

Dementia-R1: Reinforced Pretraining and Reasoning from Unstructured Clinical Notes for Real-World Dementia Prognosis

Choonghan Kim^{1,*} Hyunmin Hwang^{1,*} Hangeol Chang^{1,*} Jaemin Kim^{1,*}

Jinse Park^{2,†} Jae-Sung Lim^{3,†} Jong Chul Ye^{1,†}

¹Graduate School of AI, KAIST, Republic of Korea

²Haeundae Paik Hospital, Inje University, Republic of Korea

³Asan Medical Center, University of Ulsan College of Medicine, Republic of Korea

{choonghankim, hyunmin_hwang, hangeol, kjm981995}@kaist.ac.kr

loca99@paik.ac.kr jaesunglim@amc.seoul.kr jong.ye@kaist.ac.kr

Abstract

While Large Language Models (LLMs) have shown strong performance on clinical text understanding, they struggle with longitudinal prediction tasks such as dementia prognosis, which require reasoning over *complex, non-monotonic symptom trajectories* across multiple visits. Standard supervised training lacks explicit annotations for symptom evolution, while direct Reinforcement Learning (RL) is hindered by sparse binary rewards. To address this challenge, we introduce Dementia-R1, an RL-based framework for longitudinal dementia prognosis from unstructured clinical notes. Our approach adopts a Cold-Start RL strategy that pre-trains the model to predict verifiable clinical indices extracted from patient histories, enhancing the capability to reason about disease progression before determining the final clinical status. Extensive experiments show that Dementia-R1 achieves the best overall performance on a real-world unstructured cohort, reaching an AUROC of 84.02% and outperforming models up to 10× larger. The framework also generalizes to Parkinson’s disease dementia prediction in an independent hospital cohort, achieving an AUROC of 78.37%. On the ADNI benchmark, our 7B model attains the highest AUROC among all LLM baselines at 83.17%, demonstrating strong longitudinal reasoning over fluctuating cognitive trajectories. Code is available at <https://github.com/hangeol/dementiar1>.

1 Introduction

The digitalization of healthcare and the widespread adoption of Electronic Health Records (EHRs)

*Equal contribution.

†Corresponding author.

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

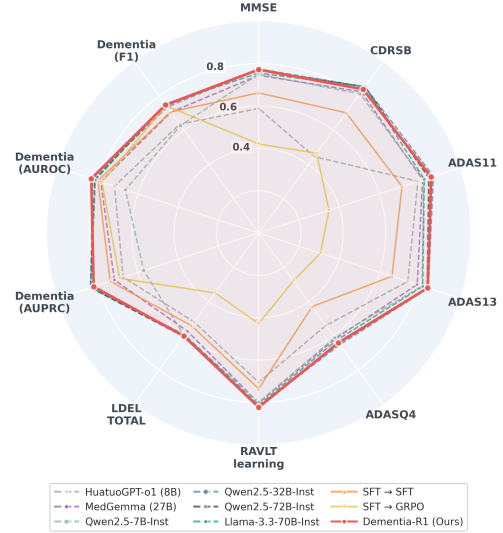


Figure 1: **Multi-dimensional Performance Profile.** Dementia-R1 demonstrates a consistent and balanced performance gain across all dimensions, including intermediate clinical reasoning tasks (e.g., MMSE, CDR-SB, ADAS-Cog) and the final dementia prognosis (F1, AUROC, AUPRC).

have resulted in massive amounts of longitudinal patient data that capture individuals’ clinical histories across months or years. However, approximately 80% of EHR data remains unstructured (Kong, 2019; Jensen et al., 2012), and temporal changes are documented implicitly rather than in structured form. Most existing longitudinal modeling frameworks target structured representations and thus struggle to incorporate unstructured clinical narratives (Waxler et al., 2025; Steinberg et al., 2024; Shmatko et al., 2025).

Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in understanding unstructured medical text for clinical decision support (Wachter and Brynjolfsson, 2024; Silcox et al., 2024). Yet current benchmarks such as MedQA (Jin et al., 2021) evaluate only static, snapshot-style scenarios (Singhal et al., 2025), largely ignoring longitudinal progression.

This is critical for dementia, where diagnosis requires integrating cognitive and functional decline across multiple encounters (Grand et al., 2011; Borson et al., 2013; Knopman and Petersen, 2014) along often non-monotonic trajectories. These longitudinal signals are predominantly documented in unstructured notes, making dementia a particularly challenging testbed (Kruse et al., 2025).

To address this challenge, we introduce **Dementia-R1**, a framework designed for longitudinal reasoning using LLMs through Reinforcement Learning (RL). Dementia prognosis is a representative task for complex disease progression: it requires tracking cognitive and functional changes over months or years from narratives that are hard to quantify (Borson et al., 2013; Knopman and Petersen, 2014). While standard fine-tuning optimizes directly for final labels, RL enables models to learn the underlying reasoning processes before prediction (DeepSeek-AI, 2025; Shao et al., 2024), making it a natural fit for longitudinal clinical inference. However, directly applying RL to a high-level binary prognosis task is challenging due to sparse rewards and implicit reasoning.

We address this issue through a Cold-Start RL strategy with verifiable clinical rewards. RLVR has so far targeted snapshot tasks such as mathematics and code, where the answer is verifiable at prediction time. Longitudinal prognosis differs: its only label is sparse and delayed. We therefore supervise the intermediate trajectory with clinical indices measured at each visit, such as the *Mini-Mental State Examination* (MMSE) (Folstein et al., 1975), *Global Deterioration Scale* (GDS) (Reisberg et al., 1982), and *Clinical Dementia Rating* (CDR) (Morris, 1993). These indices are denser than the final outcome and already recorded in the notes.

The standard SFT-based cold start instead needs distilled or expert-written rationales (Chen et al., 2025a; DeepSeek-AI, 2025). Such rationales are costly for dementia, since they require temporally consistent analysis across visits and expert validation (Kruse et al., 2025). Replacing SFT with RL on verifiable indices removes that requirement. The model acquires reasoning primitives on its own, and a second stage refines them for the final prediction.

Using clinical indices as the Stage 1 target also lets the framework compare which clinical assessments are most predictive of long-term outcomes. MMSE-family signals contribute most on AMC

and HPH, while CDR-SB dominates on ADNI (Sec. 4.8). Outcome-only RL cannot provide this, since it never scores intermediate measurements.

We validate our approach on real-world unstructured clinical notes from the Asan Medical Center (AMC) cohort, an independent cross-disease cohort from Haeundae Paik Hospital (HPH) for Parkinson’s disease dementia (PDD) prediction, and the structured ADNI benchmark (Jack Jr et al., 2008) (Figure 1). Our contributions are as follows:

- We propose Dementia-R1, an RL-based framework that enables explicit temporal reasoning on unstructured clinical notes to predict dementia prognosis.
- We introduce a Cold-Start RL method that replaces rationale-based SFT with RL on clinical indices already present in the record, and show that estimating these indices is crucial for accurate prognosis.
- We show that this design also measures *biomarker informativeness*, identifying which clinical assessments carry the most prognostic signal in each cohort.
- We validate our approach on three cohorts (two hospital systems and a public benchmark) and demonstrate cross-disease generalization to PDD, consistently outperforming general and medical-specialized LLMs.

2 Related Work

Longitudinal Clinical Modeling. Traditional longitudinal modeling has focused on structured electronic health records (EHRs) using RNNs (Choi et al., 2016) and Transformer-based models for tasks like time-to-event prediction (Steinberg et al., 2024), disease trajectory modeling (Shmatko et al., 2025), and medical event forecasting (Waxler et al., 2025). However, these models often overlook the rich behavioral and symptomatic descriptions in unstructured clinical notes. While NYUTron (Jiang et al., 2023) and CARE-AD (Li et al., 2025) have shown the potential of LLMs for unstructured clinical text, they primarily optimize for final outcomes rather than intermediate disease trajectories or temporal patterns. Consequently, current frameworks still lack mechanisms for explicit longitudinal reasoning (Kruse et al., 2025), which motivates our approach.

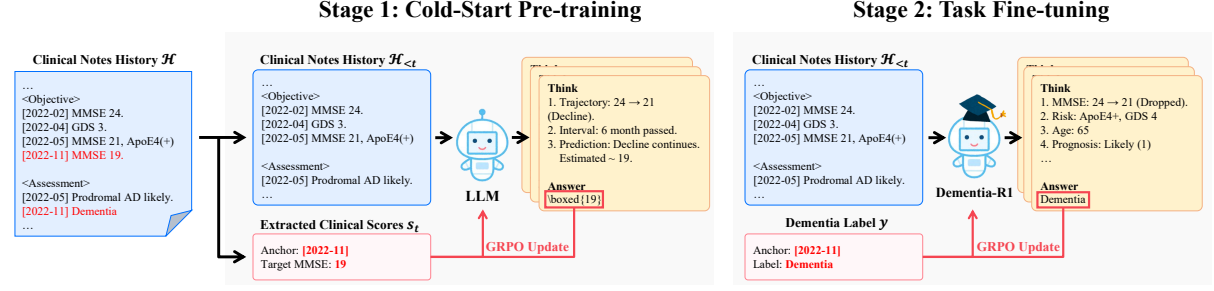


Figure 2: **Overview of the Dementia-R1 Framework.** Stage 1 pre-trains longitudinal reasoning via GRPO on clinical index forecasting; Stage 2 fine-tunes for the final dementia prognosis.

Reasoning Capabilities of Medical LLMs.

Medical LLM reasoning has been enhanced through Chain-of-Thought prompting (Wei et al., 2022) and RL-based approaches such as HuatuoGPT-o1 (Chen et al., 2025a). Recent work has shifted toward RL with Verifiable Rewards (RLVR) (DeepSeek-AI, 2025), yet applying this to clinical tasks remains challenging due to sparse rewards. C-Reason (Kim et al., 2025) uses GRPO (Shao et al., 2024) for sepsis management via masked value prediction, but does not address long-term disease progression. We extend this line of work to longitudinal dementia prediction by training the model to estimate clinical indices before determining the final prognosis.

Verifiable Rewards and Staged RL. Two lines of work inform our design. The first concerns *what* is supervised: process-based feedback rewards intermediate steps rather than only the final answer (Uesato et al., 2022; Lightman et al., 2024), later formalized as RL with Verifiable Rewards (Lambert et al., 2025) and scaled to general reasoning (DeepSeek-AI, 2025) and clinical tasks (Pan et al., 2025). The second concerns *how* training is staged: multi-stage RL pipelines (DeepSeek-AI, 2025; Wen et al., 2025) usually begin from an SFT cold start on distilled rationales (Wei et al., 2025; Chen et al., 2025b), sometimes as a curriculum over difficulty (Pang et al., 2025), since SFT memorizes while RL generalizes (Chu et al., 2025). Both assume a verifiable signal at prediction time. Longitudinal prognosis has none, which motivates our intermediate clinical indices.

3 Methodology: Dementia-R1

Input: from clinical notes to a longitudinal query. Each visit x_i is a free-text note written by the attending clinician. AMC notes follow

the SOAP convention, where the *Objective* field records measurements taken at that visit and the *Assessment* field records the clinician’s impression. HPH notes use structured S/O/A/P fields with separate index, imaging, and gait records. Clinical indices appear only at visits where the corresponding test was performed, so the longitudinal signal is irregular and largely narrative. Appendix Figures 16 and 18 show full anonymized examples.

We order a patient’s notes chronologically as $\mathcal{H} = \{x_1, \dots, x_t\}$. Given a *target anchor* T_{anchor} , all preceding notes $\mathcal{H}_{<T} = \{x_i \mid i < T\}$ are concatenated into one document, which together with the prediction horizon forms the model query q (Appendix A.20).

Task. We formulate the task as determining the final clinical status $y \in \{0, 1\}$ at T_{anchor} , conditioned on $\mathcal{H}_{<T}$. Because trajectories are non-monotonic, this requires distinguishing temporary fluctuations from persistent decline. We employ a two-stage RL framework using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with verifiable clinical rewards (Figure 2).

3.1 Constructing Verifiable Pretraining Data

Raw notes lack ground truth for longitudinal reasoning, so we build a pre-training dataset by extracting verifiable clinical indices via an auxiliary LLM $\mathcal{E}$: $s_t = \mathcal{E}(x_t)$, $s_t \in \mathcal{S}$, where $\mathcal{S}$ covers MMSE (0–30), GDS (1–7), and CDR (0–3). A human expert audit over 600 notes confirmed 98.17% slot-level accuracy (Appendix A.15), and the downstream prognosis is robust to far higher extraction noise (Appendix A.16). The resulting dataset $\mathcal{D}_{pre} = \{(\mathcal{H}_{<t}, s_t)\}$ trains the model to forecast s_t from preceding history. Stage 2 test patients are strictly excluded to prevent leakage.

3.2 Stage 1: Cold-Start Pre-training

In this stage, we align the model to reason about clinical trajectories by optimizing it to predict the extracted clinical indices s_t from $\mathcal{D}_{pre}$ using GRPO, which estimates the baseline from a group of outputs without a separate value function.

Verifiable Reward Function (R_{cold}) To accommodate the varying granularity of clinical scales, we define a tolerance-aware reward function. For a predicted index $\hat{s}_t$, ground truth s_t , and error margin δ , we set $\delta = 2$ for MMSE (0–30). This matches the reliable change index of at least 2 points for MMSE (Stein et al., 2012; Hensel et al., 2007). For coarser scales (GDS, CDR), we enforce exact matching ($\delta = 0$). The reward is:

$$R_{cold} = \mathbb{I}(|\hat{s}_t - s_t| \leq \delta), \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the tolerance condition is met and 0 otherwise.

Optimization Objective For each input query q_t with clinical history $\mathcal{H}_{<t}$, we sample a group of G outputs $\{o_t^1, o_t^2, \dots, o_t^G\}$ from the old policy $\pi_{\theta_{old}}$. The policy is optimized to maximize the following:

$$\mathcal{L}(\theta) = \mathbb{E}_{q_t, \{o_t^i\}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(\frac{\pi_{\theta}(o_t^i | q_t)}{\pi_{\theta_{old}}(o_t^i | q_t)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_t^i | q_t)}{\pi_{\theta_{old}}(o_t^i | q_t)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL} \right]. \quad (2)$$

Here, $\beta \mathbb{D}_{KL}$ is the KL-regularization term, ϵ is the clipping range, and A_i is the group-normalized advantage:

$$A_i = \frac{R_{cold}(o_t^i) - \text{mean}(\{R_{cold}(o_t^j)\}_{j=1}^G)}{\text{std}(\{R_{cold}(o_t^j)\}_{j=1}^G)}. \quad (3)$$

This encourages the model to generate reasoning paths that outperform its own group average.

3.3 Stage 2: Task Fine-tuning

After Cold-Start pre-training (Stage 1), we fine-tune the model on the binary prognostic task (Dementia vs. Non-Dementia) using the same GRPO framework in Eq. (2).

Sparse Reward Function (R_{task}) Unlike the granular indices in Stage 1, the final diagnosis is binary:

$$R_{task} = \begin{cases} 1, & \text{if prediction is correct} \\ 0, & \text{if prediction is incorrect} \end{cases} \quad (4)$$

Although this reward is sparse, the reasoning primitives acquired in Stage 1 enable effective longitudinal reasoning for the final prognosis.

4 Experimental Setup

4.1 Datasets

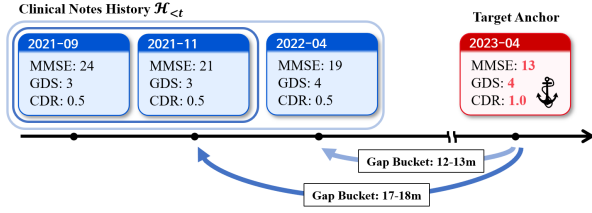
We validate Dementia-R1 on three cohorts: unstructured clinical notes from AMC, the HPH cohort for cross-disease generalization, and the structured ADNI benchmark. The original clinical notes were written in Korean and translated into English using an on-premise auxiliary LLM (Qwen2.5-14B-Instruct); see Appendix A.3 for details. We discuss the implications of this translation step in the Limitations section.

4.1.1 Data Sources and Processing

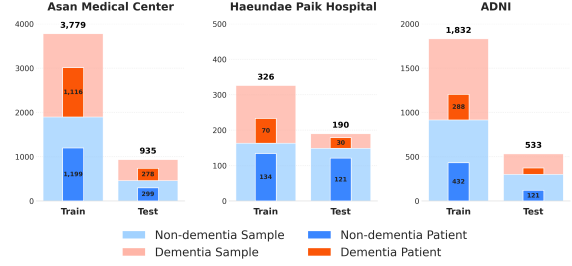
Unstructured Cohort (AMC). De-identified SOAP-formatted clinical notes from ~3,000 neurocognitive disorder patients at AMC (2021–2023). Clinical indices (MMSE, CDR, GDS) were extracted via the LLM-based pipeline (Sec 3.1) as verifiable rewards (details in Appendix A.3.1). **Label.** A patient is labeled as dementia when the Global Deterioration Scale (GDS) reaches ≥ 4 , which indicates cognitive decline severe enough to impair function. Dementia specialists assign GDS staging during routine care, integrating caregiver interview, functional history, MMSE, neuropsychological testing where available, and other clinical information.

Cross-Disease Cohort (HPH). Longitudinal EMRs of Parkinson’s disease patients (ICD-10: G20, 2010–2024) from HPH, including multimodal data (brain MRI, PET, gait analysis, MMSE, clinical notes). Patients were categorized into PD with and without dementia for cross-disease generalization to PDD. **Label.** A patient is labeled as PDD only when three criteria hold: (i) cognitive decline with impairment of activities of daily living that meets PDD criteria, (ii) an ICD-10 F02.3 assignment, and (iii) a cognitive-enhancer prescription.

Structured Benchmark Cohort (ADNI). The ADNI dataset (Jack Jr et al., 2008) serves as a structured benchmark (Appendix A.2). Tabular records were linearized into chronological textual logs, with seven indices (Gelir et al., 2025) (e.g., MMSE, CDR-SB) as verifiable rewards (Appendix A.4.2).



(a) **Longitudinal Sample Construction.** Patient history is retrospectively sliced around a Target Anchor, following a unified protocol for both unstructured (AMC, HPH) and structured (ADNI) datasets.



(b) **Dataset Overview.** Visualization of sample and patient counts. Training sets are balanced to prevent bias, while test sets retain natural patient prevalence.

Figure 3: **Experimental setup:** sample construction and cohort composition.

4.1.2 Longitudinal Sample Construction

We applied a unified protocol across both modalities (Figure 3a): a **Target Anchor** at the patient’s last confirmed assessment with all prior history as input, a **Prediction Target** (clinical indices for Stage 1; binary diagnosis for Stage 2), and a **Gap Bucket** discretizing the prediction horizon into temporal bins (1-month for Stage 1; 6-month for Stage 2, excluding $<6m$; $>18m$ consolidated for ADNI). All splits are patient-level to prevent leakage: Stage 2 test patients are excluded from Stage 1, and only notes before the anchor are used. Training sets are balanced (1:1); test sets retain natural prevalence (Figure 3b).

