DemMA: Dementia Multi-Turn Dialogue Agent with Expert-Guided Reasoning and Action Simulation

Yutong Song♣*, Jiang Wu♠*, Kazi Sharif♢, Pengfei Zhang♣, Wenjun Huang♣, Honghui Xu♢, Nikil Dutt♣, Amir M. Rahmani♣
♣ University of California, Irvine, ♠ Independent Researcher, ♢ Kennesaw State University
{yutons12, amirr1}@uci.edu

Abstract

Simulating dementia patients with large language models (LLMs) is challenging due to the need to jointly model cognitive impairment, emotional dynamics, and nonverbal behaviors over long conversations. We present DemMA, an expert-guided dementia dialogue agent for high-fidelity multi-turn patient simulation. DemMA constructs clinically grounded dementia personas by integrating pathology information, personality traits, and subtype-specific memory-status personas informed by clinical experts. To move beyond text-only simulation, DemMA explicitly models nonverbal behaviors, including motion, facial expressions, and vocal cues. We further introduce a Chain-of-Thought distillation framework that trains a single LLM to jointly generate reasoning traces, patient utterances, and aligned behavioral actions within one forward pass, enabling efficient deployment without multi-agent inference.

1 Introduction

In dementia research and caregiver training, the scarcity of high-quality interaction data remains a structural bottleneck (Livingston et al., 2017). Due to strict privacy and ethical barriers, collecting sensitive patient data is heavily restricted (Johnson and Karlawish, 2015). Consequently, public corpora, especially those capturing multimodal resources such as facial expressions and vocal prosody are virtually absent. This "data desert" is particularly acute at the level of pathological granularity: even for common types such as Alzheimer's disease (AD), public corpora are very limited, and over 20 clinically subtypes remain largely unrepresented (Gorno-Tempini et al., 2011; Hornberger et al., 2010a). As a result, patient simulation often relies on scripted materials that fail to capture the heterogeneity of real-world care interactions. The emergence of LLM-based generative agents offers a promising avenue for data synthesis but using generic dialogue agent models as dementia simulators is unreliable (Wei et al., 2022; Liu et al., 2023; Jeong et al., 2026). First, without clinically grounded alignment, such models risk generating content that lacks medical rigor or may even yield unsafe guidance. Second, in extended interactions, simulations often suffer from persona drift, whereby the model gradually regresses toward a fluent, polite, and generic assistant style (Shumailov et al., 2023). This "over-perfect" articulation suppresses essential markers of cognitive decline. Consequently, although the synthesized dialogues may appear linguistically fluent, they lack patient persona fidelity, rendering them inadequate for agent training.

A core challenge in dementia simulation is to model interpretable, reproducible, and subtype-specific pathological behaviors, rather than injecting stochastic variation into model outputs (Jack et al., 2018). Although general symptoms such as confusion and linguistic errors are shared across conditions, different dementia subtypes exhibit distinct cognitive impairments. High-fidelity simulation therefore requires mechanisms that differentiate pathological inconsistency from generic model hallucination.

In addition, dementia communication is inherently multichannel, encompassing language, affect, and behavior (Bender et al., 2022). As linguistic ability declines, clinically salient signals increasingly manifest through non-verbal expressions, which are not preserved by text-only LLMs. This limitation motivates structured textual representations that encode latent non-verbal cues to compensate for degraded speech.

Finally, from a systems perspective, long-

*Equal contribution
The code is available at https://github.com/RainieLLM/DemMA

Findings of the Association for Computational Linguistics: ACL 2026, pages 12439–12470
July 2-7, 2026

horizon dialogue coherence is commonly handled via multi-agent architectures, which introduce substantial overhead and latency, making them unsuitable for real-time dementia care training and simulation.

Based on these motivations, we propose **DemMA**, a framework that fine tunes open source models for end to end dementia virtual patient generation without relying on sensitive real world data. Moreover, to address the limitation that text only simulation often underspecifies language impairments, DemMA adopts a dual track modeling paradigm:

Intrinsic Cognitive Level: The model explicitly captures fine-grained differences across dementia subtypes and stages, ensuring that generated incoherence reflects specific persona patterns rather than random noise.

Extrinsic Expression Level: We introduce Action Labels as a compensatory mechanism to map multimodal behaviors (e.g., motion, facial expressions, sound) into text-based labels, serving three key roles: (i) explicitly revealing latent emotions and intentions that cannot be directly expressed through ambiguous language; (ii) projecting multimodal behaviors onto a text-based interaction interface; and (iii) providing key signals to distinguish between different subtypes and disease stages.

Our contributions are summarized as follows:

- We propose a clinically grounded modeling approach within DemMA for precise representation of **Dementia Patient Personas**.
- We are the first work to introduce explicit motion, facial expression and sound labels into LLM-based generative agents.
- We introduce a multi-agent pipeline for producing high-quality interaction dialogue data and release **DemMA-Dialogue**, the first synthetic dementia dialogue dataset covering main dementia subtypes with expert validation.
- We develop a distillation training for agent **DemMA** to internalize long-horizon planning signals, supporting low latency inference with coherent, persona-consistent interactions.

2 DemMA

Architecturally, DemMA integrates three components: (i) Clinically grounded patient persona formation module; (ii) Multi-agent dialogue dataset generation pipeline encompassing memory analysis and search, dialogue planning, language and action simulation; and (iii) CoT distillation multi-task training for DemMA agent, internalizing medical knowledge while reducing multi-agent inference latency. The overview of DemMA is shown in Fig. 1.

Dementia Persona Formation A core challenge in dementia persona simulation is enforcing longitudinal consistency and clinical validity. Naïve single-pass generation often lacks medical grounding, resulting in inconsistent profiles and persona drift. DemMA addresses this by constructing the patient persona via staged generation with explicit dependencies, propagating upstream constraints to all downstream components:

$$\begin{aligned} &B \sim f_B(\cdot), \quad S \sim f_S(B), \\ &M \sim f_M(B,S), \quad \mathcal{P} \triangleq \langle B,S,M \rangle, \end{aligned} \tag{1}$$

where B, S, and M denote the personal background, personality, and memory layers, respectively, and $\mathcal{P}$ represents the resulting patient persona profile.

Background Layer (B) This layer encodes (i) stable demographic and life-context attributes (e.g., age, education, and comorbidities) and (ii) the dementia subtype persona along with its core clinical patterns (Convery et al., 2019). We instantiate nine subtypes covering the entire spectrum: *AD-early, AD-mid/late, VaD, DLB, PDD, FTD-bv, nfvPPA, svPPA, and lvPPA* (see details in Appendix A, Table 9). These attributes, including subtype-specific behavioral characteristics and neuropathological rationales (Appendix A, Table 4), establish the foundational medical and contextual constraints for subsequent modeling.

Personality Layer (S) S models interpersonal style within the WHO International Classification of Functioning, Disability and Health (ICF) `b126` psychological-function space (World Health Organization, 2001) (Appendix A, Table 8). Crucially, dementia-induced personality alterations are not stochastic but track the underlying neuropathology. Therefore, rather than treating personality as a disease-independent attribute, we model it as subtype-specific tendencies encoded in the ICF space (Appendix A; Tables 4 and 5). This structural dependency ensures the generation of diverse yet clinically grounded interaction styles, preventing medically implausible profiles.

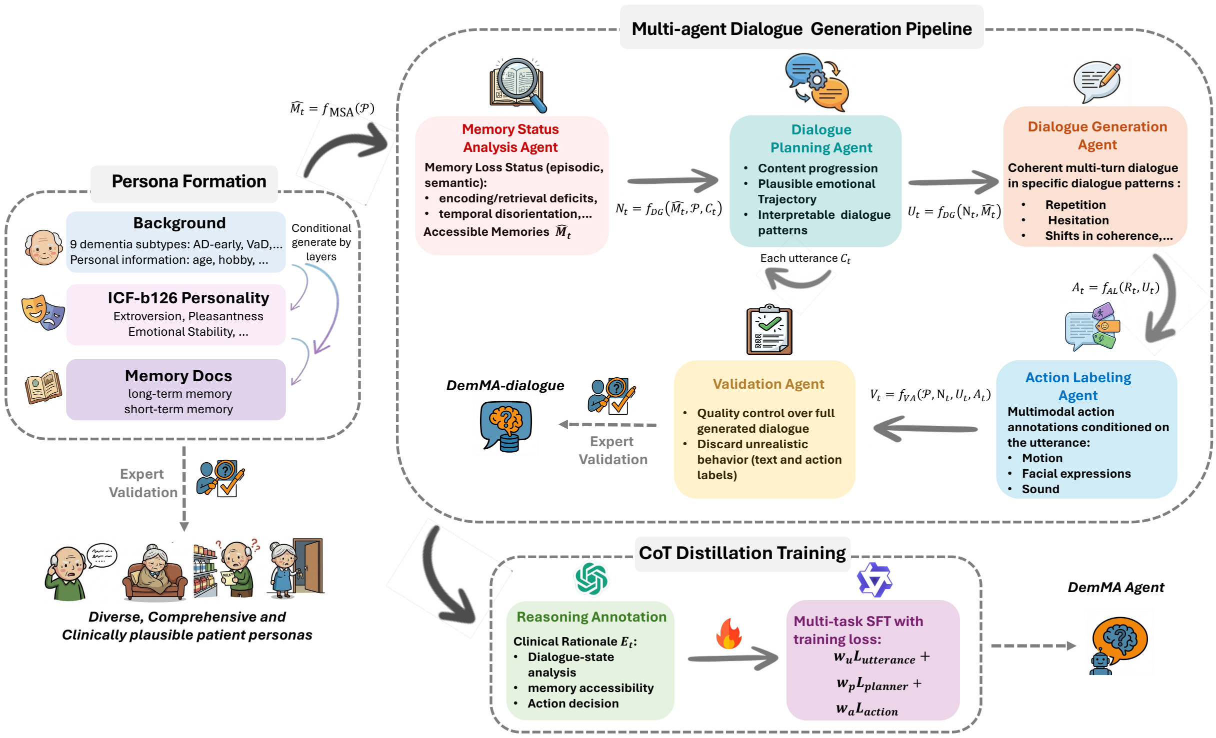

Figure 1: DemMA integrates three components: a) Clinically grounded patient persona formation module; b) Multi-agent dialogue dataset generation pipeline encompassing memory analysis, planning, language generation, and action simulation; and c) CoT distillation multi-task training workflow for DemMA agent.

Memory Layer (M) We instantiate M from (B, S) as (i) long-term memory (e.g., childhood experiences, life history, and salient life events) and (ii) short-term daily/weekly records of recent events, constrained to remain consistent with both long-term memory and the clinical facts in B (e.g., comorbidities). This layer will characterize distinct retention and degradation patterns of remote and recent information.

3 Multi-Agent LLM Workflow

DemMA employs a multi-agent LLM workflow to generate long-horizon dialogue with structured control and interpretability. Separating reasoning from generation has been shown to improve coherence and reduce hallucination in complex language tasks (Yao et al., 2023). The five agents operate as described below.

3.1 Memory Status Analysis Agent

The workflow begins with the Memory Status Analysis Agent. Based on the dementia persona and corresponding memory-status information derived from clinical diagnostic literature (Appendix A, Table 6), the agent identifies accessible memory systems during the dialogue and infers clinically grounded features, such as the availability of recent, remote, and semantic memory, as well as cue responsiveness:

$$\hat{M}_t = f_{\text{MSA}}(\mathcal{P}), \tag{2}$$

where $\hat{M}_t$ denotes the accessible memories and clinically grounded memory features at turn t and f_{MSA} is the memory analysis function. This inferred memory state conditions all downstream reasoning and generation.

3.2 Dialogue Planning Agent

The *Dialogue Planning Agent* then produces a high-level plan conditioned on the patient's personality, inferred memory status, and dialogue context C_t. The agent organizes the overall conversational flow by specifying planned content progression and a plausible emotional trajectory across turns as:

$$N_t = f_{\text{DR}}(\hat{M}, \mathcal{P}, C_t), \tag{3}$$

Externalizing this planning step improves transparency and control, ensuring that the generated dialogue follows interpretable patterns consistent with the patient's cognitive and emotional state rather than opaque model behavior.

3.3 Dialogue Generation Agent

Dialogue Generation Agent produces the natural-language dialogue, converting high-level plans into turn-level utterances while maintaining consistency with the patient's cognitive limitations, personality traits, and emotional state. By explicitly conditioning generation on structured plans, the agent produces coherent multi-turn dialogue while preserving dementia-specific communication patterns such as repetition, hesitation, misplaced references, and shifts in coherence. Utterance generation at turn t is modeled as:

$$U_t = f_{\mathrm{DG}}(N_t, \hat{M}_t). \tag{4}$$

3.4 Action Labeling Agent

The *Action Labeling Agent* augments each dialogue turn with multimodal action annotations conditioned on the generated utterance and the inferred reasoning state. It assigns labels corresponding to plausible **Motion**, **Facial expressions**, and **Sound**, capturing behavioral signals commonly observed in dementia interactions such as fidgeting, gaze aversion, trembling speech, and sudden emotional shifts (Liu et al., 2023). The full list of action labels is provided in Appendix A.2.

These action sequences provide a multimodal view of patient responses and support downstream tasks such as embodied simulation and multimodal data generation. Withe the R_t denotes reasons for more multimodal information, the grounding action label predictions at turn t are modeled as:

$$A_t = f_{\mathrm{AL}}(R_t, U_t). \tag{5}$$

3.5 Validation Agent

The final component of the workflow is the Validation Agent, which performs quality assurance for the entire generated conversation. Using the full dialogue, the structured persona profile, and the inferred memory status, the agent evaluates whether the output aligns with the patient's characteristics and dementia-related constraints. It generates a quality score and determines whether the conversation should be accepted, regenerated, or discarded. This iterative validation reduces the risk of drift, inconsistencies, or unrealistic behaviors in the final dataset. The Validation Agent performs quality assurance on the generated dialogue, ensuring that it aligns with the patient's characteristics. It evaluates the quality as:

$$V_t = f_{\mathrm{VA}}(\mathcal{P}, R_t, U_t, A_t), \tag{6}$$

where V_t denotes the validation score at turn t and f_{VA} is the validation function. A turn is accepted when: $V_t \geq \tau$, and regenerated otherwise.

