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

Anonymous ACL submission

## 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 demonstrate that Dementia-R1 achieves an **F1 score of 77.03%** on real-world unstructured clinical datasets. Notably, on the ADNI benchmark, our 7B model rivals GPT-4o, effectively capturing fluctuating cognitive trajectories. Code is available at https://anonymous.4open.science/r/dementiar1-CDB5.

## 1 Introduction

The digitalization of healthcare and the widespread adoption of Electronic Health Records (EHRs) have resulted in massive amounts of longitudinal patient data that capture individuals' clinical histories across months or years. However, approximately 80% of EHR data is recorded as unstructured text, including physician notes and imaging reports (Kong, 2019; Jensen et al., 2012). These narratives contain rich descriptions of symptom evolution and clinical assessments, yet temporal changes are often documented implicitly rather than in structured form. Since many clinical outcomes are defined retrospectively based on how a patient's condition evolves over time, effective modeling requires longitudinal analysis rather than reliance on information from a single visit. Despite this need, most existing longitudinal disease modeling frameworks are designed for structured data representations and therefore struggle to systematically incorporate unstructured clinical narratives (Waxler et al., 2025; Steinberg et al., 2024; Shmatko et al., 2025).

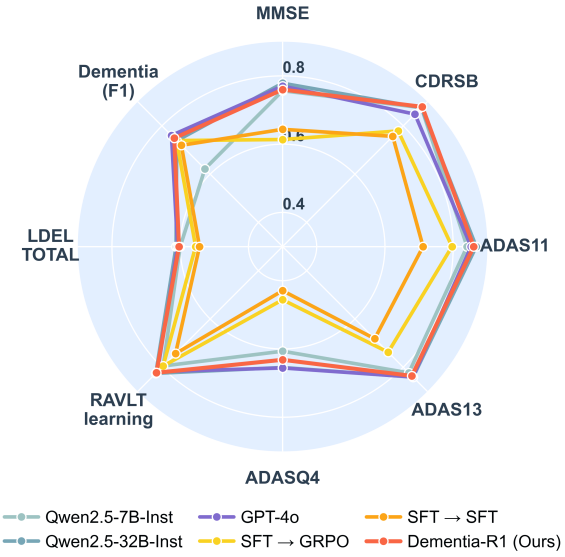


Figure 1: **Multi-dimensional Performance Profile.** Dementia-R1 demonstrates a consistent and balanced performance gain across all dimensions, including intermediate clinical reasoning tasks (e.g., MMSE, CDR-SB, ADAS-Cog) and the final dementia prognosis (F1-score)

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). In particular, LLM-based methods achieve impressive performance on static, snapshot-style benchmarks such as MedQA (Jin et al., 2021), where inputs represent isolated clinical scenarios (Singhal et al., 2025). However, such benchmarks largely ignore longitudinal disease progression. This limitation is critical for diseases

characterized by slow and cumulative progression, such as dementia, where diagnosis requires integrating evidence of cognitive and functional decline across multiple clinical encounters (Grand et al., 2011; Borson et al., 2013; Knopman and Petersen, 2014). Crucially, these trajectories are often non-monotonic; clinical status may fluctuate or temporarily improve, necessitating a holistic evaluation of the patient's condition rather than simple onset detection. In real-world practice, these longitudinal signals are predominantly documented in unstructured clinical notes rather than standardized fields, making dementia a particularly challenging testbed for longitudinal reasoning over clinical text (Kruse et al., 2025a).

To address this challenge, we introduce **Dementia-R1**, a framework designed for longitudinal reasoning using LLMs through Reinforcement Learning (RL). We focus on dementia prognosis as a representative task of complex longitudinal disease progression. Unlike acute diseases, dementia diagnosis requires tracking longitudinal cognitive and functional changes over months or years. These signals are described in clinical narratives, yet they are difficult to quantify explicitly (Borson et al., 2013; Knopman and Petersen, 2014). While standard Supervised Fine-Tuning (SFT) optimizes models to directly predict final labels, RL-based fine-tuning enables the model to learn reasoning processes before making a 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 (e.g., Dementia vs. Non-Dementia) is challenging due to the sparsity of the reward signal and the implicit nature of the underlying reasoning.

We address this issue through a Cold-Start RL strategy with verifiable clinical rewards. Prior work typically relies on SFT to introduce step-wise rationales explicitly (Chen and et al., 2024; DeepSeek-AI, 2025). However, in the context of dementia prognosis, constructing rational trajectories is particularly challenging. Longitudinal reasoning requires temporally consistent analysis across multiple visits and substantial effort from clinical experts to validate them (Kruse et al., 2025a). To mitigate these challenges, we adopt an RL-based pre-training stage using clinically established indices as reward signals rather than explicit reasoning annotations. Specifically, we train the model to predict scores measured at each visit, such as the *Mini-Mental State Examination* (MMSE) (Folstein et al., 1975), *Global Deterioration Scale* (GDS) (Reisberg et al., 2022), and *Clinical Dementia Rating* (CDR) (Morris, 1993). By inferring these indices from longitudinal unstructured notes, the model autonomously acquires essential reasoning primitives, which are subsequently refined in a second stage for the final dementia prediction task.

We validate our approach on both real-world unstructured clinical notes from the Asan Medical Center (AMC) real-world cohort and the structured benchmark (ADNI) (Jack Jr et al., 2008). As illustrated in Figure 1, our model demonstrates comprehensive multi-dimensional reasoning capabilities compared to baselines. 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 using verifiable rewards, demonstrating that learning to estimate intermediate clinical scores is crucial for an accurate dementia prognosis.
- We validate our approach on both private real-world unstructured datasets and a public structured benchmark, demonstrating consistent improvements over the strong baselines, including general-purpose LLMs and medical-specialized reasoning models.