4.2 Baselines

We compare Qwen2.5-7B-Instruct (Qwen Team, 2024) under six training configurations (**Zero-shot CoT**; single-stage **SFT/GRPO**; two-stage **SFT** $\rightarrow$ **SFT** and **SFT** $\rightarrow$ **GRPO**; **Random Forest** on ADNI), as well as external LLMs: medical-specialized (HuatuoGPT-o1, MedGemma) and general-purpose at larger scales (Qwen2.5-32B/72B, Llama-3.3-70B). Our method is **Dementia-R1 (GRPO $\rightarrow$ GRPO)**. Proprietary API models were excluded due to IRB restrictions on external data transmission.

4.3 Implementation Details

SFT. SFT baselines use knowledge distillation from Qwen2.5-32B-Instruct-AWQ, which generates CoT rationales from ground-truth labels. The student is fine-tuned on concatenated (Question, Note, CoT, Answer) sequences for three epochs.

Dementia-R1. We train with GRPO ($G=8$, effective batch size 8). Full training details are in Appendix A.17.

Evaluation protocol. All results are mean $\pm$ std over five seeds. We report AUROC and AUPRC derived from token-level logits (Schick and Schütze, 2021; Kadavath et al., 2022; Zhao et al., 2021) alongside Accuracy and F1. Significance is tested via DeLong’s (DeLong et al., 1988), McNemar’s (McNemar, 1947), and bootstrap permutation ($B=10,000$).

4.4 Real-World Unstructured Data Results

Dementia prognosis. On the AMC cohort (Table 1), Dementia-R1 achieves the best overall performance, reaching 84.02% AUROC and 77.31% F1. It outperforms every specialized 7B baseline, including direct GRPO without Stage 1 and the hybrid SFT $\rightarrow$ GRPO pipeline, with a significant AUROC gain in each case ($p < 0.01$). It also surpasses general-purpose LLMs up to $10\times$ larger, significantly on every metric for models up to 70B ($p < 0.01$). Against Qwen2.5-72B-Instruct, the gains in Accuracy, Precision, Recall, and AUPRC are significant ($p < 0.01$), while the F1 (77.31 vs. 77.30) and AUROC (84.02 vs. 83.41) differences are within noise. Stage 1 clinical-index pretraining therefore provides a substantial advantage on real-world unstructured notes.

Cross-disease generalization (PDD prediction).

On the HPH cohort for Parkinson’s disease dementia (PDD) prediction (Table 2), Dementia-R1 again achieves the best AUROC (78.37%) and AUPRC (51.03%). It outperforms Qwen2.5-72B-Instruct (72.37% AUROC, $p < 0.01$) and the strongest 7B baseline, SFT $\rightarrow$ GRPO (74.75% AUROC). Against SFT $\rightarrow$ GRPO only the AUPRC gain is significant ($+11.78$, $p < 0.01$), so the advantage here lies mainly in ranking the positive class under low prevalence. The framework thus transfers beyond

Table 1: **Results on AMC.** **Bold/underline** = best/second-best. Mean $\pm$ std over 5 seeds. */*: $p < 0.05/0.01$ vs. Dementia-R1.

Method	Size	Accuracy ($\uparrow$)	Precision ($\uparrow$)	Recall ($\uparrow$)	F1 score ($\uparrow$)	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
<i>External LLMs</i>							
HuatuoGPT-o1	8B	67.53 $\pm$ 1.1**	71.33 $\pm$ 1.7**	60.63 $\pm$ 1.2**	65.53 $\pm$ 1.1**	73.89 $\pm$ 0.5**	74.81 $\pm$ 0.8**
MedGemma	27B	67.27 $\pm$ 0.6**	62.42 $\pm$ 0.6**	89.79 $\pm$ 0.4**	73.64 $\pm$ 0.3**	74.14 $\pm$ 0.7**	71.06 $\pm$ 1.1**
Qwen2.5-7B-Inst	7B	72.81 $\pm$ 0.8**	<u>74.37</u> $\pm$ 1.2**	71.13 $\pm$ 0.6**	72.71 $\pm$ 0.7**	78.59 $\pm$ 0.8**	79.24 $\pm$ 0.8**
Qwen2.5-32B-Inst	32B	62.03 $\pm$ 0.5**	57.67 $\pm$ 0.3**	95.63 $\pm$ 0.4**	71.95 $\pm$ 0.2**	79.93 $\pm$ 1.0**	79.33 $\pm$ 1.6**
Qwen2.5-72B-Inst	72B	74.46 $\pm$ 0.6**	70.60 $\pm$ 0.7**	85.42 $\pm$ 0.7**	<u>77.30</u> $\pm$ 0.4	<u>83.41</u> $\pm$ 0.4	<u>83.44</u> $\pm$ 0.6**
Llama-3.3-70B-Inst	70B	67.27 $\pm$ 0.3**	62.14 $\pm$ 0.3**	<u>91.39</u> $\pm$ 0.5**	73.98 $\pm$ 0.3**	80.70 $\pm$ 0.3**	81.11 $\pm$ 0.6**
<i>Specialized Models</i>							
SFT w/o Stage 1	7B	64.68 $\pm$ 0.9**	60.97 $\pm$ 0.6**	85.13 $\pm$ 1.1**	71.05 $\pm$ 0.8**	71.65 $\pm$ 1.7**	69.82 $\pm$ 2.0**
GRPO w/o Stage 1	7B	72.15 $\pm$ 0.9**	70.19 $\pm$ 0.7**	78.74 $\pm$ 1.3**	74.21 $\pm$ 0.9**	79.88 $\pm$ 0.6**	80.46 $\pm$ 0.9**
SFT w/o Stage 2	7B	74.16 $\pm$ 0.4**	71.96 $\pm$ 0.3**	80.67 $\pm$ 0.8**	76.07 $\pm$ 0.4*	76.31 $\pm$ 0.9**	69.32 $\pm$ 1.7**
GRPO w/o Stage 2	7B	73.28 $\pm$ 0.6**	70.42 $\pm$ 0.8**	81.97 $\pm$ 0.4**	75.75 $\pm$ 0.4**	81.68 $\pm$ 0.3**	81.27 $\pm$ 0.2**
SFT $\rightarrow$ SFT	7B	<u>74.97</u> $\pm$ 0.8**	73.65 $\pm$ 0.8**	79.16 $\pm$ 1.0**	76.30 $\pm$ 0.8	77.15 $\pm$ 1.7**	72.41 $\pm$ 2.1**
SFT $\rightarrow$ GRPO	7B	73.60 $\pm$ 1.0**	71.00 $\pm$ 1.1**	81.58 $\pm$ 1.0**	75.92 $\pm$ 0.9**	77.75 $\pm$ 1.5**	72.69 $\pm$ 2.0**
Dementia-R1 (Ours)	7B	77.45 $\pm$ 0.2	79.26 $\pm$ 0.4	75.46 $\pm$ 0.7	77.31 $\pm$ 0.3	84.02 $\pm$ 0.3	85.73 $\pm$ 0.4

Table 2: **Results on HPH.** Notation follows Table 1.

Method	Size	Accuracy ($\uparrow$)	Precision ($\uparrow$)	Recall ($\uparrow$)	F1 score ($\uparrow$)	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
<i>External LLMs</i>							
HuatuoGPT-o1	8B	69.68 $\pm$ 2.2**	35.11 $\pm$ 3.2**	47.32 $\pm$ 5.0**	40.23 $\pm$ 3.5**	63.58 $\pm$ 3.2**	32.93 $\pm$ 1.9**
MedGemma	27B	70.21 $\pm$ 3.0**	38.69 $\pm$ 4.0**	64.39 $\pm$ 5.5	48.31 $\pm$ 4.6	73.13 $\pm$ 2.6**	41.06 $\pm$ 4.4**
Qwen2.5-7B-Inst	7B	71.26 $\pm$ 1.6**	35.91 $\pm$ 3.7**	43.41 $\pm$ 8.5**	39.19 $\pm$ 5.6**	67.90 $\pm$ 3.2**	39.28 $\pm$ 4.1**
Qwen2.5-32B-Inst	32B	74.00 $\pm$ 1.3	41.10 $\pm$ 2.3	46.83 $\pm$ 1.8**	43.76 $\pm$ 1.9**	68.54 $\pm$ 1.9**	39.20 $\pm$ 1.9**
Qwen2.5-72B-Inst	72B	<u>74.63</u> $\pm$ 1.4	42.57 $\pm$ 2.6	50.24 $\pm$ 3.6**	46.07 $\pm$ 2.9**	72.37 $\pm$ 1.1**	<u>44.78</u> $\pm$ 2.6*
Llama-3.3-70B-Inst	70B	57.58 $\pm$ 2.0**	30.28 $\pm$ 1.6**	74.15 $\pm$ 5.2**	42.98 $\pm$ 2.3**	67.41 $\pm$ 1.4**	34.91 $\pm$ 0.7**
<i>Specialized Models</i>							
SFT w/o Stage 1	7B	65.79 $\pm$ 1.1**	32.96 $\pm$ 1.4**	56.59 $\pm$ 2.8	41.65 $\pm$ 1.7**	63.75 $\pm$ 1.8**	30.39 $\pm$ 3.2**
GRPO w/o Stage 1	7B	72.95 $\pm$ 1.8	38.37 $\pm$ 3.9**	41.95 $\pm$ 6.2**	39.98 $\pm$ 4.5**	69.96 $\pm$ 3.4**	38.69 $\pm$ 2.4**
SFT w/o Stage 2	7B	41.68 $\pm$ 2.7**	21.93 $\pm$ 2.5**	<u>66.83</u> $\pm$ 9.6	33.01 $\pm$ 4.0**	52.95 $\pm$ 2.7**	25.75 $\pm$ 3.6**
GRPO w/o Stage 2	7B	74.32 $\pm$ 1.7	41.35 $\pm$ 3.2	44.88 $\pm$ 3.6**	42.99 $\pm$ 3.0**	69.73 $\pm$ 0.9**	38.53 $\pm$ 2.6**
SFT $\rightarrow$ SFT	7B	38.95 $\pm$ 3.8**	20.80 $\pm$ 1.5**	64.88 $\pm$ 3.3	31.49 $\pm$ 2.0**	49.14 $\pm$ 1.2**	22.47 $\pm$ 1.7**
SFT $\rightarrow$ GRPO	7B	<u>74.63</u> $\pm$ 1.2	44.11 $\pm$ 1.8	65.37 $\pm$ 2.4	52.66 $\pm$ 1.9	74.75 $\pm$ 2.6	39.25 $\pm$ 3.3**
Dementia-R1 (Ours)	7B	75.79 $\pm$ 1.6	45.67 $\pm$ 2.6	63.41 $\pm$ 4.1	53.06 $\pm$ 2.8	78.37 $\pm$ 1.8	51.03 $\pm$ 2.7

Table 3: **Results on ADNI Benchmark.** Includes Random Forest as ML baseline. Notation follows Table 1.

Model Method	Size	Accuracy ($\uparrow$)	Precision ($\uparrow$)	Recall ($\uparrow$)	F1 score ($\uparrow$)	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
<i>ML Baseline</i>							
Random Forest	—	83.46 $\pm$ 0.6	83.57 $\pm$ 0.7	77.13 $\pm$ 1.1	80.22 $\pm$ 0.7	90.96 $\pm$ 0.2	89.68 $\pm$ 0.3
<i>External LLMs</i>							
HuatuoGPT-o1	8B	62.51 $\pm$ 2.1**	55.30 $\pm$ 1.8**	74.59 $\pm$ 2.2**	63.51 $\pm$ 1.9**	71.69 $\pm$ 1.8**	67.34 $\pm$ 2.9**
MedGemma	27B	73.47 $\pm$ 1.0**	68.54 $\pm$ 1.2**	72.70 $\pm$ 0.8**	70.56 $\pm$ 1.0**	78.48 $\pm$ 1.0**	71.71 $\pm$ 1.1**
Qwen2.5-7B-Inst	7B	62.48 $\pm$ 1.3**	55.41 $\pm$ 1.1**	72.62 $\pm$ 1.4**	62.85 $\pm$ 1.2**	66.23 $\pm$ 1.7**	57.17 $\pm$ 2.1**
Qwen2.5-32B-Inst	32B	76.47 $\pm$ 0.7	72.16 $\pm$ 0.8	75.19 $\pm$ 1.3**	73.64 $\pm$ 0.9	<u>82.51</u> $\pm$ 0.5	<u>83.50</u> $\pm$ 0.5
Qwen2.5-72B-Inst	72B	80.00 $\pm$ 0.7**	82.80 $\pm$ 1.4**	68.50 $\pm$ 0.6**	<u>74.97</u> $\pm$ 0.8	81.45 $\pm$ 0.5*	83.56 $\pm$ 0.6
Llama-3.3-70B-Inst	70B	<u>78.91</u> $\pm$ 1.2*	<u>77.27</u> $\pm$ 1.9**	73.39 $\pm$ 1.1**	75.27 $\pm$ 1.3	81.17 $\pm$ 0.4*	83.48 $\pm$ 0.5
<i>Specialized Models</i>							
SFT w/o Stage 1	7B	69.49 $\pm$ 2.0**	62.39 $\pm$ 2.0**	76.22 $\pm$ 1.7	68.61 $\pm$ 1.9**	73.40 $\pm$ 1.7**	63.03 $\pm$ 2.1**
GRPO w/o Stage 1	7B	66.34 $\pm$ 1.3**	59.34 $\pm$ 1.4**	73.22 $\pm$ 0.7**	65.55 $\pm$ 1.1**	72.97 $\pm$ 0.6**	67.18 $\pm$ 1.3**
SFT w/o Stage 2	7B	76.59 $\pm$ 0.9	71.89 $\pm$ 1.3	<u>76.31</u> $\pm$ 0.6*	74.03 $\pm$ 0.8	81.74 $\pm$ 1.9	76.12 $\pm$ 2.4**
GRPO w/o Stage 2	7B	76.44 $\pm$ 1.2	74.05 $\pm$ 1.6*	70.99 $\pm$ 1.6**	72.48 $\pm$ 1.4*	79.07 $\pm$ 0.8**	77.65 $\pm$ 1.3**
SFT $\rightarrow$ SFT	7B	75.57 $\pm$ 0.5	74.27 $\pm$ 1.1*	67.55 $\pm$ 1.1**	70.74 $\pm$ 0.6**	79.90 $\pm$ 0.3**	73.77 $\pm$ 1.1**
SFT $\rightarrow$ GRPO	7B	77.34 $\pm$ 0.6	73.56 $\pm$ 0.9	75.19 $\pm$ 0.5**	74.37 $\pm$ 0.5	77.88 $\pm$ 1.5**	69.09 $\pm$ 2.3**
Dementia-R1	7B	76.92 $\pm$ 0.8	71.48 $\pm$ 1.3	78.63 $\pm$ 0.7	74.87 $\pm$ 0.6	83.17 $\pm$ 1.3	82.00 $\pm$ 1.1

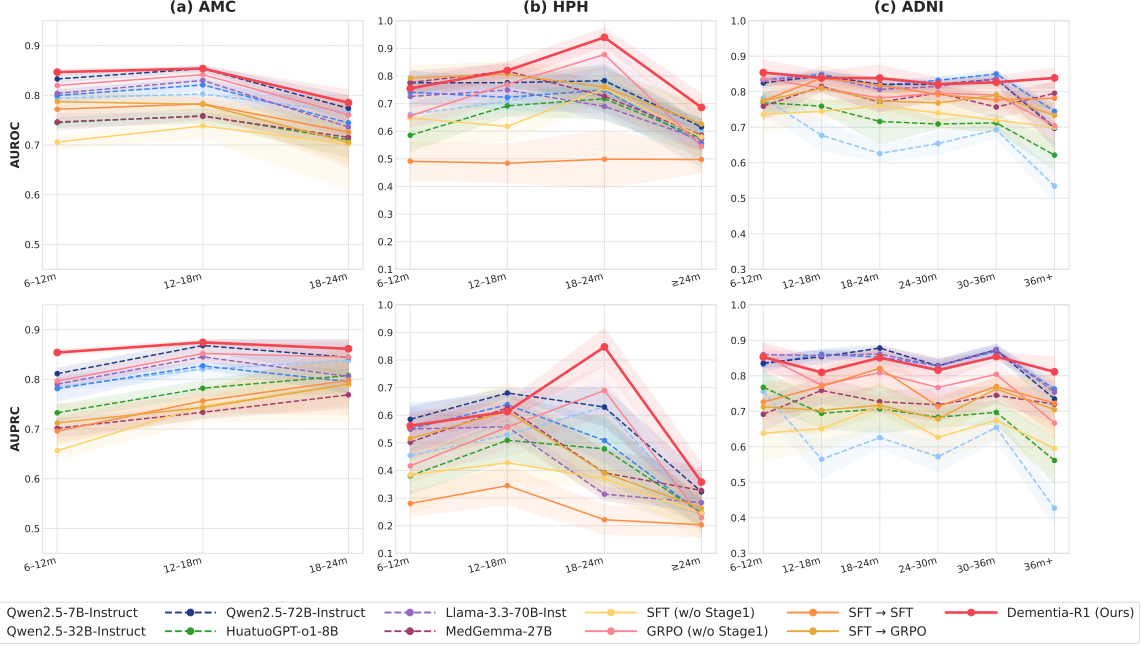


Figure 4: **Performance across time gaps.** AUROC (top) and AUPRC (bottom) by prediction horizon for (a) AMC, (b) HPH, and (c) ADNI. Dementia-R1 achieves the highest AUROC at the longest prediction horizons across all three cohorts.

AMC to a distinct cross-disease, cross-hospital setting.

4.5 Generalization to Structured Benchmarks

On the structured ADNI benchmark (Table 3), Dementia-R1 achieves the highest AUROC among all LLM baselines at 83.17%, while maintaining an F1 score (74.87%) comparable to substantially larger models. In addition, its Stage 1 performance on fine-grained intermediate targets such as CDRSB and ADAS remains competitive with strong larger-scale baselines (Figure 1; Appendix Table 14). These results suggest that the proposed framework generalizes beyond unstructured hospital notes and remains effective on structured longitudinal benchmark data.

As expected, the Random Forest baseline dominates on ADNI, which is well suited to classical tabular models; our framework instead targets unstructured longitudinal narratives and remains competitive even in this structured regime. This ordering is not ADNI-specific: ClinicalBench (Chen et al., 2026) evaluates 22 general and medical LLMs from 0.5B to 70B under prompting, prompt engineering, and fine-tuning, and finds that none surpasses traditional ML on tabular clinical prediction. Nor is the gap attributable to our linearization: substituting the best-performing serialization format from each of two established references leaves AUROC unchanged or lower (Appendix A.4.2, Table 9).

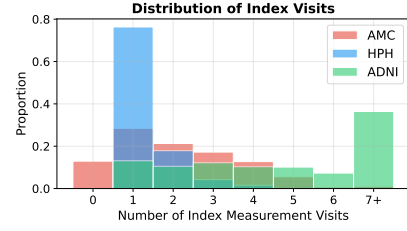


Figure 5: **Clinical Index Density.** ADNI ($\sim 4.3\times$) and AMC ($\sim 1.6\times$) have more index visits per patient than HPH.

4.6 Temporal Robustness Analysis

Figure 4 shows Dementia-R1 remains competitive across prediction horizons, achieving the best AUROC at the longest horizons on AMC (78.51% at 18–24m) and ADNI (83.90% at 36m+). On ADNI, the margin over Qwen2.5-32B/72B widens significantly in the longest-term setting, highlighting Stage 1’s benefit for extended reasoning. In HPH, Dementia-R1 performs best overall, though performance drops more noticeably in the $\geq 24m$ bucket. This aligns with HPH’s lower density of clinical indices relative to AMC and ADNI (Figure 5), suggesting that sparse measurements complicate long-range prediction.