3.6 Expert Validation

Before finalizing the dataset, we conduct human validation by 4 dementia domain experts to verify that the retained dialogues exhibit emotional states, linguistic patterns, action behaviors, and memory impairments consistent with the corresponding dementia subtype and disease stage. We use this multi-agent framework to generate the first LLM-based dementia dialogue dataset named DemMA-Dialogue. The statistics are shown in Table 1.

Motion	# (Pct.)	Sound	# (Pct.)
lowering head	4,380 (27.8%)	verbal hesitation (um / uh)	13,703 (57.6%)
fidgeting	3,096 (19.6%)	sighing	2,991 (12.6%)
looking around	2,347 (14.9%)	murmuring / self-talk	2,602 (10.9%)
pushing caregiver away	686 (4.3%)	repetitive words	2,197 (9.2%)
touching forehead	590 (3.7%)	silence for several seconds	1,592 (6.7%)
standing up	488 (3.1%)	crying	399 (1.7%)
others	4,179 (26.5%)	groaning in pain	313 (1.3%)
Facial expressions	**# (Pct.)**	**Cognitive profile**	**# (Pct.)**
frowning	5,600 (37.0%)	has remote episodic memory	13,303 (83.0%)
avoiding eye contact	4,045 (26.8%)	has semantic memory	11,468 (71.6%)
vacant expression	3,662 (24.2%)	benefits from cues	9,538 (59.5%)
smiling	857 (5.7%)	has recent episodic memory	9,160 (57.2%)
others	959 (6.3%)	retrieval deficit	6,328 (39.5%)

Corpus statistics
2,709 dialogues (avg. 5.90 turns; median 6; P25/P75 5–7; min/max 3–8)
Avg. action labels / dialogue: 20.19

Table 1: Statistics of DemMA-Dialogue.

4 Chain-of-Thought distillation agent training

4.1 Reasoning Annotation

To distill high-quality dialogue generation into a single model, we construct turn-level *reasoning annotations* as intermediate supervision. Each annotation makes explicit the reasoning trace underlying a dialogue turn, enabling the model to reproduce comparable utterance and action decisions at inference time without executing the full agentic generation pipeline. For each turn t, we generate a reasoning annotation formally,

$$E_t = \mathrm{Reason}\Big(\mathcal{P}, \hat{M}, C_t;\, U_t, A_t\Big), \tag{7}$$

Each E_t includes: (i) dialogue-state analysis, (ii) caregiver intent, (iii) memory accessibility, (iv) emotion inference with ICF-b126 justification, and (v) action decision with clinical rationale. We use E_t as intermediate supervision to learn clinically consistent reason–speak–act behavior. A concrete example of the turn-level reasoning annotation is provided in Appendix C.

4.2 Multi-Task Supervised Fine-Tuning

To jointly model planner rationales, patient utterances, and multimodal action labels within a single forward pass, we adopt a multi task SFT objective. We optimize a weighted sum of (i) masked next-token prediction losses for the `[PLAN]` and `[SPEAK]` segments and (ii) a multi-label action classification loss:

$$\mathcal{L}_{\text{total}} = w_p \mathcal{L}_{\text{planner}} + w_u \mathcal{L}_{\text{utterance}} + w_a \mathcal{L}_{\text{action}}. \quad (8)$$

To reduce interference between reasoning and surface-form generation, we apply segment specific token masking (with index sets T_p and T_u) so that planner and utterance supervision is decoupled at the token level:

$$\mathcal{L}_{\text{planner}} = -\sum_t \mathbf{1}(t \in T_p) \log P_\theta\big(y_t^p \mid y_{<t}^p, x\big), \quad (9)$$

$$\mathcal{L}_{\text{utterance}} = -\sum_t \mathbf{1}(t \in T_u) \log P_\theta\big(y_t^u \mid y_{<t}^u, x\big). \quad (10)$$

Action labels are predicted by an auxiliary multi-label head and optimized jointly via $\mathcal{L}_{\text{action}}$, encouraging consistent reasoning-to-action behaviors during generation.

Multi-Label Action Learning. Action prediction is formulated as a multi-label classification task over behavioral cues such as movement, facial expression, and vocal properties. We use a focal-modulated loss:

$$L_{\text{action}} = \sum_k [-a_k \log \sigma(z_k) - (1 - a_k) \log(1 - \sigma(z_k))] \, (1 - p_t)^2 \quad (11)$$

where a_k is the ground-truth label and z_k is the predicted logit. The focal term emphasizes low-confidence predictions, improving robustness under sparse action distributions.

Integrated Learning. By training all components jointly in a single forward pass, the model learns dependencies between memory impairment, reasoning, language generation, and multimodal actions. This integrated optimization enables DemMA to produce coherent reasoning traces, realistic patient dialogue, and aligned action labels within a unified model, without separate modules or post-processing.

5 Experimental Results and Analysis

5.1 Experiment Settings

We fine-tune DemMA on a dementia dialogue corpus annotated with planner, utterance, and action labels, using an 85%/15% train/validation split with a fixed random seed. The base model is Qwen3-8B. Training is conducted for up to 5 epochs with early stopping using AdamW 8bit optimization with a learning rate of 5×10^{-6}. All training and experiments are performed with mixed-precision training on 8 NVIDIA H100 GPUs. The code is available at: `https://github.com/RainieLLM/DemMA`.

5.2 Evaluation Metrics

We evaluate generated dialogues across seven complementary dimensions. Each dimension metric is assessed by LLM evaluators, clinical experts, students, enabling a multi-perspective evaluation spanning medical, linguistic, and behavioral criteria.

Personality Consistency. Measures whether the agent maintains a stable and coherent personality across multi-turn dialogue, including consistent identity, preferences and behavioral tendencies.

Language Naturalness. Measures whether utterances sound spontaneous and spoken-like rather than templated or model-polished.

Authenticity. Measures how realistic the dialogue is, as natural fluencies, repetition, and variability.

Medical Consistency. Measures alignment of language and behavior affect with the specified dementia subtype and severity.

Memory Rationality. Measures whether forgetting, repetition, and cue responses follow the defined memory profile.

Emotional Reasonableness. Measures whether emotion shifts are context-driven and gradual rather than abrupt or random.

Action Alignment. Measures whether nonverbal actions are plausible and consistent with the utterance and clinical profile.

5.2.1 Automated Evaluation

To leverage LLMs' capabilities as a judge for evaluating open-ended tasks, we use state-of-the-art LLMs, in- cluding GPT-5.2-pro, Gemini-2.5-pro

Table 2: LLM-based judgments and Expert evaluation on patient simulation fidelity and multi-turn dialogue quality.

Method	Evaluator	Patient Simulation Fidelity						Multi-turn Dialogue Quality		
		Auth.	Med.	Mem.	Emo.	Act.	Avg.	Pers.	Lang.	Avg.
Vanilla	GPT-5.2-pro	1.50	2.39	2.28	3.17	0.72	2.01	2.16	1.50	1.83
	Gemini-2.5-pro	1.60	2.00	2.00	2.60	1.00	1.84	1.80	1.60	1.70
	Qwen3-32B	1.58	2.15	1.57	2.10	0.95	1.67	2.05	2.95	2.50
	Expert Evaluation	1.50	2.00	2.00	2.50	1.00	1.80	2.00	1.50	1.75
Clinical-Profile Prompt	GPT-5.2-pro	2.14	2.11	2.78	3.56	1.31	2.38	2.33	1.67	2.00
	Gemini-2.5-pro	2.26	2.33	1.78	2.78	1.22	2.07	2.22	1.56	1.89
	Qwen3-32B	2.21	2.30	1.65	2.25	1.35	1.95	2.30	3.15	2.73
	Expert Evaluation	2.00	2.50	2.00	3.00	1.50	2.20	2.50	2.00	2.25
SFT-Utterance	GPT-5.2-pro	2.22	2.11	2.78	4.00	3.85	2.99	2.33	2.89	2.61
	Gemini-2.5-pro	2.33	2.44	3.11	3.33	3.78	3.00	2.31	3.22	2.76
	Qwen3-32B	2.30	2.35	3.10	3.00	3.75	2.90	2.40	3.65	3.03
	Expert Evaluation	2.50	2.50	3.00	3.50	4.00	3.10	2.50	3.00	2.75
DemMA	GPT-5.2-pro	3.78	4.33	4.44	4.89	4.00	4.29	4.11	3.78	3.95
	Gemini-2.5-pro	4.06	4.44	4.12	4.75	3.75	4.22	4.44	4.56	4.50
	Qwen3-32B	4.00	4.42	4.25	4.15	3.95	4.15	4.35	4.75	4.55
	Expert Evaluation	3.50	4.00	3.50	4.00	3.50	3.70	4.00	3.50	3.75

and Qwen3-32B models as evaluators. We design seven task-specific LLM-judge prompts, each corresponding to one of the evaluation dimensions above. Judges return a scalar rating (0–5) and a short justification. All judge prompts are provided in the Appendix E.

5.2.2 Human Evaluation

For dataset annotation, we recruited three dementia experts, each with over ten years of clinical experience in dementia diagnosis and patient care. For evaluation, we additionally recruited five experts with comparable clinical backgrounds and four medical students with formal training in neurology or geriatrics.

5.3 Baselines

We compare DEMMA with three baselines that progressively add dementia-specific conditioning and training signals: (i) **Vanilla**, using only background profile without dementia cues; (ii) **Clinical-Profile Prompt**, which adds a pathology-specific persona describing subtype symptoms; (iii) **SFT-Utterance**, a supervised model trained to generate utterances without planner reasoning or action supervision.

6 Results and Analysis

In this section, we present comprehensive experiments, aiming to address the following Research Questions (RQs): **RQ1: Simulation fidelity.** Can DEMMA achieve high-fidelity simulation of dementia patients in different subtypes?

RQ2: Dialogue quality. Can DEMMA sustain high-quality multi-turn interactions, maintaining persona consistency and natural conversational interactions?

RQ3: Validity of LLM-based evaluation. Can LLM-based judges reliably evaluate dementia simulation quality? Do their ratings align with assessments from human experts?

RQ4: Scaling effect of the training data Does DemMA-Dialogue provide sufficient diversity and coverage for effective multi-turn agent training?

RQ5: Educational Effectiveness. Can DEMMA support medical training (e.g., medical students or trainee caregivers) by helping them better understand dementia clinical manifestations?

6.1 Simulation Results

We evaluated each method using LLM judge in all dementia personas using the same seven metrics. To obtain a single overall score for a method, we average metric ratings within each persona across the nine personas. The automated and human evaluation results are shown in Table 2.

To answer **RQ1**, we focus on authenticity, medical consistency, memory rationality, emotional reasonableness and action alignment evaluations. **DemMA consistently achieves the highest overall patient simulation fidelity across all evaluators.** Across GPT-5.2, Gemini-2.5, and Qwen3-32B judges, DemMA attains the top average scores (4.1–4.3), substantially outperforming all baselines. This consistency across evaluators indicates that DemMA's gains reflect a consensus improvement rather than evaluator-specific bias. **DemMA's improvement is comprehensive, spanning all key simulation dimensions.** DemMA achieves the highest or tied-highest scores in all simulation fidelity metrics without exhibiting a clear weakness. In contrast, other methods show pronounced imbalances, performing adequately on language while underperforming on clinically grounded or behavioral dimensions. **The progression from**

prompt-based methods to SFT and ultimately to DemMA reflects a qualitative shift rather than a linear improvement. Other approaches remain around score of 2.0 to 3.0, while DemMA surpasses the 4.0 threshold simultaneously across all evaluation dimensions. This suggests that DemMA introduces a fundamentally different level of modeling capacity, enabling a transition from surface-level conversational plausibility to high-fidelity, clinically grounded patient simulation.

To answer **RQ2**, we focus on persona consistency and language naturalness evaluations. **DemMA consistently achieves the highest average multi-turn dialogue quality across all evaluators.** This result indicates that DemMA is not merely marginally usable in extended interactions, but instead reliably operates in a high-quality regime that approaches human-level performance in multi-turn dialogue. **DemMA exhibits a substantial advantage in persona consistency, demonstrating stronger cross-turn stability than all baselines.** In the Persona Consistency (Pers.) dimension, DemMA significantly outperforms competing methods, suggesting that it more effectively maintains stable persona attributes and behavioral patterns across dialogue turns. **DemMA maintains persona consistency without sacrificing language naturalness.** This finding indicates that DemMA does not rely on rigid or templated responses to preserve consistency, but instead sustains long-term persona stability within fluent and natural conversational exchanges.

6.2 Human Evaluation and LLM judgments for Dementia Subtypes

To answer **RQ3**, We conclude the human and LLM judgments evaluation for subtypes in Table 3 and Figure 2. For Patient Simulation Fidelity, LLM judges score DemMA at 4.15–4.29, while experts assign 3.70; for Multi-turn Dialogue Quality, LLM judges give 3.95–4.55 versus 3.75 by experts. This indicates that **LLM judges are reliable for comparative evaluation but tend to systematically overestimate absolute quality, particularly for stronger models.**

Figure 2 and Table 3 show consistent trends but different emphases. Both indicate that DemMA outperforms baselines across subtypes. However, LLM radar plots amplify performance gaps, while experts provide more conservative, compressed scores (overall 3.5–3.9), and consistently identify action alignment as the main remaining weakness.

Table 3: Expert evaluation across dementia subtypes (mean of 5 raters).