## 2 Related Work

**Longitudinal Clinical Modeling.** Traditional approaches for longitudinal disease modeling have primarily focused on structured electronic health records (EHRs), utilizing Recurrent Neural Networks (RNNs) to process temporal sequences of medical codes (Choi et al., 2016). Recent Transformer-based models have advanced longitudinal forecasting by leveraging large-scale structured records for tasks such as time-to-events prediction (Steinberg et al., 2024), disease trajectory modeling (Shmatko et al., 2025), and medical events modeling (Waxler et al., 2025). While these models show effectiveness for structured data, they fail to capture the nuanced behavioral and symptomatic descriptions found in unstructured clinical notes, which constitute the majority of EHR data. Recent works such as NYUTron (Jiang et al., 2023) and CARE-AD (Li et al., 2025) have demonstrated

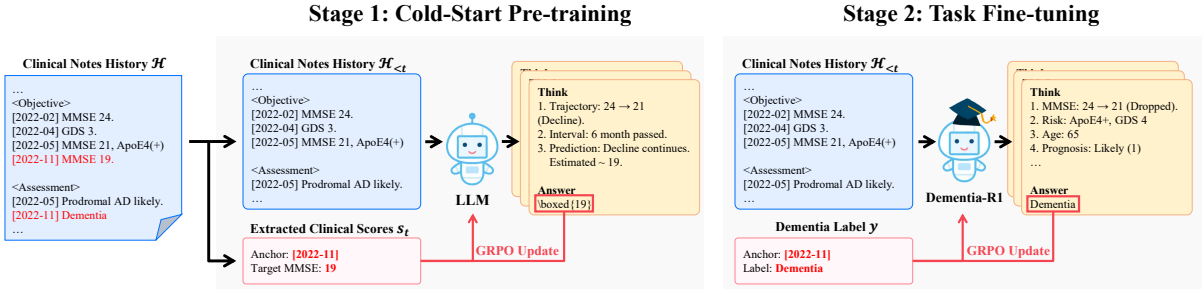


Figure 2: **Overview of the Dementia-R1 Framework.** The pipeline consists of two phases: **Stage 1: Cold-Start Pre-training**, where the base model learns longitudinal reasoning via GRPO on forecasting tasks; and **Stage 2: Task Fine-tuning**, where the reasoning-aligned model is adapted for the final dementia prediction task.

the potential of LLMs for longitudinal prediction using unstructured clinical text. However, these approaches primarily optimize for final clinical outcomes and do not explicitly train models to reason over intermediate disease trajectories or temporal progression patterns. As a result, current frameworks for unstructured clinical text still lack mechanisms for explicit longitudinal reasoning (Kruse et al., 2025b), motivating our approach.

**Reasoning Capabilities of Medical LLMs.** The reasoning capabilities of LLMs in the medical domain have been largely enhanced through Chain-of-Thought (CoT) prompting, which encourages models to generate intermediate rationales (Wei et al., 2022). HuatuoGPT-o1 (Chen and et al., 2024) further improves medical reasoning by combining Supervised Fine-Tuning (SFT) on reasoning trajectories with Reinforcement Learning (RL). In the general domain, recent advances have shifted from SFT to RL with Verifiable Rewards (RLVR), demonstrating that models can learn reasoning when the reward is easily verifiable (DeepSeek-AI, 2025). However, applying this paradigm to clinical tasks remains challenging due to the sparsity of the reward signal and the implicit nature of the required reasoning steps. C-Reason (Kim et al., 2025) partially addresses this challenge using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for sepsis management via masked value prediction; however, it does not address long-term disease progression. We extend this line of work to longitudinal dementia prediction by training the model to track disease progression by estimating clinical scores before determining the final prognosis.

## 3 Methodology: Dementia-R1

Given a sequence of unstructured clinical notes $\mathcal{H} = \{x_1, x_2, \ldots, x_t\}$, we formulate the task as determining the final clinical status $y \in \{0, 1\}$ at a *target anchor* $T_{anchor}$, conditioned on the patient's history $\mathcal{H}_{<T} = \{x_i | i < T, x_i \in \mathcal{H}\}$. This approach requires distinguishing temporary fluctuations from persistent decline across the trajectory, rather than assuming simple linear progression. To enable explicit reasoning over disease progression, we employ a two-stage reinforcement learning framework utilizing Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with verifiable clinical rewards (see Figure 2).

### 3.1 Constructing Verifiable Pretraining Data

Since raw unstructured text lacks explicit ground truth for longitudinal reasoning, we construct a pre-training dataset paired with verifiable clinical indices. We employ a strong auxiliary LLM as an extractor $\mathcal{E}$ to parse unstructured notes into structured clinical scores:

$$s_t = \mathcal{E}(x_t), \quad s_t \in \mathcal{S} \tag{1}$$

where $\mathcal{S}$ represents the set of target indices: MMSE (0–30), GDS (1–7), and CDR (0–3). Using these extracted values as ground truth, we generate a pre-training dataset $\mathcal{D}_{pre} = \{(\mathcal{H}_{<t}, s_t)\}$ where the model is trained to forecast the score $s_t$ at the target visit based on the preceding history $\mathcal{H}_{<t}$. To prevent data leakage, patients reserved for the final dementia prognosis test set are strictly excluded from this phase.

### 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 scores $s_t$ from $\mathcal{D}_{pre}$. We utilize GRPO, which eliminates the need for a value function by estimating the baseline from a group of outputs.

**Verifiable Reward Function ($R_{cold}$)** To accommodate the varying granularity of clinical scales, we define a tolerance-aware reward function. Let

$\hat{s}_t$ be the predicted score from the output $o_t$ of the LLM, $s_t$ be the ground truth, and $\delta$ be the allowable error margin. Considering the range of the MMSE score (0–30), we set a tolerance of $\delta = 2$, treating predictions within this range as correct. For coarser scales like GDS and CDR, we enforce exact matching by setting $\delta = 0$. The reward is defined as:

$$R_{cold} = \mathbb{I}\left(|\hat{s}_t - s_t| \leq \delta\right), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the score 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 $\{o_t^1, o_t^2, \ldots, o_t^G\}$ from the old policy $\pi_{\theta_{old}}$. The policy is optimized to maximize the following:

$$\mathcal{L}(\theta) = \mathbb{E}_{q_t,\{o_t^i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(o_t^i|q_t)}{\pi_{\theta_{old}}(o_t^i|q_t)}A_i,\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(o_t^i|q_t)}{\pi_{\theta_{old}}(o_t^i|q_t)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\right]. \tag{3}$$

Here, $\beta\mathbb{D}_{KL}$ controls the KL-regularization term, $\epsilon$ is the clipping hyperparameter and $A_i$ is the advantage computed by group-based normalization:

$$A_i = \frac{R_{cold}(o_t^i) - \text{mean}(\{R_{cold}(o_t^j)\}_{j=1}^G)}{\text{std}(\{R_{cold}(o_t^j)\}_{j=1}^G)}. \tag{4}$$

This stabilizes training and encourages the model to generate reasoning paths that outperform the average of its own samples.

### 3.3 Stage 2: Task Fine-tuning

After Cold-Start pre-training (Stage 1), we then fine-tune the model on the downstream prognostic classification task (Dementia vs. Non-Dementia) using the same GRPO framework in Eq. (3).

