

LUMINA : LongitUdinal MultImodal reasoNing Agent for Medical Imaging

Zhaoqi An, Yunchang Xie, Xiaobing Yu, Yuyi Yang, Ahmad Chowdhury, Peijie Qiu*

Washington University in St. Louis

{a.zhaoqi, x.yunchang, xiaobing, y.yuyi, a.chowdhury, peijie.qiu}@wustl.edu

Website: <https://visualreasoner.github.io/LUMINA>

Code: <https://github.com/VisualReasoner/LUMINA>

Correspondence: Peijie Qiu (peijie.qiu@wustl.edu)

Abstract

Multimodal large language models (MLLMs) have shown strong performance in perception and question-answering, but many clinical decisions require reasoning across observations collected over time rather than from a single input. We study this problem in longitudinal modeling, where evidence evolves across visits. We introduce **LUMINA** (LongitUdinal MultImodal reasoNing Agent), a single-agent framework that incrementally builds an auditable evidence state before making predictions. Instead of predicting directly from all observations, LUMINA organizes current evidence, compares observations across prior visits, maintains longitudinal memory, optionally retrieves relevant cross-subject summaries, and verifies whether the accumulated evidence supports the final decision. We evaluate LUMINA on two subject-level benchmarks: Alzheimer’s disease classification and 36-month progression forecasting from observed visit prefixes. Across nine MLLM backbones, LUMINA consistently outperforms MLLM baselines and an adapted medical reasoning agent, achieving 13.8%–18.5% relative balanced accuracy gains on classification and 16.8%–24.8% gains on progression forecasting over the strongest non-agent baseline. Additional evaluations preserve the same method ordering on Parkinsonian-state classification and breast cancer treatment-response prediction.

1 Introduction

Recent studies have shown strong capabilities of multimodal large language models (MLLMs) in perception, instruction following, and reasoning over static inputs (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023; Zhang et al., 2024). However, many real-world tasks require reasoning

across time: maintaining state, comparing historical evidence, selecting relevant observations, and deciding when evidence is sufficient for a final prediction. We refer to this setting as *longitudinal reasoning*. Such problems arise when observations are collected over repeated visits, modalities are heterogeneous or partially observed, and temporal change is critical to decision-making, especially in medical imaging. Disease progression modeling is a representative example, as biomarkers and clinical evidence evolve across longitudinal trajectories (Jack et al., 2010, 2018).

Longitudinal clinical reasoning combines comparison with recent prior observations and integration of longer-term trajectory evidence. This motivates an agentic formulation in which evidence selection, memory update, and decision sufficiency are represented explicitly rather than left implicit in a long prompt. In contrast, most current MLLMs collapse memory management, evidence selection, and prediction into a single prompt over the current input and loosely formatted history, making the reasoning process difficult to control and interpret.

Long contexts can impair performance even with perfect retrieval (Du et al., 2025). For tasks requiring grounding, evidence selection, memory, and verification, we therefore expose these operations explicitly rather than rely on a single long prompt. Recent agentic systems show that LLMs can be scaffolded with structured planning, memory, control, and verification modules, allowing them to handle problems that are harder to solve with direct prompting alone (Zhou et al., 2023; Sumers et al., 2024; Masterman et al., 2024; Liu et al., 2024; Xu, 2026). Across these systems, the key idea is not simply to add more steps, but to organize reasoning into reusable components that guide what the model attends to, stores, checks, and acts on

*This work was done prior to joining Amazon.

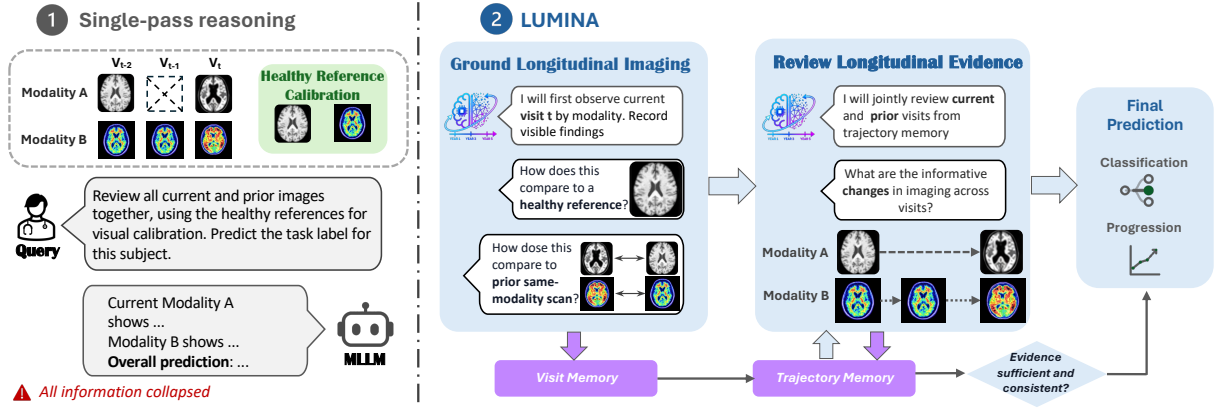


Figure 1: Comparison between collapsed static prompting (left) and LUMINA’s structured longitudinal reasoning (right). Both paradigms receive the same routed multimodal subject history; when supplied by the benchmark protocol, both also receive the same reference comparators for interpretation calibration.

(Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024; Fu et al., 2025). Most current agentic frameworks are designed for general tool use, open-ended task solving, or staged self-correction rather than specifically for evidence distributed across irregular visits, partially observed modalities, and temporally evolving trajectories. Longitudinal reasoning therefore calls for an agent design in which memory construction, temporal comparison, and evidence sufficiency are explicit first-class operations rather than implicit side effects of a long prompt.

In this paper, we formulate longitudinal multimodal reasoning as an *agentic memory-and-control problem*. To this end, we propose **LUMINA** (LongitUdinal MultImodal reasoniNg Agent), which implements this idea through a staged inference loop. Figure 1 introduces the central contrast at a conceptual level. LUMINA differs primarily in how subject-specific longitudinal evidence is organized before prediction: the visible history is converted into an explicit comparison-and-memory process rather than collapsed into a single prompt. LUMINA additionally supports optional retrieval from a frozen cross-subject memory bank constructed from disjoint subjects, whose contribution is isolated in the ablation analysis. It first inspects the subject’s history to determine which modalities and prior anchors to query, grounds the current visit, calibrates visual interpretation using shared reference comparators when supplied by the benchmark protocol, compares selected modalities against subject-specific same-modality anchors, and reconciles the resulting findings into a visit-level evidence state. LUMINA maintains a com-

pact trajectory memory that preserves the most relevant events and can optionally retrieve analogous trajectories from a frozen cross-subject bank disjoint from the evaluation cohort. Before producing a final prediction, the agent builds a task-specific belief state and runs a bounded audit-and-repair step when the current evidence is incomplete or internally inconsistent. The backbone MLLM remains the reasoning engine, but it now operates over a structured, auditable agent state rather than an unstructured concatenation of longitudinal observations.

Our contributions are threefold: *i)* We formulate longitudinal multimodal reasoning for medical imaging as a memory-and-control problem and propose LUMINA, which constructs a compact evidence state via anchor comparison, trajectory memory, and optional cross-subject retrieval. *ii)* We develop a multi-timescale longitudinal memory for local change and long-term trajectories, together with a bounded controller for grounding, comparison, reconciliation, and belief formation. *iii)* We evaluate LUMINA on disease classification and progression forecasting tasks across diverse MLLM backbones, consistently outperforming MLLM baselines and an adapted medical reasoning agent.

2 Related Work

LUMINA builds on the emerging view that LLM-powered agents are structured systems with explicit memory, control, and verification (Zhou et al., 2023; Sumers et al., 2024; Masterman et al., 2024; Liu et al., 2024; Xu, 2026). Our work is particularly aligned with the argument that memory design and

evidence control are first-class modeling choices.

Reasoning-action and self-correction. ReAct established the usefulness of interleaving reasoning and action (Yao et al., 2023). Reflexion, Self-Refine, and CRITIC showed that staged feedback, verification, and repair can improve reliability without parameter updates (Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024). We inherit the motivation for bounded control and repair, but our contribution is the longitudinal evidence-construction architecture built around those ideas.

Medical multimodal reasoning. Recent biomedical MLLMs and generalist medical AI systems have expanded multimodal reasoning over medical images and biomedical context (Tu et al., 2024; Li et al., 2023a; Moor et al., 2023; Sellergren et al., 2025; Pan et al., 2025). Agentic medical workflows such as MedAgent-Pro further show that multi-step evidence organization can improve medical reasoning over multimodal inputs (Wang et al., 2026). Many multimodal medical systems are nevertheless designed around cross-sectional inputs. LUMINA instead constructs explicit evidence states across irregular visits. The central challenge is to compare informative prior anchors, compress longer-horizon history, and turn those observations into a task-specific judgment without exposing direct label-defining clinical fields to the agent.

Reference-guided vision-language reasoning. Recent work has studied reference image-guided comparative diagnosis, where a medical MLLM receives both a query image and a matched healthy reference image to improve medical image interpretation (Jin et al., 2025). This line of work is closely related to our use of reference comparators for interpretation calibration. A complementary line of MLLM calibration work argues that visual uncertainty and reasoning uncertainty should be treated separately, because an incorrect answer can arise from a perceptual failure or from a reasoning error given correct perception (Xiao et al., 2026).

Sequential Monte Carlo. Sequential Monte Carlo (SMC) has a long history as a principled method for approximate inference under sequential uncertainty (Gordon et al., 1993; Doucet et al., 2001). We use it here as a trajectory-memory manager, not as a generic search procedure. Cross-subject memory complements this within-subject selection by retrieving analogous longitudinal summaries from a frozen, subject-disjoint bank under leakage-controlled evaluation (Lewis et al., 2020; Kapoor and Narayanan, 2023). This framing motivates the

concrete design choices in LUMINA, which we describe next.

3 Method

LUMINA is designed for longitudinal multimodal reasoning rather than direct prediction from collapsed history. Its core contribution is to turn an irregular subject-history prefix into an explicit evidence state before final prediction.

We consider a subject s with an irregular visit sequence $\mathcal{V}_s = \{1, \dots, T_s\}$. Each visit t contains an arbitrary subset of imaging modalities,

$$X_{s,t} = \{x_{s,t,m} : m \in \mathcal{M}_{s,t}\}, \quad \mathcal{M}_{s,t} \subseteq \mathcal{M},$$

where $\mathcal{M}$ denotes the full modality set. At target visit t , the model observes only the subject-history prefix $X_{s,\leq t}$ and associated pre-target context $c_{s,t}$. The goal is to construct a decision state from this partially observed longitudinal history and map it to the active benchmark label space.

For each subject-history prefix, LUMINA constructs an evidence state

$$\begin{aligned} \omega_{s,t} &= (C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}}, \mathcal{T}_{s,t}, \\ &\quad M_{s,t}^{\text{cross}}, E_{s,t}, S_{s,t}), \\ B_{s,t} &= f_{\theta}(\omega_{s,t}), \quad \hat{y}_{s,t} = \Pi(B_{s,t}). \end{aligned}$$

Here $C_{s,t}$ stores the grounded current-visit evidence, $M_{s,t}^{\text{visit}}$ stores local same-modality comparison evidence, $M_{s,<t}^{\text{traj}}$ stores compact prior within-subject trajectory evidence, and $M_{s,t}^{\text{cross}}$ stores optional cross-subject context. $E_{s,t}$ and $S_{s,t}$ denote the reconciled evidence and decision context used to form the belief state $B_{s,t}$. Current-visit observations may be visually calibrated using benchmark-supplied reference comparators $\mathcal{R}_{s,t}$ when available. The audit state $A_{s,t}$ is generated from the evidence and belief states, while Π deterministically normalizes the final belief into the label space.

The overall workflow of LUMINA is illustrated in Figure 2, with the corresponding procedure detailed in Algorithm 1.