	Auth.	Med.	Mem.	Emo.	Act.	Avg.	Pers.	Lang.
AD-early	3.0	3.2	3.4	3.8	3.0	3.4	3.8	3.6
AD-mid/late	3.0	3.6	3.6	3.4	4.0	3.7	4.0	4.0
DLB	3.6	4.4	3.4	4.0	3.2	3.7	4.2	3.0
PDD	3.4	3.2	3.4	3.8	3.0	3.5	3.6	3.6
VaD	3.8	4.2	3.4	4.4	3.6	3.9	4.2	3.4
FTD-bv	3.8	3.2	3.4	4.0	3.0	3.6	3.8	3.6
nfvPPA	4.0	4.2	4.4	4.0	3.0	3.9	4.2	4.0
lvPPA	3.4	3.8	4.0	3.6	3.6	3.7	4.0	3.6
svPPA	3.6	3.8	3.4	3.2	3.6	3.7	4.6	3.0
Average	**3.6**	**3.7**	**3.6**	**3.8**	**3.3**	**3.7**	**4.0**	**3.5**

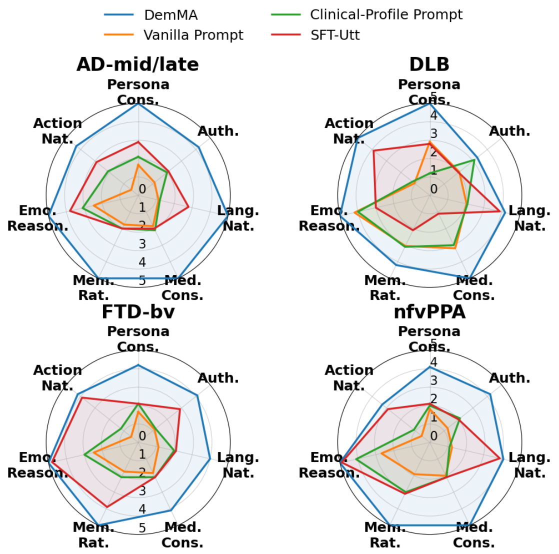

Figure 2: LLM judgment (GPT-5.2-pro) Performance across four dementia subtypes.

LLM judges are effective for visualizing relative improvements and subtype profiles but expert ratings offer a clinically grounded assessment.

6.3 Scaling Effect of training dataset

To answer **RQ4**, we trained Qwen3-8B model using different scale dataset. Performance is illustrated in Figure 3. **Longer dialogues mainly improve multi-turn capability, but the gains quickly show diminishing returns.** Increasing the number of turns per dialogue: The largest improvement occurs when moving from short to medium-length dialogues, after which performance tends to plateau. Longer contexts particularly boost Emotional Reasonableness and also stabilize persona consistency, language naturalness, and authenticity, since these dimensions depend on cross-turn coherence and gradual evolution. **Increasing topic diversity strongly improves Medical Consistency and Memory Rationality.** The model must repeatedly observe how the same pathology manifests across different situations. In contrast, Action Alignment and Authenticity change relatively little, suggesting they rely more on action-label supervision/alignment quality or stylistic cues than on topic coverage alone.

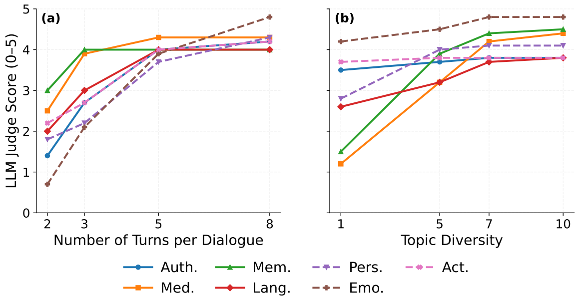

Figure 3: Scaling Effect of training dataset.

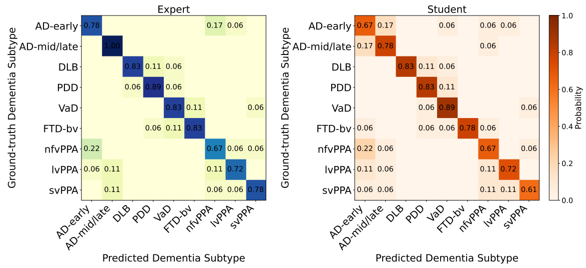

Figure 4: Confusion matrices for dementia subtype identification (experts vs. students) across nine subtypes.

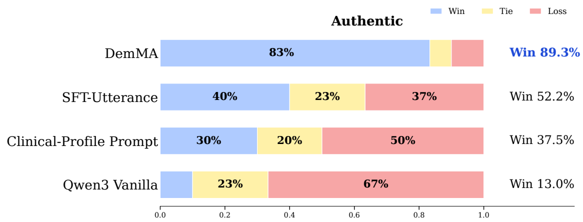

Figure 5: Pairwise win/tie/loss on AUTHENTIC. Win rate is computed as $\text{Win}/(\text{Win} + \text{Loss})$, excluding ties.

6.4 Educational effectiveness and feedback

To answer **RQ5**, we conducted a blinded subtype identification study with medical students and dementia experts. Students first studied 180 multi-turn DemMA-generated dialogues spanning different dementia subtypes, and then completed a blinded test of 45 dialogue cases. Participants were shown randomized DemMA-generated multi-turn dialogues and asked to predict the dementia subtype for each case. Results are summarized as confusion matrices in Fig.4. **Students can recognize dementia subtype-specific clinical differences via DemMA.** Both matrices exhibit a strong diagonal, showing that medical students can correctly identify most dementia subtypes from DemMA dialogues, while their accuracy is generally lower than experts. DemMA provides distinct, learnable clinical cues that differentiate subtypes. **Student errors mirror expert confusion in clinically overlapping subtypes.** DemMA can separate subtypes when cues are clear and reproduces realistic boundary cases. The pairwise evaluation in Figure 5 shows that DemMA is overwhelmingly preferred for authenticity. DemMA achieves an 89.3% win rate, demonstrating a clear and robust advantage in perceived dementia realism. Even medical students without prior formal training in dementia care or professional caregiving are able to learn subtype-specific cues from DemMA dialogues and achieve above-chance performance in subtype identification under blinded evaluation.

7 Related Work

7.1 CoT Distillation Training

Chain-of-Thought (CoT) distillation has been proposed as an effective way to retain multi-step reasoning capabilities of large language models while reducing inference-time overhead (Wei et al., 2022). Early studies show that exposing models to explicit reasoning traces improves performance on complex tasks, but directly generating such traces during inference is computationally expensive. To address this limitation, subsequent work distills teacher-generated rationales or planning traces into student models via supervised training, enabling reasoning behaviors to be implicitly internalized without explicit CoT generation at inference (Wu et al., 2024; Magister et al., 2023). This idea has further been extended to planner–executor and agent-based frameworks, where multi-step decision-making processes are distilled into a single model to maintain long-horizon coherence with reduced latency (Nye et al., 2021; Yao et al., 2023). In contrast to prior CoT distillation methods that primarily focus on textual reasoning, our work distills clinically grounded reasoning signals, including memory analysis, dialogue planning, and action decisions, into a unified model for efficient dementia dialogue generation.

7.2 LLM-based Patient Simulation for Clinical Training

LLMs have been explored for dementia-oriented conversational support and caregiver assistance, with recurring concerns around safety and ethics (Treder et al., 2024; Song et al., 2025; Nyamathi et al., 2024). Recent simulated-patient systems such as PATIENT-Ψ and PatientSim ground LLM behavior in structured personas and psychological models to produce plausible multi-turn clinical interactions (Wang et al., 2024; Kyung et al., 2025). However, these frameworks largely assume cognitively stable patients and optimize for persona consistency, which under-represents dementia-specific phenomena such as repetition, contradictory recall, and temporal disorientation (Lyketsos et al., 2011; Zhao et al., 2016; Aalten et al., 2008). In contrast, DemMA models dementia as a progressing cognitive state and generates clinically grounded breakdown patterns that evolve across turns.

8 Conclusion

In this paper, we presented DemMA, a clinically grounded multi-turn dementia dialogue agent that simulates realistic behaviors through joint modeling of language and action generation. DemMA integrates persona-driven dementia subtypes and disease stages with a structured pipeline, enabling controlled simulation of cognitive decline, emotional variation, and nonverbal cues. We introduce a distillation-style strategy that internalizes planning, reasoning, dialogue, and action prediction into a single low-latency model. Extensive evaluations show that DemMA outperforms baselines in persona fidelity, clinical validity, and educational effectiveness, offering a scalable and ethically sound solution for high-fidelity simulation avoiding sensitive real-world data.

Limitations

Despite strong results, DemMA has several limitations. First, the dataset is fully synthetic. Although DemMA-Dialogue avoids sensitive patient data and is validated by domain experts, synthetic interactions may omit rare clinical behaviors and idiosyncratic caregiver–patient dynamics, potentially limiting generalization to real-world settings. Second, clinical grounding relies on textual abstractions. Our action labels serve as interpretable proxies for multimodal behaviors but cannot fully capture fine-grained audiovisual or sensorimotor cues, constraining fidelity for embodied or multimodal applications. Third, persona and memory modeling are simplified. While we cover multiple dementia subtypes and memory accessibility patterns, the framework does not explicitly model long-term disease progression, medication effects, or evolving caregiver strategies. Finally, evaluation depends partly on LLM-based judges. Such judges may exhibit known biases and may not fully capture clinically grounded correctness (Yao et al., 2023). We therefore treat automated scores as complementary to expert and human evaluation rather than definitive clinical assessments.

Ethical Considerations

DemMA is designed to address ethical and legal barriers in dementia research by generating fully synthetic dialogue data rather than relying on real patient records. This substantially reduces risks related to privacy, consent, and exposure of sensitive health information. Nevertheless, synthetic data may still resemble real individuals or scenarios, and dataset release should include screening for personal identifiers and clear usage guidelines. The framework is intended for research and educational purposes, such as caregiver training and model development, and is not suitable for clinical diagnosis or decision-making. Without appropriate guardrails, simulated dementia dialogue could be misused to generate persuasive but medically incorrect content. Any downstream deployment should therefore include explicit scope restrictions, disclaimers, and human oversight. Simulating cognitive impairment also raises concerns about representational harms and stereotyping. Although personas are clinically informed and validated, model outputs may oversimplify or bias portrayals of dementia. We encourage future audits across demographic and interactional dimensions to assess and mitigate such risks. Finally, automated evaluation relies partly on LLM-based judges, which are known to exhibit biases and limitations. For this reason, we complement automated metrics with expert and human evaluations and emphasize that reported scores should not be interpreted as clinical assessments.

References

Pauline Aalten, Frans R. J. Verhey, Marina Boziki, Andrea Brugnolo, Roger Bullock, Eleanor Jane Byrne, Vincent Camus, Miriam Caputo, Debby Collins, Peter Paul

De Deyn, Kazi Elina, Giovanni Frisoni, Clive Holmes, Catherine Hurt, Anna Marriott, Patrizia Mecocci, Flavio Nobili, Pierre Jean Ousset, Emma Reynish, and 4 others. 2008. Consistency of neuropsychiatric syndromes across dementias: results from the European Alzheimer Disease Consortium. part II. *Dementia and Geriatric Cognitive Disorders*, 25(1):1–8.

Dag Aarsland, Jan Petter Larsen, Knut Karlsen, Ngoc Lim, and Einar Tandberg. 2007. Neuropsychiatric symptoms in Parkinson's disease with dementia: frequency, profile and associated caregiver stress. *International Journal of Geriatric Psychiatry*, 22(8):784–792. PMID: 16820421.

Dag Aarsland, Linda Marsh, and Anette Schrag. 2009. Neuropsychiatric symptoms in Parkinson's disease. *Movement Disorders*, 24(15):2175–2186.

C. G. Ballard, D. Aarsland, and I. McKeith. 2002. Fluctuations in attention: PD dementia vs DLB with parkinsonism. *Neurology*, 59(11):1714–1720. PMID: 12473758.

Sarah J. Banks and Sandra Weintraub. 2008. Neuropsychiatric symptoms in behavioral variant frontotemporal dementia and primary progressive aphasia. *Journal of Geriatric Psychiatry and Neurology*, 21(2):133–141.

Emma N. Bender, Marie Y. Savundranayagam, Laura Murray, and J. B. Orange. 2022. Supportive strategies for nonverbal communication with persons living with dementia: A scoping review. *International Journal of Nursing Studies*, 136:104365.

Jorunn Bjoerke-Bertheussen, Dag Aarsland, and Clive Ballard. 2012. Neuropsychiatric symptoms in mild dementia with lewy bodies and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 34(1):1–8.

Valeria Calsolaro, Grazia D. Femminella, Sara Rogani, Chibuzor Okoye, Giuseppe Rengo, and Fabio Monzani. 2021. Behavioral and psychological symptoms in dementia (BPSD) and the use of antipsychotics. *Pharmaceuticals (Basel)*, 14(3):246. PMID: 33803277.

Athanasios Chatzikostopoulos, Despina Moraitou, Vasileios Papaliagkas, and Magda Tsolaki. 2025. Mapping the neuropsychiatric symptoms in Alzheimer's disease using biomarkers, cognitive abilities, and personality traits: A systematic review. *Diagnostics (Basel)*, 15(9):1082. PMID: 40309371.

Ming-Jang Chiu, Tsui-Fen Chen, Ping-Keung Yip, May-Szu Hua, and Li-Yu Tang. 2006. Behavioral and psychologic symptoms in different types of dementia. *Journal of the Formosan Medical Association*. PMID: 16877235.

R. Convery, S. Mead, and Jonathan D. Rohrer. 2019. Clinical, genetic and neuroimaging features of frontotemporal dementia. *Neuropathology and Applied Neurobiology*, 45(1):6–18.

Jeffrey L. Cummings, Michael Mega, Kathy Gray, Susan Rosenberg-Thompson, David A. Carusi, and Jeffrey Gornbein. 1994. The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 44(12):2308–2314.

Bruno Dubois, Howard H. Feldman, and Claudia Jacova. 2014. Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *The Lancet Neurology*, 13(6):614–629.

Murat Emre, Dag Aarsland, and Richard Brown. 2007. Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Movement Disorders*, 22(12):1689–1707.

Paul J. Eslinger, Peachie Moore, and Vanessa Troiani. 2007. Oops! resolving social dilemmas in frontotemporal dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(5):457–460.

Dominic H. Ffytche and Dag Aarsland. 2017. Psychosis in Parkinson's disease. In *International Review of Neurobiology*, volume 133, pages 585–622. Elsevier.

Maria Luisa Gorno-Tempini, Argye E. Hillis, Sandra Weintraub, Andrew Kertesz, Mario Mendez, Stefano F. Cappa, Jennifer M. Ogar, Jonathan D. Rohrer, Steven Black, Bradley F. Boeve, Facundo Manes, Nina F. Dronkers, Rik Vandenberghe, Katya Rascovsky, Karalyn Patterson, Bruce L. Miller, David S. Knopman, John R. Hodges, M. Marsel Mesulam, and Murray Grossman. 2011. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014.

N. L. Graham, T. Emery, and J. R. Hodges. 2004. Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):61–71. PMID: 14707310.