**Sparse Reward Function ($R_{task}$)** Unlike the granular scores in Stage 1, the final diagnosis is binary. Therefore, the reward is defined as:

$$R_{task} = \begin{cases} 1, & \text{if prediction is correct} \\ 0, & \text{if prediction is incorrect} \end{cases} \tag{5}$$

Although this reward signal is sparse, the reasoning capabilities acquired in Stage 1 allow the model to enhance the capability to reason about longitudinal disease progression. In this training stage, the model is optimized for the final prognostic accuracy by generating reasoning traces.

## 4 Experimental Setup

### 4.1 Datasets

We validate the efficacy of Dementia-R1 on two distinct cohorts: the real-world unstructured clinical notes from Asan Medical Center (AMC) and the structured Alzheimer's Disease Neuroimaging Initiative (ADNI) benchmark.

#### 4.1.1 Data Sources and Processing

**Real-World Unstructured Cohort (AMC).** We constructed a large-scale longitudinal dataset using raw clinical notes from Asan Medical Center (AMC). Clinical data were retrospectively collected from approximately 3,000 patients diagnosed with neurocognitive disorders between January 1, 2021, and September 30, 2023. Inclusion criteria were based on ICD-10 codes covering Alzheimer's disease, vascular dementia, and mild cognitive impairment. Electronic Medical Records (EMRs) covering initial and follow-up visits were reviewed to extract SOAP-formatted notes. To ensure privacy, all personally identifiable information was anonymized. Since the target clinical indices (MMSE, CDR, GDS) are predominantly embedded within the free-text "Objective" section, we utilized the LLM-based extraction pipeline (described in Sec 3.1) to isolate these values as verifiable rewards.

**Structured Benchmark Cohort (ADNI).** To demonstrate generalizability, we employed the ADNI dataset (Jack Jr et al., 2008), a widely recognized benchmark for Alzheimer's research. Unlike AMC, ADNI consists of structured tabular records. To adapt this for our LLM-based framework, we applied linearization, transforming tabular rows into chronological textual logs. For verifiable rewards, we selected seven clinically significant indices (e.g., MMSE, CDR-SB) via neurological consultation and feature analysis (Gelir et al., 2025), applying standardized proportional tolerance thresholds (details in Appendix A.3.2).

#### 4.1.2 Longitudinal Sample Construction

To handle the fluctuating nature of cognitive decline across both modalities, we applied a unified construction protocol defined by three key components (illustrated in Figure 3):

- **Target Anchor:** The patient's last clinical visit with a confirmed assessment. The model utilizes the full aggregated history prior to

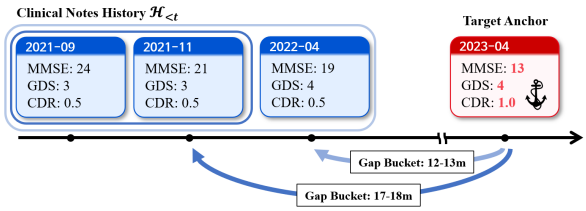


Figure 3: **Examples of Longitudinal Sample Construction.** Patient history is retrospectively sliced relative to a Target Anchor, applying the unified protocol across both unstructured (AMC) and structured (ADNI) data.

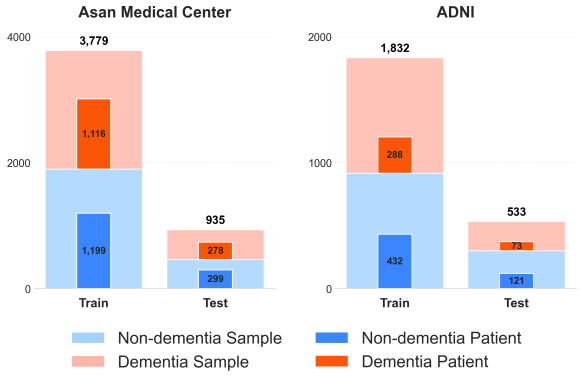


Figure 4: **Dataset Overview.** Visualization of sample and patient counts. Training sets are balanced to prevent bias, while test sets retain natural patient prevalence.

this anchor to distinguish between persistent deterioration and temporary fluctuations.

- **Prediction Target:** The ground-truth outcome varies by training stage:
  - *Stage 1 (Pre-training):* Verifiable clinical indices (extracted scores for AMC; standardized metrics for ADNI).
  - *Stage 2 (Fine-tuning):* The final binary diagnosis, defined as neurologist-adjudicated labels for AMC and standardized DX outcomes (Dementia vs. Non-Dementia) for ADNI.
- **Gap Bucket:** To model temporal sensitivity, we discretized the interval between the last input note and the target anchor:
  - *Stage 1:* Fine-grained 1-month increments (e.g., 0–1m, . . . , 23–24m).
  - *Stage 2:* Coarser intervals (e.g., 6–12m) to ensure clinical utility, excluding short-term gaps (<6m).
  - *ADNI Adaptation:* Adopting AMC's strategy, intervals beyond 24 months were consolidated into a single bucket (>24m) to accommodate longer observation periods.

#### 4.1.3 Data Splitting and Leakage Prevention

To prevent data leakage, we implemented a strict Patient-Level splitting protocol governed by three principles:

1. **Patient-Level Isolation:** Data is split by Patient ID to strictly prevent overlap between training and test sets.
2. **Holistic Test Set Exclusion:** Patients reserved for the Stage 2 test set are excluded from Stage 1 pre-training to ensure full blindness.
3. **Future Information Exclusion:** We aggregate all notes recorded prior to the target anchor, ensuring predictions rely solely on historical symptom trajectories.

Under this protocol, we use a balanced training set (1:1) while retaining natural prevalence in the test set (see Figure 4).

### 4.2 Baselines

We evaluate six configurations based on Qwen2.5-7B-Instruct (Team, 2024) to validate the efficacy of our pure RL pipeline:

- **Zero-shot CoT:** Base model prompted with Chain-of-Thought to elicit reasoning without training.
- **SFT on single stage:** Standard Supervised fine-tuning directly at each stage. Training utilizes Chain-of-Thought rationales distilled from a teacher model.
- **GRPO on single stage:** GRPO applied directly to the prediction task at each stage.
- **SFT → SFT:** A multi-stage SFT pipeline consisting of pre-training on clinical indices followed by fine-tuning on diagnosis, serving as a supervised counterpart to our method.
- **SFT → GRPO:** The conventional RLHF pipeline consisting of SFT warm-up on clinical indices followed by GRPO fine-tuning.

