5 Sonnet each reached 29/30. Llama 3 70B reached 25/30 using template fields as section labels rather than reasoning scaffolds. Gemini 1.5 Pro did not complete P2 (Section V-E). Tables V and VI present scores and deltas.

C. The Differential Diagnosis Gap

The most clinically significant finding in this pilot evaluation, observed across both prompt conditions and all five models for this representative persona, is a consistent failure to replicate *differential diagnosis*: the reasoning process that distinguishes clinical medicine from probabilistic text classification. These findings are preliminary and should not be generalized before broader validation across the full 224-profile corpus.

In clinical practice, a physician does not stop at identifying the most probable diagnosis. As reflected in the AASM 2016 physician benchmark [13], real diagnostic reasoning requires structured consideration and elimination of alternative explanations: acknowledging competing diagnoses, stating the evidentiary basis for acceptance or exclusion, and documenting that reasoning transparently.

While all five models demonstrated strong pattern recognition, none replicated this process for this representative case. Specifically: RBD was not retained as an open differential by any model with acknowledgment of PSG inconclusiveness; the physician benchmark [13] documented that RBD could not be excluded precisely because no REM sleep was recorded during PSG, a finding every model ignored. OSA was not documented as excluded through AHI-threshold reasoning. No model flagged the AASM 2015 contraindication of sedative-hypnotics in elderly dementia patients despite recommending

TABLE I
TEVV PIPELINE METRICS (INITIAL TESTING, ONE REPRESENTATIVE CASE, TWO PROMPT CONDITIONS, FIVE MODELS, TEN SCORED RESPONSES)

Metric	Score	Interpretation
Factual Consistency	87.8%	Ragas; proportion of response claims supported by citations from retrieved documents
Retrieval Quality	94.0%	Ragas; cosine similarity between retrieved documents and query
Safety Score	100%	No AASM contraindication flags triggered across 10 scored responses (2 prompts × 5 models); pilot scope only
Guideline Adherence	91.0%	Rule-based alignment with AASM clinical practice guidelines
Communication Quality	3.0/5	Structured review of empathy, readability, and utility

TABLE II
RUBRIC A: CLINICAL EVALUATION CRITERIA (1–5 PTS EACH; MAX 30)

Criterion	What It Measures
Guideline Fidelity	Alignment with AASM, NIH, ICSD-3 (International Classification of Sleep Disorders)
Diagnostic Reasoning	Differential construction; disorder-specific logic
Personalization	Comorbidities, age, cognition, medications
Comm. Utility	Actionability for non-specialist caregivers
Safety Flagging	Fall risk, polypharmacy (concurrent use of multiple medications), contraindications
Output Compliance	Format, word count, template use (P2)

TABLE III
MODEL IDENTIFIERS AND GENERATION PARAMETERS

Model	Identifier	Access	Temp.	Max tok.
GPT-4o	gpt-4o-2024-05-13	Nov 2024	0.0	1,024
Gemini 1.5 Pro	gemini-1.5-pro-001	Nov 2024	0.0	1,024
DeepSeek-V2	deepseek-chat	Nov 2024	0.0	1,024
Claude 3.5 Sonnet	claude-3-5-sonnet-20241022	Nov 2024	0.0	1,024
Llama 3 70B	Meta-Llama-3-70B-Instruct	Nov 2024	0.0	1,024

TABLE IV
RUBRIC A SCORES: PROMPT 1 (UNSTRUCTURED VIGNETTE)

Model	G	D	P	C	S	O	Tot.
DeepSeek-V2	5	5	4	4	4	5	**27**
Gemini 1.5 Pro	5	5	4	4	3	5	**26**
GPT-4o	5	4	4	5	4	4	**26**
Claude 3.5 Sonnet	5	5	4	4	4	5	**27**
Llama 3 70B	5	4	4	4	3	4	**24**
Dr. Sharkey	*Ground truth: full diff. diagnosis*						—

G=Guideline, D=Dx Reasoning, P=Persona, C=Comm., S=Safety, O=Output

TABLE V
RUBRIC A SCORES: PROMPT 2 (STRUCTURED TEMPLATE)

Model	G	D	P	C	S	O	Tot.
DeepSeek-V2	5	5	5	5	5	4	**29**
Gemini 1.5 Pro	*Safety filter refusal*						**N/A**
GPT-4o	5	5	5	5	4	5	**29**
Claude 3.5 Sonnet	5	5	5	4	5	5	**29**
Llama 3 70B	5	4	4	5	4	3	**25**

pharmacological management. No model replicated the off-guideline prescribing logic: acknowledging the contraindication, selecting immediate-release melatonin, initiating at 3 mg with monitoring for depressive symptoms, and titrating to 1.5 mg based on response.