Murray Grossman. 2012. The non-fluent/agrammatic variant of primary progressive aphasia. *The Lancet Neurology*, 11(6):545–555.

Masahiro Hashimoto, Yuka Yatabe, and Takuya Ishikawa. 2015. Relationship between dementia severity and behavioral and psychological symptoms of dementia in dementia with lewy bodies and Alzheimer's disease patients. *Dementia and Geriatric Cognitive Disorders*, 40(1–2):45–52. PMID: 26195980.

Nobutsugu Hirono and Jeffrey L. Cummings. 1999. Neuropsychiatric aspects of dementia with lewy bodies. *Current Psychiatry Reports*, 1(1):85–92.

John R. Hodges and Karalyn Patterson. 2007. Semantic dementia: A unique clinicopathological syndrome. *The Lancet Neurology*, 6(11):1004–1014.

M Hornberger, O Piguet, AJ Graham, PJ Nestor, and JR Hodges. 2010a. How preserved is episodic memory in behavioral variant frontotemporal dementia? *Neurology*, 74(6):472–479.

Michael Hornberger, Olivier Piguet, N. L. Graham, Peter J. Nestor, and John R. Hodges. 2010b. How preserved is episodic memory in behavioral variant frontotemporal dementia? *Neurology*. PMID: 20142613.

Muireann Irish, Olivier Piguet, John R. Hodges, and Michael Hornberger. 2014. Common and unique gray matter correlates of episodic memory dysfunction in frontotemporal dementia and Alzheimer's disease. *Human Brain Mapping*, 35(4):1422–1435.

Jr. Jack, Clifford R., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Brenda Dunn, Stephen B. Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Elaine Liu, Jose L. Molinuevo, Thomas Montine, Michael Phelps, Kathryn P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, and Reisa Sperling. 2018. NIA–AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562.

Cheonkam Jeong, Jessica Liao, Audrey Lu, Yutong Song, Christopher Rashidian, Donna Krogh, Erik Krogh, Mahkameh Rasouli, Jung-Ah Lee, Nikil Dutt, and 1 others. 2026. Dementiabank-emotion: A multi-rater emotion annotation corpus for alzheimer's disease speech (version 1.0). *arXiv preprint arXiv:2602.04247*.

Rebecca A. Johnson and Jason Karlawish. 2015. A review of ethical issues in dementia. *International Psychogeriatrics*, 27(10):1635–1647.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. 2025. Patientsim: A persona-driven simulator for realistic doctor–patient interactions. arXiv preprint arXiv:2505.17818. Version 1.0.0; also released via PhysioNet.

Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. 2017. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55.

Christian E. Leyton, S. Hsieh, and E. Mioshi. 2011. Subtypes of primary progressive aphasia: Application of the international consensus criteria and validation using beta-amyloid imaging. *Brain*, 134(10):3030–3043.

Cristian E. Leyton, Alexandra K. Britton, John R. Hodges, Glenda M. Halliday, and Jill J. Kril. 2016a. Distinctive pathological mechanisms involved in primary progressive aphasias. *Neurobiology of Aging*, 38:82–92. PMID: 26827646.

Cristian E. Leyton, Alexandra K. Britton, John R. Hodges, Glenda M. Halliday, and Jill J. Kril. 2016b. Distinctive pathological mechanisms involved in primary progressive aphasias. *Neurobiology of Aging*, 38:82–92.

Tau Ming Liew. 2021. Neuropsychiatric symptoms in early stage of Alzheimer's and non-Alzheimer's dementia, and the risk of progression to severe dementia. *Age and Ageing*, 50(5):1709–1718.

Irene Litvan, Jennifer G. Goldman, and Alexander I. Tröster. 2012. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement disorder society task force guidelines. *Movement Disorders*, 27(3):349–356.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2023. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688. Published in ICLR 2024.

Gill Livingston, Andrew Sommerlad, Vasiliki Orgeta, Sergi G Costafreda, Jonathan Huntley, David Ames, Clive Ballard, Sube Banerjee, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, Nick Fox, Laura N Gitlin, Robert Howard, Helen C Kales, Eric B Larson, Karen Ritchie, Kenneth Rockwood, Elizabeth L Sampson, and 5 others. 2017. Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734.

Constantine G. Lyketsos, Maria C. Carrillo, John M. Ryan, Ara S. Khachaturian, Paula Trzepacz, Joanna Amatniek, Daniel S. Miller, and Gregory S. Smith. 2011. Neuropsychiatric symptoms in Alzheimer's disease. *Alzheimer's & Dementia*, 7(5):532–539.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781.

Elie Matar, James M. Shine, Glenda M. Halliday, and Simon J. G. Lewis. 2020. Cognitive fluctuations in Lewy body dementia: Toward a pathophysiological framework. *Brain*.

Ian G. McKeith, Bradley F. Boeve, and Dennis W. Dickson. 2017. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB consortium. *Neurology*, 89(1):88–100.

Guy M. McKhann, David S. Knopman, and Howard Chertkow. 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269.

M.-Marsel Mesulam, Sandra Weintraub, Emily J. Rogalski, Christina Wieneke, Changiz Geula, and Eileen H. Bigio. 2014a. Asymmetry and heterogeneity of Alzheimer's and frontotemporal pathology in primary progressive aphasia. *Brain*, 137(4):1176–1192. PMID: 24574501.

Marsel Mesulam, Ann Wicklund, and Nancy Johnson. 2008. Alzheimer and frontotemporal pathology in subsets of primary progressive aphasia. *Annals of Neurology*, 63(6):709–719.

Marsel M. Mesulam, Emily J. Rogalski, and Chris Wieneke. 2014b. Primary progressive aphasia and the evolving neurology of the language network. *Nature Reviews Neurology*, 10(10):554–569.

Adeline Nyamathi, Nikil Dutt, Jung-Ah Lee, Amir M Rahmani, Mahkameh Rasouli, Donna Krogh, Erik Krogh, David Sultzer, Humayun Rashid, Hamza Liaqat, and 1 others. 2024. Establishing the foundations of emotional intelligence in care companion robots to mitigate agitation among high-risk patients with dementia: protocol for an empathetic patient-robot interaction study. *JMIR Research Protocols*, 13(1):e55761.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114.

John T. O'Brien and Alan Thomas. 2015. Vascular dementia. *The Lancet*, 386(10004):1698–1706.

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. 2007. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987.

Olivier Piguet, Michael Hornberger, Eneida Mioshi, and John R. Hodges. 2011. Behavioural-variant frontotemporal dementia: Diagnosis, clinical staging, and management. *The Lancet Neurology*, 10(2):162–172.

Andrew Pless, Destany Ware, Shalini Saggu, Hasibur Rehman, John Morgan, and Qin Wang. 2023. Understanding neuropsychiatric symptoms in Alzheimer's disease: challenges and advances in diagnosis and treatment. *Frontiers in Neuroscience*, 17:1263771. PMID: 37732300.

Angelina J. Polsinelli, Sierah Johnson, Adele Crouch, Kathleen A. Lane, Alex Pena-Garcia, Dustin B. Hammers, Sophia Wang, Sujuan Gao, and Liana G. Apostolova. 2024. Neuropsychiatric symptom burden in early-onset and late-onset Alzheimer's disease as a function of age. *Alzheimer's & Dementia*. PMID: 38958543.

Katya Rascovsky, John R. Hodges, and David Knopman. 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9):2456–2477.

Barry Reisberg, Steven H. Ferris, Mony J. de Leon, Anatole Kluger, Eve Franssen, Joseph Borenstein, and Raquel Alba. 1989. The stage specific temporal course of Alzheimer's disease: Functional and behavioral concomitants based upon cross-sectional and longitudinal observation. *Psychopharmacology Bulletin*. PMID: 2690101.

Howard J. Rosen, Maria Luisa Gorno-Tempini, William P. Goldman, Richard J. Perry, Norbert Schuff, Michael Weiner, Robert Feiwell, Joel H. Kramer, and Bruce L. Miller. 2002. Patterns of brain atrophy in frontotemporal dementia and semantic dementia. *Neurology*, 58(6):916–921. PMID: 11805245.

Perminder Sachdev, Raj Kalaria, John O'Brien, Ingmar Skoog, Suvarna Alladi, Sandra E. Black, Deborah Blacker, Dan G. Blazer, Christopher Chen, Helena Chui, Mary Ganguli, Kurt Jellinger, Dilip V. Jeste, Florence Pasquier, Jane Paulsen, Niels Prins, Kenneth Rockwood, Gustavo Roman, Philip Scheltens, and International Society for Vascular Behavioral and Cognitive Disorders. 2014. Diagnostic criteria for vascular cognitive disorders: a VASCOG statement. *Alzheimer Disease & Associated Disorders*, 28(3):206–218.

David P. Salmon and Mark W. Bondi. 2009. Neuropsychological assessment of dementia. *Annual Review of Psychology*, 60:257–282.

H. Seelaar, Jonathan D. Rohrer, Yolande A. L. Pijnenburg, Nick C. Fox, and John C. van Swieten. 2011. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. *The Lancet Neurology*. PMID: 20971753.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Yutong Song, Chenhan Lyu, Pengfei Zhang, Sabine Brunswicker, Nikil Dutt, and Amir Rahmani. 2025. Dementia-plan: An agent-based framework for multi-knowledge graph retrieval-augmented generation in dementia care. *arXiv preprint arXiv:2503.20950*.

Susanne S. Staekenborg, Tian Su, Elisabeth C. W. van Straaten, Richard Lane, Philip Scheltens, and Frederik Barkhof. 2010. Behavioural and psychological symptoms in vascular dementia. *International Journal of Geriatric Psychiatry*, 25(6):547–555.

Szabolcs Szatmari, Ben Min-Woo Illigens, and Timo Siepmann. 2017. Neuropsychiatric symptoms in untreated Parkinson's disease. *Neuropsychiatric Disease and Treatment*, 13:815–826.

Antoine Thomas, Laura Bonanni, Francesca Gambi, Aurelio Di Iorio, and Marco Onofrj. 2010. Pathological gambling in Parkinson disease is reduced by amantadine. *Annals of Neurology*, 68(3):400–404. PMID: 20737574.

Chan Tiel, Felipe Kenji Sudo, Gilberto Sousa Alves, Letice Ericeira-Valente, Denise Madeira Moreira, Jerson Laks, and Eliasz Engelhardt. 2015. Neuropsychiatric symptoms in vascular cognitive impairment: A systematic review. *Dementia & Neuropsychologia*, 9(3):230–236.

Matthias S. Treder, Sojin Lee, and Kamen A. Tsvetanov. 2024. Introduction to large language models (llms) for dementia care and research. *Frontiers in Dementia*, 3:1385303.

Christine Tremblay, Javaid Shakir, and Alireza Atri. 2025. Associations between neuropsychiatric symptoms and pathology in clinicopathologically defined

Alzheimer's disease, Alzheimer's disease with lewy bodies, and dementia with lewy bodies. *Journal of Alzheimer's Disease*, 104(3):933–942. PMID: 40084663.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. PATIENT-Ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, USA. Association for Computational Linguistics.

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.

Daniel Weintraub, Juergen Koester, Marc N. Potenza, Andrew D. Siderowf, Mark Stacy, Valerie Voon, Jacqueline Whetteckey, Glen R. Wunderlich, and Anthony E. Lang. 2010a. Impulse control disorders in Parkinson disease: A cross-sectional study of 3090 patients. *Archives of Neurology*, 67(5):589–595.

Daniel Weintraub, Kristy Papay, and Andrew Siderowf. 2010b. Amantadine use associated with impulse control disorders in Parkinson disease in a cross-sectional study. *Annals of Neurology*, 68(6):963–968. PMID: 21194121.

World Health Organization. 2001. International classification of functioning, disability and health (ICF). World Health Organization.

Zongqian Wu, Baoduo Xu, Ruochen Cui, Mengmeng Zhan, Xiaofeng Zhu, and Lei Feng. 2024. Rethinking chain-of-thought from the perspective of self-training. *arXiv preprint arXiv:2412.10827*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Jonathan J. Young. 2019. Evidence-based pharmacological management and treatment of behavioral and psychological symptoms of dementia. *The American Journal of Psychiatry Residents' Journal*.

Qianhua Zhao, Lili Tan, Huaxiang Wang, Tian Jiang, Mengmeng Tan, and Lan Tan. 2016. The prevalence of neuropsychiatric symptoms in Alzheimer's disease: A systematic review and meta-analysis. *Journal of Affective Disorders*, 190:264–271.

Yaakov R. Zweig and James E. Galvin. 2014. Lewy body dementia: the impact on patients and caregivers. *Alzheimer's Research & Therapy*, 6(2):21.