Table 1: **Experimental Results on Asan Medical Center Dataset.** We compare Dementia-R1 against general-purpose LLMs and medical-specific models. **Bold** and underline indicate the best and second-best performance. All results represent mean ± standard deviation across five random seeds.

| Method | Size | Accuracy (↑) | Precision (↑) | Recall (↑) | F1 score (↑) |
|---|---|---|---|---|---|
| *External LLMs* | | | | | |
| HuatuoGPT-o1 | 8B | 67.19 ± 1.3 | 71.55 ± 1.5 | 58.99 ± 1.6 | 64.67 ± 1.5 |
| Qwen2.5-7B-Inst | 7B | 71.94 ± 0.8 | 72.82 ± 0.7 | 71.60 ± 1.1 | 72.20 ± 0.8 |
| Qwen2.5-32B-Inst | 32B | 61.99 ± 0.7 | 57.65 ± 0.4 | **95.46** ± 0.7 | 71.89 ± 0.4 |
| *Specialized Models* | | | | | |
| SFT w/o Stage 1 | 7B | 74.01 ± 1.0 | 72.21 ± 1.0 | 79.58 ± 1.0 | 75.72 ± 0.9 |
| GRPO w/o Stage 1 | 7B | 74.10 ± 0.9 | 70.96 ± 0.9 | 83.15 ± 1.0 | 76.57 ± 0.8 |
| SFT w/o Stage 2 | 7B | 65.60 ± 0.8 | 61.55 ± 0.6 | 86.43 ± 1.1 | 71.90 ± 0.6 |
| GRPO w/o Stage 2 | 7B | 72.47 ± 0.9 | 70.28 ± 0.8 | 79.58 ± 0.9 | 74.64 ± 0.8 |
| SFT → SFT | 7B | **75.14** ± 0.6 | **73.43** ± 0.6 | 80.21 ± 0.6 | 76.67 ± 0.6 |
| SFT → GRPO | 7B | 73.26 ± 0.6 | 70.39 ± 0.6 | 82.24 ± 0.8 | 75.85 ± 0.6 |
| **Dementia-R1** | 7B | 74.93 ± 0.7 | 72.19 ± 0.6 | 82.56 ± 1.1 | **77.03** ± 0.7 |

Table 2: **Generalization Results on ADNI Benchmark.** Comparison extended to include strong ML baselines (Random Forest) and state-of-the-art proprietary models (GPT-4o). Notation and experimental settings follow Table 1 (highlighting performance within the LLM category).

| Model Method | Size | Accuracy (↑) | Precision (↑) | Recall (↑) | F1 score (↑) |
|---|---|---|---|---|---|
| *ML Baseline* | | | | | |
| Random Forest | — | 83.46 ± 0.6 | 83.57 ± 0.7 | 77.13 ± 1.1 | 80.22 ± 0.7 |
| *External LLMs* | | | | | |
| GPT-4o | — | **81.39** ± 0.7 | **86.94** ± 1.8 | 67.64 ± 1.8 | **76.05** ± 1.1 |
| GPT-4o-mini | — | 75.76 ± 0.8 | 73.04 ± 1.3 | 70.64 ± 0.0 | 71.82 ± 0.0 |
| HuatuoGPT-o1 | 8B | 63.11 ± 1.3 | 56.15 ± 1.3 | 71.59 ± 1.2 | 62.93 ± 1.0 |
| Qwen2.5-7B-Inst | 7B | 61.54 ± 0.8 | 54.53 ± 0.7 | 72.36 ± 1.6 | 62.19 ± 0.9 |
| Qwen2.5-32B-Inst | 32B | 76.47 ± 0.6 | 71.82 ± 1.1 | 76.05 ± 1.1 | 73.86 ± 0.6 |
| *Specialized Models* | | | | | |
| SFT w/o Stage 1 | 7B | 75.65 ± 1.5 | 70.90 ± 1.8 | 75.19 ± 2.3 | 72.97 ± 1.6 |
| GRPO w/o Stage 1 | 7B | 76.32 ± 0.5 | 74.08 ± 1.0 | 70.56 ± 1.1 | 72.26 ± 0.5 |
| SFT w/o Stage 2 | 7B | 69.68 ± 2.2 | 62.49 ± 2.0 | 76.65 ± 2.7 | 68.85 ± 2.2 |
| GRPO w/o Stage 2 | 7B | 67.39 ± 1.1 | 60.19 ± 1.0 | 75.02 ± 1.3 | 66.79 ± 1.1 |
| SFT → SFT | 7B | 76.32 ± 0.9 | 74.64 ± 1.3 | 69.44 ± 0.9 | 71.95 ± 1.1 |
| SFT → GRPO | 7B | 76.25 ± 0.9 | 71.10 ± 1.4 | 77.00 ± 0.6 | 73.92 ± 0.8 |
| **Dementia-R1** | 7B | 76.77 ± 1.4 | 70.99 ± 1.7 | **79.31** ± 1.8 | 74.91 ± 1.5 |

- **ML-based Baseline:** Random Forest, selected as the top-performing traditional algorithm on ADNI. Unlike LLMs, it is restricted to the most recent visit due to its inability to handle variable-length longitudinal sequences.

Our proposed method, **Dementia-R1 (GRPO → GRPO)**, represents a pure reinforcement learning approach and is compared against these baselines.

### 4.3 Implementation Details

**SFT.** We conducted Supervised Fine-Tuning (SFT) via knowledge distillation using Qwen2.5-32B-Instruct-AWQ. To construct the training dataset, we prompted the teacher model to generate Chain-of-Thought (CoT) rationales by reverse-engineering the ground-truth labels from the clinical notes. The student model was then fine-tuned on these concatenated (Question, Patient Note, CoT, and Answer) sequences for three epochs with a per-device batch size of 2.

**Dementia-R1.** We train Dementia-R1 using Group Relative Policy Optimization (GRPO) with a group size of $G = 8$ and an effective batch size of 8. Detailed training configurations and hardware specifications are provided in Appendix A.8.

**Evaluation protocol.** To ensure statistical reliability, we conducted all experiments across five distinct random seeds. Consequently, all reported results represent the mean performance ± standard deviation.