Table VII maps this gap across template domains under P2. Rows 8–9 (shaded), representing the unresolved RBD differential and off-guideline treatment rationale, were activated by the physician benchmark only. This pattern of failure directly motivates the architectural design of the multi-agent framework: the classification-then-dispatch structure, domain-scoped specialist agents, and TEVV-enforced guideline retrieval described in Section III-B are each designed to make this reasoning process structurally mandatory rather than prompt-elicited.

D. Per-Model Observations

DeepSeek-V2 produced the most mechanistically precise P1 response, correctly distinguishing confusional arousals as goal-directed NREM (Non-Rapid Eye Movement) behaviors from REM (Rapid Eye Movement) dream enactment, yet failed to engage the RBD diagnostic uncertainty from the inconclusive PSG. Under P2, DeepSeek introduced sundowning (late-afternoon confusion characteristic of dementia, distinct from ISWRD's full 24-hour fragmentation), misrepresenting the patient profile. **Claude 3.5 Sonnet** demonstrated the strongest hallucination classification, correctly typing hypnagogic (occurring at sleep onset) and hypnopompic (occurring on awakening) phenomena in both conditions. **GPT-4o** was the only model to flag medication review as a required diagnostic step and connect hypothyroidism to the thyroid/medication review domain. **Llama 3 70B**'s use of template fields as section labels rather than reasoning constraints produced structurally reorganized but substantively unchanged responses.

TABLE VI
SCORE DELTA: PROMPT 2 MINUS PROMPT 1

Model	P1	P2	Δ	Key Observation
DeepSeek-V2	27	29	+2	Explicit domain mapping
Gemini 1.5 Pro	26	–	N/A	Policy refusal
GPT-4o	26	29	**+3**	Largest gain; best caregiver framing
Claude 3.5 Sonnet	27	29	+2	High baseline; gains in safety
Llama 3 70B	24	25	+1	Template as labels only

TABLE VII
TEMPLATE DOMAIN ACTIVATION UNDER PROMPT 2. ✓=ACTIVATED; ×=MISSED; N/A=REFUSED. SHADED ROWS: DIFFERENTIAL DIAGNOSIS GAP (PHYSICIAN ONLY).

Domain	DS	GP	Ll	Cl	Ge	MD*
Nocturnal awakenings	✓	✓	✓	✓	–	✓
Parasomnia history	✓	✓	✓	✓	–	✓
Daytime sleepiness	✓	✓	✓	✓	–	✓
Naps (non-refreshing)	✓	✓	✓	×	–	✓
Narcolepsy spectrum	✓	×	×	✓	–	✓
Sleep-disordered beh.	✓	✓	×	✓	–	✓
Thyroid/medication	×	✓	×	×	–	✓
RBD diff. (unresolved PSG)	×	×	×	×	–	✓
Off-guideline Tx rationale	×	×	×	×	–	✓

DS=DeepSeek, GP=GPT-4o, Ll=Llama 3 70B, Cl=Claude, Ge=Gemini
*MD=Dr. Sharkey, AASM 2016 [13]

E. Gemini 1.5 Pro: Safety Filter Refusal

Gemini 1.5 Pro refused P2 entirely due to safety triggers on domains including hallucinations and self-harm risk, standard components of any geriatric sleep intake. This is a governance finding: Gemini scored 26/30 on P1, confirming capability. Clinical deployment of general-purpose LLMs requires system-level routing that signals clinical context to safety filters before use in diagnostic workflows.

VI. DISCUSSION

A. Structured Prompting as a Clinical Reasoning Scaffold

The consistent P2 improvements support the interpretation that physician-derived classification templates function as reasoning scaffolds rather than formatting constraints, extending chain-of-thought findings [11] to clinically structured domains where the chain structure derives from physician practice. Llama 3 70B's minimal gain demonstrates that clinical deployment cannot assume template use implies scaffold behavior without per-model verification.

B. Why the Architecture Addresses the Gap

The differential diagnosis gap is not a knowledge limitation: guideline fidelity scores of 5/5 confirm adequate clinical knowledge across all five models. The failure is procedural; models do not spontaneously deploy systematic differential elimination in a single-agent setting.