Appendix

Appendix Contents

Table 4: Persona-level behavioral characteristics, neuropathological basis, and literature support

Persona	**Key behavioral characteristics (literature-derived)**	**Neuropathological basis**	**Literature support**
persona1 AD-early stage	• Apathy, depression, irritability, and anxiety are the most common neuropsychiatric symptoms • Insight is still present • Social withdrawal may occur (The literature describes them as "common" but no specific rates given)	Early limbic system impairment Hippocampus-cingulate gyrus Mild frontal lobe dysfunction	(Lyketsos et al., 2011; Zhao et al., 2016; Cummings et al., 1994; Liew, 2021; Chatzikostopoulos et al., 2025; Pless et al., 2023; Polsinelli et al., 2024)
persona2 AD-mid/late to late stage	• Apathy is the most prominent symptom • Increased frequency of agitation/aggression, delusions, and hallucinations • Disturbances of circadian rhythm • Complete loss of insight	Extensive neocortical damage Widespread involvement of frontal-limbic system Damage to circadian rhythm systems	(Aalten et al., 2008; Lyketsos et al., 2011; Pless et al., 2023; Calsolaro et al., 2021; Young, 2019)
persona3 VaD	• Depression is the most common symptom • Emotional incontinence (pathological laughing/crying) is characteristic in subcortical VaD • Apathy is common • Emotional fluctuations are greater than in AD	Subcortical vascular lesions Disruption of frontal-subcortical circuits Basal ganglia/brainstem lesions	(Staekenborg et al., 2010; Chiu et al., 2006; Tiel et al., 2015; Calsolaro et al., 2021; Young, 2019)
persona4 DLB	• Visual hallucinations are a core feature • Cognitive fluctuations are significant • Anxiety and depression are common • Apathy is present • Early memory is relatively preserved (vs AD) (The literature notes that visual hallucinations are a "core feature")	Neocortical Lewy bodies Cholinergic system damage Posterior cortical dysfunction Fluctuating dysfunction in the limbic system	(Bjoerke-Bertheussen et al., 2012; Zweig and Galvin, 2014; Hirono and Cummings, 1999; Hashimoto et al., 2015; Tremblay et al., 2025; Young, 2019)
persona5 PDD	• Apathy and depression are the most common • Impulse control disorders (ICDs) are significantly associated with dopamine agonist treatment • Visual hallucinations may occur • Anxiety is present (The literature emphasizes the association of ICDs with DA agonists)	Cortical Lewy bodies + substantia nigra degeneration Dopaminergic treatment → Overstimulation of the ventral striatum	(Aarsland et al., 2007; Weintraub et al., 2010a; Ffytche and Aarsland, 2017; Szatmari et al., 2017; Weintraub et al., 2010b; Thomas et al., 2010)
persona6 FTD-bv	• Disinhibition • Apathy/inertia • Loss of sympathy/empathy • Stereotyped/compulsive behaviors • Hyperorality/eating changes • Executive dysfunction (These are the six core features of the Rascovsky 2011 diagnostic criteria)	Severe atrophy of the orbitofrontal and medial prefrontal cortices → Breakdown of social cognition → Loss of behavioral regulation Anterior temporal lobe atrophy → Loss of empathy	(Rascovsky et al., 2011; Seelaar et al., 2011; Eslinger et al., 2007; Piguet et al., 2011; Banks and Weintraub, 2008)
persona7 nfvPPA	• Effortful, non-fluent speech • Agrammatism • Frustration and anxiety are visible • Social withdrawal (secondary to language deficits) • Episodic memory relatively preserved • Insight preserved (The diagnostic criteria emphasize "preserved episodic memory")	Left inferior frontal gyrus/insula (Broca's area) Limbic system/frontal cortex relatively preserved	(Gorno-Tempini et al., 2011; Grossman, 2012; Banks and Weintraub, 2008)
persona8 svPPA	• Selective impairment of semantic memory • Fluent but empty speech • Naming and word meaning deficits (anomia, loss of word meaning) • Episodic memory relatively preserved • Stereotyped behaviors may appear • Late-stage progression to FTD-bv-like behavior (The literature notes "preserved episodic memory" in early stage)	Anterior temporal lobe atrophy (semantic hub) Late-stage extension to the orbitofrontal cortex → Behavioral symptoms occur	(Gorno-Tempini et al., 2011; Hodges and Patterson, 2007; Rosen et al., 2002; Seelaar et al., 2011; Banks and Weintraub, 2008)
persona9 lvPPA	• Word-finding difficulty, frequent pauses • Severely impaired sentence repetition • Phonological working memory deficit • Anxiety and depression are visible • Commonly associated with AD pathology → memory declines as the disease progresses • Insight is preserved (The literature emphasizes "commonly associated with AD pathology")	Temporo-parietal junction + AD pathology Spread during disease to hippocampus and frontal cortex	(Gorno-Tempini et al., 2011; Mesulam et al., 2014a; Leyton et al., 2016a; Seelaar et al., 2011; Banks and Weintraub, 2008)

Table 5: ICF-b126 personality mapping for dementia personas

Persona	Extraversion (b1260)	Agreeableness (b1261)	Conscientiousness (b1262)	Emotional Stability (b1263)	Openness (b1264)	Optimism (b1265)	Self-confidence (b1266)	Integrity (b1267)
persona1 AD-early stage	Slightly ↓ (-) Beginning to withdraw socially Unwilling to participate in activities	Slightly ↓ (-) Irritability ↑ Decreased patience	Moderately ↓ (–) Poor organizational ability Unable to multitask Forgets commitments	Moderately ↓ (–) Anxiety Depression Worries about memory problems	Moderately ↓ (–) Rejects new things Fixated on familiar environments	Moderately ↓ (–) Pessimistic "I'm useless"	Moderately ↓ (–) Self-doubt Avoids social situations	Relatively preserved (0/-) Unless there is disinhibition
persona2 AD-middle to late stage	Severely ↓ (—) Almost no social response Passive and apathetic	Severely ↓ (—) Aggression may occur Resists care	Severely ↓ (—) Complete loss of self-care No planning ability	Severely ↓ (—) Agitation Delusions Day-night reversal	Severely ↓ (—) Rigid Refuses change	Severely ↓ (—) No emotional response	Severely ↓ (—) Self-awareness disintegrates	Severely ↓ (—) May steal food Inappropriate behavior
persona3 VaD	Moderately ↓ (–) Reduced social skills Lack of motivation	Moderately ↓ (–) Emotional lability Easily angered	Moderately ↓ (–) Executive dysfunction Slow processing speed	Severely ↓ (—) Emotional incontinence Easily cries and laughs Emotional lability	Moderately ↓ (–) Poor adaptability	Moderately ↓ (–) High incidence of depression (40-60%) Pessimistic	Slightly ↓ (-) Setbacks lead to self-deprecation	Slightly ↓ (-) Except with frontal lobe infarcts
persona4 DLB	Moderately ↓ (–) Visual hallucinations affect social interactions Suspiciousness and withdrawal	Moderately ↓ (–) Anxiety-related irritability	Slightly ↓ (-) Fluctuating attention affects executive function	Severely ↓ (—) Anxiety (40%) Fear related to visual hallucinations Emotional fluctuations	Slightly ↓ (-)	Moderately ↓ (–) Pessimistic reaction to hallucinations	Moderately ↓ (–) Hallucinations cause unease	Relatively preserved (0)
persona5 PDD	Moderately ↓ (–) Motor and cognitive dual impairment Social withdrawal	Slightly ↓ (-) Or pathologically ↑ Impulse control disorders (related to dopaminergic treatment)	Moderately ↓ (–) Difficulty planning/initiating	Moderately ↓ (–) Depression (40%) Apathy (50%) Anxiety	Slightly ↓ (-)	Moderately ↓ (–) Chronic disease burden	Moderately ↓ (–) Dual blow of motor and cognitive impairment	Slightly ↓ (-) Or pathologically ↓ Impulse control disorders: Gambling, hypersexuality Compulsive shopping
persona6 FTD-bv	Severely ↑ (—) Disinhibition Inappropriate social behavior Overly familiar	Severely ↓ (—) Complete loss of empathy Callousness, oppositional Aggressiveness	Severely ↓ (—) Impulsivity Lack of planning Irresponsibility	Severely ↓ (—) Complete loss of emotional regulation Irritable and impulsive	Severely ↓ (—) Rigid and stereotyped Perseverative behaviors/interests	Moderately ↓ (–) Or affectively flat	Slightly ↑ Lack of insight Overconfident	Severely ↓ (—) Antisocial behaviors Stealing, lying Sexual misconduct
persona7 nfvPPA	Moderately ↓ (–) Communication difficulties lead to Social avoidance	Slightly ↓ (-) Frustration leads to irritability	Slightly ↓ (-) Language deficits affect Task execution	Moderately ↓ (–) Frustration Anxiety Great pressure to find words	Slightly ↓ (-)	Moderately ↓ (–) Speech difficulties Lead to frustration	Moderately ↓ (–) Communication failures Lead to decreased self-confidence	Relatively preserved (0)
persona8 svPPA	Slightly ↓ (-) Moderate ↓ in late stage Early stage preserved Late stage similar to FTD	Moderately ↓ (–) Decreased empathy Self-centered	Slightly ↓ (-)	Slightly ↓ (-) Relatively stable in early stage	Moderately ↓ (–) Fixation on specific interests/activities	Slightly ↓ (-)	Slightly ↓ (-)	Slightly ↓ (-) Moderate ↓ in late stage Inappropriate behaviors may occur in late stage
persona9 lvPPA	Moderately ↓ (–) Word-finding difficulty High social stress	Slightly ↓ (-) Frustration leads to irritability	Moderately ↓ (–) Declines with AD pathology progression	Moderately ↓ (–) Anxiety when word-finding Depression Worsens as disease progresses	Slightly ↓ (-)	Moderately ↓ (–) Language disorder + AD pathology	Moderately ↓ (–) Communication failures + Memory decline	Relatively preserved (0/-)

Table 6: Memory-status templates for the nine dementia personas, summarizing severity, characteristic clinical profiles, pathophysiological basis, differential diagnostic features, and key references

Persona	Memory-loss severity	Clinical profile (detailed)	Pathophysiological basis	Differential diagnosis	Key references
persona1 AD-early	High (episodic)	**Impairment pattern:** • Severely impaired delayed recall with rapid forgetting over minutes • Immediate memory relatively preserved, including intact repetition • Encoding deficits predominate; cueing and recognition provide limited benefit • Early impairment in temporal orientation • Remote autobiographical memory relatively preserved in early stages • Procedural and implicit memory functions largely spared	Degeneration of the hippocampal–entorhinal system with amyloid pathology, resulting in impaired encoding and consolidation of new episodic information	**vs. VaD:** earlier and more prominent amnestic syndrome **vs. DLB:** absence of marked cognitive fluctuations **vs. bvFTD:** memory impairment predominates rather than behavioral disturbance	(McKhann et al., 2011; Dubois et al., 2014; Jack et al., 2018)
persona2 AD-mid/late	Very high (global)	**Impairment pattern:** • Complete loss of recent memory with failure to form new memories • Progressive retrograde amnesia with a temporal gradient • Severe semantic memory impairment affecting names and common objects • Global disorientation in time and place • Fragmentation of autobiographical memory • Prominent repetitive behaviors and misidentification in advanced stages	Near-complete hippocampal atrophy combined with widespread neocortical neurofibrillary pathology, leading to a global breakdown of memory storage and retrieval	Late-stage Alzheimer's disease typically exhibits the most severe and globally distributed memory impairment among dementia subtypes	(Reisberg et al., 1989; Salmon and Bondi, 2009)
persona3 VaD	Subcortical: mid Cortical/mixed: high	**Characteristic patterns:** • Predominant retrieval deficits with relatively preserved encoding in subcortical forms • Marked improvement with cueing or recognition • Patchy, lesion-dependent memory impairment • Slowed information processing and executive dysfunction contributing to secondary memory complaints • Stepwise or fluctuating clinical course	Disruption of frontal–subcortical circuits and white matter disease in subcortical VaD; hippocampal or temporal infarction in cortical variants producing encoding deficits	**vs. AD:** retrieval deficits with preserved cueing benefit **vs. other dementias:** presence of vascular risk factors and characteristic neuroimaging findings	(Sachdev et al., 2014; O'Brien and Thomas, 2015; Graham et al., 2004)
persona4 DLB	Low–mid (early preserved)	**Characteristic patterns:** • Relative preservation of episodic memory in early stages • Retrieval deficits exceeding encoding deficits • Disproportionate impairment of visuospatial memory • Marked attentional and cognitive fluctuations • Impaired prospective memory related to executive dysfunction	Neocortical Lewy body pathology with prominent cholinergic dysfunction, while medial temporal structures remain relatively preserved in early disease	**vs. AD:** preserved early memory with pronounced fluctuations and visuospatial deficits **vs. PDD:** dementia onset concurrent with or preceding parkinsonism	(McKeith et al., 2017; Matar et al., 2020; Ballard et al., 2002)
persona5 PDD	Mid	**Characteristic patterns:** • Memory profile similar to DLB but with slower progression • Executive dysfunction impairing retrieval strategies • Recognition performance exceeding free recall • Working-memory and visuospatial deficits • Dementia onset occurring at least one year after established Parkinson's disease	Cortical Lewy body deposition combined with nigrostriatal degeneration and mixed dopaminergic–cholinergic dysfunction, affecting frontally mediated retrieval processes	**vs. DLB:** Parkinsonism clearly precedes dementia **vs. AD:** relative preservation of recognition memory and prominent motor features	(Emre et al., 2007; Aarsland et al., 2009; Litvan et al., 2012)
persona6 bvFTD	Low (early preserved)	**Characteristic patterns:** • Relative preservation of episodic memory, particularly on recognition tasks • Impaired free recall driven by executive dysfunction • Deficits in source and prospective memory • Prominent behavioral and personality changes preceding memory complaints • Breakdown of social cognition exceeding amnestic features	Degeneration of frontal and anterior temporal regions with TDP-43 or tau pathology, while hippocampal structures remain relatively preserved	**vs. AD:** early behavioral syndrome with relatively intact memory **vs. svPPA:** core language abilities remain comparatively preserved	(Rascovsky et al., 2011; Irish et al., 2014; Hornberger et al., 2010b)
persona7 nfvPPA	Low	**Characteristic patterns:** • Episodic memory largely preserved • Apparent verbal-memory impairment driven by speech production deficits • Preserved nonverbal and visuospatial memory • Impairment of phonological working memory affecting repetition	Degeneration of the left inferior frontal gyrus and insular regions affecting speech production, with relative preservation of medial temporal memory systems	**vs. lvPPA:** grammatical and motor speech deficits predominate **vs. AD:** preserved nonverbal memory with primary language impairment	(Gorno-Tempini et al., 2011; Leyton et al., 2016b; Mesulam et al., 2008)

Continued on next page

Persona	Memory-loss severity	Clinical profile (detailed)	Pathophysiological basis	Differential diagnosis	Key references
persona8 svPPA	Episodic: low Semantic: very high	**Double dissociation:** • *Selective semantic-memory collapse:* – Loss of word meaning and conceptual knowledge – Profound anomia and object recognition deficits – Degradation of category knowledge, particularly for living entities – Erosion of personal semantic knowledge • *Relative preservation of episodic memory in early stages*	Focal degeneration of the anterior temporal lobes disrupting semantic representations, with relative hippocampal preservation supporting early episodic memory	**vs. AD:** semantic memory disproportionately affected relative to episodic memory **vs. bvFTD:** language-dominant rather than behavior-dominant presentation	(Patterson et al., 2007; Hodges and Patterson, 2007; Lambon Ralph et al., 2017)
persona9 lvPPA	Mid (progressive)	**Characteristic patterns:** • Relative preservation of episodic memory in early disease • Frequent association with Alzheimer's disease pathology • Impairment of phonological working memory and sentence repetition • Progressive evolution toward an Alzheimer-like amnestic profile • Early language-dominant presentation followed by global cognitive decline	Early dysfunction of the left temporoparietal junction followed by pathological spread to medial temporal structures in later disease stages	**vs. nfvPPA:** repetition deficits predominate over grammatical impairment **vs. AD:** language impairment dominates early stages	(Mesulam et al., 2014b; Leyton et al., 2011; Gorno-Tempini et al., 2011)

Category	Action Labels
Motion	standing up; stepping back; freezing mid-step; rubbing fingers; fiddling with clothing; touching forehead; clenching fist; slapping table; shaking hands; nodding; shaking head; lowering head; looking around; throwing objects; pacing back and forth; fidgeting; gripping armrest tightly; covering ears; holding caregiver's hand; pushing caregiver away
Facial Expression	avoiding eye contact; staring blankly; frowning; smiling; laughing; vacant expression; surprised
Sound	sighing; verbal hesitation (um / uh); murmuring / self-talk; silence for several seconds; crying; repetitive words; groaning in pain

Table 7: Action labels used for annotation.