3.1 Longitudinal Evidence Construction

Current-visit grounding. The first step grounds what is visible at the target visit before adding historical evidence. Given the selected modality set $\mathcal{M}_{s,t}^{\text{sel}}$ and task-specific inspection questions $\mathcal{Q}_{s,t}$, the MLLM extracts modality-level observations $O_{s,t}$ from the currently available studies. This

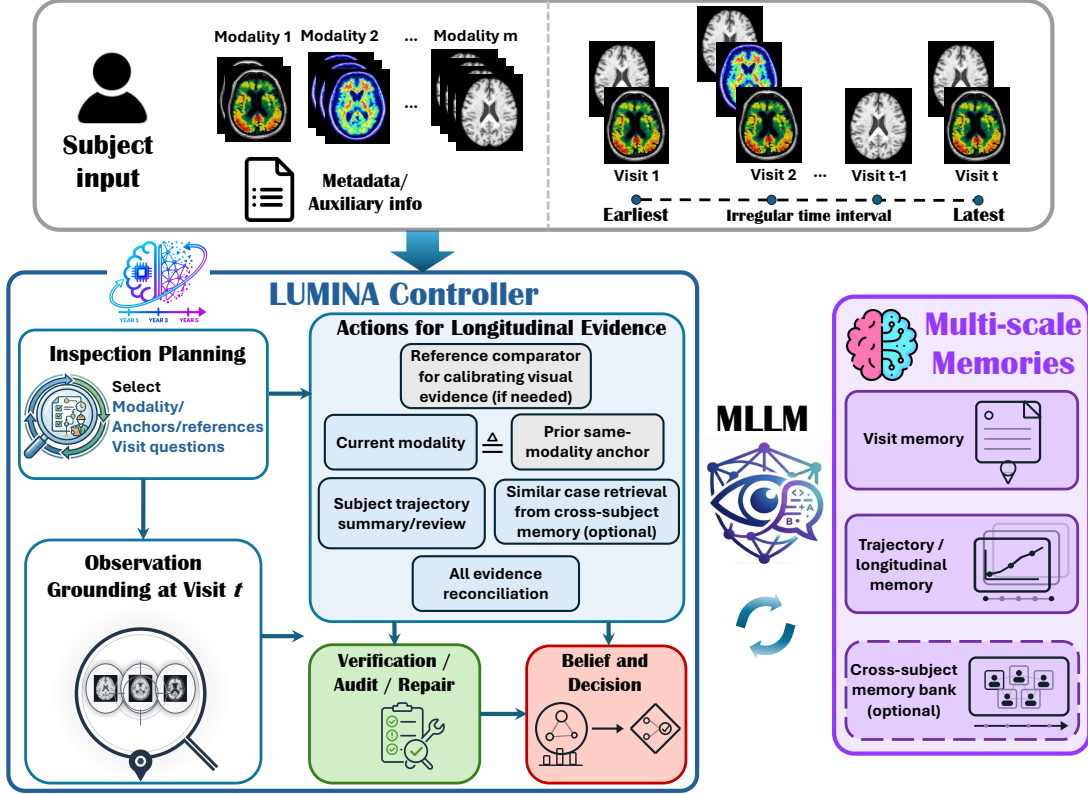


Figure 2: LUMINA as a longitudinal multimodal reasoning agent. The controller grounds the current visit, reviews past same-modality evidence, updates visit and trajectory memory, optionally retrieves cross-subject context, verifies evidence sufficiency, and then emits a task-specific belief state and benchmark prediction. When the benchmark supplies shared reference comparators, they are used only to stabilize current-visit visual interpretation and are provided to all compared methods.

grounding step records visible findings and uncertainty without yet making a longitudinal judgment.

Intra-modality evidence. For each selected modality m , LUMINA identifies the most recent valid prior study from the same subject and modality. It then compares the current evidence $C_{s,t,m}$ with this prior anchor $x_{s,a_m,m}$ and stores the resulting change direction, change magnitude, confidence, and uncertainty in $M_{s,t}^{\text{visit}}$. If no valid anchor exists, the visit memory stores the current evidence with an explicit unavailable-anchor flag rather than assigning an unsupported change estimate.

Visit memory. The visit-level evidence $M_{s,t}^{\text{visit}}$ is a short-term state for the current decision point. It keeps local evidence separated by modality, distinguishes current findings from within-subject change evidence, and records whether each comparison is supported, contradicted, or unavailable.

3.2 Multi-Timescale Memory

A straightforward pipeline would concatenate all prior visits into the prompt. LUMINA instead maintains memory at multiple timescales: visit memory

for the current decision point, compact trajectory memory $M_{s,t}^{\text{traj}}$ for longer-horizon within-subject patterns, and optional cross-subject memory for analogy-level context.

Trajectory memory. Each trajectory event is derived from grounded observations and local comparison outcomes. It summarizes what was concluded earlier: dominant modality changes, conflict patterns, unresolved uncertainty, and resulting belief shifts. Trajectory memory is therefore a compact trajectory representation built from the most informative subset of those prior events.

SMC-guided event selection. To retain informative historical events under a fixed memory budget, SMC-style compression decides which events remain active. Let $\mathcal{H}_{s,t}$ denote the set of historical events available for trajectory compression at target visit t . Each event $e \in \mathcal{H}_{s,t}$ receives a utility score

$$u(e) = \lambda_g g(e) + \lambda_c c(e) + \lambda_r r(e),$$

where $g(e)$ is the event-level change grade, $c(e)$ is event confidence, and $r(e)$ is an exponential recency term, $r(e) = \exp(-\ln 2 \cdot \Delta t(e)/\tau)$.

Algorithm 1: LUMINA subject-level inference

Input: Multimodal history prefix $X_{s,\leq t}$, context $c_{s,t}$, optional references $\mathcal{R}_{s,t}$, prior trajectory state $M_{s,<t}^{\text{traj}}$, memory bank $\mathcal{B}_{s,t}$, task profile

Result: Prediction $\hat{y}_{s,t}$ and auditable longitudinal trace

- 1 **Initialize:** $M_{s,t}^{\text{visit}} \leftarrow \emptyset$; $M_{s,t}^{\text{cross}} \leftarrow \emptyset$; $\mathcal{T}_{s,t} \leftarrow \emptyset$; $A_{s,t} \leftarrow \emptyset$;
 $\triangleright$ Plan selected modalities, evidence needs, and valid anchors.
- 2 $(M_{s,t}^{\text{sel}}, \mathcal{A}_{s,t}, \mathcal{Q}_{s,t}) \leftarrow \text{Inspect}(X_{s,\leq t}, c_{s,t})$;
- 3 $O_{s,t} \leftarrow \text{Observe}(X_{s,t}, \mathcal{Q}_{s,t})$;
- 4 **if** $\mathcal{R}_{s,t}$ is available **then**
- 5 $C_{s,t} \leftarrow \text{CalibrateVisualRead}(O_{s,t}, \mathcal{R}_{s,t})$;
- 6 **else**
- 7 $C_{s,t} \leftarrow O_{s,t}$;
- 8 $\triangleright$ Build local comparison evidence.
- 9 **foreach** $m \in \mathcal{M}_{s,t}^{\text{sel}}$ **do**
- 10 $a_m \leftarrow \text{GetAnchor}(m, \mathcal{A}_{s,t})$;
- 11 **if** $a_m \in \mathcal{A}_{s,t}$ **then**
- 12 $\Delta_{s,t,m} \leftarrow \text{ComparePastEvidence}(C_{s,t}, m, x_{s,a_m,m})$;
- 13 **else**
- 14 $\Delta_{s,t,m} \leftarrow \text{CurrentOnlyEvidence}(m, C_{s,t})$;
- 15 $M_{s,t}^{\text{visit}} \leftarrow M_{s,t}^{\text{visit}} \cup \{\Delta_{s,t,m}\}$;
- 16 $\triangleright$ Optionally review compact trajectory evidence.
- 17 **if** $\text{NEEDTRAJECTORYCONTEXT}(C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}})$ **then**
- 18 $\mathcal{T}_{s,t} \leftarrow \text{REVIEWTRAJECTORY}(M_{s,<t}^{\text{traj}})$;
- 19 $\triangleright$ Optionally retrieve cross-subject analogies.
- 20 **if** $\text{NEEDRETRIEVE}(C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}}, c_{s,t})$ **then**
- 21 $M_{s,t}^{\text{cross}} \leftarrow \text{RetrieveCrossSubject}(\mathcal{B}_{s,t}, C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}}, c_{s,t})$;
- 22 $\triangleright$ Reconcile multimodal evidence and finalize belief.
- 23 $E_{s,t} \leftarrow \text{ReconcileMultimodalEvidence}(C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}}, \mathcal{T}_{s,t}, M_{s,t}^{\text{cross}})$;
- 24 $S_{s,t} \leftarrow \text{BuildDecisionContext}(C_{s,t}, M_{s,t}^{\text{visit}}, M_{s,<t}^{\text{traj}}, M_{s,t}^{\text{cross}}, c_{s,t})$;
- 25 $B_{s,t} \leftarrow \text{FinalBelief}(C_{s,t}, M_{s,t}^{\text{visit}}, E_{s,t}, S_{s,t})$;
- 26 **if** $\text{HASCONTRADICTIONS}(B_{s,t}, S_{s,t})$ **then**
- 27 $B_{s,t} \leftarrow \text{VerifyAndRepair}(B_{s,t}, S_{s,t})$;
- 28 $A_{s,t} \leftarrow \text{Audit}(C_{s,t}, M_{s,t}^{\text{visit}}, E_{s,t}, B_{s,t})$;
- 29 $\triangleright$ Update compact trajectory memory.
- 30 $M_{s,\leq t}^{\text{traj}} \leftarrow \text{UpdateTrajectoryState}_{\text{SMC}}(M_{s,<t}^{\text{traj}}, C_{s,t}, M_{s,t}^{\text{visit}}, E_{s,t}, B_{s,t})$;
- 31 $\hat{y}_{s,t} \leftarrow \Pi(B_{s,t})$;
- 32 **return** $(\hat{y}_{s,t}, A_{s,t})$;

Utilities are normalized into event weights,

$$w(e) = \frac{u(e)}{\sum_{e' \in \mathcal{H}_{s,t}} u(e')},$$

and the trajectory manager keeps the top- K active events by weight. We refer to this as SMC-style compression because it uses importance weighting in the spirit of sequential Monte Carlo while functioning here as a compact event-selection layer rather than a full latent-state particle filter. It is not a separate classifier and does not access labels.

Cross-subject memory. When the current subject history is limited or ambiguous, LUMINA can retrieve compact summaries from a frozen cross-subject memory bank with subject-disjoint construction. Retrieval uses a deterministic similarity search over numeric signatures derived from allowed observations, trajectory statistics, modality

availability, and controller-trace features. Retrieved records contain compact longitudinal summaries and audit patterns rather than target labels, future outcomes, cognitive scores, or label-defining clinical variables. This component provides analogy-level context rather than label retrieval.

3.3 Evidence Reconciliation and Final State

Evidence reconciliation. LUMINA reconciles current observations $C_{s,t}$, local same-modality change evidence $M_{s,t}^{\text{visit}}$, compact trajectory memory $M_{s,<t}^{\text{traj}}$, reviewed trajectory context $\mathcal{T}_{s,t}$, and optional cross-subject context $M_{s,t}^{\text{cross}}$ into a task-specific evidence state $E_{s,t}$. The reconciliation stage explicitly records which evidence sources agree, which conflict, and which uncertainties remain unresolved.

Final belief and verification. The reconciled evidence state is passed to a task-aware final belief prompt. Before prediction, the controller runs one bounded verification pass only when critical evidence is missing or contradictions remain.

4 Experiments

4.1 Experimental setup

Dataset. The main benchmark experiments use the Alzheimer’s Disease Neuroimaging Initiative (ADNI) research study (Mueller et al., 2005; Petersen et al., 2010; Weiner et al., 2017). We construct longitudinal subject histories from structural MRI, amyloid PET, and tau PET acquisitions, together with limited contextual metadata. The indexed cohort contains 1,512 visit rows from 685 subjects, from which task-specific target selection yields 380 AD-continuum targets and 401 progression-forecasting targets. Diagnosis fields, cognitive scores, and other variables used directly for target selection or label construction are excluded from model inputs. During inference, the agent observes only the available imaging records and limited demographic information available before making a final decision. Appendix D describes target selection and cohort filtering.

We converted raw 3D DICOM imaging series into standardized volumetric inputs and represented each scan with three canonical 2D views: axial, coronal, and sagittal center slices, which we treat as a 2.5D representation. After reorientation, intensity normalization, and resizing, the resulting three views provide a common visual representation across MRI and PET. Appendix C details

modality-specific preprocessing.

Tasks and Evaluation. We evaluate two benchmarks from the same longitudinal cohort in the Alzheimer’s disease (AD) domain. Both use the multimodal history visible up to a target visit as input, but differ in prediction target. The AD-continuum classification benchmark predicts the current state: cognitively unimpaired subjects (CU), Preclinical AD, Symptomatic AD. The progression-forecasting benchmark predicts a held-out future outcome bucket: Stable, Mild progression, Clear progression over a fixed 36-month post-target horizon. Future follow-up is used only for label assignment in progression forecasting and is hidden from the agent during inference.