4.7 Neurologist Evaluation

Two board-certified neurologists conducted a blinded pairwise evaluation against Qwen2.5-32B-Instruct across five clinical dimensions and two independent cohorts (AMC and HPH), yielding 100 judgment pairs. Inter-rater agreement was moder-

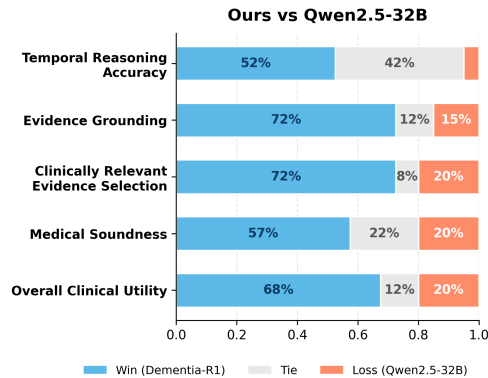


Figure 6: **Neurologist Blind Pairwise Evaluation.** Results from two cohorts (AMC and HPH) comparing Dementia-R1 against Qwen2.5-32B-Instruct across five clinical dimensions.

ate (Gwet’s $AC1 = 0.54$). Dementia-R1 achieved a 68% win rate in Overall Clinical Utility, 72% in both Evidence Grounding and Clinically Relevant Evidence Selection, and a 95% Win/Tie rate in Temporal Reasoning Accuracy (Figure 6). We also conducted the same evaluation against Qwen2.5-72B-Instruct; however, inter-rater agreement was insufficient, so we report only the 32B comparison.

4.8 Ablation Study

Effect of Stage 1 Pre-training. Figure 7 shows that incorporating Stage 1 consistently improves convergence stability and peak AUROC across AMC, HPH, and ADNI. However, the gain on HPH is smaller than that on AMC. Since clinical-index annotations are substantially sparser in HPH than in AMC (Figure 5), we hypothesize that this performance gap stems from the limited availability of intermediate clinical-index supervision. To test this hypothesis, we conduct an ablation study on the AMC cohort by progressively reducing the density of clinical-index labels. Figure 8 shows that decreasing supervision density monotonically degrades downstream prognosis performance on AMC. Taken together, these results indicate that richer intermediate supervision leads to stronger final longitudinal prediction.

Choice of Stage 1 Target. Using a fixed 1,000-sample exploratory ablation, GDS-only pretraining slightly outperforms the combined-target setting on AMC and HPH (Appendix Table 30). We do not interpret this as a universal ranking of index importance. Rather, the usefulness of a pretraining target appears to depend on scale granularity, annotation density, and cohort-specific availability. A coarser target such as GDS may offer a cleaner op-

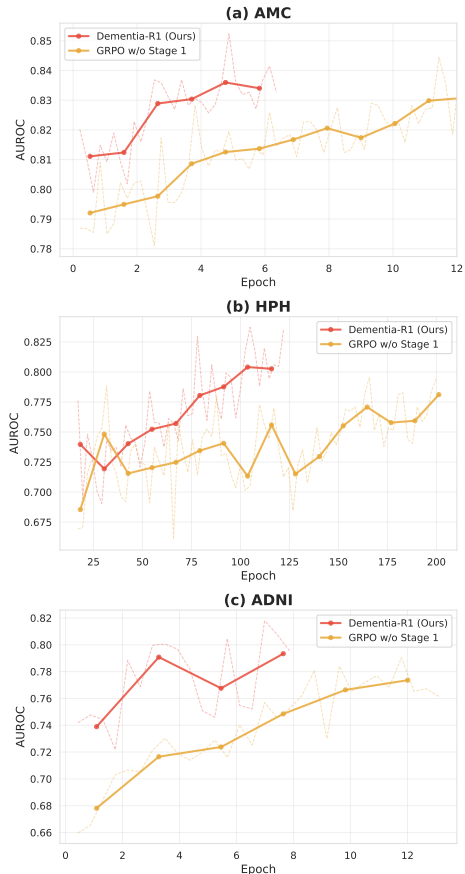


Figure 7: **Stage 1 Pre-training Impact.** AUROC over training epochs on (a) AMC, (b) HPH, and (c) ADNI. Solid lines show 500-step binned averages; dashed lines show raw checkpoint values.

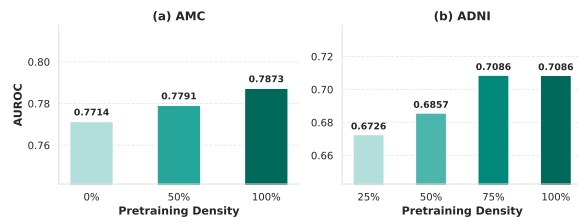


Figure 8: **Pretraining Density Ablation.** Best overall AUROC at varying Stage 1 index density levels on (a) AMC and (b) ADNI. Performance degrades monotonically with reduced density.

timization signal during Stage 1, whereas finer or sparser targets can be harder to exploit consistently. For example, CDR is substantially sparser than MMSE and GDS in AMC (Appendix Figure 14), which may reduce its utility during Stage 1 despite its clinical relevance at inference. The most effective pretraining target is therefore not always the one that contributes most at inference, as the masking analysis below shows. Together with those results, this suggests that different clinical indices play distinct roles in pretraining and in downstream prognosis.

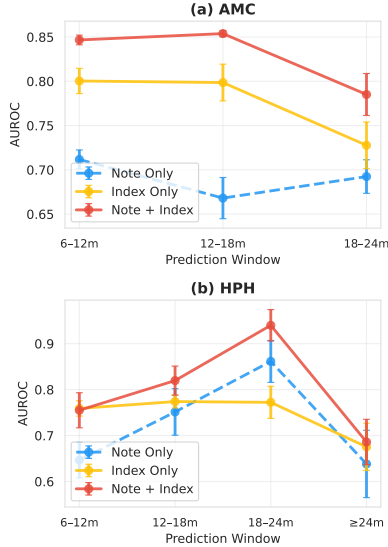


Figure 9: **Note vs. Index Ablation.** AUROC by prediction window on (a) AMC and (b) HPH. Three input settings are compared: **Note Only** (raw clinical text), **Index Only** (structured MMSE/GDS/CDR), and **Note + Index** (combined).

Role of Clinical Indices at Inference. We compare three input settings: **Note Only** (raw clinical text), **Index Only** (structured MMSE/GDS/CDR), and **Note + Index** (combined). Note + Index performs best on both AMC and HPH (Figure 9), showing that structured indices and free-text notes provide complementary prognostic information. The benefit of index input tracks measurement density: AMC has far more index-bearing visits per patient than HPH (Figure 5), so explicit indices give a stronger signal there. We also mask one or more indices at inference, writing **−M** for MMSE, **−G** for GDS, and **−C** for CDR (CDRSB on ADNI). MMSE-family signals contribute most on AMC and HPH, while CDRSB has the largest effect on ADNI (Figure 10). Performance nevertheless remains substantial even when all indices are masked, indicating that the model has internalized longitudinal reasoning beyond explicit index copying. Appendix A.13 confirms this on the naturally occurring zero-index subsets. Non-monotonic trajectories are a further stress case. On the 209 AMC patients whose GDS trajectory crosses the dementia cutoff, Dementia-R1 leads all nine baselines, and its margin over the direct Stage 1 ablation widens to 7.35 AUROC points (Appendix A.14).

4.8.1 Generalization Across Backbones

Repeating all configurations on Llama-3.1-8B-Instruct yields the same conclusion: Dementia-R1 achieves the best F1, AUROC, and AUPRC among

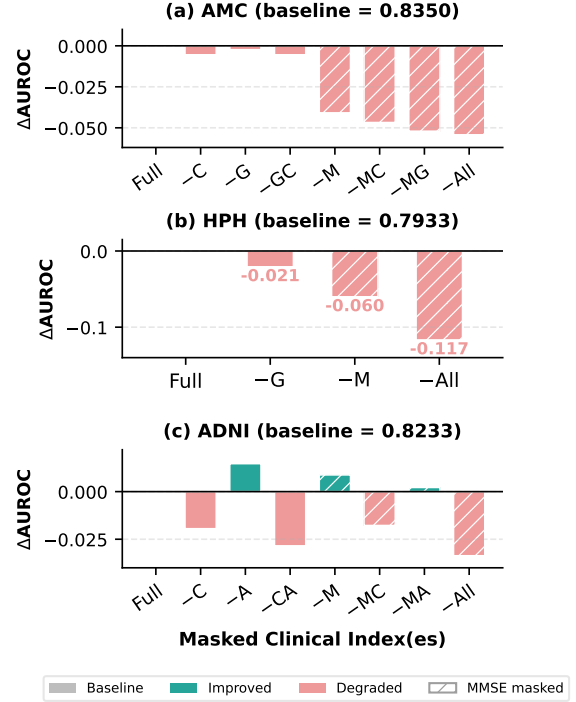


Figure 10: **Clinical Index Masking Ablation.** AUROC change (Δ) when masking clinical indices at inference. Conditions are named by the index removed: **M** = MMSE, **G** = GDS, **C** = CDR in (a) AMC and (b) HPH, and **M** = MMSE, **C** = CDRSB, **A** = ADAS-Cog/memory in (c) ADNI. *Full* retains all indices and *−All* masks every index. Hatching marks conditions in which MMSE is masked. Exact values are in Appendix A.11 (AMC) and A.10 (ADNI).

all Llama-based variants (Appendix Table 18), supporting that the benefit of the framework is not backbone-specific.

5 Conclusion

We introduced Dementia-R1, a reinforcement learning framework for longitudinal dementia prognosis from clinical histories. A Cold-Start RL stage on verifiable clinical indices teaches longitudinal reasoning under sparse rewards. Across two real-world hospital cohorts and the structured ADNI benchmark, Dementia-R1 consistently achieved strong performance, including cross-hospital generalization and the best AUROC among LLM baselines on ADNI. These results support verifiable-reward RL as an effective approach for longitudinal clinical prediction from real-world patient records.

Limitations

We acknowledge several limitations in our study.

- **Cohort and language generalization.** All unstructured data come from two Korean hospitals, so a single nation and ethnic group is represented, and generalization to other demographics, languages, or documentation styles is untested. The framework itself is not language-specific, since it depends only on the availability of verifiable clinical indices. Once a comparable extractor exists for the target language, the same pipeline transfers directly. Validation on English-language EHRs from US or UK hospital systems is the immediate next step.
- **Translation.** Converting Korean notes into English may lose subtle cues, such as syntactic irregularities, that are themselves informative for cognitive decline. Translation is a practical convenience rather than a methodological requirement. Future work should apply the method directly to native-language text.
- **Dependence on the auxiliary extractor.** Stage 1 rewards come from an auxiliary LLM, so extraction errors could in principle propagate into RL training. A human expert audit over 600 notes measured 98.17% slot-level accuracy, with case-level rather than systematic errors (Appendix A.15). A noise-injection test shows that the downstream prognosis is insensitive to this failure mode: corrupting 70% of the Stage 1 indices changes the final AUROC by less than half a point (Appendix A.16). The audit used a single clinician, so inter-rater agreement remains unquantified.
- **Reliance on quantifiable clinical indices.** Our approach requires indices that can be verified numerically. In practice this is rarely a barrier, since most chronic and progressive diseases are monitored through quantifiable assessments such as UPDRS, PHQ-9, NYHA class, or HbA1c, and our Cold-Start RL is scale-agnostic. Where no established scale fits, the same design becomes a discovery tool, because comparing candidate indices as Stage 1 targets reveals which signals carry the most prognostic information (Sec. 4.8). Con-

ditions defined mainly by subjective markers remain out of reach.

- **Nature of the reference labels.** The dementia and PDD labels are clinically derived from staging instruments, ICD-10 coding, and prescription records (Sec. 4.1.1). They are not biomarker-confirmed or neuropathologically verified. Some diagnostic heterogeneity is therefore possible, and absolute performance should be read against clinical reference standards.
- **Residual prodromal signal.** Explicit leakage is prevented by construction, since the input ends at the Target Anchor and all splits are patient-level (Sec. 4.1.2). Some residual signal nevertheless persists, as in any longitudinal clinical task. We believe *prodromal signal* is the more precise framing. Gradual cognitive decline, mood changes, or subtle functional impairment in earlier notes are clinically meaningful precursors that a model should use. Removing them would require a single-visit input, which defeats the task. We therefore frame our task as *prognosis* rather than screening.
- **Binary formulation.** We represent the outcome as dementia versus non-dementia. This is a reasonable first design, but it collapses clinically meaningful distinctions. Two extensions follow naturally. Mild cognitive impairment could be added as a third class to capture the prodromal stage. Subtype classification within dementia is also feasible, given the cross-disease generalization to PDD shown here.

Ethics Statement

This retrospective study was approved by the Institutional Review Board (IRB No. 2023-1628), which waived the requirement for informed consent due to the use of de-identified medical records. All methods were performed in accordance with the relevant guidelines and regulations of the AMC Ethics Committee and the Declaration of Helsinki. The HPH study was approved by its Institutional Review Board (IRB No. 2023-11-042), and all procedures were conducted in accordance with the Declaration of Helsinki and relevant institutional guidelines.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work was supported by the National Research Foundation of Korea under Grant RS-2024-00336454. This research was supported by the National Research Foundation of Korea (NRF) (RS-2023-00262527).

References

Soo Borson, Lori Frank, Peter J Bayley, Malaz Boustani, Marge Dean, Pei-Jung Lin, J Riley McCarten, John C Morris, David P Salmon, Frederick A Schmitt, Richard G Stefanacci, Marta S Mendiola, Susan Peschin, Eric J Hall, Howard Fillit, and J Wesson Ashford. 2013. [Improving dementia care: The role of screening and detection of cognitive impairment](#). *Alzheimer’s & Dementia*, 9(2):151–159.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Shuang Zhou, Yuan Luo, Rui Zhang, Danielle Bitterman, Fei Wang, and Kai Shu. 2026. [Clinical-Bench: Can LLMs beat traditional ML models in clinical prediction?](#) In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, volume 2.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025a. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025b. [Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning](#). *arXiv preprint arXiv:2506.04207*.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. [Doctor AI: Predicting clinical events via recurrent neural networks](#). In *Machine Learning for Healthcare Conference (MLHC)*, pages 301–318. PMLR.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. [SFT memorizes, RL generalizes: A comparative study of foundation model post-training](#). In *International Conference on Machine Learning (ICML)*.

DeepSeek-AI. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. [Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach](#). *Biometrics*, 44(3):837–845.

Marshall F Folstein, Susan E Folstein, and Paul R McHugh. 1975. [“Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician](#). *Journal of Psychiatric Research*, 12(3):189–198.

Fatih Gelir, Taymaz Akan, Sait Alp, Emrah Gecili, Md Shenuarin Bhuiyan, Elizabeth A Disbrow, Steven A Conrad, John A Vanchiere, Christopher G Kevel, and Mohammad Alfrad Nobel Bhuiyan. 2025. [Machine learning approaches for predicting progression to Alzheimer’s disease in patients with mild cognitive impairment](#). *Journal of Medical and Biological Engineering*, 45(1):63–83.

Jacob HG Grand, Sienna Caspar, and Stuart WS MacDonald. 2011. [Clinical features and multidisciplinary approaches to dementia care](#). *Journal of Multidisciplinary Healthcare*, 4:125–147.

- Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 5549–5581.
- Anke Hensel, Matthias C Angermeyer, and Steffi G Riedel-Heller. 2007. [Measuring cognitive change in older adults: Reliable change indices for the Mini-Mental State Examination](#). *Journal of Neurology, Neurosurgery & Psychiatry*, 78(12):1298–1303.
- Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, Anders M Dale, Joel P Felmlee, Jeffrey L Gunter, Derek L G Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, and 16 others. 2008. [The Alzheimer’s Disease Neuroimaging Initiative \(ADNI\): MRI methods](#). *Journal of Magnetic Resonance Imaging*, 27(4):685–691.
- Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. [Mining electronic health records: towards better research applications and clinical care](#). *Nature Reviews Genetics*, 13(6):395–405.
- Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, and 9 others. 2023. [Health system-scale language models are all-purpose prediction engines](#). *Nature*, 619(7969):357–362.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? A large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Junu Kim, Chaeun Shim, Sungjin Park, Su Yeon Lee, Gee Young Suh, Chae-Man Lim, Seong Jin Choi, Song Mi Moon, Kyoung-Ho Song, Eu Suk Kim, Hong Bin Kim, Sejoong Kim, Chami Im, Dong-Wan Kang, Yong Soo Kim, Hee-Joon Bae, Sung Yoon Lim, Han-Gil Jeong, and Edward Choi. 2025. Enhancing LLMs’ clinical reasoning with real-world data from a nationwide sepsis registry. *arXiv preprint arXiv:2505.02722*.
- David S Knopman and Ronald C Petersen. 2014. [Mild cognitive impairment and mild dementia: A clinical perspective](#). *Mayo Clinic Proceedings*, 89(10):1452–1459.
- Hyoun-Joong Kong. 2019. [Managing unstructured big data in healthcare system](#). *Healthcare Informatics Research*, 25(1):1–2.
- Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025. [Large language models with temporal reasoning for longitudinal clinical summarization and prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20715–20735.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafford, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Conference on Language Modeling (COLM)*.
- Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025. [CARE-AD: A multi-agent large language model framework for Alzheimer’s disease prediction using longitudinal clinical notes](#). *npj Digital Medicine*, 8(1):541.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *International Conference on Learning Representations (ICLR)*.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- John C Morris. 1993. [The clinical dementia rating \(CDR\): Current version and scoring rules](#). *Neurology*, 43(11):2412–2414.
- 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 via reinforcement learning. *arXiv preprint arXiv:2502.19634*.
- Bo Pang, Deqian Kong, Silvio Savarese, Caiming Xiong, and Yingbo Zhou. 2025. Reasoning curriculum: Bootstrapping broad LLM reasoning from math. *arXiv preprint arXiv:2510.26143*.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Barry Reisberg, Steven H Ferris, Mony J de Leon, and Thomas Crook. 1982. [The global deterioration scale for assessment of primary degenerative dementia](#). *American Journal of Psychiatry*, 139(9):1136–1139.