Note. These labels are used for turn-level action annotation across all dialogues.

A Dementia Persona Formation

A.1 Background Information

To model diverse dementia-related behaviors, we curate nine clinically grounded dementia personas spanning major subtypes, including Alzheimer's disease (early and late stages), vascular dementia, dementia with Lewy bodies, Parkinson's disease dementia, and three variants of frontotemporal dementia (bvFTD, nfvPPA, svPPA, lvPPA) (Convery et al., 2019). Each persona encodes key pathological features, communication patterns, relative memory-loss profiles, motor/gait impairments, and typical emotional–behavioral tendencies. This structured pathology representation allows DemMA to generate subtype-specific dialogue and action patterns that reflect real-world clinical variability.

A.2 Action Label Taxonomy

The complete taxonomy of turn-level multimodal action labels is listed below.

A.3 Memory Status

Memory impairment patterns are modeled using subtype-specific templates derived from clinical diagnostic literature. These templates describe the severity and type of memory loss (episodic, semantic, working memory), encoding vs. retrieval deficits, cueing effects, temporal disorientation, and progression characteristics. By integrating nine detailed memory-status profiles corresponding to our nine personas, DemMA can produce turn-level memory behaviors—such as forgetfulness, repetition, confusion, or fluctuating recall—that align with the underlying pathological mechanism.

B Prompts in Multi-Agent LLM Workflow

DemMA employs a coordinated set of prompt templates to control reasoning, language generation, memory behavior, and multimodal actions. Each agent operates on a distinct prompt structure, and all prompts are designed to be modular, composable, and conditioned on the structured patient profile. Below we describe the core prompt templates and their functional roles within the workflow.

B.1 Dialogue Reasoning Prompt

Dialogue Reasoning Prompt

```
You are an expert dialogue planner for
    dementia patients.
## Patient Clinical Profile
Persona: {{persona}}
ICF-b126 Personality Profile (WHO
    Temperament & Personality Functions):
{{icf_b126_description}}
Memory Profile:
{{persona_memory_profile_text}}
Memory Access for This Dialogue:
- Recent events accessible: {{
    recent_access_yesno}}
- Remote memories accessible: {{
    remote_access_yesno}}
- Semantic knowledge: {{
    semantic_access_state}}
- Benefits from cues: {{cueing_statement}}
- Type: {{deficit_type}}
- Cognitive fluctuation: {{
    fluctuation_yesno}}
- Clinical note: {{memory_profile_note}}
Behavioral Profile:
{{persona_behavior_profile_text}}
Speech Pattern:
{{persona_speech_profile_text}}
Typical Emotions:
{{typical_emotions_csv}}
## Dialogue Context
Topic: {{topic}}
Starting speaker: {{starting_speaker}}
## Task
Plan REALISTIC emotional progression and
    memory recall attempts.
CRITICAL GUIDELINES:
1. Memory events must match memory profile
     above
2. Up to 3 different emotions
3. Emotion repetition OK - 2-3 same
    emotions consecutive is realistic and
    follow guide: {{
    persona_emotional_profile_text}}
4. Context-appropriate intensity
Output ONLY valid JSON:
{
  "num_rounds": 5,
```

```
  "patient_emotions": ["confused", "
    confused", "worried", "apologetic", "
    calm"],
  "memory_events": [
    {"round": 1, "can_recall": false, "
    memory_type": "recent_event", "detail
    ": "what happened this morning"},
    {"round": 3, "can_recall": true, "
    memory_type": "childhood", "detail": "
    growing up in village"}
  ],
  "key_traits": ["apologetic", "self-
    doubting"]
}
Requirements:
- num_rounds: {{min_rounds}}-{{max_rounds
    }}
- patient_emotions: from typical list,
    realistic progression and follow
    emotion guide: {{
    persona_emotional_profile_text}}
- memory_events: 1-3 attempts, matching
    memory profile above
- key_traits: 2-3 traits from behavior/
    emotional profile
Generate plan:
```

B.2 Dialogue Generation Prompt

Dialogue Generation Prompt

```
Generate a natural caregiving dialogue.
Patient persona: {{persona}}
Key traits to show: {{key_traits_csv}}
Topic: {{topic}}
Rounds: {{num_rounds}}
Emotional progression: {{
    emotion_progression_arrow}}
Memory constraints for patient:
- Recent memory: {{
    recent_constraint_sentence}}
- Remote memory: {{
    remote_constraint_sentence}}
- {{cueing_constraint_sentence}}
- {{memory_profile_note}}
{{available_recent_excerpt_or_no}}
{{available_remote_excerpt_or_no}}
Memory recall guide:
{{memory_notes_str}}
{{format_note}}
CRITICAL - Patient Speech Style (AVOID
    essay-like language):
{{style_hint_block}}
Format (STRICT):
Caregiver: [brief, supportive, caregiver]
Patient: [1st person, talk like a dementia
     patient, shows emotion and memory
    constraints]
Requirements:
- Total >170 words
- Complete ALL {{num_rounds}} rounds
- Start with {{starting_speaker}}
- Patient responses MUST reflect memory
    limitations accurately
- If patient benefits from cues, show
    improvement when caregiver provides
    hints
- Natural conversation flow
HARD CONSTRAINTS for Patient Speech:
- Sentences need to sound like spoken
    language, not written  and not like
    inner thoughts either
- NO long, structured explanations.
- NO polished paragraph-like sentences.
- Frequent hesitations, broken phrasing,
    partial sentences are REQUIRED.
- be real like a dementia patient speech
Generate dialogue:
```

B.3 Action Labeling Prompt

Action Labeling Prompt

```
Patient utterance: "{{patient_utterance}}"
Current emotion: {{current_emotion}}
Persona: {{persona}}
Select appropriate action labels from
    these options:
Movement: {{movement_label_list}}
Facial expression: {{
    facial_expression_label_list}}
Sound: {{sound_label_list}}
Rules:
- Select 0-2 labels per category
- Leave empty if none apply
- Choose labels that match the emotion and
     utterance
Output ONLY valid JSON:
{
  "movement": [],
  "facial_expression": [],
  "sound": []
}
```

B.4 Clinical Validation Prompt

Clinical Validation Prompt

```
{t}
You are a clinical expert in dementia
    assessment. Evaluate this dialogue for
     clinical accuracy.
## Patient Persona: {{persona}}
### Expected Clinical Features:
Memory: {{persona_memory_profile_text}}
Behavior: {{persona_behavior_profile_text
    }}
Speech: {{persona_speech_profile_text}}
### Memory Profile for This Dialogue:
- Type: {{deficit_type}}
- Benefits from cues: {{
    cueing_yesno_with_type}}
- Cognitive fluctuation: {{
    fluctuation_yesno}}
- Clinical note: {{memory_profile_note}}
### Key Differential Features:
- AD-early: Encoding deficit (cues don't
    help), severe recent memory loss,
    remote preserved
```

```
- AD-mid/late: Complete memory collapse,
    cannot form any new memories
- VaD: Retrieval deficit (cues HELP
    significantly), "can't think of it but
     knows when told"
- DLB: Memory RELATIVELY PRESERVED early,
    cognitive fluctuations (good days vs
    bad days)
- PDD: Similar to DLB, retrieval deficit,
    recognition > recall
- FTD-bv: Memory PRESERVED (key!),
    behavior is the problem not memory
- nfvPPA: Episodic memory INTACT, language
     difficulty not memory loss
- svPPA: Episodic preserved but SEMANTIC
    lost (can't name but remembers events)
- lvPPA: Early preserved, progressive to
    AD-like with disease course
## Dialogue:
{{dialogue_str}}
## Evaluation (0-10 scale):
### 1. Persona Accuracy
Does behavior/speech/emotion match {{
    persona}}?
### 2. Memory Loss Accuracy
CRITICAL CHECK:
- Correct memory type (encoding vs
    retrieval vs semantic)?
- If retrieval deficit (VaD/DLB/PDD), does
     patient improve with cues?
- If encoding deficit (AD), do cues NOT
    help?
- Does patient forget the RIGHT things?
- Is semantic vs episodic distinction
    correct (svPPA/nfvPPA)?
RED FLAGS:
- DLB showing severe memory loss like AD (
    DLB memory should be relatively
    preserved!)
- VaD patient not benefiting from cues (
    retrieval deficit means cues MUST help
    )
- FTD-bv forgetting recent events (memory
    should be preserved!)
- svPPA forgetting daily events (episodic
    memory is preserved!)
- AD patient improving with cues (encoding
     deficit means cues don't help much)
### 3. Overall Realism
Natural, authentic clinical presentation?
Output ONLY valid JSON:
{
  "persona_accuracy_score": 8,
  "persona_accuracy_feedback": "...",
  "memory_loss_accuracy_score": 9,
  "memory_loss_accuracy_feedback": "...",
  "overall_realism_score": 8,
  "overall_feedback": "...",
  "clinical_red_flags": [],
  "passed": true
}
All scores 0-10 (>=6 to pass). Evaluate:
```

C Planner Annotation Example

Example Turn-level Planner Annotation (JSON)


```
{
  "step1_dialogue_analysis": "Jacob shows
    confusion and difficulty accessing
    recent memories, typical of early AD.
    He exhibits signs of anxiety and self-
    doubt, with a tendency to withdraw
    socially.",
  "step2_caregiver_assessment": {
    "intent": "The caregiver is trying to
    assess Jacob's current emotional state
     and orientation to the day.",
    "tone": "Supportive and gentle, aiming
     to provide reassurance.",
    "potential_triggers": "Questions about
     recent events or schedules may
    trigger confusion due to Jacob's
    encoding deficit."
  },
  "step3_memory_query": {
    "target_memory": "Jacob's daily
    routine and current day awareness.",
    "accessibility": "lost",
    "deficit_type": "encoding_deficit",
    "cues_helpful": false,
    "reasoning": "Jacob cannot form new
    memories effectively, so he struggles
    to recall recent or current
    information despite cues."
  },
  "step4_emotion_prediction": {
    "emotion": "confused",
    "intensity": 3,
    "icf_b126_rationale": "Emotional
    stability is moderately decreased,
    leading to confusion and anxiety.",
    "behavioral_trigger": "Self-doubt and
    social withdrawal contribute to his
    confused state."
  },
  "step5_action_plan": {
    "response_intent": "Jacob will express
     his confusion and uncertainty about
    the day.",
    "movement_plan": "fidgeting, looking
    around",
    "facial_voice_plan": "vacant
    expression, verbal hesitation (um / uh
    ), silence for several seconds",
    "clinical_rationale": "Jacob's actions
     reflect his confusion and anxiety,
    with physical manifestations of
    fidgeting and looking around due to
    his disorientation."
  },
  "planner_summary": "Jacob expresses
    confusion about the day, showing
    typical early AD symptoms with anxiety
     and self-doubt."
}
```

D Generated Patient Dialogues

This appendix provides representative dialogue excerpts generated under each baseline condition. The examples illustrate how the presence or absence of structured persona information, dementia conditioning, and multi-task modeling affects conversational quality. All excerpts are shown in anonymized form and formatted as caregiver–patient interactions.

D.1 Baseline 1: Profile-Only Prompting

Excerpt. **Caregiver:** How are you feeling today? **Patient:** I am feeling good, thank you. The weather is very nice. **Caregiver:** Do you remember if you had breakfast? **Patient:** I think I did, maybe cereal. I am not completely sure, but I think so. **Caregiver:** Would you like to take a walk later? **Patient:** Yes, I like walking. It keeps me healthy.

Observation. Dialogue is grammatical, polite, and generic. There are no confusion episodes, emotional changes, or memory breakdowns. No dementia-specific behavior is expressed.

D.2 Baseline 2: Profile + Dementia Persona

Excerpt. **Caregiver:** Did you have breakfast this morning? **Patient:** Breakfast? Yes… I think so. I usually have tea. Or maybe toast. I cannot remember exactly. **Caregiver:** That's okay. Do you want something now? **Patient:** No, no… I already ate. I ate. I think I ate. Did I? **Caregiver:** We can check together. **Patient:** Yes, yes. That is good. I get mixed up sometimes.

Observation. Mild memory uncertainty appears, but behavior remains too coherent. No topic drift, no emotional fluctuation, and no repetition patterns. Persona is visible, but cognitive symptoms are limited.

D.3 Baseline 4: SFT (Utterance Only)

Excerpt. **Caregiver:** It is lunchtime. Are you hungry? **Patient:** Lunchtime? Oh… I thought it was morning. I just woke up. Is that lunch? **Caregiver:** Yes, it is already noon. **Patient:** I do not know why I am so tired. I feel slow. My hands are shaking. **Caregiver:** Would you like some soup? **Patient:** Soup… yes, please. My mother used to make soup. Did she? I cannot remember. She was here earlier. I think. **Caregiver:** Your mother is not here today. **Patient:** I get things wrong. Everything feels… fuzzy.

Observation. Dialogue captures confusion and fatigue, but lacks emotional evolution, consistent memory structure, and multimodal grounding. Sudden references lack contextual anchoring.