Reference comparators for interpretation calibration. We use shared reference comparators as a benchmark-level interpretation-calibration protocol rather than as a LUMINA-specific reasoning module. The comparator pool is constructed offline from cognitively unimpaired subjects and frozen before evaluation. Within this fixed pool, target-specific comparators are matched separately by modality, using age as the primary variable and sex as the secondary variable. Matching for an evaluated subject does not use that subject’s diagnosis fields, cognitive scores, target label, progression label, or future outcomes. The comparators provide shared visual anchors for interpreting protocol-consistent MRI anatomy and low-burden PET uptake and are supplied under the same protocol to all compared methods.

Model Comparison. All methods are evaluated on the same targets with the same 2.5D study views, the same shared reference comparators, and the same visible metadata. Reference selection is frozen before inference. MedAgent-Pro is adapted to the same routed longitudinal prefix and shared reference comparators, but without LUMINA’s controller and memory design. When cross-subject memory is used, retrieval is restricted to a frozen subject-disjoint bank, and online updates are disabled during evaluation to prevent label leakage (Kapoor and Narayanan, 2023; Rosenblatt et al., 2024). Balanced accuracy is the primary metric due to multi-class imbalance, with Macro-F1 reported as a secondary metric. Main tables report subject-level bootstrap 95% confidence intervals for balanced accuracy, with full Macro-F1 uncertainty in the appendix.

4.2 Main Results

Across the nine evaluated backbones, LUMINA is the strongest method on both benchmarks (Table 1); Figure 3 summarizes the GPT-4o-mini comparison. Relative to the best non-agent baseline, LUMINA improves balanced accuracy by 9.46–11.89 points on AD-continuum classification and by 10.63–14.27 points on progression forecasting. The gains are not restricted to a single backbone family: they appear for Qwen3-VL thinking and instruct variants, medical MLLMs, and GPT-4o-mini. MedAgent-Pro is generally stronger than the non-agent baselines, but remains below LUMINA, suggesting that generic multi-step medical reasoning is helpful but does not replace explicit longitudinal evidence construction.

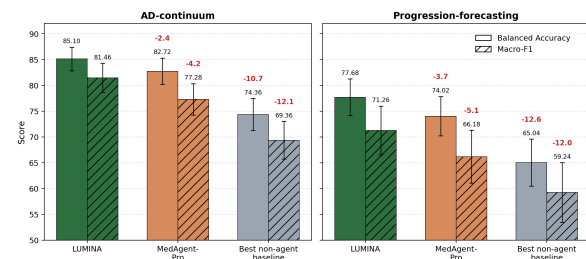


Figure 3: LUMINA improves over both MedAgent-Pro and the best non-agent baseline on balanced accuracy and Macro-F1 across the two tasks in the GPT-4o-mini setting. Red labels indicate the score gap relative to LUMINA for the corresponding method and metric.

The best non-agent baseline is selected metric-wise from *MLLM Visit Sequence* and *MLLM Bundle*. To avoid conflating this selection with method improvement, Appendix Tables 12 and 13 report the full balanced-accuracy values for both non-agent baselines.

4.2.1 AD-Continuum Classification

LUMINA improves performance on this task across both balanced accuracy and Macro-F1. With GPT-4o-mini, LUMINA reaches 85.10 balanced accuracy and 81.46 Macro-F1, compared with 82.72 and 77.28 for MedAgent-Pro and 74.36 and 69.36 for the best non-agent baseline. The per-class analysis in Appendix Table 15 shows that the improvement is not limited to the easiest categories. LUMINA performs well on CU and Symptomatic AD, while the Preclinical AD class remains the hardest class, consistent with its intermediate biological and clinical profile.

Interpretation calibration is particularly important for this task. Appendix Table 20 shows that

Table 1: Main benchmark results on the AD-continuum and progression-forecasting tasks. Balanced accuracy is reported as the point estimate with the 95% bootstrap confidence-interval half-width. For each backbone, task, and metric, the reported best non-agent value is the higher value between *MLLM Visit Sequence* and *MLLM Bundle*. Future follow-up labels are excluded from the agent input in progression forecasting. Full Macro-F1 uncertainty is reported in Appendix Tables 9 and 10.

Backbone	Balanced Accuracy			Macro-F1		
	LUMINA	MedAgent-Pro	Best non-agent	LUMINA	MedAgent-Pro	Best non-agent
AD-continuum						
Qwen3-VL-2B-Thinking	75.49 $\pm$ 3.08	72.94 $\pm$ 3.21	66.03 $\pm$ 3.62	71.58	67.36	59.40
Qwen3-VL-2B-Instruct	75.95 $\pm$ 2.91	71.60 $\pm$ 3.36	64.08 $\pm$ 3.58	70.84	64.88	57.82
Qwen3-VL-4B-Thinking	80.38 $\pm$ 2.68	77.28 $\pm$ 2.92	69.42 $\pm$ 3.41	77.10	72.42	64.06
Qwen3-VL-4B-Instruct	77.64 $\pm$ 2.86	75.76 $\pm$ 3.05	66.51 $\pm$ 3.69	73.06	70.88	61.94
Qwen3-VL-8B-Thinking	83.21 $\pm$ 2.39	79.75 $\pm$ 2.77	72.06 $\pm$ 3.22	79.42	74.02	66.18
Qwen3-VL-8B-Instruct	82.73 $\pm$ 2.54	79.46 $\pm$ 2.81	70.84 $\pm$ 3.35	78.12	73.46	65.02
MedGemma-4B	80.59 $\pm$ 2.63	77.85 $\pm$ 2.89	70.82 $\pm$ 3.36	76.54	72.90	63.88
MedVLM-R1-2B	80.03 $\pm$ 3.01	75.11 $\pm$ 3.20	68.24 $\pm$ 3.63	74.06	69.12	62.14
GPT-4o-mini	85.10 $\pm$ 2.28	82.72 $\pm$ 2.55	74.36 $\pm$ 3.11	81.46	77.28	69.36
Progression-forecasting						
Qwen3-VL-2B-Thinking	69.05 $\pm$ 4.41	66.85 $\pm$ 4.55	58.42 $\pm$ 5.10	60.72	57.16	50.84
Qwen3-VL-2B-Instruct	69.98 $\pm$ 4.34	65.44 $\pm$ 4.66	56.06 $\pm$ 5.28	59.84	55.42	50.22
Qwen3-VL-4B-Thinking	73.52 $\pm$ 4.01	70.88 $\pm$ 4.18	61.18 $\pm$ 4.86	66.08	61.34	54.82
Qwen3-VL-4B-Instruct	72.05 $\pm$ 4.20	69.72 $\pm$ 4.38	59.62 $\pm$ 5.04	63.22	60.78	53.64
Qwen3-VL-8B-Thinking	76.42 $\pm$ 3.67	72.96 $\pm$ 3.93	62.15 $\pm$ 4.72	69.10	64.80	57.96
Qwen3-VL-8B-Instruct	75.39 $\pm$ 3.86	71.82 $\pm$ 4.09	61.28 $\pm$ 4.89	67.44	62.16	55.48
MedGemma-4B	74.89 $\pm$ 3.91	72.47 $\pm$ 4.13	64.12 $\pm$ 4.61	68.06	64.22	56.22
MedVLM-R1-2B	72.44 $\pm$ 4.27	69.10 $\pm$ 4.46	59.28 $\pm$ 5.18	62.58	59.94	54.12
GPT-4o-mini	77.68 $\pm$ 3.55	74.02 $\pm$ 3.81	65.04 $\pm$ 4.55	71.26	66.18	59.24

without reference comparators, the balanced accuracy of *MLLM Visit Sequence* and LUMINA falls to 43.3 and 48.7, respectively; under the standard interpretation-calibration protocol, the corresponding scores are 74.4 and 85.1. The prediction-distribution analysis in Appendix Table 21 shows that without reference comparators, *MLLM Visit Sequence* strongly over-predicts Preclinical AD. These results support the use of reference comparators as a benchmark-level interpretation-calibration protocol, especially for stabilizing PET interpretation; LUMINA’s main contribution remains the organization of the calibrated multimodal history into longitudinal evidence states.

4.2.2 Progression Forecasting

Progression forecasting is harder for all methods because the target label is defined from held-out future follow-up, while the model only observes the prefix available at the target visit. Absolute balanced accuracy and Macro-F1 are therefore lower than in AD-continuum classification. This task also requires a different kind of evidence organization: the model must use available modality patterns and visible longitudinal change to predict one of three held-out future-outcome categories: Stable, Mild progression, or Clear progression.

LUMINA therefore attains lower absolute scores than on AD-continuum classification, but larger gains over the comparison baselines. With GPT-4o-mini, LUMINA achieves 77.68 balanced accuracy and 71.26 Macro-F1, compared with 74.02 and 66.18 for MedAgent-Pro and 65.04 and 59.24 for the best non-agent baseline. The paired bootstrap comparison in Appendix Table 14 shows that the gain over the best non-agent baseline remains robust under subject-level resampling. The gain over MedAgent-Pro is smaller and less certain for progression forecasting, which is consistent with the greater difficulty of the task.

Here, macro-F1 is lower than balanced accuracy because precision errors remain concentrated in the harder progression classes. Appendix Table 16 shows that LUMINA improves recall across Stable, Mild progression, and Clear progression, but the Clear progression class remains precision-limited. This pattern suggests that LUMINA improves trajectory evidence construction without eliminating the intrinsic difficulty of predicting future progression from an incomplete prefix.

4.3 Ablation Studies

Table 2 ablates the main LUMINA components on the AD-continuum task for GPT-4o-mini and

Table 2: Ablations on the AD-continuum task. Balanced accuracy and Macro-F1 are reported with 95% bootstrap confidence-interval half-widths. Δ denotes the change in balanced accuracy relative to the full LUMINA system.

Variant	GPT-4o-mini			Qwen3-VL-8B-Thinking		
	Bal. Acc.	Δ Bal. Acc.	Macro-F1	Bal. Acc.	Δ Bal. Acc.	Macro-F1
LUMINA (full)	85.10 ± 2.28	—	81.46 ± 2.82	83.21 ± 2.39	—	79.42 ± 3.01
$\times$ anchored comparison	80.42 ± 2.91	-4.68	75.86 ± 3.62	76.83 ± 3.06	-6.38	73.64 ± 3.78
$\times$ trajectory memory	81.73 ± 2.76	-3.37	77.92 ± 3.34	79.04 ± 2.92	-4.17	75.18 ± 3.55
$\times$ cross-subject memory	82.44 ± 2.65	-2.66	78.16 ± 3.21	79.19 ± 2.78	-4.02	76.62 ± 3.36
$\times$ SMC compression	83.02 ± 2.55	-2.08	78.82 ± 3.16	80.16 ± 2.70	-3.05	76.04 ± 3.42
$\times$ evidence verification/repair	83.92 ± 2.34	-1.18	81.12 ± 2.91	81.36 ± 2.47	-1.85	79.06 ± 3.08

Table 3: Call-budget comparison with GPT-4o-mini. Static self-consistency repeats the bundled prompt and aggregates predictions by majority vote under the same average call budget as LUMINA.

Method	Avg. LLM calls	AD-continuum		Progression forecasting	
		Bal. Acc.	Macro-F1	Bal. Acc.	Macro-F1
MLLM Bundle	1.0	72.9	66.9	63.7	56.8
Static self-consistency	5.7	76.1	75.3	66.2	62.5
MLLM Visit Sequence	4.3	74.4	69.4	65.0	59.2
MedAgent-Pro	5.1	82.7	77.3	74.0	66.2
LUMINA	5.7	85.1	81.5	77.7	71.3

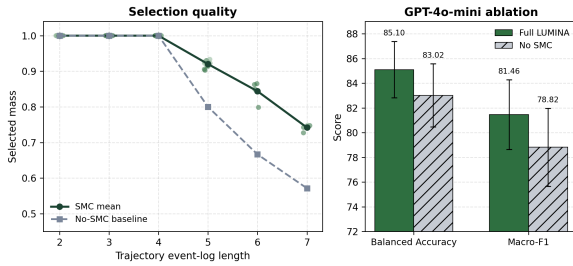


Figure 4: Selection quality of SMC-style trajectory compression. Left: selected mass captured by the active set versus event-log length, compared with a recency-only baseline using the same event budget. Right: GPT-4o-mini full-system and no-SMC ablations, with bootstrap 95% CI half-widths.