- Timo Schick and Hinrich Schütze. 2021. [Exploiting cloze-questions for few-shot text classification and natural language inference](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 255–269.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. DeepSeek-Math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. 2025. [Learning the natural history of human disease with generative transformers](#). *Nature*, 647:248–256.
- Christina Silcox, Eyal Zimlichmann, Katie Huber, Neil Rowen, Robert Saunders, Mark McClellan, Charles N Kahn III, Claudia A Salzberg, and David W Bates. 2024. [The potential for artificial intelligence to transform healthcare: perspectives from international health leaders](#). *npj Digital Medicine*, 7(1):88.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Janine Stein, Melanie Lupp, Wolfgang Maier, Michael Wagner, Steffen Wolfsgruber, Martin Scherer, Mirjam Köhler, Marion Eisele, Siegfried Weyerer, Jochen Werle, Horst Bickel, Edelgard Mösch, Birgitt Wiese, Jana Prokein, Michael Pentzek, and Angela Fuchs. 2012. [Assessing cognitive changes in the elderly: Reliable change indices for the Mini-Mental State Examination](#). *Acta Psychiatrica Scandinavica*, 126(3):208–218.
- Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. 2024. [MOTOR: A time-to-event foundation model for structured medical records](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets LLM: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 645–654.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.
- Robert M Wachter and Erik Brynjolfsson. 2024. [Will generative artificial intelligence deliver on its promise in health care?](#) *JAMA*, 331(1):65–69.
- Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhurst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. 2025. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*.
- Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Tang Lifu, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. [Light-R1: Curriculum SFT, DPO and RL for long CoT from scratch and beyond](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Industry Track)*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

A Appendix

A.1 Temporal Density Ablation

We uniformly subsample longitudinal measurement time points to simulate reduced testing frequencies and assess the impact on prognosis (Figure 15). On AMC (456 patients, ≥ 2 visits; densities: 100/50/0%), ADNI (342 patients, ≥ 4 visits; densities: 100/75/50/25%), and HPH (0% vs. 100% only, due to sparse index coverage), AU-ROC degrades monotonically with reduced density: -7.3% on AMC ($85.83 \rightarrow 78.50$), -4.8% on HPH ($78.37 \rightarrow 73.59$), and -11.9% on ADNI ($88.73 \rightarrow 76.80$). The steeper ADNI decline reflects greater reliance on structured measurements in the absence of free-text narratives. Notably, even at 0% density on AMC, the model retains 78.50% AU-ROC, confirming that it has internalized reasoning from unstructured text beyond explicit scores.

A.2 ADNI Data Description

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).

A.3 Preprocessing Data Statistics

A.3.1 AMC

To prevent patient-level leakage, we exclude all patients reserved for downstream fine-tuning from the Stage 1 pretraining corpus. After removing 577 fine-tuning test patients from the original cohort of 11,163 patients, we obtain 46,746 longitudinal samples for intermediate clinical index forecasting (MMSE/GDS/CDR). We perform a patient-level split with a test ratio of 0.20, resulting in 3,568 training patients (37,112 samples) and 892 test patients (9,634 samples). For evaluation efficiency, the test split is task-stratified and reduced to 800 samples (401 patients). Finally, samples exceeding 8,000 tokens under the Qwen2.5-7B-Instruct tokenizer are removed, yielding 32,681 training samples and 722 test samples. Table 4 summarizes

Table 4: **Table A1: Stage 1 (Pretraining) dataset statistics (AMC).** Fine-tuning test patients are fully excluded to prevent leakage. A patient-level split is applied for Stage 1.

Item	#Patients	#Samples
Original cohort (raw)	11,163	–
Excluded: FT test patients	577	–
Pretraining (after exclusion, before split)	–	46,746
Train split (patient-level)	3,568	37,112
Test split (patient-level)	892	9,634
After token filter ($\leq 8,000$ tokens)	–	–
Train kept / removed	–	32,681 / 4,431
Test kept / removed	–	722 / 78

Table 5: **Table A2: Task distribution for Stage 1 pre-training (AMC).**

Task	Train	Test
MMSE	17,131	4,593
GDS	15,787	3,972
CDR	4,194	1,069
Total	37,112	9,634

the overall dataset composition, and Table 5 reports the task-wise distribution.

A.3.2 ADNI

For Stage 1 pretraining on the ADNI benchmark, we construct longitudinal next-visit prediction samples across six cognitive targets (MMSE, CDRSB, ADAS11, ADAS13, ADASQ4, and RAVLT_learning) from linearized structured records (Sec. A.4.2). To prevent leakage, we exclude all participants reserved for the downstream fine-tuning test set. After removing DX targets (not used in our pretraining), we obtain 11,319 candidate samples before filtering and keep 9,953 samples after excluding fine-tuning test participants. No samples are removed by the token-length constraint ($\leq 8,000$ tokens) or parsing/ID issues in our pipeline. We bucket the time gap to the prediction target into 1-month bins up to 6 months and an additional $>6m$ bin (see Sec. A.5 for the bucket definition). Finally, we perform a patient-level split to create Stage 1 train/test sets, resulting in 7,958 training samples and 1,995 test samples. Table 6 summarizes the overall dataset composition, and Table 7 reports task-wise statistics.

To adapt distinct data modalities for our unified reasoning framework, we developed specialized preprocessing pipelines for both unstructured clinical notes (Asan) and structured tabular records (ADNI). We applied a consistent protocol consisting of Data Transformation followed by Dataset Construction.

Table 6: **Table A3: Stage 1 (Pretraining) dataset statistics (ADNI).** Fine-tuning test participants are fully excluded to prevent leakage.

Item	Value
Candidate samples (before filtering; 6 tasks)	11,319
Kept samples ($\leq 8,000$ tokens & excl. FT-test)	9,953
Excluded: fine-tuning test participants	1,366
Excluded: token length / parsing / ID issues	0 / 0 / 0
Stage 1 split (samples)	train 7,958; test 1,995

Table 7: **Table A4: Task-wise distribution for Stage 1 pretraining (ADNI).** “Input” counts are computed before excluding fine-tuning test participants; “Kept” counts are used for Stage 1 training/evaluation.

Task	Input	Kept	Train	Test
MMSE	1,899	1,671	1,331	340
CDRSB	1,882	1,656	1,322	334
ADAS11	1,891	1,663	1,311	352
ADAS13	1,865	1,637	1,307	330
ADASQ4	1,897	1,669	1,340	329
RAVLT_learning	1,885	1,657	1,347	310
Total	11,319	9,953	7,958	1,995

A.4 Data Preprocessing Details

A.4.1 AMC (Unstructured Clinical Notes)

Data Transformation (Translation & Extraction). We transformed raw Korean clinical notes into English reasoning contexts using a secure pipeline. We utilized Qwen2.5-14B-Instruct as an auxiliary LLM to translate notes and extract clinical indices (MMSE, GDS, and CDR) to serve as verifiable ground truth targets. Crucially, all inference processes were conducted in a strictly isolated on-premise environment to prevent any external data transmission.

Stage 1 Construction Pipeline. We constructed the pre-training dataset with the following criteria:

1. **Tolerance-Aware Labeling:** We defined the prediction targets as extracted clinical indices. Recognizing extraction variability, we applied a tolerance of ± 2 for MMSE, treating predictions within this range as correct. Exact matching was enforced for coarser scales (GDS, CDR).
2. **Token Filtering:** Using the Qwen2.5 tokenizer, we filtered out samples exceeding 8,000 tokens to fit context constraints.
3. **Evaluation Set:** The dataset was split into training and test sets at a **patient level** (80:20) to evaluate Stage 1 performance.

A.4.2 ADNI (Structured Tabular Data)

Data Transformation (Linearization). We transformed structured tabular records into longitudinal textual logs suitable for LLM input. For each visit, we aggregated key biomarkers—including cognitive scores (MMSE, CDR-SB, ADAS-Cog), CSF biomarkers ($A\beta$, Tau), and MRI measures—into a structured text block (e.g., “2011-05-12: <<VISIT 1>> CDRSB: 0.5, MMSE: 28...”). These blocks were concatenated chronologically to form the patient history.

Stage 1 Construction Pipeline. We applied a construction protocol parallel to the AMC dataset but adapted for the continuous nature of ADNI biomarkers:

1. **Target Indices:** We selected seven key indicators: MMSE, CDR-SB, ADAS-Cog (11, 13, Q4), RAVLT (Learning), and LDELTOTAL.
2. **Proportional Tolerance-Aware Labeling:** Unlike categorical labels, these indices vary widely in range. To standardize difficulty, we defined a relative tolerance ratio $\rho \approx 6.7\%$ (derived from the standard allowance of ± 2 points on the 30-point MMSE scale). For each index, the allowable error margin δ was calculated as $\lceil \text{Range} \times \rho \rceil$. The specific thresholds are detailed in Table 8.
3. **Token Filtering:** Samples exceeding 8,000 tokens were filtered out using the tokenizer constraints.
4. **Evaluation Set:** Consistent with the AMC protocol, we applied a stratified **patient-level** split (80:20). Due to the high computational cost of longitudinal reasoning, the final evaluation was conducted on a stratified 50% sub-sample of the test set.

Table 8: **Tolerance Thresholds for ADNI Indices.** Error margins (δ) were scaled proportionally to the range of each metric.

Clinical Index	Range	Tolerance (δ)
MMSE	0–30	± 2
CDRSB	0–18	± 1.0
ADAS-Cog 11	0–70	± 5
ADAS-Cog 13	0–85	± 6
ADAS-Cog Q4	0–10	± 1
RAVLT (Learning)	-20–20	± 3
LDELTOTAL	0–25	± 2

Choice of Linearization Format. Because linearizing tabular records into text could in principle discard signal, we compared our format against the best-performing serialization reported by two established references, rather than against arbitrary variants: the row-level *Text Template* of TabLLM (Hegselmann et al., 2023) and the table-level *HTML* markup of *Table Meets LLM* (Sui et al., 2024). Our linearization corresponds to TabLLM’s *List Template*. We re-ran Stage 2 fine-tuning separately for each variant, reusing the same Stage 1 checkpoint and hyperparameters. As shown in Table 9, neither alternative meaningfully exceeds our baseline: Text Template is statistically indistinguishable from it, consistent with TabLLM’s observation that format differences vanish given sufficient training data (ADNI provides 1,832 training samples), while HTML markup actively hurts, indicating that the markup-familiarity gains reported for structural-understanding tasks do not transfer to longitudinal clinical reasoning.

Table 9: **ADNI Linearization Format Comparison.** Test AUROC (mean $\pm$ std over 5 seeds) after re-running Stage 2 for each serialization format.

Linearization	AUROC	Δ vs. List
List Template (ours)	0.8317 $\pm$ 0.0130	—
Text Template (Hegselmann et al., 2023)	0.8269 $\pm$ 0.0098	−0.48 pp
HTML (Sui et al., 2024)	0.7808 $\pm$ 0.0112	−5.09 pp

The gap between Dementia-R1 and the Random Forest baseline on ADNI (Table 3) is therefore attributable to the intrinsically tabular nature of the benchmark rather than to signal loss in our linearization.

A.5 Detailed Temporal Distributions

To validate the model’s capability in long-term prediction, we analyze the time intervals between the input data and the prediction target. Figure 11 details the distribution of these time gaps for the test sets. The two real-world cohorts (AMC and HPH) are predominantly concentrated in shorter intervals (6–18m), reflecting dense clinical follow-up prior to diagnosis. In contrast, the ADNI cohort presents a substantially more challenging scenario, with a large proportion of samples falling into the >18 m bucket, reflecting the longitudinal nature of MCI progression monitoring. This diversity ensures that our evaluation covers both immediate screening and long-term prognostic scenarios.

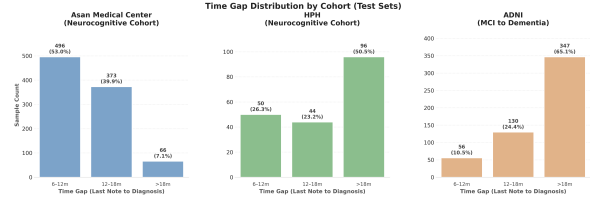


Figure 11: **Time Gap Distribution by Cohort (Test Sets).** The histograms show the interval between the last available clinical note and the diagnosis date across three cohorts. Intervals ≥ 18 months are consolidated into a single >18 m bucket. The AMC and HPH cohorts are concentrated in shorter intervals (6–18m), reflecting relatively dense clinical follow-up prior to diagnosis. In contrast, the ADNI cohort displays a substantially larger proportion of long-term samples (>18 m), reflecting the longitudinal nature of MCI progression monitoring.

A.6 Stage 1: Performance of Intermediate Clinical Indices

We analyze the model’s capability to predict intermediate clinical indices. On the structured ADNI benchmark (Tables 13, 14), Dementia-R1 demonstrates superior performance in long-term forecasting (>18 months), particularly on the critical CDR-SB metric. Similarly, on the unstructured Asan cohort (Tables 12), the model consistently outperforms baselines in short-to-mid term intervals, achieving notable gains in GDS and CDR prediction.

Figures 12 and 13 visualize the multi-dimensional performance profiles for the AMC and HPH cohorts, respectively. Dementia-R1 achieves the most balanced profile across both intermediate clinical reasoning and final prognosis on each cohort.

Beyond time-gap analysis, Table 10 reports the overall clinical index prediction accuracy on the AMC cohort. Dementia-R1 achieves the highest average accuracy (55.1%), outperforming Qwen2.5-72B-Instruct (56.3%) on GDS and matching it on MMSE—while being 10 $\times$ smaller.

Table 11 reports the clinical index prediction accuracy on the HPH cohort (MMSE and GDS only, as CDR was not available). Qwen2.5-72B-Instruct achieves the highest overall accuracy (45.2%), while Dementia-R1 (7B) ranks third (42.1%), notably outperforming all other 7B configurations and matching Qwen2.5-32B-Instruct (42.6%).

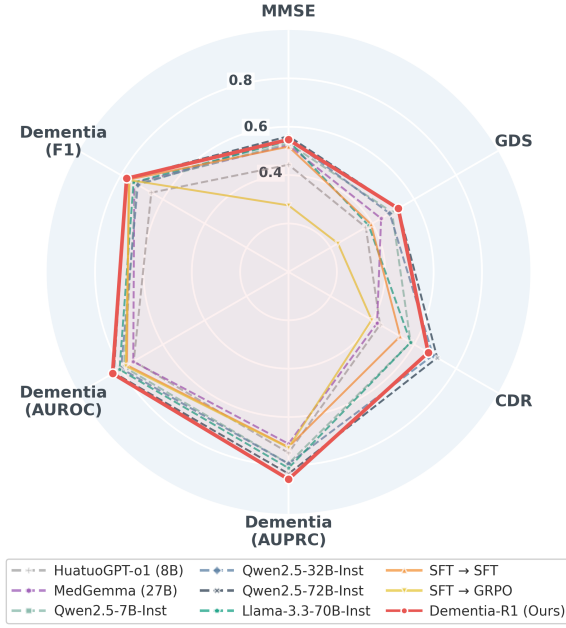


Figure 12: **Multi-dimensional Performance Profile (AMC)**. Radar chart combining Stage 1 clinical index accuracy (MMSE, GDS, CDR) and Stage 2 dementia prognosis (F1, AUROC, AUPRC). Dementia-R1 achieves the most balanced profile overall.

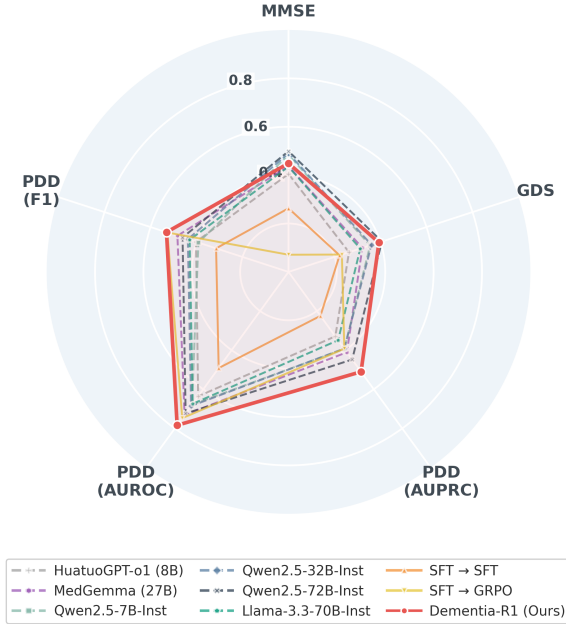


Figure 13: **Multi-dimensional Performance Profile (HPH)**. Radar chart combining Stage 1 clinical index accuracy (MMSE, GDS) and Stage 2 PDD prognosis (F1, AUROC, AUPRC). Dementia-R1 dominates on all prognosis metrics.

A.7 Stage 2: Performance of Final Binary Prognosis

Building on the clinical indicators established in Stage 1, Stage 2 assesses the model’s ultimate ca-

Table 10: **Performance on Clinical Index Prediction for the AMC cohort**. We evaluate the accuracy of predicting MMSE, GDS, and CDR scores. **Bold/underline** = best/second-best. Mean $\pm$ std over 3 seeds.

Model Method	Size	MMSE	GDS	CDR	Average
<i>External LLMs</i>					
HuatuoGPT-o1	8B	44.3 $\pm$ 0.5	36.9 $\pm$ 2.5	43.7 $\pm$ 8.7	41.1 $\pm$ 0.6
MedGemma	27B	52.4 $\pm$ 0.3	44.2 $\pm$ 0.5	42.3 $\pm$ 2.1	47.8 $\pm$ 0.3
Qwen2.5-7B-Inst	7B	53.1 $\pm$ 1.6	49.4 $\pm$ 1.2	58.3 $\pm$ 3.2	52.1 $\pm$ 1.5
Qwen2.5-32B-Inst	32B	55.2 $\pm$ 1.2	48.3 $\pm$ 2.2	<u>69.0</u> $\pm$ 3.6	53.9 $\pm$ 0.2
Qwen2.5-72B-Inst	72B	56.1 $\pm$ 1.3	<u>52.4</u> $\pm$ 1.8	71.0 $\pm$ 1.0	56.3 $\pm$ 1.0
Llama-3.3-70B-Inst	70B	54.2 $\pm$ 2.2	38.3 $\pm$ 1.4	58.3 $\pm$ 2.1	48.0 $\pm$ 0.7
<i>Specialized Models</i>					
SFT $\rightarrow$ SFT	7B	51.8 $\pm$ 2.2	39.4 $\pm$ 0.7	53.3 $\pm$ 1.5	46.8 $\pm$ 1.4
SFT $\rightarrow$ GRPO	7B	27.5 $\pm$ 2.6	23.6 $\pm$ 3.2	39.7 $\pm$ 3.1	27.3 $\pm$ 1.8
Dementia-R1	7B	<u>54.7</u> $\pm$ 0.6	52.4 $\pm$ 0.9	66.7 $\pm$ 2.1	<u>55.1</u> $\pm$ 0.4

Table 11: **Performance on Clinical Index Prediction for the HPH cohort**. MMSE and GDS accuracy. **Bold/underline** = best/second-best. Mean $\pm$ std over 3 seeds.

Model Method	Size	MMSE	GDS	Average
<i>External LLMs</i>				
HuatuoGPT-o1	8B	40.5 $\pm$ 4.3	26.3 $\pm$ 1.0	33.4 $\pm$ 1.7
MedGemma	27B	43.7 $\pm$ 1.9	32.1 $\pm$ 1.0	37.9 $\pm$ 0.9
Qwen2.5-7B-Inst	7B	48.4 $\pm$ 2.1	35.5 $\pm$ 1.2	41.9 $\pm$ 0.5
Qwen2.5-32B-Inst	32B	49.0 $\pm$ 0.9	36.1 $\pm$ 0.6	42.6 $\pm$ 0.2
Qwen2.5-72B-Inst	72B	49.9 $\pm$ 2.6	40.6 $\pm$ 1.6	45.2 $\pm$ 1.2
Llama-3.3-70B-Inst	70B	43.5 $\pm$ 2.3	31.2 $\pm$ 0.8	37.4 $\pm$ 1.5
<i>Specialized Models</i>				
SFT $\rightarrow$ SFT	7B	26.3 $\pm$ 3.9	22.2 $\pm$ 2.7	24.3 $\pm$ 1.2
SFT $\rightarrow$ GRPO	7B	7.1 $\pm$ 1.1	23.2 $\pm$ 1.3	15.2 $\pm$ 0.7
Dementia-R1	7B	<u>44.9</u> $\pm$ 1.4	<u>39.3</u> $\pm$ 0.8	<u>42.1</u> $\pm$ 1.0

capacity to determine the final binary prognosis. As summarized in Table 15, Dementia-R1 achieves the highest overall AUROC on the Asan dataset, demonstrating notable consistency across all time intervals (6–12m, 12–18m, 18–24m). Similarly, the results on the ADNI benchmark (Table 17) and the HPH cohort (Table 16) further highlight the model’s robustness across diverse prediction horizons. In particular, Dementia-R1 achieves the highest AUROC in the longest-term bucket (36m+ for ADNI; $\geq 24m$ for HPH), confirming its effectiveness in modeling longitudinal disease trajectories.