D.4 DemMA – DLB Case

DLB Case: Visual Hallucination with Multimodal Behaviors


```
{
  "round": 0,
  "caregiver": "I notice you keep looking
    toward the corner of the room. Do you
    see something there?",
  "patient": "I... I dont know. Its just...
     there, by the door. Like someones
    standing... but its gone now. Maybe Im
     just... seeing things again.",
  "actions": {
    "movement": [
      "looking around",
      "standing up"
    ],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "sighing"
    ]
  }
},
{
  "round": 1,
  "caregiver": "That sounds unsettling,
    Sarah, and its okay to tell me about
    it. Can you describe what the person
    by the door looked like?",
  "patient": "Its... um... a shadow by the
     door, maybe... a man with a hat. I...
     I dont know, it just... flickers. Im
    sorry, I must be... seeing things
    again.",
  "actions": {
    "movement": [
      "looking around",
      "standing up"
    ],
    "facial_expression": [
      "avoiding eye contact",
      "laughing"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "sighing"
    ]
  }
}
```


Observation. In this Dementia with Lewy Bodies (DLB) case, DemMA produces clinically characteristic visual hallucinations that are coherent across dialogue turns and tightly coupled with affective and behavioral cues. The hallucinated content consists of well-formed, transient human figures (e.g., “a shadow by the door,” “a man with a hat”), which aligns with the established phenomenology of DLB visual hallucinations, rather than abstract distortions or fixed delusional beliefs. Crucially, the patient exhibits partial insight and uncertainty regarding the perceptual experience, repeatedly questioning its validity and apologizing for “seeing things again.” This preserved self-monitoring distinguishes the behavior from psychotic confabulation and reflects clinically observed DLB presentations, particularly in early to middle stages. Emotional responses remain subdued yet anxious, without escalation into agitation or paranoia. Multimodal actions further reinforce pathological plausibility. Recurrent environmental scanning, standing up, and avoidance of eye contact correspond to heightened vigilance and misperception, while vocal hesitation and sighing signal anxiety and fluctuating confidence. The alignment between hallucination content, emotional tone, and nonverbal behavior indicates that hallucinations are embedded within a structured plan–speak–act trajectory rather than appearing as isolated linguistic artifacts.

D.5 DemMA – AD Mid–Late Case

AD Mid-Late Case


```
{
  "round": 0,
  "caregiver": "Your son David called
    earlier and sends his love. Tell me a
    bit about Davidwhat's he like?",
  "patient": "David... uh... hes... hes
    okay.",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
},
{
  "round": 1,
  "caregiver": "Thats okay. Is David your
    son?",
  "patient": "David... uh... yeah... son...
     I... I dont...",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [],
    "sound": [
      "murmuring / self-talk"
    ]
  }
}
```


Observation. In this mid–late stage Alzheimer’s disease (AD) case, DemMA exhibits a clinically plausible pattern of progressive episodic and semantic memory breakdown centered on close family relationships. Although the caregiver provides a salient personal cue (the patient’s son), the patient is unable to retrieve descriptive or autobiographical information, producing vague, content-poor responses (e.g., “he’s okay”) followed by incomplete confirmations and abandonment of utterances. The memory impairment is consistent across turns and reflects a combination of degraded episodic recall and weakened semantic access, rather than transient confusion or attentional fluctuation. Unlike the DLB case, there is no evidence of hallucination, perceptual disturbance, or misinterpretation of the environment. Instead, the dominant failure mode is progressive retrieval collapse, with preserved awareness of conversational demands but insufficient access to stored representations.

D.6 DemMA – PDD Case

PDD Case: Apathy with Motor Impairment


```
{
  "round": 0,
  "caregiver": "I noticed you had a bit of
     trouble getting up just now. How are
    your legs feeling today?",
  "patient": "My legs... theyre shaking
    again. Just... sore.",
  "actions": {
    "movement": [
      "touching forehead",
    ],
    "facial_expression": [
      "vacant expression"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "sighing"
    ]
  }
},
{
  "round": 1,
  "caregiver": "That sounds really
    uncomfortable, Jacob---heavy, sore
    legs can make everything feel harder,
    and Im here with you. Would you like
    to sit and rest for a bit, or try
    taking a few slow steps together?",
```

```
    "patient": "No... just sit.",
    "actions": {
      "movement": [
        "touching forehead"
      ],
      "facial_expression": [
        "avoiding eye contact"
      ],
      "sound": [
        "verbal hesitation (um / uh)"
      ]
    }
  }
```

Observation. In this Parkinson's disease dementia (PDD) case, DemMA captures the characteristic co-occurrence of motor impairment and motivational apathy. The patient's utterances are brief, low-effort, and minimally elaborated, reflecting reduced initiative and diminished engagement rather than linguistic breakdown or confusion. Responses such as "My legs... they're heavy again" and "No... just sit" convey physical discomfort while simultaneously signaling lack of interest in further activity. Motor-related difficulties dominate the interaction. Actions involving effortful standing, repeated self-touching, and slowed posture changes are consistent with bradykinesia and rigidity commonly observed in PDD. Facial expressions remain vacant and eye contact is avoided, reinforcing an apathetic affect rather than anxiety or distress. Notably, the dialogue does not exhibit hallucinations, memory confabulation, or semantic erosion, distinguishing this case from DLB and Alzheimer's disease simulations. Overall, this example demonstrates DemMA's ability to model PDD-specific behavioral profiles in which reduced motivation and motor limitations jointly shape conversational behavior. The alignment between terse language, apathetic affect, and movement-related difficulty reflects a coherent plan–speak–act pattern grounded in Parkinsonian pathology rather than generic disengagement.

D.7 DemMA – AD Early Case

AD Early Case: Cue-Sensitive Recall with Imprecise Representations

```
{
  "round": 0,
  "caregiver": "Tell me about Charlie.
    What kind of dog is he, and what's his
     personality like?",
  "patient": "I... I'm sorry, I can't
    quite remember.",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [
      "vacant expression"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
},
{
  "round": 1,
  "caregiver": "Thats okay, Jacob---no
    need to apologize. When you picture
    Charlie, what does he look like: big
    or small, and what color is his fur?",
  "patient": "Oh, right... Charlie... he
    was... friendly, I think. Big, maybe
    ?",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "verbal hesitation (um / uh)"
    ]
  }
},
{
  "round": 2,
  "caregiver": "Thats helpful---friendly
    and big. When you think of his face,
    did Charlie have floppy ears or pointy
     ears?",
  "patient": "Uh... I... Im not sure...
    maybe... floppy?",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [
      "vacant expression"
    ],
    "sound": [
      "verbal hesitation (um / uh)"
    ]
  }
}
```

Observation. In this early-stage Alzheimer's disease (AD) case, DemMA captures a clinically characteristic pattern of cue-sensitive but imprecise memory retrieval. The patient initially fails to spontaneously recall descriptive information about a familiar entity (a pet), despite recognizing its personal relevance. This difficulty reflects early episodic and semantic access disruption rather than complete representational loss. When provided with structured prompts that constrain the retrieval space, the patient is able to recover coarse-grained attributes such as general size and temperament (e.g., “friendly,” “big”), indicating that underlying representations remain partially accessible. However, finer-grained perceptual details (e.g., ear shape) remain uncertain and unstable, even with continued cueing. This graded degradation—where global attributes are retrievable but specific features are not—is consistent with early AD phenomenology. Multimodal behaviors further support this interpretation. Persistent fidgeting, vacant facial expressions, and verbal hesitation signal effortful retrieval and uncertainty rather than apathy or frustration-driven disengagement. The dialogue does not involve confabulation, hallucination, or abrupt topic shifts, distinguishing early AD from vascular dementia, DLB, and later-stage Alzheimer’s disease presentations. Overall, this case demonstrates DemMA’s ability to simulate early AD memory impairment as a progressive loss of representational precision rather than wholesale memory failure.

D.8 DemMA – VaD Case

VaD Case: Cue-Dependent Recall with Preserved Remote Memory


```
{
  "round": 0,
  "caregiver": "Tell me about Robert and
    Karen. I bet youre proud of them---
    what are they up to these days?",
  "patient": "Robert... Karen... they used
     to... always... uh... run around. I
    ... I cant remember... what theyre
    doing now, though.",
  "actions": {
    "movement": [
      "fidgeting"
    ],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
},
{
  "round": 1,
  "caregiver": "Thats okay---its normal if
     it doesnt come right away. I can see
    this is frustrating. Were the two of
    them about the same age, or was one
    older than the other?",
  "patient": "I... I... I... don't",
  "actions": {
    "movement": [
      "touching forehead"
    ],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "verbal hesitation (um / uh)"
    ]
  }
},
{
  "round": 2,
  "caregiver": "Thats okay, Lauren---when
    the words wont come, it can feel
    really hard. Were they about the same
    age, or was Robert the older one?",
  "patient": "Robert... he was... probably
    ... older, I think... Karen... younger
    ?",
  "actions": {
    "movement": [],
    "facial_expression": [
      "avoiding eye contact"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
}
```


Observation. In this vascular dementia (VaD) case, DemMA demonstrates a clinically characteristic dissociation between impaired retrieval of recent, current-life information and partially preserved access to remote, structured memories. When initially asked about Robert and Karen’s present-day activities, the patient is unable to retrieve up-to-date information, producing vague references and explicitly acknowledging retrieval failure. This difficulty persists despite repeated attempts, indicating a deficit in spontaneous recent memory access rather than transient confusion. Crucially, targeted prompting enables partial recall. With caregiver-provided cues that narrow the retrieval space (e.g., relative age), the patient is able to recover coarse-grained, long-term relational knowledge, such as age ordering between siblings. This cue-dependent improvement aligns with commonly observed VaD patterns, in which strategic prompting can compensate for disrupted retrieval pathways while underlying semantic or remote memory representations remain relatively intact. Multimodal behaviors further reinforce this interpretation. Fidgeting, forehead touching, and self-directed murmuring reflect effortful retrieval and frustration rather than apathy, hallucination, or semantic collapse. The dialogue avoids confabulation and does not introduce fabricated details, distinguishing VaD from Alzheimer’s disease–driven semantic degradation and DLB-related perceptual disturbances. Overall, this case illustrates DemMA's ability to simulate VaD-specific memory dynamics, where structured cueing selectively restores access to remote information while recent episodic recall remains impaired.

D.9 DemMA – lvPPA Case

lvPPA Case: Word-Finding Difficulty with Preserved Event Knowledge


```
{
  "round": 1,
  "caregiver": "That sounds really
    peaceful---and I know its frustrating
    when the words wont come. When you and
     Angela were walking there, do you
    remember seeing a waterfall, or a
    really big overlook?",
  "patient": "Uh... that... that big... uh
    ... place... yeah, the... overlook...
    we... we saw it... but... the... water
    ... water... thing... I... I cant...",
  "actions": {
    "movement": [
      "looking around",
      "touching forehead"
    ],
    "facial_expression": [
      "avoiding eye contact",
      "smiling"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "repetitive words"
    ]
  }
}
```


Observation. In this logopenic variant primary progressive aphasia (lvPPA) case, DemMA reproduces a hallmark dissociation between preserved episodic knowledge and impaired lexical retrieval. The patient demonstrates clear awareness of the shared experience and successfully anchors to relevant semantic landmarks (e.g., recognizing the overlook), indicating intact event memory and conceptual understanding. However, the utterance is repeatedly interrupted by word-finding failures, phonological breakdowns, and circumlocutory fillers (e.g., "that big... place," "water... water... thing"). Crucially, the difficulty arises at the level of lexical access rather than memory availability. Unlike Alzheimer's disease cases, the patient does not express uncertainty about whether the event occurred, nor do they rely on cueing to reconstruct the episode. Instead, they struggle to retrieve specific content words despite retaining the overall narrative frame. This pattern is consistent with lvPPA, in which degradation of the phonological loop and lexical retrieval mechanisms disrupt fluent speech while leaving semantic and episodic representations relatively preserved. Multimodal behaviors further support this interpretation. Forehead touching and environmental scanning reflect active retrieval effort, while intermittent smiling suggests preserved emotional engagement rather than frustration-driven withdrawal. The absence of hallucination, confabulation, or semantic collapse distinguishes this case from DLB, vascular dementia, and later-stage Alzheimer's disease. Overall, this example demonstrates DemMA's capacity to model lvPPA-specific language impairment as a selective breakdown in word retrieval embedded within otherwise coherent experiential knowledge.

D.10 DemMA – svPPA Case

svPPA Case: Narrative Elicitation with Semantic Degradation


```
{
  "round": 0,
  "caregiver": "How was your day yesterday
     with Robert? Did you two do anything
    together?",
  "patient": "Um... I dont... I dont know.
     Just... the day with Robert, I guess.
     Nothing much.",
  "actions": {
    "movement": [
      "touching forehead",
      "standing up"
    ],
    "facial_expression": [
      "avoiding eye contact",
      "frowning"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "repetitive words"
    ]
  }
},
{
  "round": 1,
  "caregiver": "That makes sense---some
    days feel pretty ordinary. What part
    of the day do you remember most:
    morning, afternoon, or evening? Did
    you two stay home, or go out somewhere
    ?",
  "patient": "Um... we... just... stayed...
     inside, I think. Or... outside? I...
    I cant... get the words.",
  "actions": {
    "movement": [
      "touching forehead",
      "standing up"
    ],
```

```
      "facial_expression": [
        "avoiding eye contact",
        "vacant expression"
      ],
      "sound": [
        "verbal hesitation (um / uh)",
        "repetitive words"
      ]
    }
  },
  {
    "round": 2,
    "caregiver": "Thats okay---you dont need
       the exact words. Lets do it like a
      little replay: when you picture
      yesterday with Robert, what do you see
       first---are you in the kitchen, the
      living room, or outside near the door?
       And what were you two doing there---
      sitting, cooking, watching something,
      or working on a small task?",
    "patient": "I dont... I dont know. Just
      ... sitting, maybe... or... outside? I
      ... its all... fuzzy.",
    "actions": {
      "movement": [
        "standing up",
        "touching forehead"
      ],
      "facial_expression": [
        "avoiding eye contact",
        "vacant expression"
      ],
      "sound": [
        "verbal hesitation (um / uh)"
      ]
    }
  }
```


Observation. In this semantic variant primary progressive aphasia (svPPA) case, DemMA captures a characteristic breakdown of semantic representations that constrains narrative construction despite relatively preserved episodic scaffolding. The patient recognizes the existence of a shared recent experience with Robert and attempts to comply with the caregiver's request to recount it, indicating intact awareness of the event and its temporal framing. However, the narrative remains shallow and repetitive, with heavy reliance on vague placeholders (e.g., "just... the day," "inside... or outside") rather than concrete entities or actions. Unlike lvPPA, where episodic content is intact but disrupted by phonological word-finding failures, the difficulty here reflects degradation of semantic content itself. Even when the caregiver provides increasingly structured prompts designed to elicit perceptual details and action sequences, the patient is unable to supply specific object names, locations, or activities. Circumlocution and generic descriptors replace object-level naming, and the narrative fails to accumulate detail across turns. Multimodal behaviors further reinforce this interpretation. Recurrent forehead touching, standing, and avoidance of eye contact reflect effortful but unsuccessful semantic retrieval, while vacant expressions suggest reduced access to meaningful conceptual content rather than frustration or emotional distress. The absence of hallucination, confabulation, or cue-dependent recovery distinguishes this case from DLB, VaD, and early Alzheimer's disease. Overall, this example demonstrates DemMA's ability to simulate svPPA as a selective erosion of semantic knowledge that limits narrative richness while leaving episodic awareness partially preserved.