Qwen3-VL-8B-Thinking. Removing anchored comparison causes the largest degradation, followed by trajectory memory. This ordering supports the central design claim that LUMINA benefits most from making within-subject longitudinal comparison explicit. Cross-subject memory and SMC-style trajectory compression provide smaller but consistent gains, while evidence verification and repair have the smallest average effect in the current benchmark.

The SMC ablation measures the effect of memory quality under a fixed active-event budget. Figure 4 shows that the SMC-selected set captures more weighted history mass than a recency-only

baseline. Removing SMC causes a smaller but consistent drop across both backbones, indicating that SMC-style compression improves which historical events remain active, while anchored comparison and trajectory memory contribute the largest gains.

4.4 Inference-Budget Analysis

Table 3 compares methods under their observed inference budgets. Static self-consistency (Wang et al., 2023) improves over one-shot bundle prompting, confirming that repeated inference contributes to performance. Under the same average 5.7-call budget, however, LUMINA remains 9.0 balanced-accuracy points higher on AD-continuum classification and 11.5 points higher on progression forecasting. MLLM Visit Sequence reaches 74.4 and 65.0 balanced accuracy with chronological multi-call processing, while MedAgent-Pro reaches 82.7 and 74.0 with generic multi-step medical reasoning. LUMINA reaches 85.1 and 77.7, showing an additional gain from structured longitudinal evidence-state construction.

4.5 Credibility and Shortcut Analyses

Appendix I.1 evaluates paired subject-level differences, output normalization, metric-wise baseline selection, and possible shortcuts from metadata or modality availability.

Paired subject-level bootstrap tests give observed balanced-accuracy differences over the best non-

Table 4: External-transfer results with GPT-4o-mini. Values are balanced accuracy (%). Subjects are randomly sampled after task-specific eligibility filtering. PPMI is a three-class task and ACRIN-6698 is binary, so scores are interpreted within each dataset.

Dataset	Task	Subjects	Self-cons.	MedAgent-Pro	LUMINA
PPMI	HC / Prodromal / PD	100	71.8	78.5	79.7
ACRIN-6698	T0+T1 → pCR / non-pCR	100	71.4	79.3	83.9

agent baseline of +10.74 for AD-continuum classification and +12.64 for progression forecasting. The paired gains over MedAgent-Pro are smaller, especially for progression forecasting, where the confidence interval includes zero.

Shortcut controls show that metadata-only and modality-availability-only inputs remain far below full LUMINA. The unordered-image control is stronger than these non-visual controls but remains below full LUMINA. The gap is smaller for AD-continuum classification, where current imaging already carries strong disease-state signal, and larger for progression forecasting, where visit order and trajectory-level evidence are more central.

4.6 External-Transfer Evaluation

We evaluate LUMINA and matched baselines on 100 randomly sampled eligible subjects from each of PPMI (Marek et al., 2011) and ACRIN-6698 (Partridge et al., 2018; Newitt et al., 2021). PPMI provides a three-class classification setting spanning healthy controls, prodromal participants, and participants with Parkinson’s disease. ACRIN-6698 provides a longitudinal treatment-response setting in which pretreatment T0 and early-treatment T1 imaging are used to predict eventual pathologic complete response; the sampled cohort contains 46 pCR and 54 non-pCR subjects. Across the two external datasets, the controller action set, evidence schemas, within-subject multi-timescale memory, SMC-style compression, evidence reconciliation, and verification procedure remain unchanged. Cross-subject retrieval is disabled in both evaluations, so they test transfer of the within-subject evidence-construction core without external subject context. Only the dataset-specific modality/input mapping, task rubric, output label space, and reference policy where applicable are configured for each setting. Appendix E provides cohort, imaging, preprocessing, and input-exclusion details.

As shown in Table 4, LUMINA achieves 79.7 balanced accuracy on PPMI, compared with 71.8

for static self-consistency and 78.5 for MedAgent-Pro. On ACRIN-6698, LUMINA reaches 83.9, compared with 71.4 and 79.3, respectively. The resulting gains over static self-consistency are 7.9 and 12.5 points, while the gains over MedAgent-Pro are 1.2 and 4.6 points. LUMINA therefore preserves the method ordering observed on ADNI across both external-transfer settings while reusing the same within-subject evidence-construction structure and without using cross-subject retrieval.

5 Conclusion

We introduced LUMINA, a reasoning agent that constructs an explicit evidence state from longitudinal multimodal histories before final prediction. Its contribution is architectural and empirical: LUMINA makes evidence grounding, same-modality comparison, multi-timescale memory, multimodal reconciliation, and verification explicit operations within an auditable longitudinal state. On AD-continuum classification and progression forecasting, it improves over backbone-matched non-agent baselines and an adapted multi-step medical reasoning baseline. The matched-call analysis shows that this improvement is not explained by repeated inference alone, while the PPMI and ACRIN-6698 evaluations reuse the same within-subject evidence-construction core with dataset-specific task configuration and no cross-subject retrieval.

Limitations

There are three main limitations of the current work that can motivate future investigations. First, the primary benchmarks use ADNI, with external transfer to PPMI and ACRIN-6698 using 100 eligible subjects per dataset. Each setting uses dataset-specific modalities, labels, rubrics, and references; broader external validation remains future work. Second, current progression forecasting is prefix-based: future follow-up is withheld and used only for labels. Thus, it evaluates longitudinal reasoning rather than prospective clinical prognosis, with results also limited by class im-

balance. Third, each 3D study is represented by three center slices, which may miss volumetric patterns, though LUMINA could be paired with volume-aware encoders. Cross-subject memory requires subject-disjoint construction, and staged inference is slower than one-shot prompting. Reference calibration may also need cohort-specific designs. Finally, performance depends on the underlying MLLM, and the system is not clinically validated.

Ethical considerations

This work uses de-identified research-cohort data and evaluates retrospective disease-state classification and progression-forecasting benchmarks. LUMINA is not a clinical diagnostic system and should not be used for patient care without prospective validation, expert review, and site-specific governance. Although the system excludes direct label-defining clinical variables from prompts, demographic covariates, modality-availability patterns, and benchmark-supplied reference comparators may still encode cohort biases and therefore require explicit monitoring. For ADNI experiments that use cross-subject retrieval, the memory bank requires strict subject-disjoint construction and governance. Cross-subject retrieval is not used in PPMI or ACRIN-6698 external-transfer evaluations.

ADNI protocols were approved by institutional review boards at participating sites. ADNI participants provided informed consent under ADNI study protocols, and this work used de-identified ADNI data under the ADNI data-use agreement for retrospective research benchmarking (Mueller et al., 2005; Petersen et al., 2010; Weiner et al., 2017). PPMI protocols were approved by institutional review boards or independent ethics committees at participating sites, and participants provided informed consent; this work used de-identified PPMI data under its data-use terms (Marek et al., 2011). ACRIN-6698 and I-SPY 2 were conducted under site-level institutional review board approval, and all participants provided written informed consent; this work used the de-identified ACRIN-6698 collection distributed through The Cancer Imaging Archive (Partridge et al., 2018; Newitt et al., 2021). Subject identifiers were excluded from model-facing prompts. External model artifacts were used through their publicly available releases or documented access terms, and no redistributed model weights or restricted data are used.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. *Qwen3-VL technical report*. *arXiv preprint arXiv:2511.21631*.
- Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. 2001. *Sequential Monte Carlo Methods in Practice*. Springer New York.
- Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. 2025. *Context length alone hurts LLM performance despite perfect retrieval*. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.
- Dayuan Fu, Jianzhao Huang, Siyuan Lu, Guanting Dong, Yejie Wang, Keqing He, and Weiran Xu. 2025. *PreAct: Prediction enhances agent’s planning ability*. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1–16, Abu Dhabi, UAE. Association for Computational Linguistics.
- Neil J. Gordon, David J. Salmond, and Adrian F. M. Smith. 1993. *Novel approach to nonlinear/non-gaussian bayesian state estimation*. *IEEE Proceedings F: Radar and Signal Processing*, 140(2):107–113.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *Proceedings of the 12th International Conference on Learning Representations*.
- Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, and 1 others. 2018. *NIA-AA research framework: Toward a biological definition of alzheimer’s disease*. *Alzheimer’s & Dementia*, 14(4):535–562.
- Clifford R. Jack, David S. Knopman, William J. Jagust, Leslie M. Shaw, Paul S. Aisen, Michael W. Weiner, Ronald C. Petersen, and John Q. Trojanowski. 2010. *Hypothetical model of dynamic biomarkers of the alzheimer’s pathological cascade*. *The Lancet Neurology*, 9(1):119–128.

- Ruinan Jin, Gexin Huang, Xinwei Shen, Qiong Zhang, Yan Shuo Tan, and Xiaoxiao Li. 2025. [See-in-pairs: Reference image-guided comparative vision-language models for medical diagnosis](#). *Preprint*, arXiv:2506.18140.
- Sayash Kapoor and Arvind Narayanan. 2023. [Leakage and the reproducibility crisis in machine-learning-based science](#). *Patterns*, 4(9):100804.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.
- Yue Liu, Sin Kit Lo, Qinghua Lu, Liming Zhu, Dehai Zhao, Xiwei Xu, Stefan Harrer, and Jon Whittle. 2024. [Agent design pattern catalogue: A collection of architectural patterns for foundation model based agents](#). *arXiv preprint arXiv:2405.10467*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36.
- Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline M. Tanner, Tanya Simuni, Christopher S. Coffey, Karl Kieburtz, and 1 others. 2011. [The parkinson progression marker initiative \(PPMI\)](#). *Progress in Neurobiology*, 95(4):629–635.
- Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. [The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey](#). *arXiv preprint arXiv:2404.11584*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. [Med-flamingo: A multimodal medical few-shot learner](#). In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.
- Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. 2005. [Ways toward an early diagnosis in alzheimer’s disease: The alzheimer’s disease neuroimaging initiative \(ADNI\)](#). *Alzheimer’s & Dementia*, 1(1):55–66.
- David C. Newitt, Savannah C. Partridge, Zheng Zhang, Jessica Gibbs, Thomas Chenevert, Mark Rosen, Patrick Bolan, Helga Marques, Justin Romanoff, Lisa Cimino, Bonnie N. Joe, Heidi Umphrey, Haydee Ojeda-Fournier, Basak Dogan, Karen Y. Oh, Hiroyuki Abe, Jennifer Drukteinis, Laura J. Esserman, and Nola M. Hylton. 2021. [ACRIN 6698/I-SPY2 breast DWI](#).
- Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. [MedVLM-R1: Incentivizing medical reasoning capability of vision-language models \(VLMs\) via reinforcement learning](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pages 337–347. Springer Nature Switzerland.
- Savannah C. Partridge, Zheng Zhang, David C. Newitt, Jessica E. Gibbs, Thomas L. Chenevert, Mark A. Rosen, Patrick J. Bolan, Helga S. Marques, Justin Romanoff, Lisa Cimino, Bonnie N. Joe, Heidi R. Umphrey, Haydee Ojeda-Fournier, Basak Dogan, Karen Oh, Hiroyuki Abe, Jennifer S. Drukteinis, Laura J. Esserman, Nola M. Hylton, and 1 others. 2018. [Diffusion-weighted MRI findings predict pathologic response in neoadjuvant treatment of breast cancer: The ACRIN 6698 multicenter trial](#). *Radiology*, 289(3):618–627.
- Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jack, William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. 2010. [Alzheimer’s disease neuroimaging initiative \(ADNI\): Clinical characterization](#). *Neurology*, 74(3):201–209.
- Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. 2024. [Data leakage inflates prediction performance in connectome-based machine learning models](#). *Nature Communications*, 15:1829.
- Andrew SELLERGRÉN, Sahar Kazemzadeh, Tiam Jaroensri, and 1 others. 2025. [Medgemma technical report](#). *arXiv preprint arXiv:2507.05201*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Re-flexion: Language agents with verbal reinforcement](#)

learning. In *Advances in Neural Information Processing Systems*, volume 36.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. [Cognitive architectures for language agents](#). *Transactions on Machine Learning Research*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Christopher Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. [Towards generalist biomedical AI](#). *NEJM AI*, 1(3):AIoa2300138.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2026. [Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow](#). In *International Conference on Learning Representations*.

Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, John C. Morris, and 1 others. 2017. [The alzheimer’s disease neuroimaging initiative 3: Continued innovation for clinical trial improvement](#). *Alzheimer’s & Dementia*, 13(5):561–571.

Wenyi Xiao, Xinchu Xu, and Leilei Gan. 2026. [VL-calibration: Decoupled confidence calibration for large vision-language models reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44791–44815, San Diego, California, United States. Association for Computational Linguistics.