## 5 Results

### 5.1 Real-World Unstructured Data Results

**Dementia prognosis.** Table 1 presents the comparative performance on the Asan Medical Cen-

Table 3: **Performance on Clinical Index Prediction for the AMC cohort.** We evaluate the accuracy of predicting MMSE, GDS, and CDR scores.

| Model Method | MMSE | GDS | CDR | Average |
|---|---|---|---|---|
| Qwen2.5-32B-Inst | **57.9** $\pm$ 0.3 | 46.1 $\pm$ 0.3 | 69.8 $\pm$ 0.3 | 57.9 $\pm$ 0.1 |
| Qwen2.5-7B-Inst | 56.1 $\pm$ 0.7 | 45.1 $\pm$ 0.1 | 62.8 $\pm$ 0.7 | 54.7 $\pm$ 0.3 |
| SFT $\rightarrow$ SFT | 52.2 $\pm$ 0.2 | 38.9 $\pm$ 0.5 | 64.1 $\pm$ 1.6 | 51.7 $\pm$ 0.5 |
| SFT $\rightarrow$ GRPO | 54.3 $\pm$ 0.5 | 43.5 $\pm$ 0.6 | 69.7 $\pm$ 0.8 | 55.8 $\pm$ 0.5 |
| **Dementia-R1** | 57.3 $\pm$ 0.3 | **47.7** $\pm$ 0.4 | **73.9** $\pm$ 1.1 | **59.6** $\pm$ 0.5 |

ter (AMC) dataset, which consists of real-world, unstructured clinical narratives. Dementia-R1 achieves the highest F1 score of 77.03%, highlighting the effectiveness of our framework. Specifically, Dementia-R1 outperforms the GRPO baseline (GRPO w/o Stage 1: 76.57%), indicating that avoiding the sparse reward problem with verifiable clinical indices effectively contributes to performance improvement. Furthermore, our pipeline exceeds the standard hybrid approach (SFT $\rightarrow$ GRPO, 75.85%), indicating that active exploration in RL-based pre-training (Stage 1) facilitates more effective modeling of symptom trajectories than supervised fine-tuning.

**Clinical index prediction.** Beyond categorical dementia classification, we further evaluate the model's reasoning capability through quantitative clinical index prediction on the AMC cohort. As shown in Table 3, Dementia-R1 achieves the highest average accuracy (59.61%), surpassing the 7B baselines. Notably, it outperforms the 32B model on GDS and CDR–rigorous metrics used by neurologists for precise disease staging–while maintaining competitive performance on the simpler MMSE screening tool. This capability to infer fine-grained severity demonstrates the model's alignment with expert clinical judgment.

### 5.2 Generalization to Structured Benchmarks

To demonstrate the generalizability of our framework across different data modalities, we applied the Dementia-R1 methodology to the structured ADNI benchmark. By training on linearized tabular records as described in Sec 4.1.1, we verify whether our reinforcement learning approach remains effective on structured data. Table 2 summarizes the performance. Dementia-R1 achieves an F1 score of 74.91%, demonstrating that our framework successfully adapts to structured clinical logs. This performance is comparable to substantially larger models such as GPT-4o (76.05%) and Qwen2.5-32B (73.86%).

To further probe fine-grained reasoning beyond dementia-level classification, we visualize the multi-dimensional performance in Figure 1. Despite having only 7B parameters, Dementia-R1 matches or closely approaches the best-performing models on CDRSB and ADAS scores (see Appendix Table 11). This confirms that our methodology – reinforcement learning with verifiable clinical rewards – is not limited to unstructured text but generalizes effectively to structured data representations.

### 5.3 Neurologist Evaluation

To validate the clinical utility and reasoning quality of Dementia-R1, we conducted a blinded human evaluation involving two board-certified neurologists. We adopted a pairwise comparison protocol on a subset of test cases to analyze the alignment of the models' internal logic with clinical standards. The experts assessed responses across six dimensions: (1) Temporal Reasoning Accuracy, (2) Evidence Grounding, (3) Clinically Relevant Evidence Selection, (4) Medical Soundness, (5) Completeness of Key Findings, and (6) Overall Clinical Utility. For each comparison, evaluators selected the superior response (Win) or marked them as equal (Tie), restricted to cases where both models provided the correct final prognosis. To assess the reliability of the human evaluation, we measured inter-rater agreement, resulting in a Cohen's Kappa score of 0.56, indicating moderate agreement.

For this comparative assessment, we selected Qwen2.5-32B-Instruct as the baseline. Although significantly larger than our 7B backbone, this model demonstrated the second-best performance in the quantitative clinical index prediction task (Section 5.1), surpassing other 7B baselines. This selection enables a rigorous investigation into whether our reasoning-aligned framework generates more clinically valid trajectories than simply scaling model parameters.

As shown in Figure 5, Dementia-R1 recorded a 55% win rate in Overall Clinical Utility. Regarding evidence usage, the model obtained 60% win rates in both Evidence Grounding and Clinically Relevant Evidence Selection. These results suggest that the proposed two-stage RL framework, which incorporates Cold-Start pre-training on clinical indices, enables more clinically grounded longitudinal reasoning compared to parameter scaling alone. In terms of Temporal Reasoning Accuracy, the model achieved a combined Win/Tie rate of 95% (40% Win, 55% Tie) against the 32B base-

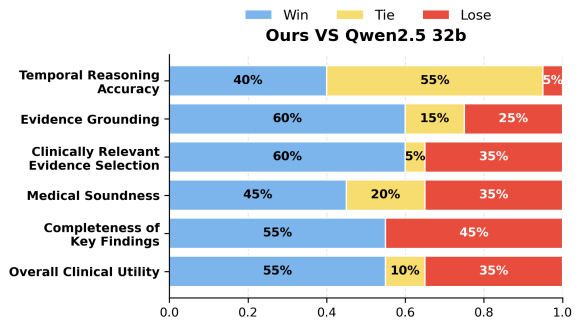


Figure 5: **Neurologist Blind Pairwise Evaluation**. Comparison between Dementia-R1 and the baseline model across six clinical dimensions.

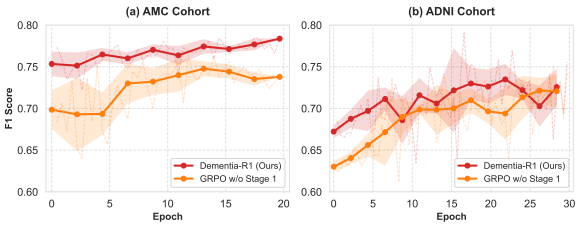


Figure 6: **Training Dynamics.** F1 score trajectories on (a) AMC cohort and (b) ADNI cohort. The inclusion of Stage 1 leads to significantly faster convergence and higher stability across both unstructured and structured domains.

line. These findings suggest that the Cold-Start pre-training can yield clinical reasoning capabilities comparable to those of larger models.