A.8 Qualitative Analysis: Comparative Reasoning

To demonstrate the impact of our proposed method on reasoning quality, we compare the outputs of Dementia-R1 against the Qwen2.5-32B model using a representative longitudinal case from the AMC cohort. Figure 16 illustrates the input clinical note, which follows a semi-structured SOAP format. In this record, critical signals such as cognitive scores (MMSE, GDS) and medication changes are embedded within the free-text Objective and Plan sections across multiple visits spanning from 2020 to 2023. This presents a complex reasoning

challenge, requiring the model to aggregate scattered clinical indicators and correctly reconstruct the patient’s disease trajectory from the unstructured narrative.

While both models correctly predict the final diagnosis, their reasoning processes diverge significantly. Figure 17 presents the reasoning outputs generated by both models. Dementia-R1 effectively structures the longitudinal information by organizing the output into distinct sections for cognitive assessment history, diagnosis, and current status. This structural clarity allows clinicians to rapidly verify the evidence. Notably, the model accurately reconstructs the temporal trajectory of cognitive decline and correctly identifies the drop in MMSE scores from 23 down to 17 alongside the plateau in the most recent visits. Furthermore, it correctly identifies the medication switch involving the discontinuation of Bearcept and the addition of Ebixa, demonstrating precise grounding in the clinical text.

In contrast, despite arriving at the correct label, Qwen2.5-32B produces a dense, unstructured block of text that is difficult to audit for clinical decision-making. More critically, it exhibits significant failures in medical soundness and domain knowledge. First, the baseline hallucinates a pharmacological equivalence by describing "Ebixa (donepezil)" even though Ebixa is memantine, an NMDA receptor antagonist distinct from the cholinesterase inhibitor donepezil. Such hallucinations pose potential safety risks in clinical settings. Second, the baseline misinterprets the Global Deterioration Scale (GDS) scores of 3-4 as indicators of "mild depression" and confuses the dementia staging scale with a depression inventory. These findings underscore that general-purpose reasoning capabilities, even in larger models, do not automatically translate to clinical accuracy. Our results demonstrate that the domain-specific alignment integrated into Dementia-R1 is essential for correcting such misconceptions and ensuring the high reliability required for longitudinal dementia prognosis.

A.9 Llama-3.1 Backbone Results

To verify that our framework generalizes beyond the Qwen backbone, we evaluate all training configurations using Llama-3.1-8B-Instruct as the base model. As shown in Table 18, Dementia-R1 achieves the highest F1 (77.90%), AUROC (85.29%), and AUPRC (85.58%) among all Llama-based variants, with significant improvements ($p <$

0.01).

A.10 Leave-One-Out Index Ablation (ADNI)

To quantify the contribution of each clinical index learned during Stage 1 pre-training, we perform a leave-one-out ablation on the ADNI dataset. At inference time ($T=0$), we remove one index from the input and evaluate the full Stage 2 dementia prognosis performance. Table 19 summarizes the results. Performance remained stable across all removals (F1: 0.74–0.77, AUROC: 0.83–0.85), demonstrating that the model does not over-rely on any single index. Among all indices, removing CDR-SB caused the sharpest decline (F1: 0.74, AUPRC: 0.80), identifying it as the most influential prognostic indicator. Notably, high AUROC was maintained even without memory-specific scales (RAVLT, LDELTOTAL), confirming strong cross-index redundancy.

A.11 Clinical Index Masking Ablation (AMC, Full Results)

Table 20 reports the full results of the clinical index masking ablation on the AMC cohort, including overall and bucket-wise performance breakdowns.

A.12 Temporal Density Ablation (Full Results)

Table 21 reports the full results of the temporal density ablation on the HPH cohort, including overall and bucket-wise performance breakdowns.

A.13 Robustness under Naturally Sparse Clinical Indices

The masking ablation (Figure 10) removes clinical indices artificially at inference time. To complement it, we evaluate the naturally occurring sparse regime by stratifying the test sets according to how many MMSE/GDS/CDR mentions are actually present in the input history, and reporting performance on the *zero-index* subset, in which the model must rely entirely on free-text narrative.

AMC. Of the 935 AMC test samples, 93 contain no clinical index mention at all. Table 22 reports AUROC on this subset. Dementia-R1 leads all nine baselines, including the direct Stage 1 ablation (GRPO w/o Stage 1, -2.66) and every general-purpose model up to 72B.

HPH. HPH is naturally the sparsest cohort: 166 of its 190 test samples (87%) contain no clinical index mention, making it the most direct real-world

Table 12: **Stage 1 Performance by Time Gap (AMC).** **Dementia-R1** achieves superior accuracy in short-to-mid term intervals (0–18 months), validating the effectiveness of the Cold-Start strategy.

Method	Overall	Accuracy by Time Gap to Prediction Target			
	Acc ($\uparrow$)	0–6m	6–12m	12–18m	>18m
<i>External LLMs</i>					
Qwen2.5-32B-Inst	57.90 $\pm$ 0.05	60.66 $\pm$ 0.43	54.91 $\pm$ 0.53	55.15 $\pm$ 1.21	50.43 $\pm$ 5.08
Qwen2.5-7B-Inst	54.68 $\pm$ 0.32	57.56 $\pm$ 0.69	51.19 $\pm$ 1.16	52.76 $\pm$ 0.99	46.32 $\pm$ 2.28
<i>Specialized Models</i>					
SFT $\rightarrow$ SFT	51.73 $\pm$ 0.53	55.07 $\pm$ 0.87	48.37 $\pm$ 0.75	46.95 $\pm$ 0.59	46.90 $\pm$ 5.02
SFT $\rightarrow$ GRPO	55.81 $\pm$ 0.46	59.60 $\pm$ 0.69	51.21 $\pm$ 1.15	52.64 $\pm$ 0.41	49.62 $\pm$ 5.43
Dementia-R1	59.61 $\pm$ 0.46	63.38 $\pm$ 0.45	55.25 $\pm$ 0.69	56.41 $\pm$ 0.58	49.49 $\pm$ 2.84

Table 13: **Stage 1 Performance by Time Gap (ADNI).** **Dementia-R1** demonstrates superior long-term reasoning capability (>18 months) compared to larger baselines (32B), validating the efficacy of the proposed RL framework.

Method	Overall	Accuracy by Time Gap to Prediction Target			
	Acc ($\uparrow$)	0–6m	6–12m	12–18m	>18m
<i>External LLMs</i>					
Qwen2.5-32B-Inst	77.91 $\pm$ 0.18	81.26 $\pm$ 0.27	77.38 $\pm$ 0.25	76.77 $\pm$ 0.52	<u>80.30</u> $\pm$ 0.67
Qwen2.5-7B-Inst	75.67 $\pm$ 0.31	77.87 $\pm$ 0.46	75.09 $\pm$ 0.43	75.46 $\pm$ 0.58	78.19 $\pm$ 1.05
<i>Specialized Models</i>					
SFT $\rightarrow$ SFT	64.42 $\pm$ 0.23	66.60 $\pm$ 0.28	63.58 $\pm$ 0.69	63.26 $\pm$ 0.68	68.10 $\pm$ 1.74
SFT $\rightarrow$ GRPO	67.61 $\pm$ 0.33	71.26 $\pm$ 1.16	67.47 $\pm$ 0.62	65.05 $\pm$ 0.51	70.33 $\pm$ 1.34
Dementia-R1	<u>77.04</u> $\pm$ 0.28	<u>79.44</u> $\pm$ 0.92	<u>76.30</u> $\pm$ 0.51	<u>76.23</u> $\pm$ 0.52	80.62 $\pm$ 0.99

Table 14: **Stage 1 Accuracy by Clinical Index (ADNI).** **Bold/underline** = best/second-best. Mean $\pm$ std over 3 seeds.

Method	Size	Overall Acc ($\uparrow$)	CDRSB Acc ($\uparrow$)	ADAS11 Acc ($\uparrow$)	ADAS13 Acc ($\uparrow$)	RAVLT Acc ($\uparrow$)	MMSE Acc ($\uparrow$)	ADASQ4 Acc ($\uparrow$)	LDEL Acc ($\uparrow$)
<i>External LLMs</i>									
HuatuoGPT-o1	8B	62.17 $\pm$ 0.94	44.70 $\pm$ 1.34	79.10 $\pm$ 1.44	74.08 $\pm$ 1.12	70.55 $\pm$ 0.53	58.91 $\pm$ 1.89	54.00 $\pm$ 0.40	51.96 $\pm$ 2.00
MedGemma	27B	74.37 $\pm$ 0.38	82.52 $\pm$ 0.36	81.45 $\pm$ 0.68	78.76 $\pm$ 0.42	80.34 $\pm$ 0.39	74.55 $\pm$ 0.58	61.72 $\pm$ 1.14	57.32 $\pm$ 1.39
Qwen2.5-7B-Inst	7B	74.87 $\pm$ 0.48	81.35 $\pm$ 0.43	81.95 $\pm$ 0.91	81.22 $\pm$ 0.87	79.99 $\pm$ 1.22	75.05 $\pm$ 0.69	61.09 $\pm$ 1.78	59.93 $\pm$ 0.93
Qwen2.5-32B-Inst	32B	77.94 $\pm$ 0.19	85.48 $\pm$ 0.77	86.82 $\pm$ 0.44	84.28 $\pm$ 0.55	81.83 $\pm$ 0.80	77.53 $\pm$ 0.32	65.24 $\pm$ 0.07	60.19 $\pm$ 0.48
Qwen2.5-72B-Inst	72B	77.35 $\pm$ 0.18	85.40 $\pm$ 0.07	84.51 $\pm$ 0.54	83.68 $\pm$ 0.12	82.66 $\pm$ 0.15	76.98 $\pm$ 0.63	63.77 $\pm$ 0.98	60.46 $\pm$ 0.60
Llama-3.3-70B-Inst	70B	76.03 $\pm$ 0.24	85.52 $\pm$ 0.25	82.28 $\pm$ 0.21	81.50 $\pm$ 0.44	81.26 $\pm$ 0.35	75.60 $\pm$ 0.77	62.64 $\pm$ 0.50	59.45 $\pm$ 0.33
<i>Specialized Models</i>									
SFT $\rightarrow$ SFT	7B	63.82 $\pm$ 0.70	70.27 $\pm$ 1.43	71.21 $\pm$ 0.44	66.22 $\pm$ 1.41	73.61 $\pm$ 1.78	66.25 $\pm$ 0.94	42.98 $\pm$ 1.02	54.09 $\pm$ 2.38
SFT $\rightarrow$ GRPO	7B	37.20 $\pm$ 0.55	46.60 $\pm$ 1.59	34.94 $\pm$ 0.12	30.71 $\pm$ 1.51	42.81 $\pm$ 3.94	42.10 $\pm$ 1.01	27.92 $\pm$ 1.12	35.03 $\pm$ 1.84
Dementia-R1	7B	<u>77.38</u> $\pm$ 0.14	83.98 $\pm$ 0.42	<u>85.71</u> $\pm$ 0.31	<u>84.00</u> $\pm$ 0.07	<u>82.35</u> $\pm$ 0.72	<u>77.27</u> $\pm$ 0.86	<u>64.15</u> $\pm$ 0.88	60.14 $\pm$ 0.37

Table 15: **Stage 2 Fine-tuning Performance by Time Gap (AMC).** We report the breakdown of AUROC and AUPRC across prediction horizons. **Dementia-R1** consistently outperforms all baselines across all time gaps. All results represent mean $\pm$ standard deviation across five random seeds.

Method	Overall		6–12m		12–18m		18–24m	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
<i>External LLMs</i>								
HuatuoGPT-o1	73.89 $\pm$ 0.5	74.81 $\pm$ 0.9	74.50 $\pm$ 1.0	73.28 $\pm$ 1.6	75.90 $\pm$ 1.4	78.21 $\pm$ 1.6	71.08 $\pm$ 3.5	80.77 $\pm$ 3.6
Qwen2.5-7B-Inst	78.59 $\pm$ 0.8	79.24 $\pm$ 0.9	79.41 $\pm$ 0.9	78.48 $\pm$ 0.7	80.27 $\pm$ 1.8	82.07 $\pm$ 1.8	76.62 $\pm$ 2.6	83.80 $\pm$ 3.2
Qwen2.5-32B-Inst	79.93 $\pm$ 1.2	79.33 $\pm$ 1.7	79.95 $\pm$ 1.3	78.13 $\pm$ 2.6	82.09 $\pm$ 1.1	82.71 $\pm$ 0.6	74.47 $\pm$ 3.3	79.54 $\pm$ 2.4
Qwen2.5-72B-Inst	<u>83.41</u> $\pm$ 0.4	<u>83.44</u> $\pm$ 0.7	<u>83.29</u> $\pm$ 0.7	<u>81.15</u> $\pm$ 1.1	<u>85.37</u> $\pm$ 0.6	<u>86.80</u> $\pm$ 1.1	<u>77.38</u> $\pm$ 2.6	84.43 $\pm$ 3.6
Llama-3.3-70B-Inst	80.70 $\pm$ 0.3	81.11 $\pm$ 0.7	80.39 $\pm$ 0.5	79.03 $\pm$ 0.9	83.00 $\pm$ 0.7	84.52 $\pm$ 0.7	73.61 $\pm$ 1.9	80.58 $\pm$ 3.2
MedGemma-27B	74.14 $\pm$ 0.8	71.06 $\pm$ 1.2	74.65 $\pm$ 1.6	70.17 $\pm$ 1.9	75.78 $\pm$ 0.7	73.37 $\pm$ 1.3	71.56 $\pm$ 1.6	76.88 $\pm$ 2.6
<i>Specialized Models</i>								
SFT w/o Stage 1	71.65 $\pm$ 1.9	69.82 $\pm$ 2.2	70.59 $\pm$ 1.2	65.63 $\pm$ 1.7	73.83 $\pm$ 2.6	74.63 $\pm$ 3.1	70.24 $\pm$ 9.4	78.91 $\pm$ 7.5
GRPO w/o Stage 1	81.68 $\pm$ 0.4	81.27 $\pm$ 0.3	81.97 $\pm$ 0.8	79.67 $\pm$ 1.2	84.16 $\pm$ 0.9	85.22 $\pm$ 1.8	76.07 $\pm$ 2.1	<u>84.47</u> $\pm$ 2.8
SFT $\rightarrow$ SFT	77.15 $\pm$ 1.9	72.41 $\pm$ 2.4	77.20 $\pm$ 2.1	69.59 $\pm$ 2.8	78.28 $\pm$ 1.8	75.66 $\pm$ 2.2	72.59 $\pm$ 7.1	79.77 $\pm$ 7.0
SFT $\rightarrow$ GRPO	77.75 $\pm$ 1.7	72.69 $\pm$ 2.3	78.72 $\pm$ 2.5	71.23 $\pm$ 3.8	78.18 $\pm$ 1.2	74.34 $\pm$ 2.0	70.52 $\pm$ 4.3	79.03 $\pm$ 4.1
Dementia-R1	84.02 $\pm$ 0.4	85.73 $\pm$ 0.4	84.67 $\pm$ 0.6	85.40 $\pm$ 0.6	85.39 $\pm$ 0.4	87.44 $\pm$ 0.5	78.51 $\pm$ 2.7	86.15 $\pm$ 1.7

Table 16: **Stage 2 Fine-tuning Performance by Time Gap (HPH)**. We report the breakdown of AUROC and AUPRC across prediction horizons. All results represent mean $\pm$ standard deviation across five random seeds.

Method	Overall		6-12m		12-18m		18-24m		$\geq 24m$	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
<i>External LLMs</i>										
HuatuogPT-o1	63.58 $\pm$ 3.2	32.93 $\pm$ 1.9	58.51 $\pm$ 5.9	38.06 $\pm$ 6.5	69.18 $\pm$ 3.6	50.94 $\pm$ 7.3	71.76 $\pm$ 7.3	47.81 $\pm$ 11.9	57.44 $\pm$ 6.9	24.62 $\pm$ 4.5
Qwen2.5-7B-Inst	67.90 $\pm$ 3.2	39.28 $\pm$ 4.1	66.12 $\pm$ 4.4	45.42 $\pm$ 5.3	70.79 $\pm$ 5.4	52.81 $\pm$ 7.6	78.70 $\pm$ 9.0	63.06 $\pm$ 10.2	60.07 $\pm$ 6.5	24.60 $\pm$ 2.5
Qwen2.5-32B-Inst	68.54 $\pm$ 1.9	39.20 $\pm$ 1.9	74.12 $\pm$ 4.0	55.05 $\pm$ 9.3	72.26 $\pm$ 4.6	63.83 $\pm$ 3.8	74.92 $\pm$ 8.3	50.85 $\pm$ 7.9	55.68 $\pm$ 8.7	24.42 $\pm$ 5.1
Qwen2.5-72B-Inst	72.37 $\pm$ 1.1	44.78 $\pm$ 2.6	77.52 $\pm$ 4.1	58.52 $\pm$ 4.6	77.57 $\pm$ 1.9	68.05 $\pm$ 1.7	78.28 $\pm$ 5.9	62.97 $\pm$ 6.9	61.52 $\pm$ 1.4	32.25 $\pm$ 6.1
Llama-3.3-70B-Inst	67.41 $\pm$ 1.4	34.91 $\pm$ 0.7	72.57 $\pm$ 3.9	55.01 $\pm$ 7.7	74.91 $\pm$ 2.4	55.84 $\pm$ 3.2	68.99 $\pm$ 5.2	31.43 $\pm$ 2.4	56.96 $\pm$ 3.8	28.34 $\pm$ 3.7
MedGemma-27B	73.13 $\pm$ 2.6	41.06 $\pm$ 4.4	<u>77.63</u> $\pm$ 4.1	50.20 $\pm$ 5.9	<u>81.76</u> $\pm$ 2.5	62.69 $\pm$ 5.3	72.65 $\pm$ 5.6	39.15 $\pm$ 8.4	58.55 $\pm$ 5.7	<u>32.75</u> $\pm$ 8.2
<i>Specialized Models</i>										
SFT w/o Stage 1	63.75 $\pm$ 1.8	30.39 $\pm$ 3.2	64.78 $\pm$ 9.8	38.41 $\pm$ 13.0	61.74 $\pm$ 5.3	42.81 $\pm$ 7.1	75.50 $\pm$ 4.9	37.11 $\pm$ 3.0	57.92 $\pm$ 13.2	24.50 $\pm$ 9.6
GRPO w/o Stage 1	69.96 $\pm$ 3.4	38.69 $\pm$ 2.4	65.99 $\pm$ 9.7	44.93 $\pm$ 9.5	71.32 $\pm$ 4.7	55.65 $\pm$ 6.8	<u>84.96</u> $\pm$ 4.7	<u>64.54</u> $\pm$ 4.3	<u>65.48</u> $\pm$ 5.4	27.15 $\pm$ 4.0
SFT $\rightarrow$ SFT	49.14 $\pm$ 1.2	22.47 $\pm$ 1.7	49.10 $\pm$ 6.8	28.04 $\pm$ 4.5	48.41 $\pm$ 7.1	34.50 $\pm$ 7.0	49.87 $\pm$ 10.4	22.16 $\pm$ 5.4	49.73 $\pm$ 4.8	20.33 $\pm$ 4.2
SFT $\rightarrow$ GRPO	<u>74.75</u> $\pm$ 2.6	39.25 $\pm$ 3.3	79.23 $\pm$ 4.6	51.57 $\pm$ 7.6	80.62 $\pm$ 4.3	61.43 $\pm$ 10.1	75.97 $\pm$ 6.7	39.07 $\pm$ 9.4	62.61 $\pm$ 6.7	26.32 $\pm$ 7.2
Dementia-R1	78.37 $\pm$ 1.8	51.03 $\pm$ 2.7	75.50 $\pm$ 3.8	<u>56.18</u> $\pm$ 4.6	81.96 $\pm$ 3.2	61.35 $\pm$ 3.5	93.99 $\pm$ 3.4	84.76 $\pm$ 6.7	68.60 $\pm$ 4.9	35.79 $\pm$ 6.9

Table 17: **Stage 2 Fine-tuning Performance by Time Gap (ADNI)**. While large-scale general models (e.g., Qwen2.5-32B) show strong performance in certain horizons, **Dementia-R1** demonstrates superior robustness in the longest-term reasoning (>36 months). All results represent mean $\pm$ standard deviation across five random seeds.