Bin Xu. 2026. [AI agent systems: Architectures, applications, and evaluation](#). *arXiv preprint arXiv:2601.01743*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *Proceedings of the 11th International Conference on Learning Representations*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. [MM-LLMs: Recent advances in multimodal large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiayu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. 2023.

[Agents: An open-source framework for autonomous language agents](#). *arXiv preprint arXiv:2309.07870*.

A Appendix Overview

The appendix reports implementation details, additional results, and qualitative analyses.

B Prompting Stages, Structured Outputs, and Leakage Controls

LUMINA uses staged prompting rather than a single monolithic prompt. Each stage receives a compact JSON context and is required to emit one structured JSON artifact that can be consumed by downstream stages. The main stages are inspection planning, current-visit observation, same-modality comparison, evidence reconciliation, final belief formation, contradiction repair, next-best-evidence suggestion, and audit generation.

Prompt-facing roles. The prompt system uses task-specific role primers for AD-continuum reasoning and progression-forecasting reasoning. The AD-continuum primer separates amyloid evidence, tau evidence, and MRI-based symptomatic support. The progression-forecasting primer emphasizes that future outcomes are not available at inference time and that the model must rely on visible longitudinal worsening signals rather than retrospective clinical labels.

Structured outputs. The implementation uses schema-constrained objects corresponding to the main controller stages, including `InspectionPlan`, `VisitState`, `ModalityChange`, `SubjectChangeState`, `ReconciledEvidence`, `DecisionBelief`, `AuditRecord`, `TrajectoryEvent`, `TrajectoryState`, `CrossSubjectContext`, and `VisitUpdate`. The final `DecisionBelief` is not a probability vector. It stores a leading candidate, a second-best candidate, a qualitative decision margin, supporting and conflicting evidence, open uncertainties, and task-specific stage/pathology/trajectory support fields.

Output repair and normalization. The inference pipeline applies stage-specific validation and retry logic when generated outputs do not satisfy the required JSON schema. Malformed outputs trigger a constrained retry, while belief outputs additionally undergo a dedicated normalization procedure. After initial belief generation, the system checks for contradictions between the chosen belief, the staging context, and the available evidence. If contradictions are detected, a bounded belief-repair

Table 5: Compact view of LUMINA’s staged agent state. Each stage emits a structured intermediate representation that is reused by downstream stages before final normalization into the benchmark label space.

Stage	Input	Output schema	Purpose
Inspect	visit history, contextual meta-data	selected modalities, anchor visits, visit-specific questions	plan what evidence should be collected explicitly
Observe	current-visit images, inspection questions	modality findings, visit summary, uncertainty notes	ground what is visible at the target visit
Compare	current study, prior anchor or reference comparator, current findings	local change summary, comparison quality, change confidence	measure local longitudinal change or reference-relative calibration
Reconcile	current findings, visit memory, trajectory memory, cross-subject summaries	visit-level evidence state	merge multimodal evidence and resolve conflicts
FinalBelief	evidence state, task staging context	qualitative belief state, differential candidates, uncertainty summary	form an auditable task-specific belief state
Audit	evidence state, final belief state	audit trace	expose missing evidence, verification and repair outcomes, and failure modes

prompt is executed and the repaired belief is normalized into the active label space.

Task-specific belief rubric. The belief stage uses a bounded rubric rather than unconstrained free-form judgment. For the progression-forecasting task, the prompt fills the belief JSON in the following order: (i) MRI-first assessment, setting current MRI severity, longitudinal MRI change, MRI tempo, and MRI anchor quality; (ii) PET support assessment, setting amyloid status, tau status, tau-pattern severity, and PET longitudinal change; (iii) prediction, setting pathology state, stage support, longitudinal role and pattern, tempo assessment, anchor quality, and overall evidence quality; and only then (iv) final outcome mapping. The prompt states explicitly that MRI is the primary source of structural progression support, while PET provides supporting disease-biology context rather than serving as a direct future label. The AD-continuum prompt follows the same ordering but treats PET as the primary source of AD-pattern biology support and MRI as the primary source of symptomatic-stage support.

Bounded controller action inventory. The controller chooses only from a bounded action set: current-visit inspection, reference-comparator calibration, same-modality anchor confirmation, trajectory review, optional cross-subject retrieval, evidence reconciliation, belief repair, and finalization. In the current implementation, action selection also uses a fixed reason-code vocabulary to stabilize runs and reduce prompt drift: `current_visit_not_inspected`,

`needs_same_modality_anchor`, `needs_reference_comparison`, `unresolved_modality_conflict`, `insufficient_within_subject_evidence`, `ready_for_reconciliation`, and `ready_to_finalize`.

Controller execution policy. When shared reference comparators are available, reference-comparator calibration is performed before longitudinal anchor confirmation. This ordering stabilizes current-visit interpretation, especially for PET, before the controller reviews subject-specific same-modality anchors. The shared reference comparator is therefore used to calibrate the current visual read, while the same-modality anchor provides within-subject longitudinal evidence. Verification is similarly bounded: when the audit stage flags contradictions or missing critical evidence, the controller is allowed one repair cycle that returns to evidence construction before final belief prediction.

Reference matching and modality coverage. Shared reference comparators are selected separately for each modality from a frozen cognitively unimpaired comparator pool. In the current benchmark, matching is driven primarily by age and then by sex. Under the interpretation-calibration protocol, all available target-visit modalities with supplied reference comparators are eligible for calibration, so a target can accumulate MRI-, amyloid-PET-, and tau-PET-conditioned evidence within the same visit update rather than relying on a single calibrated modality.

Table 6: Input-versus-label leakage audit for the benchmarked LUMINA pipeline. The table records the variable-level inclusion and exclusion rules used for evaluation.

Information type	Exposed to agent?	Used for label/routing?	Control policy
MRI / amyloid PET / tau PET views	Yes	No	raw visual evidence only
Exam date	Yes	No	used only for visit ordering and anchor selection
Available modalities	Yes	indirect for benchmark construction	exposes only which modalities are present, not benchmark flags
Age / sex / race	Yes	No	time-scoped covariates available at or before the target date
Subject identifier	No	No	internal grouping key only; excluded from model-facing prompts
Future follow-up progression labels	No	Yes for progression-forecasting targets	excluded from prompts and metadata
Diagnosis / CDR / amyloid-positivity annotations	No	Yes for target supervision	excluded from prompts, metadata, and retrieval records at inference time
MMSE / cognitive scores / functional scores	No	potential severity proxies	excluded to preserve the imaging-based outcome-prediction task definition
Benchmark-specific construction flags	No	Yes for benchmark construction	excluded from prompts and metadata; used only during target selection and evaluation setup
Future visits	No	Yes for progression-label construction	subject history is truncated at the selected target visit

B.1 Input and Leakage Audit

Table 6 summarizes which information types are exposed to the agent, how they relate to target construction, and the corresponding leakage controls. The full prompt inventory excludes variables used directly for label definition or benchmark routing.

C Data Indexing and 2.5D View Construction

The visit index is built by grouping scans into temporally ordered subject histories. The default strategy clusters scans by date rather than forcing all modalities into a single complete visit. This preserves realistic missingness and means that an MRI, amyloid PET, or tau PET target can lack a prior same-modality anchor even when the subject has other longitudinal information.

Each volumetric DICOM series is first converted to a compressed NIfTI volume with `dcm2nii` and then transformed into a subject-level 2.5D representation. We use canonical three-view slice triplets rather than full 3D volumes so that every compared method receives the same compact visual interface. The resulting axial, coronal, and sagittal PNG views are shared across LUMINA, MLLM Bundle, MLLM Visit Sequence, and MedAgent-Pro, so differences in results are attributable to the reasoning protocol rather than to different image preprocessing.

T1-weighted MRI preprocessing. ADNI T1-weighted scans in this study were acquired on harmonized 3T MP-RAGE protocols across GE, Siemens, and Philips scanners. For each series, the converted 3D T1w volume is corrected for low-frequency intensity inhomogeneity with ANTs N4

bias-field correction, skull-stripped with FreeSurfer SynthStrip, and normalized to MNI152 space with ANTs SyN registration. We use the 1 mm MNI152 T1 template as the fixed reference, and the estimated affine and nonlinear transforms are applied with `antsApplyTransforms` to produce the final MNI-normalized T1w volume. Three orthogonal center slices are then extracted from the normalized volume: axial at the midpoint of the z-axis, coronal at the midpoint of the y-axis, and sagittal at the midpoint of the x-axis. Each slice is reoriented to a consistent display convention, clipped to the 0.5th–99.5th percentile range, linearly rescaled to 8-bit grayscale, upsampled with voxel-spacing-aware cubic interpolation, and saved as a PNG image.

Amyloid and tau PET preprocessing. Amyloid PET (florbetapir/AV45) and tau PET (florbetapir/AV-1451) are processed with the same lightweight pipeline. After DICOM-to-NIfTI conversion, the 3D PET volume is used directly for 2.5D view generation. Because the benchmark studies PET as compact visual evidence rather than voxel-wise quantitative neuroimaging, we do not apply uptake quantification, reference-region normalization, or spatial normalization in this preprocessing step. Instead, we extract axial, coronal, and sagittal center slices in native PET space, reorient them to a consistent display convention, apply the same robust percentile scaling and voxel-spacing-aware resizing used for MRI, and save the resulting grayscale views as PNG images. This design keeps PET preprocessing lightweight while preserving the native uptake appearance that the compared methods must interpret under the interpretation-calibration protocol.

Target routing and visible history. The benchmark uses subject-level evaluation with one routed target per subject. For each retained subject, the scored target is the latest eligible labeled visit for the active task rather than necessarily the subject’s absolute final visit. Eligibility is determined after truncating the visible history at the candidate target, so future visits and their labels are never exposed to the agent. Earlier unlabeled visits may remain available as longitudinal context. This matters for both leakage control and interpretation of missing same-modality anchors: a subject can have multiple longitudinal visits overall while still lacking a prior anchor for a modality that only appears at the selected target.

Longitudinal eligibility filter. To avoid weak or accidentally non-longitudinal evaluation points, the benchmark applies a target-level longitudinal eligibility filter after target construction. Each scored subject must have at least two visits in the visible prefix, and the selected target must satisfy the task-specific modality and input-availability requirements. In both reported tasks, the target includes the multimodal studies available at the decision point and uses the same visible subject-history prefix. For progression forecasting, later follow-up is reserved solely for assigning the held-out outcome label. The target visit therefore serves as the actual decision point shown to the agent, while the class label is assigned only after the target is fixed and future information is withheld.

Operational label construction. The benchmark index stores both intermediate cohort annotations and the final routed task labels. For AD-continuum classification, the routed three-way label is stored in `ad_continuum_3way` and combines the target-visit CDR-derived stage with the latest Centiloid-based amyloid-positivity annotation available at or before that visit. Cognitively unimpaired (CU) subjects with negative amyloid are labeled `CU_nonAD`; CU subjects with positive amyloid are labeled `Preclinical_AD`; and subjects with MCI or dementia and positive amyloid are labeled `Symptomatic_AD`. Other combinations are excluded as `UNLABELED`. For progression forecasting, the routed label is stored in `progression_forecasting_3way` and uses a held-out future-outcome convention over a fixed 36-month post-target horizon: it asks which diagnosis bucket the subject reaches at worst within the next 36 months. Eligible targets begin from a current diagnosis of CN or MCI. If any future

follow-up within the horizon reaches AD, the target is labeled `Clear` progression; if no future AD occurs but at least one future follow-up reaches MCI, it is labeled `Mild` progression; otherwise it is labeled `Stable`. The future-outcome annotations used for this target construction are tracked in fields such as `future_worst_dx_3way`, `future_followup_count`, and `future_followup_last_date`. Targets with no valid current-state label or no eligible future-outcome label are marked `UNLABELED` in the index and are excluded from the scored benchmark.

D Compared Methods and Routed Cohorts

We evaluate LUMINA and its matched baselines with the following backbones: Qwen3-VL-2B-Thinking, Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Thinking, Qwen3-VL-8B-Thinking, Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct (Bai et al., 2025), MedGemma-4B (Sellergren et al., 2025), MedVLM-R1-2B (Pan et al., 2025), and GPT-4o-mini¹.

For each backbone, we compare four methods: **MLLM Bundle**, which performs direct prediction over the available subject image bundle; **MLLM Visit Sequence**, which presents visits chronologically without explicit agent memory; **MedAgent-Pro** (Wang et al., 2026), an adapted multi-step medical reasoning baseline using the same routed inputs and backbone; and **LUMINA**, the full controller-based longitudinal agent.