### 5.4 Ablation Study

We investigate the impact of Stage 1 pre-training on learning dynamics. Figure 6 displays the F1 score trajectories evaluated on the test set for Dementia-R1 and a baseline model trained without Stage 1. As observed, the model utilizing Stage 1 shows earlier convergence and higher final F1 scores compared to the baseline across both datasets. These results suggest that alignment with verifiable clinical rewards aids in stabilizing the reinforcement learning process in sparse-reward environments.

### 5.5 Temporal Robustness Analysis

We evaluate the model's robustness across varying temporal intervals between the last clinical note and diagnosis, as visualized in Figure 7. Detailed numerical results are provided in Tables 12 and 13. In the AMC cohort, Dementia-R1 shows consistent performance, peaking at the 12–18 month interval with an F1 score of 79.28%, exceeding the hybrid baseline (SFT $\rightarrow$ GRPO: 78.00%) and the 32B model (74.38%). A similar trend is observed on

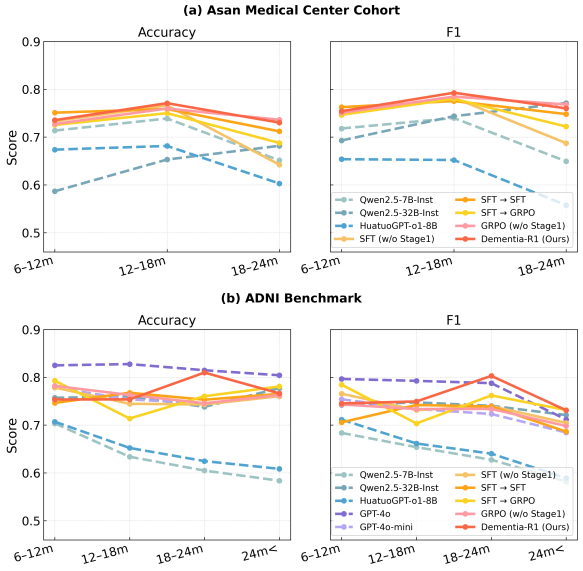


Figure 7: **Performance across time gaps.** Dementia-R1 demonstrates consistent stability, especially in long-term predictions, compared to baselines.

the ADNI cohort. Specifically, based on F1 scores, Dementia-R1 outperforms GPT-4o in the 18–24 month interval (80.30% vs. 78.78%) and maintains higher performance in the >24 month horizon (73.11% vs. 71.18%). These findings suggest that aligning with longitudinal trajectories through verifiable rewards contributes to sustained reasoning capabilities in long-term forecasting scenarios.

## 6 Conclusion

In this work, we presented Dementia-R1, a Reinforcement Learning framework designed to infer longitudinal disease progression from unstructured clinical narratives. Addressing the limitations of sparse rewards in prognostic tasks, we introduced a Cold-Start RL strategy that aligns the model with verifiable clinical indices before fine-tuning for the final diagnosis. Empirical results on both the real-world AMC cohort and the structured ADNI benchmark demonstrate that our approach enables a 7B parameter model to achieve performance comparable to, or exceeding, that of significantly larger baselines. Furthermore, qualitative evaluations by neurologists indicate that explicit training on intermediate clinical scores fosters more grounded and transparent reasoning trajectories. We hope this work inspires further research into reinforcement learning with verifiable rewards for complex, long-horizon clinical decision-making.

## Limitations

We acknowledge several limitations in our study.

- First, regarding data generalization, our unstructured dataset comes from a single institution (Asan Medical Center). This may limit the model's ability to generalize to other demographics or documentation styles. Future validation on diverse, multi-center datasets is necessary.
- Second, linguistic limitations may arise from the translation process. Converting Korean notes into English might result in the loss of subtle nuances, such as syntax errors, which are important for assessing cognitive decline. Future work should apply our method directly to native-language texts.
- Third, our framework relies on the performance of the auxiliary Large Language Models (LLMs). We utilized the Qwen2.5 series for data preprocessing, including the translation of clinical notes and the extraction of clinical scores. Consequently, our reward mechanism depends on the accuracy of these models; since we use the extracted clinical scores as rewards, any extraction errors or hallucinations could introduce noise into the reinforcement learning process.
- Finally, our approach relies on quantifiable clinical indices (e.g., MMSE) for rewards. This limits immediate application to diseases that lack standardized numerical records. Extending this framework to conditions with subjective or qualitative markers remains a challenge for future work.

## 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 Asan Medical Center Ethics Committee and the Declaration of Helsinki.

## 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, and 1 others. 2013. Improving dementia care: the role of screening and detection of cognitive impairment. *Alzheimer's & Dementia*, 9(2):151–159.

Junying Chen and et al. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

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*, pages 301–318. PMLR.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Marshal 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 Kevil, 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*, 45:63–83.

Jacob HG Grand, Sienna Caspar, and Stuart WS MacDonald. 2011. Clinical features and multidisciplinary approaches to dementia care. *Journal of multidisciplinary healthcare*, pages 125–147.

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, and 1 others. 2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 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, and 1 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.

Junu Kim, Chaeeun Shim, Sungjin Park, Su Yeon Lee, Gee Young Suh, Chae-Man Lim, Seong Jin Choi, Song Mi Moon, Kyoung-Ho Song, Sejoong Kim, and 1 others. 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. In *Mayo clinic proceedings*, volume 89, pages 1452–1459. Elsevier.

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. 2025a. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, volume 2025, pages 20715–20735.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025b. Zero-shot large language models for long clinical text summarization with temporal reasoning. *medRxiv*, pages 2025–07.

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.

John C Morris. 1993. The clinical dementia rating (cdr) current version and scoring rules. *Neurology*, 43(11):2412–2412.

Barry Reisberg, Ramu Vadukapuram, and Sunnie Kenowsky. 2022. The global deterioration scale (gds). *World Alzheimer Report 2022*, page 44.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, and 1 others. 2025. Learning the natural history of human disease with generative transformers. *Nature*.

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, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam H Shah. 2024. Motor: A time-to-event foundation model for structured medical records. In *ICLR*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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, and 1 others. 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.

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 pretraining (AMC).**

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

# A Appendix

## A.1 Data Preprocessing Details

## A.2 Pretraining Data Statistics (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 score 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 the overall dataset composition, and Table 5 reports the task-wise distribution.

## A.3 Pretraining Data Statistics (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.3.2). 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.4 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.

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 |

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.

### A.3.1 Asan Medical Center (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 $\pm 2$ 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.3.2 ADNI Benchmark (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 Asan 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 $\pm 2$ points on the 30-point MMSE scale). For each index, the allowable error margin $\delta$ 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 Asan 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% subsample of the test set.