Method	Overall		6-12m		12-18m		18-24m		24-30m		30-36m		36m+	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
<i>External LLMs</i>														
HuatuogPT-o1	71.69 $\pm$ 2.0	67.34 $\pm$ 3.2	76.93 $\pm$ 2.6	76.76 $\pm$ 4.5	75.92 $\pm$ 1.6	69.42 $\pm$ 4.2	71.64 $\pm$ 6.4	70.70 $\pm$ 6.7	70.91 $\pm$ 1.2	68.35 $\pm$ 2.8	71.23 $\pm$ 2.4	69.70 $\pm$ 3.3	62.15 $\pm$ 4.1	56.16 $\pm$ 6.4
Qwen2.5-7B-Inst	66.23 $\pm$ 1.9	57.17 $\pm$ 2.4	78.26 $\pm$ 4.5	75.61 $\pm$ 6.6	67.74 $\pm$ 4.3	56.48 $\pm$ 5.3	62.61 $\pm$ 2.3	62.58 $\pm$ 3.0	65.42 $\pm$ 3.1	57.23 $\pm$ 4.3	69.28 $\pm$ 2.3	65.42 $\pm$ 4.0	53.42 $\pm$ 4.5	42.69 $\pm$ 4.7
Qwen2.5-32B-Inst	<u>82.51</u> $\pm$ 0.6	<u>83.50</u> $\pm$ 0.6	83.34 $\pm$ 1.6	83.24 $\pm$ 2.1	84.90 $\pm$ 1.2	86.09 $\pm$ 1.1	81.70 $\pm$ 1.4	85.24 $\pm$ 1.7	83.22 $\pm$ 1.0	82.90 $\pm$ 1.6	85.00 $\pm$ 1.1	86.92 $\pm$ 1.9	74.48 $\pm$ 1.7	76.27 $\pm$ 2.1
Qwen2.5-72B-Inst	81.45 $\pm$ 0.6	83.56 $\pm$ 0.7	82.49 $\pm$ 1.3	83.60 $\pm$ 1.3	84.47 $\pm$ 1.2	85.38 $\pm$ 2.1	82.11 $\pm$ 1.9	87.81 $\pm$ 0.8	82.05 $\pm$ 0.9	82.80 $\pm$ 1.8	83.56 $\pm$ 1.7	82.31 $\pm$ 1.5	69.74 $\pm$ 5.5	73.47 $\pm$ 3.2
Llama-3.3-70B-Inst	81.17 $\pm$ 0.5	83.48 $\pm$ 0.6	83.26 $\pm$ 0.9	85.98 $\pm$ 1.0	<u>84.47</u> $\pm$ 0.8	<u>85.71</u> $\pm$ 1.4	80.65 $\pm$ 0.9	<u>86.12</u> $\pm$ 0.5	81.51 $\pm$ 1.1	82.79 $\pm$ 1.1	<u>83.68</u> $\pm$ 0.6	87.37 $\pm$ 0.8	70.20 $\pm$ 1.5	75.41 $\pm$ 0.8
MedGemma-27B	78.48 $\pm$ 1.1	71.71 $\pm$ 1.2	75.94 $\pm$ 3.7	69.16 $\pm$ 4.5	81.50 $\pm$ 1.6	75.88 $\pm$ 3.7	77.08 $\pm$ 2.8	72.71 $\pm$ 2.8	79.65 $\pm$ 2.1	71.85 $\pm$ 4.7	75.70 $\pm$ 3.1	74.52 $\pm$ 2.3	<u>79.57</u> $\pm$ 3.2	72.01 $\pm$ 4.9
<i>Specialized Models</i>														
SFT w/o Stage 1	73.40 $\pm$ 1.9	63.03 $\pm$ 2.4	73.68 $\pm$ 5.3	63.82 $\pm$ 7.7	74.47 $\pm$ 2.0	65.14 $\pm$ 5.9	76.01 $\pm$ 4.9	71.43 $\pm$ 6.4	74.02 $\pm$ 3.9	62.70 $\pm$ 6.6	71.99 $\pm$ 3.4	67.46 $\pm$ 5.3	69.98 $\pm$ 7.0	59.53 $\pm$ 7.4
GRPO w/o Stage 1	79.07 $\pm$ 0.9	77.65 $\pm$ 1.5	<u>83.72</u> $\pm$ 2.7	<u>85.75</u> $\pm$ 3.5	81.03 $\pm$ 1.2	77.37 $\pm$ 2.9	78.23 $\pm$ 3.8	80.89 $\pm$ 4.4	79.60 $\pm$ 1.4	76.71 $\pm$ 3.7	79.08 $\pm$ 2.0	80.42 $\pm$ 3.8	70.45 $\pm$ 4.9	66.69 $\pm$ 3.8
SFT $\rightarrow$ SFT	79.90 $\pm$ 0.3	73.77 $\pm$ 1.2	77.43 $\pm$ 2.3	72.61 $\pm$ 4.8	83.89 $\pm$ 3.6	77.04 $\pm$ 5.1	81.89 $\pm$ 1.4	82.10 $\pm$ 5.5	79.22 $\pm$ 2.1	71.38 $\pm$ 2.5	77.71 $\pm$ 1.8	76.96 $\pm$ 2.8	78.18 $\pm$ 4.7	72.28 $\pm$ 7.8
SFT $\rightarrow$ GRPO	77.88 $\pm$ 1.7	69.09 $\pm$ 2.6	77.41 $\pm$ 5.3	71.21 $\pm$ 7.6	80.99 $\pm$ 3.6	70.29 $\pm$ 6.4	77.42 $\pm$ 2.3	71.70 $\pm$ 2.6	76.84 $\pm$ 5.6	67.88 $\pm$ 5.7	79.02 $\pm$ 3.8	76.30 $\pm$ 3.3	73.35 $\pm$ 3.5	70.50 $\pm$ 8.1
Dementia-R1	83.17 $\pm$ 1.5	82.00 $\pm$ 1.2	85.39 $\pm$ 3.5	85.29 $\pm$ 4.1	83.87 $\pm$ 1.9	80.96 $\pm$ 3.5	83.85 $\pm$ 3.7	85.09 $\pm$ 3.2	<u>82.13</u> $\pm$ 1.9	81.59 $\pm$ 3.3	82.66 $\pm$ 1.6	85.41 $\pm$ 1.1	83.90 $\pm$ 2.8	81.13 $\pm$ 4.1

Table 18: **Experimental Results on AMC Dataset (Llama-3.1 Backbone)**. We compare Dementia-R1 against Llama-3.1-8B-Instruct and its specialized variants. **Bold** and underline indicate the best and second-best performance. All results represent mean $\pm$ standard deviation across five random seeds. Statistical significance compared to Dementia-R1 is denoted by * ($p < 0.05$) and ** ($p < 0.01$).

Method	Size	Accuracy ($\uparrow$)	Precision ($\uparrow$)	Recall ($\uparrow$)	F1 score ($\uparrow$)	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
<i>External LLMs</i>							
Llama-3.1-8B-Inst	8B	54.37 $\pm$ 2.0**	68.94 $\pm$ 1.8**	58.91 $\pm$ 1.8**	63.53 $\pm$ 1.8**	66.66 $\pm$ 1.4**	63.44 $\pm$ 1.8**
<i>Specialized Models</i>							
SFT w/o Stage 1	8B	63.49 $\pm$ 0.8**	60.15 $\pm$ 0.5**	<u>83.78</u> $\pm$ 2.3	70.01 $\pm$ 1.0**	67.19 $\pm$ 1.4**	66.74 $\pm$ 1.4**
GRPO w/o Stage 1	8B	61.30 $\pm$ 1.8**	59.09 $\pm$ 1.3**	79.31 $\pm$ 2.4**	67.72 $\pm$ 1.6**	70.71 $\pm$ 1.7**	70.97 $\pm$ 1.3**
SFT w/o Stage 2	8B	73.43 $\pm$ 0.6**	70.56 $\pm$ 0.7**	82.06 $\pm$ 1.0**	75.87 $\pm$ 0.6**	75.72 $\pm$ 0.9**	69.56 $\pm$ 1.4**
GRPO w/o Stage 2	8B	74.31 $\pm$ 0.8**	70.11 $\pm$ 0.6**	86.55 $\pm$ 1.1**	<u>77.46</u> $\pm$ 0.7**	84.01 $\pm$ 0.6**	<u>84.05</u> $\pm$ 0.9**
SFT $\rightarrow$ SFT	8B	73.86 $\pm$ 1.0**	72.46 $\pm$ 1.0**	78.49 $\pm$ 0.9**	75.35 $\pm$ 0.9**	74.30 $\pm$ 1.0**	69.41 $\pm$ 0.6**
SFT $\rightarrow$ GRPO	8B	<u>76.02</u> $\pm$ 0.5**	<u>77.52</u> $\pm$ 0.9**	74.54 $\pm$ 0.7**	75.99 $\pm$ 0.4**	<u>84.04</u> $\pm$ 0.4**	83.89 $\pm$ 0.6**
Dementia-R1	8B	77.39 $\pm$ 0.6	78.09 $\pm$ 0.7	77.72 $\pm$ 0.9	77.90 $\pm$ 0.6	85.29 $\pm$ 0.2	85.58 $\pm$ 1.2

Table 19: **Leave-One-Out Index Ablation (ADNI, $T=0$)**. Each row removes the specified index from the input at inference. CDR-SB removal causes the largest performance drop, while overall stability confirms cross-index redundancy.

Removed Index	Acc	Prec	Rec	F1	AUROC	AUPRC
MMSE	.79	.74	.80	.77	.85	.83
CDRSB	.76	.70	.78	.74	.83	.80
ADAS11	.79	.74	.79	.76	.84	.85
ADAS13	.79	.77	.73	.75	.84	.83
ADASQ4	.78	.75	.74	.74	.84	.85
RAVLT	.80	.76	.78	.77	.85	.83
LDELTOTAL	.79	.75	.78	.76	.85	.86

analogue of the extremely sparse setting. Table 23

shows that Dementia-R1 again leads all nine base-lines. The margin over the direct Stage 1 ablation widens to -8.76 , indicating that Cold-Start pre-training is most valuable precisely when explicit indices are unavailable and the model must infer trajectory from narrative alone.

Taken together with the masking ablation, these results indicate that Dementia-R1 does not merely copy explicit clinical indices: it retains its advantage over every baseline in the sparsest regime on both cohorts.

Table 20: **Clinical Index Masking Ablation (AMC)**. Performance of Dementia-R1 when clinical indices are masked at inference time. “None” denotes the unmasked baseline. The largest drops are consistently associated with MMSE removal.

Masked Index(es)	Overall		6–12m		12–18m		18–24m	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
None (baseline)	.8350	.8456	.8480	.8473	.8458	.8606	.7894	.8641
CDR	.8295	.8371	.8378	.8309	.8406	.8614	.7990	.8818
GDS	.8326	.8415	.8380	.8349	.8492	.8684	.7981	.8742
GDS + CDR	.8295	.8350	.8391	.8350	.8424	.8542	.7808	.8492
MMSE	.7941	.8065	.8065	.8054	.8160	.8356	.6731	.7644
MMSE + CDR	.7882	.7783	.8022	.7772	.8082	.8061	.7067	.7659
MMSE + GDS	.7829	.7834	.8056	.7885	.7903	.7993	.6923	.7691
ALL	.7807	.7825	.7965	.7840	.7941	.8040	.7135	.7854

Table 21: **Temporal Density Ablation (HPH)**. Performance at 0% (all indices masked) vs. 100% (all indices retained). The model retains meaningful predictive power even without explicit clinical indices.

Density	Overall		6–12m		12–18m		18–24m		≥24m	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
100%	.7837 ± .0199	.5103 ± .0306	.7550 ± .0427	.5618 ± .0515	.8196 ± .0353	.6135 ± .0387	.9399 ± .0384	.8476 ± .0744	.6860 ± .0553	.3579 ± .0770
0%	.7359 ± .0245	.5465 ± .0207	.7088 ± .0412	.5365 ± .0213	.7774 ± .0256	.7113 ± .0244	.8487 ± .1203	.7381 ± .1070	.6360 ± .0183	.3658 ± .0664

Table 22: **AMC Zero-Index Subset** ($n=93$). AUROC (mean ± std over 5 seeds) on test samples containing no MMSE/GDS/CDR mention.

Model	Size	AUROC	Δ vs. Ours
Dementia-R1 (Ours)	7B	72.24 ± 1.76	—
GRPO w/o Stage 1	7B	69.58 ± 2.07	−2.66
SFT → GRPO	7B	64.40 ± 2.97	−7.84
SFT → SFT	7B	61.82 ± 6.07	−10.42
Qwen2.5-7B-Instruct	7B	66.97 ± 2.70	−5.27
Qwen2.5-32B-Instruct	32B	64.55 ± 2.71	−7.69
Qwen2.5-72B-Instruct	72B	63.60 ± 1.91	−8.64
Llama-3.3-70B-Instruct	70B	62.48 ± 1.97	−9.76
HuatuoGPT-o1 (medical)	8B	65.27 ± 6.01	−6.97
MedGemma (medical)	27B	65.82 ± 3.41	−6.42

Table 23: **HPH Zero-Index Subset** ($n=166$). AUROC (mean ± std over 5 seeds) on test samples containing no MMSE/GDS mention.

Model	Size	AUROC	Δ vs. Ours
Dementia-R1 (Ours)	7B	75.24 ± 2.62	—
GRPO w/o Stage 1	7B	66.48 ± 4.96	−8.76
SFT → GRPO	7B	74.56 ± 3.08	−0.68
SFT → SFT	7B	48.91 ± 3.10	−26.33
Qwen2.5-7B-Instruct	7B	63.50 ± 3.01	−11.74
Qwen2.5-32B-Instruct	32B	63.62 ± 2.99	−11.62
Qwen2.5-72B-Instruct	72B	67.36 ± 1.29	−7.88
Llama-3.3-70B-Instruct	70B	65.74 ± 2.24	−9.50
HuatuoGPT-o1 (medical)	8B	60.50 ± 4.19	−14.74
MedGemma (medical)	27B	69.70 ± 3.12	−5.54

A.14 Robustness to Conflicting Longitudinal Signals

Real clinical trajectories are frequently non-monotonic: a patient may cross the dementia

threshold at one visit and fall back below it at the next. We isolate this regime explicitly.

Stratification. We stratify the 935 AMC test samples by whether the patient’s GDS trajectory crosses the dementia cutoff ($GDS \geq 4$). Excluding patients with fewer than two GDS readings ($n=231$), the remaining 704 samples split into **Stable Non-Dementia** ($n=331$), **Stable Dementia** ($n=164$), and **Crossing** ($n=209$), the last of which contains directly contradictory signals across visits. HPH cannot be stratified this way because 87.4% of its patients lack repeated GDS readings. No model was retrained for this analysis.

Results on the crossing subset. Table 24 reports AUROC on the 209 crossing samples. Dementia-R1 leads all nine baselines, and the gap over the direct Stage 1 ablation (−7.35) is substantially wider than on the full test set, indicating that Cold-Start pre-training is what allows the model to weigh a transient fluctuation against the surrounding trajectory rather than reacting to the most recent value.

Case studies. To characterize *how* the models differ, we examine three crossing patients (seed 1), comparing Dementia-R1 against Qwen2.5-32B-Instruct (the comparator used in the neurologist evaluation, Sec. 4.7) and against GRPO w/o Stage 1 (which isolates the effect of Stage 1). We report the predicted probability p of dementia.

- **Case A** — transient spike (gold = 0; GDS

Table 24: **AMC Crossing Subset** ($n=209$). AUROC (mean $\pm$ std over 5 seeds) on patients whose GDS trajectory crosses the dementia cutoff, i.e. whose visits carry contradictory signals.

Model	Size	AUROC	Δ vs. Ours
Dementia-R1 (Ours)	7B	77.72 $\pm$ 1.11	—
GRPO w/o Stage 1	7B	70.37 $\pm$ 0.79	−7.35
SFT $\rightarrow$ GRPO	7B	66.34 $\pm$ 4.29	−11.38
SFT $\rightarrow$ SFT	7B	64.79 $\pm$ 3.58	−12.93
Qwen2.5-7B-Instruct	7B	68.50 $\pm$ 4.22	−9.22
Qwen2.5-32B-Instruct	32B	71.53 $\pm$ 2.84	−6.19
Qwen2.5-72B-Instruct	72B	74.10 $\pm$ 1.93	−3.62
Llama-3.3-70B-Instruct	70B	71.79 $\pm$ 1.09	−5.93
HuatuoGPT-o1 (medical)	8B	59.17 $\pm$ 1.70	−18.55
MedGemma (medical)	27B	56.24 $\pm$ 3.59	−21.48

3 $\rightarrow$ 4 $\rightarrow$ 3). Dementia-R1 correctly discounts the spike ($p=0.001$) by integrating spouse-reported memory improvement with a sustained MMSE, whereas Qwen2.5-32B ($p=1.000$) and GRPO w/o Stage 1 ($p=0.989$) both over-call on the single elevated reading.