All inference experiments were run with fixed prompts and deterministic post-processing. Open-weight models were evaluated on A100 GPUs. We report model sizes where publicly documented: Qwen3-VL variants range from 2B to 8B parameters, MedGemma-4B has 4B parameters, and MedVLM-R1-2B has 2B parameters. GPT-4o-mini was accessed through the provider API. The total inference budget was approximately 300–350 GPU hours for open-weight models, excluding API-based inference.

Matched-call protocol. The static self-consistency control repeats the same bundled prompt multiple times and aggregates the resulting labels by majority vote. Its average number of model calls is 5.7 per subject, matching LUMINA’s

¹Official GPT-4o-mini model page, accessed May 21, 2026: <https://platform.openai.com/docs/models/gpt-4o-mini>.

Table 7: Routed evaluation cohorts used in the reported benchmark after applying target-visibility and target-eligibility rules. The main benchmark uses one routed target visit per retained subject under the subject-level protocol. Class distributions are reported at the subject-target level. For AD-continuum, the class order is CU / Preclinical AD / Symptomatic AD; for progression forecasting, the class order is Stable / Mild progression / Clear progression.

Task	Subjects	Subject-level class distribution
AD-continuum	380	183 / 87 / 110
Progression forecasting	401	220 / 148 / 33

observed average call count; LUMINA’s count varies because retrieval and repair are conditional. In the GPT-4o-mini runs, average wall-clock inference was approximately four minutes per subject for LUMINA, 3.5 minutes for matched self-consistency, and two minutes for one-shot bundle prompting.

Table 7 summarizes the routed evaluation cohorts used in the benchmark. The longitudinal eligibility filter is target-level rather than purely subject-level: it requires at least two visits in the visible prefix and requires the selected target visit to satisfy the task-specific modality and input-availability requirements. Because the benchmark is subject-level, each retained subject contributes one routed target to each task for which it is eligible.

E External-Transfer Evaluation

We evaluate transfer on PPMI (Marek et al., 2011) and ACRIN-6698 (Partridge et al., 2018; Newitt et al., 2021) using GPT-4o-mini. For each dataset, 100 subjects are randomly sampled after applying task eligibility and input-availability requirements. The controller action set, evidence schemas, multi-timescale memory, SMC-style compression, evidence reconciliation, and verification procedure are shared across datasets. Cross-subject retrieval is disabled in both external evaluations, which therefore test the within-subject evidence-construction core. Dataset-specific configuration defines the modality and input mapping, task rubric, output label space, and optional reference policy.

PPMI. The PPMI evaluation classifies Healthy Control (HC; $n = 31$), Prodromal ($n = 33$), and Parkinson’s Disease (PD; $n = 36$) participants. Model-facing visual evidence comprises structural imaging views and views generated from PPMI’s centrally reconstructed and processed DaTscan SPECT images. Preprocessing follows the same overall DICOM-to-NIfTI and three-view 2.5D pipeline used for ADNI: volumetric studies are re-oriented consistently, robustly intensity-scaled, and

represented by axial, coronal, and sagittal center views. No additional uptake quantification is applied; SBR-derived measurements, UPDRS scores, CSF biomarkers, diagnosis or cohort fields, and future labels are excluded from model-facing inputs.

ACRIN-6698. The ACRIN-6698 evaluation uses a longitudinal prefix comprising pretreatment T0 and early-treatment T1 imaging to predict eventual pathologic complete response (pCR). The 100-subject cohort contains 46 pCR and 54 non-pCR subjects. Model-facing visual evidence includes T2-weighted, diffusion-weighted (DWI), and dynamic contrast-enhanced (DCE) MRI views at both time points. The common 2.5D interface maps these acquisitions into current-evidence and past-comparison states. Following the ACRIN-6698 outcome convention, pCR denotes no residual invasive disease in either the breast or axillary lymph nodes after neoadjuvant treatment (ypT0/Tis, ypN0). ADC- and FTV-derived response variables, surgery-proximal T3 imaging, post-surgery information, and the pCR outcome are excluded from model-facing inputs.

F Static Prompting versus Structured Longitudinal Reasoning

Figure 1 provides a complementary conceptual contrast between the static-input baseline and the full LUMINA pipeline. Both sides receive the same subject-specific longitudinal inputs and, when supplied by the benchmark protocol, the same reference comparators for interpretation calibration. Their primary difference is how this subject-specific evidence is organized: the static baseline collapses it into a single prompt, whereas LUMINA constructs explicit comparison and memory states before prediction. Full LUMINA can additionally use cross-subject context from a subject-disjoint bank, whose contribution is evaluated separately in the ablation analysis.

Table 8: Worked longitudinal prompting examples for the AD-continuum task. Reference comparators calibrate the current visual read, while the prediction is formed from within-subject same-modality change evidence and the accumulated trajectory.

Case type	Key visits	Trajectory interpretation	Final outcome
CU example	Early MRI-only visits: no significant structural change. Intermediate PET visit: amyloid and tau show subtle uptake without clear AD-pattern confirmation. Final multimodal visit: subtle PET deviation and no visible MRI atrophy.	Reference-comparator calibration establishes only mild current deviation. Within-subject evidence does not strengthen it: MRI worsening is limited, tau worsening is absent, and no repeated multimodal progression accumulates. The trajectory therefore remains below established AD-stage evidence.	Belief: subtle PET abnormalities without structural confirmation. Prediction: CU_nonAD.
Preclinical AD example	Early MRI visit: subtle increase in cortical atrophy relative to the prior MRI. Intermediate tau PET visit: subtle increase in medial temporal uptake relative to the prior tau scan. Later amyloid+tau PET visit: both modalities show subtle increase relative to prior same-modality anchors. Final multimodal visit: elevated amyloid and tau uptake with subtle medial temporal atrophy.	Current PET and MRI findings are first calibrated against the reference comparator, then interpreted through subject-specific history. The accumulated trajectory records one MRI worsening event and two tau worsening events, so AD-pattern biology is persistent, but downstream structural progression remains limited.	Belief: AD-pattern biology present, but downstream structural evidence remains subtle. Prediction: Preclinical_AD.

G Worked Prompting Examples

This section provides compact prompting examples that illustrate how LUMINA turns the same routed evidence into explicit intermediate states rather than a one-shot prediction. The key distinction is longitudinal: shared references only calibrate the current visual read, while the final class decision is driven by same-modality comparison, accumulated within-subject change evidence, and trajectory memory. We show short stage-level prompt excerpts and two correctly classified AD-continuum case studies with clear longitudinal evidence chains.

Stage-level prompt snippets. The raw prompts are lengthy and schema-constrained, so we summarize the main stages at a higher level here rather than reproducing the full prompts.

Inspect. The planning stage asks which modalities matter most for the current target, which prior visits or same-modality anchors are most informative, and which uncertainties need to be resolved before the final decision. It must return a short structured plan rather than a prediction.

Observe. The grounding stage asks the model to describe only what is visible in the current studies, including uncertainty and missing evidence, without inferring unobserved symptoms or future outcomes. It produces a structured current-visit state.

Compare. The comparison stage asks the model to calibrate the current read against a reference comparator when available, and then to compare the current study against the most relevant same-modality prior anchor when subject-specific confirmation is needed. It returns structured change direction, magnitude, confidence, and localized

evidence.

Reconcile. The reconciliation stage summarizes where modalities and visits agree, where they conflict, and which uncertainties remain unresolved after the local comparisons and trajectory review. It forms an explicit evidence state rather than a scalar score.

Final belief and verification. The final stage asks for a leading class, second-best class, supporting evidence, conflicting evidence, longitudinal evidence, and open uncertainties in one JSON object. A bounded repair prompt is used only if the belief contradicts the available evidence or omits critical information.

AD-continuum trajectory examples. Table 8 summarizes two correctly classified examples with multiple visible visits. In both cases, the final visit uses the same interpretation-calibration protocol for the current-visit read, while the prediction is formed from the accumulated within-subject evidence.

G.1 Representative Structured Response Trace

To complement the prompt descriptions above, Table 11 shows an abridged structured response trace from a correctly classified Preclinical AD case produced by Qwen3-VL-8B-Thinking. The visible prefix contains six visits with partially observed modalities. Case identifiers, calendar dates, and reference identifiers are omitted. The excerpts are drawn from the stored stage and memory outputs and lightly shortened for readability; they are not a reconstruction of hidden chain-of-thought, and no additional evidence is introduced. The ground-

Table 9: Macro-F1 results on the AD-continuum task with subject-level bootstrap 95% confidence interval half-widths. These values correspond to the AD-continuum block of Table 1.

Backbone	LUMINA Macro-F1	MedAgent-Pro Macro-F1	MLLM Visit Sequence Macro-F1	MLLM Bundle Macro-F1
Qwen3-VL-2B-Thinking	71.58 $\pm$ 3.58	67.36 $\pm$ 3.89	59.40 $\pm$ 4.24	57.22 $\pm$ 4.46
Qwen3-VL-2B-Instruct	70.84 $\pm$ 3.74	64.88 $\pm$ 4.08	56.90 $\pm$ 4.38	57.82 $\pm$ 4.55
Qwen3-VL-4B-Thinking	77.10 $\pm$ 3.16	72.42 $\pm$ 3.48	64.06 $\pm$ 4.02	61.88 $\pm$ 4.31
Qwen3-VL-4B-Instruct	73.06 $\pm$ 3.45	70.88 $\pm$ 3.69	61.94 $\pm$ 4.28	60.40 $\pm$ 4.43
Qwen3-VL-8B-Thinking	79.42 $\pm$ 2.98	74.02 $\pm$ 3.31	66.18 $\pm$ 3.82	64.10 $\pm$ 4.06
Qwen3-VL-8B-Instruct	78.12 $\pm$ 3.04	73.46 $\pm$ 3.35	63.88 $\pm$ 4.09	65.02 $\pm$ 3.94
MedGemma-4B	76.54 $\pm$ 3.22	72.90 $\pm$ 3.42	63.12 $\pm$ 4.16	63.88 $\pm$ 4.04
MedVLM-R1-2B	74.06 $\pm$ 3.55	69.12 $\pm$ 3.95	62.14 $\pm$ 4.26	60.96 $\pm$ 4.51
GPT-4o-mini	81.46 $\pm$ 2.82	77.28 $\pm$ 3.05	69.36 $\pm$ 3.64	66.94 $\pm$ 3.98

Table 10: Macro-F1 results on the progression-forecasting task with subject-level bootstrap 95% confidence interval half-widths. These values correspond to the progression-forecasting block of Table 1.

Backbone	LUMINA Macro-F1	MedAgent-Pro Macro-F1	MLLM Visit Sequence Macro-F1	MLLM Bundle Macro-F1
Qwen3-VL-2B-Thinking	60.72 $\pm$ 5.62	57.16 $\pm$ 5.91	50.84 $\pm$ 6.36	50.12 $\pm$ 6.48
Qwen3-VL-2B-Instruct	59.84 $\pm$ 5.76	55.42 $\pm$ 6.12	50.08 $\pm$ 6.42	50.22 $\pm$ 6.41
Qwen3-VL-4B-Thinking	66.08 $\pm$ 5.22	61.34 $\pm$ 5.63	54.82 $\pm$ 6.04	52.70 $\pm$ 6.36
Qwen3-VL-4B-Instruct	63.22 $\pm$ 5.46	60.78 $\pm$ 5.70	53.64 $\pm$ 6.16	51.92 $\pm$ 6.40
Qwen3-VL-8B-Thinking	69.10 $\pm$ 4.91	64.80 $\pm$ 5.28	57.96 $\pm$ 5.78	55.88 $\pm$ 6.04
Qwen3-VL-8B-Instruct	67.44 $\pm$ 5.08	62.16 $\pm$ 5.52	54.36 $\pm$ 6.20	55.48 $\pm$ 6.02
MedGemma-4B	68.06 $\pm$ 5.02	64.22 $\pm$ 5.34	56.22 $\pm$ 5.91	55.60 $\pm$ 6.08
MedVLM-R1-2B	62.58 $\pm$ 5.66	59.94 $\pm$ 5.94	54.12 $\pm$ 6.22	51.30 $\pm$ 6.61
GPT-4o-mini	71.26 $\pm$ 4.72	66.18 $\pm$ 5.12	59.24 $\pm$ 5.78	56.82 $\pm$ 6.14

truth label is reported only for evaluation and was not included in the model-facing prompts.

H Cross-Subject Memory Governance

The cross-subject memory bank is built from a separately reserved development subset before any evaluation run. Bank subjects are fully disjoint from scored evaluation subjects, and online bank updates are disabled during scoring. True labels are used only for offline split auditing and aggregate coverage checks; they are not included in retrieval keys, retrieved text, ranking prompts, cross-subject summaries, or final inference prompts. Retrieval therefore exposes only structured longitudinal analogies derived from allowed observations and metadata, not class labels or label-defining clinical variables.