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

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

### A.4 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 8 details the distribution of these time gaps for the test sets. Notably, the ADNI cohort (right) presents a significantly more challenging scenario, with approximately half ($\sim$50%) of the samples having a gap exceeding 24 months, including a long tail extending beyond 36 months. This contrasts sharply with the Asan cohort, which is predominantly concentrated in the short-term range (6–12 months). This diversity ensures that our evaluation covers both immediate screening and long-term prognostic scenarios.

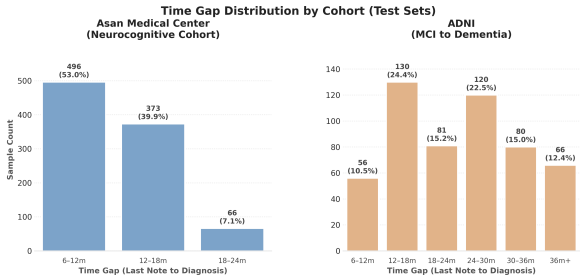


Figure 8: **Time Gap Distribution by Cohort (Test Sets).** The histograms show the interval between the last available clinical note and the diagnosis date. The Asan cohort is concentrated in shorter intervals (6–18m), reflecting relatively dense clinical follow-up prior to diagnosis. In contrast, the ADNI cohort displays a substantially wider temporal range, extending beyond 36 months, which reflects the longitudinal nature of MCI progression monitoring.

### A.5 Stage 1: Performance of Intermediate Clinical Indices

We analyze the model's capability to predict intermediate clinical indices. On the structured ADNI benchmark (Tables 10, 11), 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 9), the model consistently outperforms baselines in short-to-mid term intervals, achieving notable gains in GDS and CDR prediction.

### A.6 Stage 2: Performance of Final Binary Prognosis

Building on the clinical indicators established in Stage 1, Stage 2 assesses the model's ultimate capacity to determine the final binary prognosis. As summarized in Table 12, Dementia-R1 achieves the highest overall F1-score on the Asan dataset, demonstrating notable consistency across the term intervals. Notably, it maintains competitive performance in both short-term (6–12m) and long-term (18–24m) intervals, proving its stability across different forecasting ranges. Similarly, the results on the ADNI benchmark (Table 13) further highlight the model's enhanced robustness in long-term forecasting ($>$ 18 months). In these extended horizons, Dementia-R1 not only surpasses proprietary frontier models such as GPT-4o but also outperforms larger specialized baselines, effectively validating confirming its effectiveness in modeling longitudinal disease trajectories.

### A.7 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 9 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 10 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.8 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

Table 9: **Stage 1 Performance by Time Gap (Asan). Dementia-R1** achieves superior accuracy in short-to-mid term intervals (0–18 months), validating the effectiveness of the Cold-Start strategy.

| | Overall | Accuracy by Time Gap to Prediction Target | | | |
|---|---|---|---|---|---|
| **Method** | **Acc (↑)** | **0–6m** | **6–12m** | **12–18m** | **18–24m** |
| *External LLMs* | | | | | |
| Qwen2.5-32B-Inst | 57.90 ± 0.05 | 60.66 ± 0.43 | 54.91 ± 0.53 | 55.15 ± 1.21 | **50.43** ± 5.08 |
| Qwen2.5-7B-Inst | 54.68 ± 0.32 | 57.56 ± 0.69 | 51.19 ± 1.16 | 52.76 ± 0.99 | 46.32 ± 2.28 |
| *Specialized Models* | | | | | |
| SFT → SFT | 51.73 ± 0.53 | 55.07 ± 0.87 | 48.37 ± 0.75 | 46.95 ± 0.59 | 46.90 ± 5.02 |
| SFT → GRPO | 55.81 ± 0.46 | 59.60 ± 0.69 | 51.21 ± 1.15 | 52.64 ± 0.41 | 49.62 ± 5.43 |
| **Dementia-R1** | **59.61** ± 0.46 | **63.38** ± 0.45 | **55.25** ± 0.69 | **56.41** ± 0.58 | 49.49 ± 2.84 |

Table 10: **Stage 1 Performance by Time Gap (ADNI). Dementia-R1** demonstrates superior long-term reasoning capability (>18 months) compared to larger baselines (GPT-4o, 32B), validating the efficacy of the proposed RL framework.

| | Overall | Accuracy by Time Gap to Prediction Target | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Acc (↑)** | **0–6m** | **6–12m** | **12–18m** | **18–24m** | **>24m** |
| *External LLMs* | | | | | | |
| GPT-4o | 77.04 ± 0.19 | 79.52 ± 0.67 | 77.06 ± 0.31 | 76.22 ± 0.49 | 79.33 ± 0.99 | 77.55 ± 0.59 |
| Qwen2.5-32B-Inst | **77.91** ± 0.18 | **81.26** ± 0.27 | **77.38** ± 0.25 | **76.77** ± 0.52 | 80.87 ± 0.62 | 79.72 ± 0.72 |
| Qwen2.5-7B-Inst | 75.67 ± 0.31 | 77.87 ± 0.46 | 75.09 ± 0.43 | 75.46 ± 0.58 | 78.89 ± 0.91 | 77.48 ± 1.17 |
| *Specialized Models* | | | | | | |
| SFT → SFT | 64.42 ± 0.23 | 66.60 ± 0.28 | 63.58 ± 0.69 | 63.26 ± 0.68 | 67.65 ± 1.40 | 68.55 ± 2.02 |
| SFT → GRPO | 67.61 ± 0.33 | 71.26 ± 1.16 | 67.47 ± 0.62 | 65.05 ± 0.51 | 69.16 ± 1.07 | 71.50 ± 1.56 |
| **Dementia-R1** | 77.04 ± 0.28 | 79.44 ± 0.92 | 76.30 ± 0.51 | 76.23 ± 0.52 | **80.97** ± 1.09 | **80.27** ± 0.89 |

Table 11: **Stage 1 Accuracy by Clinical Index (ADNI). Dementia-R1** outperforms larger 32B models on **CDR-SB** while maintaining competitive performance against GPT-4o on other key metrics (e.g., ADAS-Cog, RAVLT).