- **Case B** — alternation (gold = 0; GDS alternating between 3 and 4). Dementia-R1 ($p=0.000$) anchors on a K-MMSE of 27, above the impairment threshold, and attributes the fluctuation to severe depression; both base-lines over-call ($p \approx 1.000$).
- **Case C** — apparent recovery (gold = 1; GDS 3 $\rightarrow$ 4 $\rightarrow$ 3, with a vascular dementia diagnosis). Dementia-R1 ($p=0.924$) integrates the vascular dementia diagnosis with the surrounding multimodal evidence, while GRPO w/o Stage 1 ($p=0.001$) produces a false negative by anchoring on an isolated “MMSE 24, normal” mention.

Across all three patterns — spike, alternation, and apparent recovery — Dementia-R1 integrates non-monotonic cues rather than following the most recent index value, and Case C attributes this behavior specifically to Stage 1 Cold-Start pre-training.

A.15 LLM Extraction Quality Audit

To validate the fidelity of the LLM-based extraction pipeline used for constructing verifiable rewards (Sec. 3.1), we conducted a human expert audit in two rounds.

Initial audit (200 notes / 600 slots). A clinician independently reviewed 200 randomly sampled clinical notes, verifying the extracted values for all three target indices (MMSE, GDS, CDR), yielding

600 slot-level comparisons. Table 25 summarizes the results. The overall slot-level accuracy was 98.5%, confirming that the extracted rewards are high-fidelity and sufficient for robust RL optimization.

Table 25: **LLM Extraction Audit Results (Initial)**. Slot-level accuracy of the Qwen2.5-14B-Instruct extractor on 200 randomly sampled clinical notes (600 total slots).

Clinical Index	Slots	Accuracy (%)
MMSE	200	97.5
GDS	200	98.5
CDR	200	99.5
Overall	600	98.5

Expanded audit (600 notes). To verify that this estimate is not an artifact of the sample size, we expanded the audit to 600 clinical notes, auditing 200 independently sampled notes per clinical index. The aggregate slot-level accuracy was 98.17% (589/600 correct, 11 errors), consistent with the initial estimate. Per-index results are reported in Table 26.

Table 26: **LLM Extraction Audit Results (Expanded)**. Per-index accuracy of the Qwen2.5-14B-Instruct extractor over 600 audited clinical notes (200 per index).

Clinical Index	Notes	Correct	Errors	Accuracy (%)
MMSE	200	194	6	97.00
GDS	200	196	4	98.00
CDR	200	199	1	99.50
Total	600	589	11	98.17

Error type breakdown. We categorized the 11 errors observed in the expanded audit (Table 27). All are case-level and random rather than systematic; we observed no directional bias that could be amplified during RL training.

Representative failure modes. From the initial audit we identified three representative failure modes among the nine misextracted slots:

1. **Scale confusion.** The extractor misidentified a Barthel Index score as the GDS value when both appeared in close proximity within the “Objective” section (e.g., “CDR 0.5 (3) / GDS 4 / Barthel Index 14/20 / S-IADL 11/45”).
2. **Temporal misattribution.** When clinical indices and laboratory results from different

Table 27: **Error Type Breakdown (Expanded Audit).** The 11 extraction errors are case-level random errors rather than a systematic, directional bias.

Error type	Count	Description
Fabrication	3	A value is extracted although only the index keyword appears in the note.
False negative	4	The value is missed because it falls outside the extraction boundary window.
Misattribution	4	An overloaded abbreviation is resolved to the wrong scale (e.g., “GDS” as Geriatric Depression Scale rather than Global Deterioration Scale).
Total	11	

dates co-occurred in the same note, the extractor assigned a value to the wrong visit date. For instance, “2021/11/23 MMSE 12 CDR 0.5” was erroneously attributed to a subsequent lab date (2022-04-05) listed immediately below.

3. **Contextual misparse.** In notes where the “MMSE :” field was followed by detailed neurological examination subscores (e.g., “phone-mic 8 / 30 secs”) rather than a total score, the extractor incorrectly inferred MMSE = 30 from the denominator notation.

These errors were rare (1.83% of slots in the expanded audit) and predominantly affected MMSE extraction, which operates on the finest-grained scale. The tolerance-aware reward ($\delta = \pm 2$ for MMSE) further mitigates the impact of such boundary errors during RL training. Appendix A.16 additionally shows that the downstream prognosis is insensitive to extraction noise far exceeding this observed error rate.

A.16 Robustness to Extraction Noise

Because Stage 1 rewards are produced by an auxiliary extractor, a natural concern is that extraction errors could cascade into and be amplified by RL training. We test this directly with a noise-injection stress test.

Setup. On a 1,000-sample subset, we corrupted the Stage 1 clinical indices for 30%, 50%, and 70% of samples. Each injected error was placed *above*

the reward tolerance margin of the corresponding scale, so that the corrupted label is guaranteed to be treated as incorrect by R_{cold} (Eq. 1). We then re-ran Stage 1 GRPO for each noise level and fine-tuned Stage 2 with identical hyperparameters.

Downstream performance is essentially unaffected. As shown in Table 28, the final dementia AUROC stays within ± 0.5 percentage points of the noise-free baseline even when 70% of the Stage 1 labels are corrupted beyond tolerance.

Table 28: **Downstream Robustness to Stage 1 Extraction Noise.** Best dementia AUROC after injecting above-tolerance noise into the Stage 1 clinical indices.

Injected noise	Best dementia AUROC	Δ vs. 0%
0%	0.7945	—
30%	0.7926	−0.19 pp
50%	0.7929	−0.16 pp
70%	0.7902	−0.43 pp

Mechanism: noisy samples are filtered, not learned. GRPO standardizes rewards within each prompt’s group of G completions, $A_i = (r_i - \bar{r})/\sigma_r$ (Eq. 3). A corrupted label is typically implausible given the patient’s longitudinal record, so the model’s prior steers most of the G completions away from it. When *all* G completions miss the corrupted target, every reward is 0, hence $r_i - \bar{r} = 0$ and $A_i = 0$ for all i : the gradient contribution of that prompt vanishes and the sample is effectively dropped rather than learned. Table 29 provides direct evidence: as the injected noise grows, the mean training reward falls while the fraction of auto-filtered (zero-gradient) prompts rises by roughly 13 percentage points.

Table 29: **Automatic Filtering of Noisy Prompts under GRPO.** As noise increases, more prompts collapse to a zero advantage and contribute no gradient.

Injected noise	Mean training reward	Zero-gradient prompts
0%	0.268	0.712
30%	0.198	0.770
50%	0.152	0.823
70%	0.116	0.840

Practical significance. The measured slot-level error rate of our extractor is 1.83% (Appendix A.15), roughly $38\times$ smaller than the 70% stress level at which the downstream AUROC still degrades by less than half a point. Cascading extraction error is therefore negligible in practice.

Method versus extractor. A related question is how much of the reported gain comes from the method rather than from a strong extractor. The same Qwen2.5-14B-Instruct extractor supplies the clinical indices for *all* specialized baselines in Tables 1–3 (SFT and GRPO w/o Stage 1, SFT → SFT, SFT → GRPO). With the extractor held constant, Dementia-R1 still outperforms the strongest 7B baseline (SFT → GRPO) by +6.27 AUROC on AMC (84.02 vs. 77.75) and +3.62 on HPH (78.37 vs. 74.75), isolating the gain to the training framework rather than to the extraction pipeline.

A.17 Training Implementation Details

We implemented our framework using PyTorch. All experiments were conducted on four NVIDIA H100 (80GB) GPUs.

Reinforcement Learning (Dementia-R1) For the RL stage, we utilized the Open-R1 framework. We employed DeepSpeed ZeRO-3 and vLLM (colocate mode) to optimize memory usage for processing long clinical narratives. We set the per-device batch size to 1 with a gradient accumulation of 2 and a group size of $G = 8$ (effective batch size of 8). The model was trained for 5,000 steps in Bfloat16 precision with a 2,000-token completion limit.

A.18 Pretraining Target Ablation (Figures)

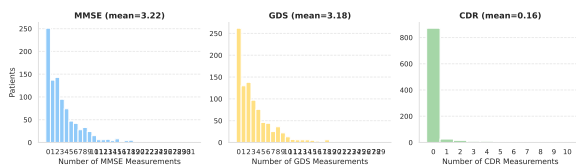


Figure 14: **Clinical Index Distribution (AMC Test Set).** Number of MMSE, GDS, and CDR measurements per patient. CDR annotations are substantially sparser than MMSE and GDS.

Table 30: **Pretraining Target Ablation.** Overall AUROC (%) when Stage 1 uses individual vs. combined targets. Best per cohort in **bold**.

Target	AMC	HPH
All (combined)	79.68	70.76
MMSE	78.82	69.95
GDS	81.03	70.88
CDR	78.88	–

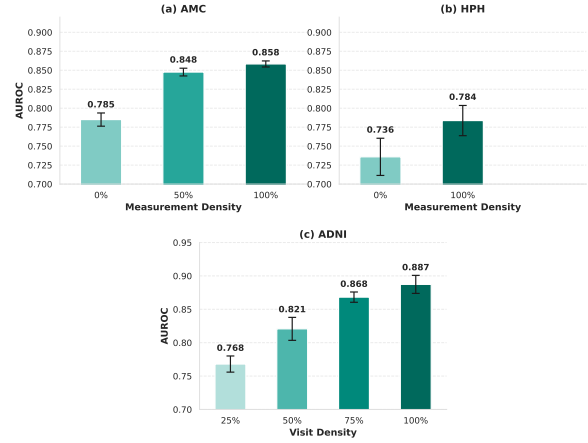


Figure 15: **Temporal Density Ablation.** Overall AUROC at varying measurement density levels across three cohorts. (a) AMC: 0/50/100%; (b) HPH: 0/100%; (c) ADNI: 25/50/75/100%. Performance degrades monotonically with reduced density.

A.19 Human evaluation protocol

To assess the clinical quality of reasoning, we conducted a blinded human evaluation with medical experts using a pairwise comparison protocol. For each case, experts were presented with two anonymized model responses (Model A and Model B) generated for the same patient record and prediction task. For each evaluation criterion, experts were asked to select one of three options: Model A, Model B, or Tie.

Each model pair was evaluated using 10 question-answer cases per comparison across two cohorts (AMC and HPH), yielding 100 total judgment pairs. Judgments were collected independently for the following five clinically motivated dimensions:

1. **Temporal Reasoning Accuracy:** Which response appropriately interprets changes in symptoms and the rate of progression by comparing earlier records with the most recent records?
2. **Evidence Grounding:** Which response cites evidence that is explicitly present in the original clinical notes and does not introduce information that is absent from the records?
3. **Clinically Relevant Evidence Selection:** Which response avoids being influenced by clinically irrelevant details or overlooking key evidence, and instead bases its reasoning on diagnostically important evidence from the clinical notes?

4. **Medical Soundness:** Which response is more medically sound with respect to dementia diagnostic criteria and clinical judgment, in terms of both the reasoning process and the final conclusion?
5. **Overall Clinical Utility:** When used as reference material in real-world clinical practice, which response is more reliable and more helpful for reducing clinical decision-making time?

Longitudinal Clinical Note Input (AMC)

2023-08-30:
 Subjective
 Follow-up with Professor **, Constipation, pletaal dosage reduced Sometimes forgetful, but there are times when it's okay Caregiver observes that there is a slight decline Handles all household chores personally Forgets what they went for when crossing the room
 Objective
 F/79y; Date of Birth (anonymized): ****/**/**
 2020/08/27 MMSE 23 GDS 4
 2022/02/03 MMSE 21 GDS 3
 2022/09/01 MMSE 17 GDS 4
 2023-08-30 MMSE; 17 (recall 2) GDS; 4
 Right-handed
 Highest Education Level; Illiterate
 Assessment
 # major neurocognitive disorder * VaD * MTA2/3 D3P3 (2022/09)
 Plan
 Add ebixa and discontinue bearcept LICA
 Pletaal tab [50mg] 1 TAB DP 1 time 35 days PO
 Lexapro tab [10mg] 1 TAB N 1 time 35 days PO
 Ebixa tab [10mg] 0.5 TAB BNP 2 times 35 days PO
 2023-10-04:
 Subjective
 Follow-up patient of Professor
 Scheduled for LICA after dementia team consultation on 2023/08/30
 Discontinued bearcept and added ebixa
 Bowel movements improved after changing the medication
 Objective
 F/79y; Date of Birth (anonymized): ****/**/**
 2020/08/27 MMSE 23 GDS 4
 2022/02/03 MMSE 21 GDS 3
 2022/09/01 MMSE 17 GDS 4
 2023/08/30 MMSE 17 (recall 2) GDS 4 (illiterate)
 2022-08-21 eGFR(CKD-EPI) (Qn), Blood 69 ml/min/1.73m²
 Assessment
 # major neurocognitive disorder
 * VaD
 * MTA2/3 D3P3 (2022/08)
 Plan
 * Reduced pletaal due to incontinence (outpatient of Professor **)
 LICA on 07/07
 Increase ebixa dosage, reduce back to half tablet if side effects occur
 Pletaal tab [50mg] 1 TAB DP 1 time 56 days PO
 Lexapro tab [10mg] 1 TAB N 1 time 56 days PO
 Ebixa tab [10mg] 1 TAB BNP 2 times 56 days PO
 2023-12-02:
 Subjective
 Post LICA visit
 No gastrointestinal side effects with current medication
 mood: so so
 Objective
 F/79y; Date of Birth (anonymized): ****/**/**
 Unlearned
 2020/08/27 MMSE 23 GDS 4
 2022/02/03 MMSE 21 GDS 3
 2022/09/01 MMSE 17 GDS 4
 2023/08/30 MMSE 17 GDS 4 (recall 2)
 2023/11/24 GDS 3 CDR 0.5 SB 1.0 BI 20 SIADL 5 NPI 2
 Note> Z score -1.5 or lower in some cognitive domains. The test results suggest a retrieval deficit in verbal memory and a deficit in visual memory. Other functions such as frontal/executive functions, attention, language and related functions, and visuoconstruction ability are all within normal levels. Therefore, bilateral frontal lobe dysfunction is suggested. The patient does not report any decline in ADL and is currently in an amnesic mild cognitive impairment state. The patient has shown cognitive decline since around '18, which has gradually worsened, and with multiple vascular risk factors and test results indicating a retrieval deficit, subcortical vascular cognitive impairment needs to be considered first. Given the patient's advanced age and the more pronounced memory impairment compared to other cognitive domains, Alzheimer's disease underlying pathology or a combined condition should also be considered, and the influence of hypothyroidism cannot be completely ruled out. A correlation with neuroimaging studies is recommended for an accurate diagnosis, and follow-up is recommended if cognitive function decline worsens. Treatment for depression is also recommended.
 2022-08-21 eGFR(CKD-EPI) (Qn), Blood 69 ml/min/1.73m²
 Assessment
 # impending dementia
 * VaD * MTA2/3 D3P3 (2022/08) Plan
 * Reduce pletaal due to fecal incontinence (outpatient visit with Professor **)
 ditto
 Continued lexapro after consultation
 Consultation to change medication if fecal incontinence persists
 Pletaal tab [50mg] 1 TAB DP 1 time 91 days PO
 Lexapro tab [10mg] 1 TAB N 1 time 91 days PO
 Ebixa tab [10mg] 1 TAB BNP 2 times 91 days PO

Figure 16: Example of longitudinal clinical notes. To protect patient privacy, all personal identifiers (e.g., names) were removed, and dates were randomly shifted.

Qualitative Comparison of Longitudinal Reasoning for 6-12 Month Dementia Prognosis

[Dementia-R1 output]

The patient's clinical notes and cognitive assessments provide a clear picture of their cognitive decline over time. Here's the step-by-step reasoning:

1. **Cognitive Assessments Over Time:**

- 2020/08/27: MMSE 23, GDS 4 (Severe cognitive decline)
- 2022/02/03: MMSE 21, GDS 3 (Moderate cognitive decline)
- 2022/09/01: MMSE 17, GDS 4 (Severe cognitive decline)
- 2023/08/30: MMSE 17 (recall 2), GDS 4 (Severe cognitive decline)

2. **Clinical Dementia Rating (CDR):**

- The CDR scores are consistently 0.5 or 1, indicating very mild to mild dementia.

3. **Symptoms and Functional Decline:**

- The patient reports forgetfulness, sometimes forgetting what they went for when crossing the room.
- Caregiver observes a slight decline.
- The patient handles all household chores personally but shows signs of memory impairment.

4. **Diagnosis and Plan:**

- The patient is diagnosed with VaD (Vascular Dementia) and MTA2/3 D3P3.
- The patient is currently in an amnesic mild cognitive impairment state, with a retrieval deficit in verbal and visual memory.
- The patient has shown cognitive decline since around 2019, which has gradually worsened.

5. **Current Status:**

- The most recent MMSE score (17) and GDS score (4) indicate severe cognitive decline.
- The patient's CDR score is 0.5 or 1, indicating very mild to mild dementia.
- The patient's condition has not improved over the past year, and the decline is consistent with a diagnosis of dementia.

Given the consistent cognitive decline, severe MMSE and GDS scores, and the diagnosis of VaD, it is clear that the patient is likely to have dementia.

[QWEN 2.5 32B output]

1. The patient's MMSE scores have been declining over time: 23 in 2020, 21 in 2022, and 17 in 2022 and 2023. A score of 17 is indicative of mild to moderate cognitive impairment.
2. The patient has been diagnosed with major neurocognitive disorder, specifically vascular dementia (VaD), and is classified as MTA2/3 D3P3.
3. The caregiver observes a slight decline in the patient's cognitive function, and the patient sometimes forgets what they went for when crossing the room, which is a sign of cognitive impairment.
4. The patient's GDS score has been consistently around 3-4, indicating mild depression, which can be associated with cognitive decline.
5. The patient's cognitive test results suggest a retrieval deficit in verbal memory and a deficit in visual memory, with other cognitive functions within normal levels.
6. The patient is in an amnesic mild cognitive impairment state, with cognitive decline since around 2019, which has gradually worsened.
7. The patient has multiple vascular risk factors, and the test results indicate a retrieval deficit, suggesting subcortical vascular cognitive impairment.
8. The patient's advanced age and more pronounced memory impairment compared to other cognitive domains suggest Alzheimer's disease underlying pathology or a combined condition.
9. The patient is being treated with Ebixa (donepezil), which is commonly used in the treatment of Alzheimer's disease and other forms of dementia

Figure 17: A comparison of the reasoning processes generated by Dementia-R1 and the Qwen2.5-32B for the clinical case shown in Figure 16. While Dementia-R1 produces a structured and grounded rationale, the baseline generates a dense narrative containing medical hallucinations.