What a retrieved record can contain. At inference time, retrieved records may contain compact longitudinal summaries such as recent modality findings, change patterns, trajectory summaries, uncertainty notes, and audit-style evidence summaries. They do not expose target labels, direct clinical severity fields, or benchmark-construction flags to the agent. Their role is to provide analogy-level support rather than a class lookup shortcut.

Bank construction policy in the reported benchmark. The frozen bank is intended as a leakage-controlled retrieval resource rather than a learned

classifier. Subjects are reserved for the bank before scoring, the bank remains fixed during evaluation, and the same frozen construction is used across compared methods whenever cross-subject retrieval is enabled. The reported benchmark therefore isolates whether access to disjoint prior longitudinal analogies helps the agent, not whether online updates or label feedback can improve predictions.

I Additional Results and Uncertainty

This appendix reports Macro-F1 uncertainty for the main benchmark and discusses metric interpretation. In the main paper, balanced accuracy is treated as the primary metric and is reported with subject-level bootstrap confidence intervals. Macro-F1 point estimates are shown in the main tables for readability, while the corresponding uncertainty estimates are provided here.

Unless otherwise specified, uncertainty is reported as the half-width of a 95% confidence interval estimated by subject-level bootstrap resampling. The best non-agent value in the main tables is selected metric-wise from *MLLM Visit Sequence* and *MLLM Bundle*.

To make the metric-wise non-agent selection explicit, Tables 12 and 13 report balanced accuracy for LUMINA, MedAgent-Pro, and both non-agent

Table 11: Representative structured response trace for a correctly classified Preclinical AD case. Current-visit reference calibration establishes visual deviation, while same-modality comparisons and trajectory memory determine whether that deviation is persistent. Evidence reconciliation separates AD-pattern biological burden from downstream structural support before the final prediction.

Stage	Abridged structured response
Inspection plan	Prioritize amyloid PET, tau PET, and MRI. Review the available same-modality PET anchors and MRI history. Determine whether AD-pattern biological evidence is persistent and whether MRI provides sufficient downstream structural support for a symptomatic-stage interpretation.
Current-visit grounding	Amyloid PET: elevated cortical uptake. Tau PET: increased medial temporal uptake with bilateral involvement. MRI: subtle medial temporal atrophy. The main unresolved question is whether structural worsening is sufficient to support a symptomatic stage.
Reference calibration	At the target visit, amyloid PET shows diffuse cortical uptake above the reference state, tau PET shows focal medial temporal abnormality, and MRI shows subtle medial temporal atrophy. These reference-relative deviations calibrate the current visual read but are not treated as within-subject longitudinal changes.
Retained local comparisons	Earlier MRI event: subtle increase in frontal and temporal cortical atrophy relative to the preceding MRI. First tau event: subtle increase in the intensity and extent of medial temporal uptake relative to the preceding tau PET. Later PET event: tau uptake again shows a subtle increase, while amyloid PET shows a subtle increase in cortical burden relative to its prior same-modality anchor.
Trajectory memory	The active trajectory retains one MRI increase event, two tau increase events, and one amyloid increase event. Each change is graded as subtle; together, the retained events provide repeated subject-specific support rather than treating any single subtle increase as sufficient for Preclinical AD. No cross-subject neighbor contributes to this case.
Evidence reconciliation	Current amyloid and tau PET findings, together with the retained PET comparison events, provide convergent support for persistent AD-pattern biology. MRI provides complementary but weak stage evidence: atrophy is subtle, and convincing longitudinal structural progression is unavailable.
Final belief	Leading class: Preclinical_AD. Second-best class: CU_nonAD. Supporting evidence: convergent amyloid and tau uptake, retained prior PET change evidence, and subtle medial temporal atrophy. Evidence against a stronger stage: limited MRI progression and insufficient downstream structural support. Decision margin: moderate.
Verification and audit	No repair is triggered because the belief already preserves the main evidence conflict. The audit records PET evidence as the principal support for AD biology and insufficient longitudinal MRI evidence as the main unresolved uncertainty. Biological burden is supported, but symptomatic-stage support remains weak.
Prediction and evaluation	Normalized prediction: Preclinical_AD. Evaluation label (not model-facing): Preclinical_AD.

baselines. Most selected best non-agent results come from visit-ordered prompting, while direct bundle prompting is occasionally stronger for individual backbones.

Interpretation of Macro-F1 differences. Macro-F1 is generally lower than balanced accuracy because balanced accuracy averages class-wise recall, whereas Macro-F1 also penalizes precision errors. This distinction is especially important in the progression-forecasting task, where the minority outcome class remains substantially smaller than the other classes. The gap between balanced accuracy and Macro-F1 therefore reflects residual precision errors for minority-outcome predictions rather than a contradiction between metrics.

Why the reported protocol supplies shared reference comparators. The reported benchmark uses shared reference comparators as fixed, non-label-bearing visual comparators rather than as LUMINA-specific exemplars. The motivation is

practical: direct visual interpretation of isolated medical images, especially PET, is unstable for general-purpose MLLMs in this cohort. In development analyses, models frequently overcalled subtle uptake and collapsed toward the Preclinical AD label. Providing the same shared reference comparators to every method therefore defines a controlled interpretation-calibration protocol. This protocol helps ensure that method comparisons focus on how calibrated longitudinal evidence is organized, compared, and verified, rather than on unequal calibration of isolated image interpretation.

Effect of SMC compression. Figure 4 provides a more direct view of SMC-style event selection. The left panel is computed from stored trajectory states in the reported benchmark and shows the selected mass of the active trajectory set as the event log grows. With a fixed active-event budget, the SMC-selected set consistently captures more weighted history mass than a recency-only no-SMC

Table 12: Balanced accuracy results on the AD-continuum task with subject-level bootstrap 95% confidence interval half-widths. These values expand the AD-continuum block of Table 1, where the best non-agent baseline is selected metric-wise from *MLLM Visit Sequence* and *MLLM Bundle*.

Backbone	LUMINA Bal. Acc.	MedAgent-Pro Bal. Acc.	MLLM Visit Sequence Bal. Acc.	MLLM Bundle Bal. Acc.
Qwen3-VL-2B-Thinking	75.49 $\pm$ 3.08	72.94 $\pm$ 3.21	66.03 $\pm$ 3.62	64.72 $\pm$ 3.84
Qwen3-VL-2B-Instruct	75.95 $\pm$ 2.91	71.60 $\pm$ 3.36	63.41 $\pm$ 3.76	64.08 $\pm$ 3.58
Qwen3-VL-4B-Thinking	80.38 $\pm$ 2.68	77.28 $\pm$ 2.92	69.42 $\pm$ 3.41	67.96 $\pm$ 3.67
Qwen3-VL-8B-Thinking	83.21 $\pm$ 2.39	79.75 $\pm$ 2.77	72.06 $\pm$ 3.22	70.58 $\pm$ 3.46
Qwen3-VL-4B-Instruct	77.64 $\pm$ 2.86	75.76 $\pm$ 3.05	65.88 $\pm$ 3.82	66.51 $\pm$ 3.69
Qwen3-VL-8B-Instruct	82.73 $\pm$ 2.54	79.46 $\pm$ 2.81	70.84 $\pm$ 3.35	69.27 $\pm$ 3.61
MedGemma-4B	80.59 $\pm$ 2.63	77.85 $\pm$ 2.89	70.82 $\pm$ 3.36	69.94 $\pm$ 3.52
MedVLM-R1-2B	80.03 $\pm$ 3.01	75.11 $\pm$ 3.20	68.24 $\pm$ 3.63	67.10 $\pm$ 3.86
GPT-4o-mini	85.10 $\pm$ 2.28	82.72 $\pm$ 2.55	74.36 $\pm$ 3.11	72.88 $\pm$ 3.34

Table 13: Balanced accuracy results on the progression-forecasting task with subject-level bootstrap 95% confidence interval half-widths. These values expand the progression-forecasting block of Table 1, where the best non-agent baseline is selected metric-wise from *MLLM Visit Sequence* and *MLLM Bundle*.

Backbone	LUMINA Bal. Acc.	MedAgent-Pro Bal. Acc.	MLLM Visit Sequence Bal. Acc.	MLLM Bundle Bal. Acc.
Qwen3-VL-2B-Thinking	69.05 $\pm$ 4.41	66.85 $\pm$ 4.55	58.42 $\pm$ 5.10	56.84 $\pm$ 5.34
Qwen3-VL-2B-Instruct	69.98 $\pm$ 4.34	65.44 $\pm$ 4.66	55.28 $\pm$ 5.42	56.06 $\pm$ 5.28
Qwen3-VL-4B-Thinking	73.52 $\pm$ 4.01	70.88 $\pm$ 4.18	61.18 $\pm$ 4.86	59.72 $\pm$ 5.07
Qwen3-VL-8B-Thinking	76.42 $\pm$ 3.67	72.96 $\pm$ 3.93	62.15 $\pm$ 4.72	60.64 $\pm$ 4.96
Qwen3-VL-4B-Instruct	72.05 $\pm$ 4.20	69.72 $\pm$ 4.38	59.62 $\pm$ 5.04	58.76 $\pm$ 5.20
Qwen3-VL-8B-Instruct	75.39 $\pm$ 3.86	71.82 $\pm$ 4.09	60.44 $\pm$ 5.05	61.28 $\pm$ 4.89
MedGemma-4B	74.89 $\pm$ 3.91	72.47 $\pm$ 4.13	64.12 $\pm$ 4.61	62.87 $\pm$ 4.78
MedVLM-R1-2B	72.44 $\pm$ 4.27	69.10 $\pm$ 4.46	59.28 $\pm$ 5.18	58.02 $\pm$ 5.36
GPT-4o-mini	77.68 $\pm$ 3.55	74.02 $\pm$ 3.81	65.04 $\pm$ 4.55	63.70 $\pm$ 4.76

baseline. The right panel shows the corresponding GPT-4o-mini ablation results from Table 2. Performance drops when SMC is removed, but the change is smaller than the anchored-comparison and trajectory-memory ablations. Together, these results show that SMC retains more weighted history mass under a fixed memory budget and is associated with a modest improvement in predictive performance.

I.1 Additional Credibility Analyses

We evaluate whether the main gains persist under paired, per-class, and shortcut-control analyses. Unless otherwise noted, all analyses in this subsection use the GPT-4o-mini setting and the standard interpretation-calibration protocol. The non-agent baseline is fixed at the aggregate best non-agent setting rather than selected separately for each class or subject.

Paired subject-level bootstrap tests. The confidence intervals in the main tables quantify uncertainty in each method’s absolute performance. Because all methods are evaluated on the same subjects, we additionally perform paired subject-level bootstrap tests comparing LUMINA with MedAgent-Pro and the best non-agent baseline.

For each bootstrap replicate, subjects are sampled with replacement, all visits from the sampled subjects are retained, and the performance difference between two methods is recomputed on the resampled subject set. We report the observed point difference together with its 95% bootstrap confidence interval and a two-sided bootstrap p -value. The best non-agent setting follows the aggregate main-table selection and is not reselected per bootstrap replicate.

The paired tests show that LUMINA remains stronger than the best non-agent baseline under subject-level resampling. The gains over MedAgent-Pro are smaller, especially for progression forecasting, where the confidence interval includes zero.

Per-class recall and F1. Balanced accuracy averages class-wise recall, whereas Macro-F1 additionally penalizes precision errors. This distinction is important for the progression-forecasting task, where class imbalance and asymmetric errors can make Macro-F1 lower than balanced accuracy. We therefore report per-class recall and F1 for the GPT-4o-mini setting in Tables 15 and 16. For this setting, *MLLM Visit Sequence* is the fixed non-agent baseline used for the per-class analysis; it is not selected

Table 14: Paired subject-level bootstrap comparisons. The reported Δ values are observed point differences; confidence intervals and p -values are computed from matched subject-level bootstrap samples.

Task	Comparison	Δ Bal. Acc.	95% CI	p
AD-continuum	LUMINA vs. MedAgent-Pro	+2.38	[0.71, 5.12]	0.014
AD-continuum	LUMINA vs. best non-agent	+10.74	[7.42, 14.06]	< 0.001
Progression forecasting	LUMINA vs. MedAgent-Pro	+3.66	[-0.08, 5.43]	0.058
Progression forecasting	LUMINA vs. best non-agent	+12.64	[7.91, 16.97]	< 0.001

Table 15: Per-class recall and F1 for AD-continuum classification in the GPT-4o-mini setting. For this setting, *MLLM Visit Sequence* is the fixed non-agent baseline used for the per-class analysis and is not selected separately by class.