| **Method** | **Overall Acc (↑)** | **CDRSB Acc (↑)** | **ADAS11 Acc (↑)** | **ADAS13 Acc (↑)** | **RAVLT Acc (↑)** | **MMSE Acc (↑)** | **ADASQ4 Acc (↑)** | **LDEL Acc (↑)** |
|---|---|---|---|---|---|---|---|---|
| *External LLMs* | | | | | | | | |
| GPT-4o | 77.04 ± 0.19 | 84.85 ± 0.51 | 85.23 ± 0.32 | 84.04 ± 0.24 | 81.83 ± 0.39 | 77.01 ± 0.47 | **65.56** ± 0.29 | 60.76 ± 0.95 |
| Qwen2.5-32B-Inst | **77.91** ± 0.18 | 87.52 ± 0.51 | **86.74** ± 0.47 | **84.09** ± 0.54 | **82.52** ± 0.30 | **77.86** ± 0.24 | 65.43 ± 0.61 | **61.21** ± 0.61 |
| Qwen2.5-7B-Inst | 75.67 ± 0.31 | 87.55 ± 0.47 | 84.06 ± 0.26 | 82.32 ± 0.65 | 79.48 ± 0.90 | 75.65 ± 1.26 | 60.65 ± 1.49 | 60.00 ± 0.75 |
| *Specialized Models* | | | | | | | | |
| SFT → SFT | 64.42 ± 0.23 | 75.66 ± 0.88 | 71.17 ± 0.84 | 68.18 ± 0.99 | 74.35 ± 0.97 | 64.35 ± 1.81 | 42.92 ± 1.86 | 54.34 ± 1.72 |
| SFT → GRPO | 67.61 ± 0.33 | 77.91 ± 0.85 | 79.65 ± 1.16 | 73.76 ± 1.42 | 79.50 ± 1.08 | 61.36 ± 1.03 | 45.58 ± 0.98 | 55.48 ± 1.21 |
| **Dementia-R1** | 77.04 ± 0.28 | **87.89** ± 0.42 | 86.02 ± 0.54 | 83.58 ± 0.94 | 82.28 ± 1.13 | 76.00 ± 0.93 | 63.17 ± 1.33 | 60.35 ± 1.06 |

Table 12: **Stage 2 Fine-tuning Performance by Time Gap (Asan). Dementia-R1** achieves the highest overall F1 score, demonstrating its robust reasoning capabilities across the entire temporal intervals.

| | Overall | F1 score by Time Gap to Prediction Target | | |
|---|---|---|---|---|
| **Method** | **F1 (↑)** | **6–12m** | **12–18m** | **18–24m** |
| *External LLMs* | | | | |
| Qwen2.5-32B-Inst | 71.89 ± 0.80 | 69.29 ± 0.70 | 74.38 ± 0.70 | **77.12** ± 1.60 |
| Qwen2.5-7B-Inst | 72.20 ± 0.80 | 71.80 ± 0.60 | 74.00 ± 1.60 | 64.94 ± 3.20 |
| *Specialized Models* | | | | |
| SFT → SFT | 76.67 ± 0.60 | **76.27** ± 0.90 | 77.55 ± 1.00 | 74.82 ± 3.50 |
| SFT → GRPO | 75.85 ± 0.60 | 74.66 ± 0.90 | 78.00 ± 1.10 | 72.21 ± 0.90 |
| **Dementia-R1** | **77.03** ± 0.72 | 75.43 ± 0.69 | **79.28** ± 0.80 | 76.02 ± 2.73 |

Bfloat16 precision with a 2,000-token completion limit.

### A.9 Human evaluation protocol

To assess the clinical quality of reasoning, we conducted a blinded human evaluation with med-

Table 13: **Stage 2 Fine-tuning Performance by Time Gap (ADNI).** While large-scale general models (e.g., GPT-4o) excel in short-term forecasting, **Dementia-R1** demonstrates superior robustness in long-term reasoning ($>$ 18 months).

| | Overall | F1 score by Time Gap to Prediction Target | | | |
|---|---|---|---|---|---|
| **Method** | **F1 (↑)** | **6–12m** | **12–18m** | **18–24m** | **>24m** |
| *External LLMs* | | | | | |
| GPT-4o | **76.05** ± 1.05 | **79.66** ± 2.51 | **79.26** ± 0.78 | 78.78 ± 1.68 | 71.18 ± 6.98 |
| Qwen2.5-32B-Inst | 73.86 ± 0.57 | 74.21 ± 2.15 | 74.76 ± 2.00 | 74.04 ± 1.63 | 72.10 ± 4.64 |
| Qwen2.5-7B-Inst | 62.19 ± 0.94 | 68.37 ± 4.76 | 65.39 ± 2.99 | 62.76 ± 2.27 | 58.18 ± 5.95 |
| *Specialized Models* | | | | | |
| SFT → SFT | 71.95 ± 1.07 | 70.58 ± 4.87 | 74.21 ± 2.78 | 73.91 ± 1.96 | 68.65 ± 6.43 |
| SFT → GRPO | 73.92 ± 0.76 | 78.46 ± 4.64 | 70.35 ± 1.47 | 76.21 ± 3.01 | **73.15** ± 5.48 |
| **Dementia-R1** | 74.91 ± 1.49 | 74.54 ± 4.43 | 74.95 ± 1.73 | **80.30** ± 3.81 | 73.11 ± 3.03 |

ical 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, and judgments were collected independently for the following six 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. **Completeness of Key Findings:** Which response reflects all important symptoms documented in the clinical notes without omitting key findings?

6. **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?

### A.10 Prompt Templates

**Stage 1 (Cold-Start Pre-training).** Figures 11 and 12 present the prompt templates for the Asan Medical Center 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 13 and 14 present the prompt templates for the cohort-specific downstream tasks of dementia detection on Asan 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 15 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.

## Longitudinal Clinical Note Input

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/1.73
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 amnestic 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/1.73
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 9: 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 amnestic 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 amnestic 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 10: A comparison of the reasoning processes generated by Dementia-R1 and the Qwen2.5-32B for the clinical case shown in Figure 9. While Dementia-R1 produces a structured and grounded rationale, the baseline generates a dense narrative containing medical hallucinations.

**Pre-training Task: Asan Medical Center (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 11: Pre-training prompt template for the Asan Medical Center 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 $\epsilon$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 12: 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: Asan Medical Center (Dementia Detection)

**[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 13: Fine-tuning prompt template for the Asan Medical Center dataset. The task requires detecting dementia presence based on unstructured clinical notes and integrated scoring indicators.

## 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 $\epsilon$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 14: Fine-tuning prompt template for the ADNI dataset. The model predicts MCI-to-dementia conversion using longitudinal trends of clinical scores 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 15: Prompt template for generating diagnostic rationales for Supervised Fine-Tuning.