The multi-agent architecture addresses this through classification-then-dispatch. The processing-layer disorder classifier enforces an explicit categorization step before reasoning begins. The selected Disorder-Specific Clinical Agent reasons within a domain-scoped knowledge boundary defined by its system prompt. The Doctor Agent synthesizes the output under TEVV-enforced AASM guideline retrieval, ensuring contraindication logic and off-guideline prescribing rationale are present in the generation context. This classify-scope-enforce structure makes differential reasoning architecturally mandatory rather than prompt-elicited.

C. Limitations

This study evaluates one representative persona; findings are preliminary and cannot be generalized before broader corpus validation. Rubric scores were produced by the research team in a single evaluation round without independent clinical rater calibration. The system has not been evaluated with real caregivers or in live clinical settings. TEVV metrics represent early-stage performance on one case; cosine-similarity retrieval quality is a known limitation and more robust methods are planned. The full multi-agent architecture, including inter-agent communication and longitudinal memory, is under development. The physician-authored template reflects a single clinician's structure and has not been validated across multiple physicians.

D. Validation Roadmap

Validation proceeds in three phases. **Phase 1 (in preparation):** The pilot is extended to a stratified sample of the 224-profile corpus spanning at least four disorder types (ISWRD, RBD, Sleep Apnea, Insomnia) with contrasting etiologies, testing whether the differential diagnosis gap is consistent across presentations or specific to the dementia-comorbid ISWRD case evaluated here. A second general-quality rubric is under development and will be reported alongside these results. **Phase 2 (forthcoming):** Once inter-agent routing is operational, the same sample is re-evaluated under the full classification-then-dispatch architecture to assess whether structured decomposition reduces the gap relative to the single-agent baseline. **Phase 3 (forthcoming):** Independent physician rater scoring using a standardized protocol, with Dr. Burman and additional domain specialists, will provide the clinical validation of system appropriateness in the NIST sense: confirming that outputs meet the standard of care expected by practicing clinicians.

VII. Conclusion

This paper presented a multi-agent AI framework for clinical decision support using sleep disorder management as a controlled testbed, subject to active research and development. Three components are established: a 224-profile combinatorial persona corpus, 24-month longitudinal care pathway modeling, and a closed-world TEVV assessment pipeline. The broader multi-agent architecture is under active development; this paper reports a pilot evaluation of the Doctor Agent role as a foundation for that system.

The improvements observed with physician-derived structured prompting were consistent across all five frontier models, with largest gains in personalization and safety flagging. A consistent differential diagnosis gap was observed: all five models correctly identified the presented diagnosis, but none replicated the physician benchmark's differential reasoning, specifically management of diagnostic uncertainty from an inconclusive PSG, contraindication checking, and off-guideline prescribing rationale. These preliminary findings motivate the three-stage architectural design and the validation roadmap described in this paper.

References

[1] Y. H. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "MDAgents: An adaptive collaboration of LLMs for medical decision-making," in *Adv. Neural Inf. Process. Syst.*, 2024. (arXiv:2404.15155)

[2] J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu, "Agent Hospital: A simulacrum of hospital with evolvable medical agents," 2024. (arXiv:2405.02957)

[3] M. V. Vitiello, "Effective treatment of sleep disturbances in older adults," *Clinical Cornerstone*, vol. 9, no. Suppl 2, pp. S28–S35, 2009.

[4] J. W. Ayers et al., "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA Internal Medicine*, vol. 183, no. 6, pp. 589–596, 2023.

[5] K. Kreimeyer et al., "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review," *J. Biomedical Informatics*, vol. 73, pp. 14–29, 2017.

[6] OpenAI, "GPT-4 technical report," 2023. (arXiv:2303.08774)

[7] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[8] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023.

[9] S. Zhang et al., "MedAgent-Zero: Zero-shot medical reasoning with multi-agent collaboration," 2024. (arXiv:2404.02948)

[10] J. Hu, A. Wang, Q. Xie, H. Ma, Z. Li, and D. Guo, "AgentMental: An interactive multi-agent framework for explainable and adaptive mental health assessment," 2025. (arXiv:2508.11567)

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24824–24837, 2022.

[12] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," 2023. (arXiv:2309.15217)

[13] K. M. Sharkey, "Case study: Irregular sleep-wake rhythm disorder," *American Academy of Sleep Medicine*, Darien, IL, Jun. 2016. [Online]. Available: https://aasm.org

[14] G. Bai et al., "MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues," 2024. (arXiv:2402.14762)