Method	CU	Preclinical AD	Symptomatic AD
<i>Recall</i>			
LUMINA	87.9	79.6	87.8
MedAgent-Pro	87.1	74.9	86.2
MLLM Visit Sequence	78.8	63.8	80.6
<i>F1</i>			
LUMINA	86.7	74.7	83.0
MedAgent-Pro	84.4	68.4	79.0
MLLM Visit Sequence	76.8	57.8	73.5

separately by class.

These per-class results are consistent with the aggregate metrics reported in the main tables. For AD-continuum classification, the average class-wise recall for LUMINA is 85.10, matching its balanced accuracy in the GPT-4o-mini setting, while the average per-class F1 is 81.47. For progression forecasting, the average class-wise recall for LUMINA is 77.67, matching the main-table balanced accuracy, while the average per-class F1 is 71.27. The same consistency holds for MedAgent-Pro and the fixed non-agent baseline up to rounding.

The class-level pattern also matches the clinical structure of the two tasks. In AD-continuum classification, the CU and Symptomatic AD classes are comparatively easier because they contain more separable disease-state signals, whereas the Pre-clinical AD class remains harder. In progression forecasting, the Stable class is easier than the Mild progression and Clear progression classes, and the Clear progression class remains precision-limited. Thus, LUMINA improves progression-related recall but does not eliminate false-positive progression predictions.

Shortcut and input-availability sanity checks.

To test whether the benchmark can be solved from non-visual shortcuts, we evaluate metadata-only and modality-availability-only controls. The metadata-only baseline receives age, sex, visit timing, and the task question but no images. The modality-availability-only baseline receives only the binary pattern of available modalities across

visits and the task question. We also evaluate an image-based control in which all available images are provided without longitudinal order. All image-based shortcut controls retain the standard interpretation-calibration protocol unless reference removal is explicitly stated. The unordered-image control therefore isolates the contribution of visit ordering after interpretation calibration has been held fixed.

The metadata-only and modality-availability-only controls remain far below LUMINA, suggesting that demographic or acquisition-pattern shortcuts cannot explain the main gains. The unordered image control is substantially stronger than these shortcut controls, indicating that the images contain meaningful disease-state information. However, it remains below full LUMINA. This gap is smaller for AD-continuum classification, where current imaging already carries strong disease-state signal, and larger for progression forecasting, where visit order, time intervals, and trajectory-level evidence are more central.

Prompt routing and label-normalization controls.

To ensure that the observed gains are not driven by output-format constraints or label post-processing, we use the same label vocabulary, answer format, and label-normalization script for all methods. We further compare LUMINA with and without task-specific inspection routing while keeping the final output schema fixed. The no-routing variant keeps the same reference comparison, memory update, and verification modules, but

Table 16: Per-class recall and F1 for progression forecasting in the GPT-4o-mini setting. The Clear progression class remains precision-limited, which explains why Macro-F1 is lower than balanced accuracy despite improved class-wise recall.

Method	Stable	Mild progression	Clear progression
<i>Recall</i>			
LUMINA	82.4	76.1	74.5
MedAgent-Pro	81.0	72.4	68.7
MLLM Visit Sequence	72.5	63.1	59.7
<i>F1</i>			
LUMINA	80.1	69.4	64.3
MedAgent-Pro	76.6	64.8	57.1
MLLM Visit Sequence	69.2	58.3	50.2

Table 17: Shortcut-control baselines in the GPT-4o-mini setting. Metadata-only and modality-availability-only inputs remain far below image-based methods, suggesting that the benchmark is not solved by demographic or acquisition-pattern shortcuts. Unordered image evidence is informative but remains below full LUMINA, supporting the importance of explicit longitudinal organization.

Input setting	AD-continuum Bal. Acc.	Progression Bal. Acc.
Metadata only	41.8	56.7
Modality availability only	36.9	52.8
Images without longitudinal order	75.8	65.8
Full LUMINA	85.1	77.7

replaces task-specific inspection selection with a fixed generic inspection policy.

The shared label-normalization control shows that *MLLM Visit Sequence* does not lose performance because of a different output vocabulary or parsing procedure. Removing task-specific routing lowers LUMINA from 85.1 to 82.1 balanced accuracy on AD-continuum classification and from 77.7 to 73.6 on progression forecasting. The no-routing variant remains above *MLLM Visit Sequence* at 74.4 and 65.0, showing that task-specific routing contributes to performance but does not account for the full gain from structured evidence construction.

J Qualitative Failure Analysis

Qualitative inspection of misclassified GPT-4o-mini cases identifies recurring errors from insufficient evidence in the visible prefix and from difficulty reconciling available multimodal evidence. Table 19 summarizes the observed patterns.

Representative Preclinical AD error. One ground-truth Preclinical AD case was predicted as CU. The visible prefix showed low but detectable amyloid burden and subtle tau uptake, while MRI remained stable and provided little downstream structural support. The model treated the weak cross-modality support as evidence against pathology rather than separating early biological burden from absent downstream change.

Representative Clear progression error. One ground-truth Clear progression case was predicted

as Stable. MRI appeared stable and tau abnormality was early, with no clear worsening visible in the prefix. The held-out follow-up subsequently entered the Clear progression outcome bucket. This error reflects a future outcome that was not visibly expressed at the target visit rather than a failure to organize an available worsening signal.

J.1 Sensitivity to Interpretation Calibration

We evaluate sensitivity to the interpretation-calibration protocol. The shared reference comparator is a fixed, non-label-bearing visual comparator available to all evaluated methods under the same input protocol. The comparator pool is constructed offline from cognitively unimpaired subjects and frozen before evaluation. Within this fixed pool, target-specific matching uses age as the primary variable and sex as the secondary variable. Matching for an evaluated subject does not use that subject’s diagnosis, cognitive scores, target or progression labels, or future outcomes. LUMINA’s contribution is therefore not access to the reference itself, but a structured reasoning protocol that uses calibrated visual evidence, prior same-modality anchors, memory, and verification to construct longitudinal evidence.

This protocol is motivated by an empirical failure mode in the current AD-continuum benchmark. Without a shared visual comparator, isolated PET interpretation by the tested MLLM baselines was unstable, with frequent overcalling of subtle uptake

Table 18: Prompt-routing and label-normalization controls in the GPT-4o-mini setting. *MLLM Visit Sequence* uses the same output vocabulary and parsing procedure as LUMINA. Task-specific routing improves LUMINA by prioritizing relevant longitudinal comparisons, but the no-routing variant remains substantially stronger than the non-agent baseline.

Setting	AD-continuum Bal. Acc.	Progression Bal. Acc.
MLLM Visit Sequence with shared label normalization	74.4	65.0
LUMINA without task-specific routing	82.1	73.6
LUMINA with task-specific routing	85.1	77.7

Table 19: Qualitative failure taxonomy from representative GPT-4o-mini errors.

Failure mode	Main affected class	Typical pattern	Interpretation
Weak downstream support	Symptomatic AD	Biological evidence is present, but structural evidence is weak, stable, or not clearly worsening.	Biological burden alone may not establish symptomatic-stage support.
Subtle PET abnormality ambiguity	CU / Preclinical AD	Reference-relative PET deviation is weak and not reinforced by the visible trajectory.	Visual uncertainty can be mistaken for early pathology or for absence of pathology.
Incomplete same-modality history	Preclinical / Symptomatic AD	Current abnormality is visible, but prior anchors are missing, sparse, or stale.	Longitudinal change is under-supported even when current burden is apparent.
Cross-modality conflict	Preclinical / Symptomatic AD	Molecular imaging suggests pathology while structural evidence remains weak or stable.	Reconciliation remains ambiguous when modalities support different disease stages.
Future outcome not visible in prefix	Mild / Clear progression	Held-out follow-up worsens, but the visible prefix contains little or no corresponding change.	Later worsening is not expressed in the evidence available at the target visit.
Current burden without near-term decline	Mild / Clear progression	Current abnormality is present, but the held-out outcome remains stable within the forecast horizon.	Disease burden does not necessarily imply near-term progression.

and a tendency to collapse predictions toward the Preclinical AD class. This failure mode also affects LUMINA when references are removed, because longitudinal reasoning can organize evidence but cannot fully correct systematic visual misinterpretation of the underlying images. Progression forecasting is less dominated by this failure mode because it depends more on visible change over time than on current-state biological interpretation alone, but it still inherits visual-interpretation noise from the underlying studies. We therefore expose the same shared reference comparators to every compared method so that method differences reflect how calibrated evidence is organized and verified, rather than differences in interpretation calibration. Although the reference pool consists of cognitively unimpaired subjects, the evaluated methods are not given target labels or outcome information for the test subject; the reference acts as a fixed visual comparator for calibration, not as a retrieved label-bearing case.

Interpretation-calibration sensitivity protocol.

To separate interpretation calibration from structured longitudinal reasoning, we compare four settings with GPT-4o-mini. In the no-reference setting, the method receives the same longitudinal visits but no shared visual comparator. In the reference-comparator setting, the method receives the standard fixed comparator selected by age as

the primary matching variable and sex as the secondary matching variable. No setting uses diagnosis, cognitive scores, target labels, or future outcomes for reference selection. Other sanity checks in Appendix I.1 use the standard interpretation-calibration protocol; the experiments in this subsection explicitly remove references only to isolate the effect of interpretation calibration.

The reference sensitivity results are consistent with the aggregate results reported elsewhere in the appendix. The reference-comparator rows correspond to the standard evaluation protocol. The balanced-accuracy difference between *MLLM Visit Sequence* and LUMINA is 10.7 points for AD-continuum classification and 12.7 points for progression forecasting, consistent with the paired bootstrap comparisons in Table 14 up to rounding.

Shared reference comparators improve both *MLLM Visit Sequence* and LUMINA, with a larger effect on AD-continuum classification than on progression forecasting. LUMINA also remains stronger after both methods receive the same interpretation calibration, separating the benefit of structured longitudinal reasoning from the interpretation-calibration protocol.

Label-collapse analysis. We further examine whether reference comparators reduce systematic prediction collapse in AD-continuum classification. For each method, we measure the proportion of

Table 20: Reference-comparator sensitivity in the GPT-4o-mini setting. Without a shared visual comparator, *MLLM Visit Sequence* performs poorly on AD-continuum classification, reflecting unstable isolated PET interpretation and collapse toward the Preclinical AD class. Removing references also substantially reduces LUMINA, showing that structured longitudinal reasoning cannot compensate for systematic visual misinterpretation when the underlying observations are poorly calibrated. Reference comparators improve both methods by calibrating visual interpretation.

Setting	AD-continuum Bal. Acc.	AD-continuum Macro-F1	Progression Bal. Acc.	Progression Macro-F1
MLLM Visit Sequence, no reference	43.3	36.8	58.9	52.1
MLLM Visit Sequence, reference comparator	74.4	69.4	65.0	59.2
LUMINA, no reference	48.7	42.9	68.6	61.9
LUMINA, reference comparator	85.1	81.5	77.7	71.3

Table 21: Prediction-distribution analysis for AD-continuum classification in the GPT-4o-mini setting. Without a shared visual comparator, *MLLM Visit Sequence* strongly over-predicts the Preclinical AD class. LUMINA without references reduces this collapse only modestly, indicating that longitudinal reasoning cannot fully compensate for miscalibrated visual interpretation. Reference comparators further stabilize the prediction distribution.

Setting	CU	Preclinical AD	Symptomatic AD
Ground-truth distribution	48.2	22.9	28.9
MLLM Visit Sequence, no reference	12.4	71.6	16.0
MLLM Visit Sequence, reference comparator	28.7	41.6	29.7
LUMINA, no reference	17.2	64.8	18.0
LUMINA, reference comparator	34.2	35.7	30.1

predictions assigned to each AD-continuum class in the GPT-4o-mini setting.

The label-distribution analysis clarifies why reference comparators are part of the benchmark protocol. Without interpretation calibration, AD-continuum prediction is partly dominated by unstable isolated visual interpretation, particularly for PET images. The no-reference *MLLM Visit Sequence* baseline collapses strongly toward the Preclinical AD class, and LUMINA without references only modestly reduces this collapse. With reference comparators, both methods produce more balanced prediction distributions, indicating that the reference acts as a shared interpretation-calibration comparator rather than a LUMINA-specific source of supervision.

The reference is not selected using target labels, diagnosis fields, cognitive scores, or future outcomes, and it is provided to all compared methods under the same interpretation-calibration protocol.