Longitudinal Clinical Note Input (HPH)

=== Initial notes ===

2019-11-05: Department[Neurology] BP(S)[*] BP(D)[86] Hb[15.3] PLT[254] AST[28] ALT[27] Glucose[81] Chief complaint[Left leg tremor onset 2 months ago] Pain presence[None] Current disease status[left finger tapping: decreased left foot tapping: decreased left side rigidity left resting tremor on foot Parkinson's dz Tremor dominant HY 1] Medical history[Hypertension] Medical history[Diabetes] Medical history[Hyperlipidemia] Diagnosis[Parkinson's disease] Treatment plan[FP CIT PET]

=== Follow-up notes ===

2019-11-20: Department[Neurology] S[] O[Explained FP CIT PET scan] A[Parkinson's disease Tremor dominant HY 1] P[repeat initial Parkinson's disease]

2019-12-04: S[Feeling nauseous after taking medication.] O[150/81 finger tapping: intact gait: normal – “on” state] A[Parkinson's disease HY stage 1 S/E madopar - vomiting] P[repeat]

2020-01-15: S[Body shakes 2–3 times a day.] O[136/89 bradykinesia (+) gait: normal] P[repeat next follow-up to confirm tremor]

2020-03-11: S[Shaking intermittently.] O[mild finger tapping decreased overall bradykinesia 123/75] A[Parkinson's disease HY stage 1 S/E madopar - vomiting] P[tepra increase trihexine increase]

2020-06-03: S[Shaking in legs due to lack of movement.] O[135/75 left foot tremor at rest, finger tapping: equivocal shuffling gait] A[Parkinson's disease HY stage 1 S/E madopar - vomiting] P[Increase madopar 100-50-100 trihexine increase explained side effects]

2020-07-01: S[Body feels a little less stiff, right leg shakes] O[127/72 right foot tremor at rest (+) gait: normal] A[Parkinson's disease HY stage 1] P[repeat]

2020-08-26: S[Leg shakes, dizzy.] O[128/69 bradykinesia (-) right foot tremor - no change] A[Parkinson's disease HY stage 1] P[tepra → indenol change]

2020-11-16: S[Shaking in arms and legs is difficult. Intermittent headaches. Feeling suicidal.] O[116/70] A[Parkinson's disease HY stage 1] P[trihexine additional ssri additional increase madopar if necessary psychiatry consultation]

2020-12-07: S[Shaking intermittently. Memory is very poor.] O[126/74 Bradykinesia (-/-) tremor (-)] A[Parkinson's disease HY stage 1] P[trihexine decrease psychiatry consultation discontinue depression medication]

2021-02-01: S[Memory decline is severe. Crying a lot, too scared to see psychiatry, mmse=28] O[119/64] P[psychiatry consultation absolutely necessary]

2021-03-09: S[Psychiatry consultation made me feel a little better.] O[left side bradykinesia (+) 132/68] A[Parkinson's disease HY stage 1] P[add azilect]

2021-05-06: S[medication only] O[140/78 bradykinesia (+) resting tremor on left hand] A[Parkinson's disease HY stage 1] P[prescribe amantadine]

2021-06-28: S[feeling tired and having difficulty working in the restaurant.] O[122/63 mild bradykinesia gait: mild slow] A[Parkinson's disease HY stage 1 S/E xanax - somnolence] P[discontinue xanax]

2021-08-23: S[feeling sleepy.] O[131/76] A[Parkinson's disease HY stage 1] P[repeat]

2021-12-07: A[Parkinson's disease HY stage 1] P[diagnosis letter]

2022-02-20: S[feeling tremors.] O[140/70 left side bradykinesia (+)] A[Parkinson's disease HY stage 1] P[add trihexyphenidyl]

2022-04-04: S[significant memory decline. Short-term memory decline] O[137/71] A[Parkinson's disease HY stage 1] P[discontinue trihexyphenidyl (memory decline) mmse]

2022-05-30: S[back pain, leg pain. Left leg pain - orthopedic] O[164/90] A[Parkinson's disease HY stage 1] P[adjust to bid, follow up with orthopedic]

2022-09-26: S[back pain fluctuates, antalgic gait, memory decline affecting work] O[139/81] A[Parkinson's disease HY stage 1] P[repeat, check for dementia symptoms]

2023-02-20: S[severe depression, significant blood pressure fluctuations] O[132/61] P[add trihexyphenidyl, add lexapro, ANS function test at next follow-up]

2023-05-08: S[feeling tremors, having difficult thoughts due to depression] P[increase lexapro, increase trihexyphenidyl]

2023-06-26: S[Feels like blood is gushing in the head. Sensory symptom on the right parietal area. Feels heavy. Poor memory.] O[110/63] A[Parkinson's disease HY stage 1] P[REPEAT New sensory symptom BRAIN MRI]

2023-08-21: S[Feels like head is flowing. Lower back pain. Dragging the feet. Memory decline.] O[mri results explained] A[Parkinson's disease HY stage 1] P[Etivil Buspar prescribed MMSE prescribed]

=== Autonomic nervous system tests ===

2019-11-28: Blood pressure fluctuation [Abnormal] Valsalva ratio [Abnormal] Conclusion [Adrenergic and cardiovagal dysfunction. No orthostatic hypotension. CASS Cardiovascular score - 1 Adrenergic score - 2] Head-up tilt table test [Normal HUT]

2023-05-08: Blood pressure fluctuation [Abnormal] Valsalva ratio [Abnormal] Conclusion [Adrenergic and cardiovagal dysfunction. No orthostatic hypotension.] Head-up tilt table test [Normal HUT]

=== K-MMSE records ===

2019-11-28: 27 2021-02-01: 28 2022-05-30: 29

=== GDS records ===

2019-11-28: 2 2021-02-01: 2 2022-05-30: 2

=== Brain MRI ===

2019-11-20: Absence of septum pellucidum. Probable small vessel disease.

=== PER-CT ===

2019-11-05: Decreased FP-CIT uptake in dorsal posterior putamen (Rt. << Lt.), probable Parkinson's disease.

=== Gait analysis ===

2019-11-20: The Cadence is within normal range. The Double Support is within normal range. The Single Support is within normal range. The Step Length is more symmetrical than average. The Step Time is within normal range. The Stride Length is within normal range. The Stride Time is more symmetrical than average. The Walking Speed is within normal range.

=== Motor symptom description ===

Speech is normal. Facial expression is slightly impaired. Rest tremor is normal in the face and chin. Rest tremor is normal in the right arm. Rest tremor is normal in the left arm. Rest tremor is normal in the right leg. Rest tremor is slightly impaired in the left leg. Action tremor is normal in the right arm. Action tremor is normal in the left arm. Stiffness is normal in the neck. Stiffness is normal in the right arm. Stiffness is normal in the left arm. Stiffness is normal in the right leg. Stiffness is normal in the left leg. Right finger tapping is normal. Left finger tapping is slightly impaired. Right hand movement is normal. Left hand movement is normal. Rapid right hand movement is normal. Rapid left hand movement is mildly impaired. Left foot tapping is slightly impaired. Right foot tapping is normal. Rising from chair is normal. Standing ability is slightly impaired. Gait is normal. Postural stability is slightly impaired. Movement speed is normal.

Figure 18: Example of longitudinal clinical notes from the HPH cohort for PDD prediction. Unlike the AMC cohort's SOAP-format notes, HPH records consist of structured fields (S/O/A/P) combined with separate clinical index records, autonomic tests, neuroimaging, gait analysis, and motor symptom descriptions. To protect patient privacy, all personal identifiers were removed, and dates were randomly shifted.

Qualitative Comparison of Longitudinal Reasoning for 6–12 Month PDD Prognosis (HPH)

[Dementia-R1 output]

1. Cognitive Assessments:

- MMSE Score: 2019-11-28: 27; 2021-02-01: 28; 2022-05-30: 29
- GDS Score: 2019-11-28: 2; 2021-02-01: 2; 2022-05-30: 2

2. Cognitive Decline Indicators:

- The MMSE scores are relatively stable and within the normal range (27–29), which suggests no significant cognitive decline.
- The GDS scores are consistently at 2, indicating very mild cognitive decline.

3. Functional Assessments:

- Gait Analysis: The gait analysis shows that the patient's gait is normal, and there are no significant impairments noted.
- Motor Symptoms: The patient's motor symptoms are generally stable, with some minor impairments noted, primarily in the left leg and hand.

4. Other Clinical Indicators:

- The patient has experienced some memory decline, but the memory decline is not severe, and the patient's overall cognitive function remains within a normal range.
- The patient has also experienced some psychological issues, including depression and suicidal thoughts, which could be contributing to the perception of cognitive decline.

5. Conclusion: Given the stable MMSE scores (27–29) and the consistent GDS scores (2), the patient does not show significant cognitive decline. The minor motor and gait impairments are not severe enough to suggest a diagnosis of PDD at this time.

Answer: 0

[QWEN 2.5 32B output]

The patient has a history of Parkinson's disease (PD) and has been followed over several years. The notes indicate that the patient has experienced tremors, bradykinesia, and other motor symptoms typical of PD. Additionally, the patient has reported memory decline and cognitive changes.

1. Cognitive Assessments: MMSE scores have been relatively stable over time: 27 in 2019, 28 in 2021, and 29 in 2022. These scores are within the range of mild cognitive impairment but not severely impaired. GDS scores have consistently been 2, indicating very mild cognitive decline.

2. Motor Symptoms: The patient has had tremors and bradykinesia. The HY stage has remained at 1, indicating mild motor impairment. Gait analysis shows normal walking speed and cadence.

3. ANS Tests: The ANS tests have shown adrenergic and cardiovagal dysfunction, which can be associated with PD but are not specific indicators of dementia.

4. Recent Notes: Recent notes (e.g., 2023) indicate significant memory decline and difficulty with work. However, the MMSE score of 29 and GDS score of 2 still indicate only very mild cognitive impairment.

Given that the cognitive assessments have not shown severe cognitive impairment, the patient is unlikely to have developed PDD yet. However, the recent memory decline and cognitive symptoms warrant further monitoring.

Answer: 0

Figure 19: A comparison of the reasoning processes generated by Dementia-R1 and Qwen2.5-32B for the HPH clinical case shown in Figure 18. Both models correctly predict non-progression to PDD, but Dementia-R1 provides a more structured analysis that systematically separates cognitive, functional, and motor assessments, while the baseline produces a less organized narrative.

A.20 Prompt Templates

Stage 1 (Cold-Start Pre-training). Figures 20 and 21 present the prompt templates for the AMC and ADNI pre-training tasks, respectively. In this stage, the model is trained to predict verifiable intermediate clinical indices (e.g., MMSE/GDS/CDR) extracted from unstructured notes or structured records. All templates enforce a unified <think> / <answer> format, enabling reliable parsing of the predicted value for training.

Stage 2 (Task Fine-tuning). Figures 22, 23, and 24 present the prompt templates for the cohort-specific downstream tasks of dementia prognosis on AMC, Parkinson’s disease dementia (PDD) prediction on HPH, and MCI-to-dementia conversion prediction on ADNI. These templates retain the same constrained output format as Stage 1 and guide the model to base its prediction on longitudinal evidence across the provided history.

Other prompts and reuse across settings. Figure 25 shows a separate prompt used to generate teacher rationales for constructing CoT-supervised data for SFT baselines. We reuse the same task prompts across all experimental pipelines, including SFT→SFT, SFT→GRPO, GRPO→GRPO, and single-stage baselines. The pipelines differ only in the optimization procedure and in whether teacher-generated rationales are included.

Pre-training Task: AMC (Future Score Prediction)

[System Prompt]

A conversation between a User and an Assistant. The User provides clinical notes and metadata describing a patient's condition before a cognitive or functional assessment. Each input explicitly specifies the target scale (GDS, MMSE, or CDR), the time interval, and the required output format. The Assistant must carefully read the provided instructions, understand which scale is being predicted, and output the correct numerical value according to the described scoring rule.

[Input Data]

Instruction: You are given longitudinal clinical notes recorded BEFORE a cutoff relative to a cognitive assessment. The most recent included note lies <TIME_INTERVAL> prior to the anchor assessment date.

Task: Predict the target score (**Example: MMSE**) for the anchor assessment.

Format: Output step-by-step reasoning in <think> tags and the final value in \boxed{} within <answer> tags.

Scoring Indicators Glossary:

- MMSE: Integer score ranging from 0 to 30 (Higher = better global cognition).
- GDS: Global Deterioration Scale from 1 to 7 (Higher = more severe impairment).
- CDR: Clinical Dementia Rating global score chosen from {0, 0.5, 1, 2, 3}.

=== Clinical note ===
<CLINICAL_NOTE>

Figure 20: Pre-training prompt template for the AMC dataset. While **MMSE** is shown as an example, the model is pre-trained to predict various global cognitive scores, including **GDS** and **CDR**, based on unstructured clinical notes.

Pre-training Task: ADNI (Future Score Prediction)

[System Prompt]

A conversation between a User and an Assistant. The User provides longitudinal structured ADNI clinical, cognitive, imaging, and biomarker data across multiple visits. The Assistant must predict the future score or diagnosis at the NEXT visit within a specified time window. Target tasks include MMSE, CDRSB, ADAS11, ADAS13, ADASQ4, RAVLT_learning, and LDELTOTAL. Respond only in the specified <think> and <answer> format.

[Input Data]

Instruction: You are given longitudinal records for a single participant. All visits occur before the target visit.

Task: Predict the target score (**Example: MMSE**) at the NEXT visit.

Constraint: Time gap bucket = 2–3 months.

Format: Output step-by-step reasoning in <think> tags and the final predicted value in \boxed{} within <answer> tags.

Variable Glossary:

- PTEDUCAT/APOE4: Education years / Number of APOE ϵ 4 alleles.
- CDRSB/ADAS13/MMSE/MOCA: Clinical severity and cognitive scores (Higher CDRSB/ADAS = worse; Higher MMSE/MOCA = better).
- RAVLT/LDELTOTAL: Memory scores (Lower = poorer memory).
- FAQ: Functional Activities Questionnaire (Higher = worse daily function).
- ABETA/TAU/PTAU: CSF biomarkers for amyloid and tau pathology.
- Ventricles/Hippocampus/WholeBrain: MRI volumetric measures (Structural atrophy).

=== Clinical Assessment Data ===

2006-12-11: «<VISIT 1/2>»

ABETA: 446.8, ADAS13: 25.0, MMSE: 27, CDRSB: 0.5, LDELTOTAL: 12, ...

—(Longitudinal history continues)—

(Prediction target: MMSE score at the next visit)

Figure 21: Pre-training prompt template for the ADNI dataset. The model predicts future indicators (e.g., MMSE, CDRSB) by analyzing longitudinal structured assessment data. The input includes a variable glossary to assist in interpreting clinical indicators.

Fine-tuning Task: AMC (Dementia Prognosis)

[System Prompt]

A conversation between a User and an Assistant. The User provides longitudinal clinical notes and metadata describing a patient's condition. The Assistant must determine whether the patient is likely to be diagnosed with dementia. Output 0 if the patient is unlikely to have dementia, and 1 if the patient is likely to have dementia. Respond only in the specified <think> and <answer> format.

[Input Data]

Instruction: You are given longitudinal clinical notes collected BEFORE a cutoff relative to a dementia diagnosis date. The interval to the diagnosis date is: <TIME_INTERVAL> (e.g., 12–18m).

Task: Predict whether the patient is likely to have dementia (0: unlikely, 1: likely).

Format: Output step-by-step reasoning in <think> tags and the final answer in \boxed{0 or 1} within <answer> tags.

Scoring Indicators Glossary:

- CDR (Global Score): 0 (No dementia), 0.5 (Very mild), 1 (Mild), 2 (Moderate), 3 (Severe). Higher = worse.
- MMSE (Total Score): Integer from 0 to 30. Higher = better cognitive function; Lower = more impairment.
- GDS (Global Deterioration Scale): 1 (No decline) to 7 (Very severe cognitive decline). Higher = worse.

=== Clinical note ===

<CLINICAL_NOTE>

Figure 22: Fine-tuning prompt template for the AMC dataset. The task requires detecting dementia presence based on unstructured clinical notes and integrated scoring indicators.

Fine-tuning Task: HPH (PDD Prediction)

[System Prompt]

A conversation between a User and an Assistant. The User provides longitudinal clinical notes and metadata describing a patient's condition. The clinical notes may contain standardized cognitive and functional assessment information. When present, these indicators should be interpreted as follows:

Scoring Indicators Glossary:

- MMSE (Total Score): Integer from 0 to 30. Higher = better cognitive function; Lower = more impairment.
- GDS (Global Deterioration Scale): 1 (No decline) to 7 (Very severe cognitive decline). Higher = worse.

The Assistant must determine whether the patient is likely to be diagnosed with dementia. Output 0 if the patient is unlikely to have dementia, and 1 if the patient is likely to have dementia. Respond only in the specified <think> and <answer> format.

[Input Data]

Instruction: You are given longitudinal clinical notes collected BEFORE a cutoff relative to the patient's last clinical assessment. The interval between the last assessment date and the most recent available note is: <TIME_INTERVAL> (e.g., ≥ 24 m).

Task: Predict whether the patient diagnosed with Parkinson's disease will develop Parkinson's disease dementia (PDD) (0: unlikely, 1: likely).

Format: Output step-by-step reasoning in <think> tags, justifying the decision using relevant clinical indicators (e.g., MMSE score, GDS, gait metrics, postural instability), and the final answer in \boxed{0 or 1} within <answer> tags.

=== Clinical notes (EN) ===

<CLINICAL_NOTE>

Figure 23: Fine-tuning prompt template for the HPH dataset. The task requires predicting Parkinson's disease dementia (PDD) progression based on longitudinal clinical notes including cognitive assessments, motor symptoms, and gait analysis.

Fine-tuning Task: ADNI Cohort (MCI Conversion Prediction)

[System Prompt]

A conversation between a User and an Assistant. The User provides longitudinal clinical assessment data and metadata describing a patient's cognitive and functional trajectory. The Assistant must determine whether the patient has progressed from a baseline status of Mild Cognitive Impairment (MCI) to dementia by the time of the final diagnosis. Output 0 if the final diagnosis is non-dementia (MCI or CN), and 1 if the patient has converted to dementia. Use trends across longitudinal data (cognition, function, severity scores) for reasoning. Respond only in the specified <think> and <answer> format.

[Input Data]

Instruction: You are given longitudinal clinical assessment data for a patient with baseline MCI. Records are collected before a cutoff set prior to the patient's last diagnostic assessment. The interval between the last diagnosis and the most recent visit is: <TIME_INTERVAL> (e.g., 6–12m).

Task: Predict whether the patient has progressed to dementia (0: non-dementia, 1: converted).

Format: Output step-by-step reasoning in <think> tags and the final answer in \boxed{0 or 1} within <answer> tags.

Variable Glossary:

- PTEDUCAT/APOE4: Education years / Number of APOE ϵ 4 alleles.
- CDRSB/ADAS13/MMSE/MOCA: Clinical severity and cognitive scores (Higher CDRSB/ADAS = worse; Higher MMSE/MOCA = better).
- RAVLT/LDELTOTAL: Memory scores (Lower = poorer memory).
- FAQ: Functional Activities Questionnaire (Higher = worse daily function).
- ABETA/TAU/PTAU: CSF biomarkers for amyloid and tau pathology.
- Ventricles/Hippocampus/WholeBrain: MRI volumetric measures (Structural atrophy).

=== Clinical Assessment Data ===

2011-11-28: «<VISIT 1/7>»

CDRSB: 0.5, ADAS13: 14.0, MMSE: 28, FAQ: 0, Hippocampus: 6521, ...

—(Longitudinal history continues)—

Figure 24: Fine-tuning prompt template for the ADNI dataset. The model predicts MCI-to-dementia conversion using longitudinal trends of clinical indices and biomarkers.

Diagnostic Rationale Generation

[System Prompt]

You are an AI assistant that generates step-by-step reasoning paths.

[User Prompt]

Problem: {problem}

Answer: {answer}

Task: Generate a clear step-by-step reasoning path that explains how to solve the problem and arrive at the answer.

Reasoning:

Figure 25: Prompt template for generating diagnostic rationales for Supervised Fine-Tuning.