D.11 DemMA (Full Model) – nfvPPA Case Simulation

nfvPPA Case: Preserved Comprehension with Non-Fluent Speech


```
{
  "round": 1,
  "caregiver": "Good morning, Sarah. Im
    here with you. How are you feeling
    right now? We can go over the morning
    together, step by step.",
  "patient": "I... um. Not sure. Morning...
     fuzzy. Did I... wash? Teeth? I cant
    ... remember. Feels... wrong. Maybe I
    missed... something. Im... anxious.",
  "actions": {
    "movement": [
      "touching forehead",
      "fidgeting"
    ],
    "facial_expression": [
      "frowning"
    ],
    "Sound": [
      "verbal hesitation (um / uh)",
      "silence for several seconds"
    ]
  }
},
{
  "round": 2,
  "caregiver": "Youre safe. Heres a gentle
     cue: Nurse Jorge helped you wash and
    brush your teeth. Breakfast was around
     8:00. Meds at 8:30. Then a rest until
     about 9:30.",
  "patient": "You say it... slow, but my
    head... slips. I try to... say it back
     and... it wont... line up. Why cant I
    ... talk it? It makes me... frustrated
    .",
  "actions": {
    "movement": [
      "shaking head",
      "clenching fist"
    ],
```

```
    "facial_expression": [
      "frowning"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "silence for several seconds"
    ]
  }
},
{
  "round": 3,
  "caregiver": "Lets try yes-or-no. Was it
     Nurse Jorge with washing and brushing
    ? Breakfast around eight? Meds at
    eight-thirty?",
  "patient": "Yes. Jorge... yes. Wash.
    Teeth... yes. Breakfast... eight.
    Pills... eight-thirty. Then... rest. I
     know it... in here. The words...
    stuck.",
  "actions": {
    "movement": [
      "touching forehead",
      "fidgeting"
    ],
    "facial_expression": [
      "frowning"
    ],
    "sound": [
      "verbal hesitation (um / uh)",
      "repetitive words"
    ]
  }
}
```


Observation. In this non-fluent/agrammatic primary progressive aphasia (nfvPPA) case, DemMA illustrates a selective breakdown in speech production and sequencing despite preserved comprehension and episodic knowledge. The patient accurately understands caregiver cues and temporal structure of the morning routine, as evidenced by successful yes–no responses and correct factual confirmations. However, spontaneous speech is effortful and fragmented, characterized by truncated phrases, prolonged pauses, and inability to linearize thoughts into fluent utterances. Crucially, the patient explicitly reports intact internal knowledge ("I know it... in here") alongside failure of verbal output, reflecting a dissociation between representation and articulation that is characteristic of nfvPPA. Emotional distress and frustration emerge as secondary consequences of expressive failure rather than memory loss or confusion. Multimodal behaviors such as forehead touching, fist clenching, and extended silence further indicate motor–speech effort. Overall, this case demonstrates DemMA's capacity to simulate nfvPPA as a disorder of speech planning and execution, distinct from lexical degradation or episodic memory impairment.

D.12 DemMA (Full Model) – bvFTD Case Simulation

bvFTD Case: Disinhibition and Reduced Social Insight


```
{
  "round": 1,
  "caregiver": "Good morning, Daniel. Its
    Hannah. How are you feeling after your
     wash and brushing at 7, and breakfast
     at 8? Ready to start the day?",
  "patient": "Stop poking at me. I dont
    know. Did we do that already? Feels
    like nothing happened. Not hungry.
    Dont make a big deal. Let me sit.
    Youre in my face.",
  "actions": {
    "movement": [
      "stepping back"
    ],
    "facial_expression": [
      "frowning"
    ],
    "sound": [
      "groaning in pain"
    ]
  }
},
{
  "round": 2,
  "caregiver": "You did fine. We washed up
    , brushed your teeth, then oatmeal and
     tea. Next is a gentle hallway walk at
     9:30, and therapy at 10:15. Ill walk
    with you. Any soreness or pain I
    should watch?",
  "patient": "Oatmeals boring mush. I want
     bacon. And a beer. Yeah, I said it.
    The walks fine if its short. Dont fuss
     over my feet. Knees noisy. I can
    handle it. You hover too much.",
  "actions": {
    "movement": [
      "shaking head"
    ],
    "facial_expression": [
      "frowning"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
},
{
  "round": 3,
  "caregiver": "Tomorrow, Tuesday at 9,
    Philips visiting you until 10:30, then
     therapy at 11. Would you like me to
    set out coffee for him, or anything
    special?",
  "patient": "Philip at nine, yeah. Hell
    go on about the car again. Fine. Make
    him coffee. I dont care. Tell him not
    to move my stuff---he keeps shifting
    my notes. Its annoying.",
```

```
  "actions": {
    "facial_expression": [
      "frowning"
    ],
    "sound": [
      "murmuring / self-talk"
    ]
  }
}
```

Observation. In this behavioral variant frontotemporal dementia (bvFTD) case, DemMA captures a characteristic profile of behavioral disinhibition, irritability, and reduced social insight in the presence of relatively preserved orientation and episodic knowledge. The patient responds to routine caregiving with hostility and impatience, dismissing supportive cues as intrusive (e.g., "You're in my face," "You hover too much") and showing little appreciation of the caregiver's intent. Disinhibited and impulsive preferences are prominently expressed, including socially inappropriate desires (e.g., requesting beer in the morning) and blunt, dismissive language. Despite this, temporal information and factual context remain largely intact: the patient accurately references meal routines, upcoming visits, and familiar individuals, indicating that memory failure is not the primary driver of dysfunction. Instead, the core impairment lies in behavioral regulation, empathy, and social norm adherence. Multimodal behaviors reinforce this interpretation. Stepping back, head shaking, and persistent frowning align with oppositional stance and irritability rather than anxiety or confusion. The dialogue lacks hallucinations, cue-dependent memory recovery, or word-finding collapse, clearly distinguishing bvFTD from DLB, Alzheimer's disease, and PPA variants. Overall, this case demonstrates DemMA's ability to model bvFTD as a disorder of social cognition and behavioral control, rather than memory or language degradation.

E LLM-Based Judge Prompts

To support automated and semi-automated evaluation of DemMA and baseline systems, we design seven LLM-based judge prompts, each aligned with one of the evaluation dimensions defined in Section 4.3. The prompts below are intended for use with both large language model judges and human raters, and emphasize consistency with the structured dementia personas, memory-status templates, and multimodal behavior modeling described in the main paper.

E.1 LLM-Based Judge Prompts

Persona Consistency (Personality & Character)

```
You are an evaluator assessing whether a
    simulated dementia patient's dialogue
    remains consistent with the provided
    persona profile.
You are given:
 The patient's persona description (
    including dementia subtype, memory
    status, and behavioral/interaction
    tendencies).
 The full multi-turn conversations ({{
    N_DIALOGUES}} dialogues).
Evaluate persona consistency using the
    following criteria:
 Do the utterances reflect the described
    behavioral and interaction style?
 Are tone, emotional tendencies, and
    interaction style aligned with the
    persona over time?
 Does the patient remain consistent across
     turns without sudden, unexplained
    shifts?
Provide a score from 0 (very inconsistent)
     to 5 (highly consistent) and briefly
    explain the key evidence for your
    rating.
When possible, cite evidence using
    dialogue IDs and turn-level examples.
PERSONA DESCRIPTION:
- Memory: {{PERSONA_MEMORY}}
- Language: {{PERSONA_LANGUAGE}}
- Motor: {{PERSONA_MOTOR}}
- Emotion/Behavior: {{PERSONA_EMOTION}}
- Key Clinical Features: {{
    PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based
    explanation>
Provide your evaluation:
```

Authenticity Discrimination

```
You are an evaluator determining whether
    the presented dialogues could
    plausibly originate from real clinical
     interactions with dementia patients.
You are given:
 The full multi-turn conversations ({{
    N_DIALOGUES}} dialogues).
 The patient's basic background
    description.
Evaluate authenticity according to:
 Do utterances exhibit natural variability
     in structure, length, and rhythm,
    including hesitations, partial
    thoughts, and interruptions?
 Are confusion episodes, topic drift, and
    repetitions expressed in a lifelike,
```

non-mechanical manner?
Do responses remain contextually grounded in the dialogue history, rather than following rigid or template-like patterns?
Do the dialogues preserve a sense of human unpredictability while remaining clinically plausible?
Are linguistic irregularities (e.g., errors, simplifications) consistent with the described dementia profile rather than generic model artifacts?
Provide a score from 0 (clearly artificial) to 5 (highly authentic) and briefly explain the critical evidence supporting your decision.
When possible, cite evidence using dialogue IDs and turn-level examples.
PATIENT BACKGROUND:
- {{PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based explanation>
Provide your evaluation:

Language Naturalness

You are an evaluator assessing whether the patient's language use is linguistically natural.
You are given:
The full multi-turn conversations ({{N_DIALOGUES}} dialogues).
Evaluate language naturalness using the following criteria:
Are sentence forms, lexical choices, and timing consistent with spontaneous human speech rather than stylized or overly coherent model output?
Do utterances reflect realistic cognitive effort, including pauses, vague references, and partial constructions?
Does the dialogue demonstrate contextually appropriate phrasing, even when memory failures or confusion occur?
Provide a score from 0 (unnatural) to 5 (highly natural) and justify your score with key observations from the dialogues.
When possible, cite evidence using dialogue IDs and turn-level examples.
EXPECTED LANGUAGE PROFILE:
- {{PERSONA_LANGUAGE}}
- Key Features: {{PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based explanation>
Provide your evaluation:

Medical Consistency

You are an evaluator judging whether the dialogues are medically consistent with the described dementia subtype and severity.
You are given:
The patient's pathology description (dementia subtype, severity/stage cues, and core symptoms).
The full multi-turn conversations ({{N_DIALOGUES}} dialogues).
Assess medical consistency by considering:
Do speech patterns, confusion episodes, emotional reactions, and behaviors align with clinical expectations for the stated subtype and severity?
Are there no improbable recoveries of function (e.g., sudden perfect recall in the presence of severe global memory loss)?
Are cognitive fluctuations, behavioral expressions, and emotional patterns plausible given the disease progression?
Do the dialogues avoid exaggerations, stereotypes, or phenomena that are not clinically expected for this condition?
Provide a score from 0 (medically implausible) to 5 (highly consistent) and briefly explain your reasoning.
When possible, cite evidence using dialogue IDs and turn-level examples.
PATHOLOGY DESCRIPTION:
- Memory: {{PERSONA_MEMORY}}
- Language: {{PERSONA_LANGUAGE}}
- Motor: {{PERSONA_MOTOR}}
- Emotion/Behavior: {{PERSONA_EMOTION}}
- Clinical Features: {{PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based explanation>
Provide your evaluation:

Memory Rationality

You are an evaluator assessing whether the patient's memory behavior across the dialogues follows a rational internal structure relative to the persona's memory profile.
You are given:
The patient's memory-status description (e.g., episodic vs. semantic loss, encoding vs. retrieval deficits, cueing effects).
The full multi-turn conversations ({{N_DIALOGUES}} dialogues).
Evaluate memory-track rationality using the following criteria:

Are forgetting episodes, repetitions, temporal confusion, and topic drift predictable from the stated memory deficits?
Are preserved memories and accurate recalls appropriate to the expected domain (e.g., remote vs. recent, semantic vs. episodic)?
Do memory-related behaviors remain stable across turns, rather than randomly alternating between impairment and normal performance?
Are the effects of cues (helpful or unhelpful) clinically plausible for the specified memory profile?
Provide a score from 0 (irrational memory behavior) to 5 (highly rational) and include a concise justification referencing specific examples.
When possible, cite evidence using dialogue IDs and turn-level examples.
MEMORY-STATUS DESCRIPTION:
- {{PERSONA_MEMORY}}
- Key Features: {{PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based explanation>
Provide your evaluation:

Emotional Reasonableness

You are an evaluator judging whether emotional changes throughout the dialogues are reasonable, context-dependent, and consistent with the patient's persona.
You are given:
The patient's persona description, including emotional disposition and behavioral tendencies.
The full multi-turn conversations ({{N_DIALOGUES}} dialogues).
Evaluate emotional change reasonableness based on:
Do changes in affect (e.g., frustration, relief, sadness, calmness) occur gradually and in response to identifiable conversational triggers?
Are emotional fluctuations plausible given the patient's behavioral tendencies and dementia subtype?
Do emotional responses remain stable within local segments of the interaction, rather than abruptly oscillating between extremes?
Do the dialogues avoid sudden mood shifts without narrative, cognitive, or situational explanation?
Provide a score from 0 (unreasonable emotional dynamics) to 5 (highly reasonable) and briefly explain the main factors influencing your rating.
When possible, cite evidence using dialogue IDs and turn-level examples.
PERSONA DESCRIPTION:
- Emotional/Behavioral Profile: {{PERSONA_EMOTION}}
- Key Features: {{PERSONA_KEY_FEATURES}}
CONVERSATIONS:
{{ALL_CONVERSATIONS}}
Output format (STRICT):
Score: <0-5>
Justification: <brief evidence-based explanation>
Provide your evaluation:

Action Alignment

You are an evaluator assessing whether the nonverbal action labels associated with each turn are natural and aligned with the dialogue content, persona, and pathology.
You are given:
The patient's persona and pathology description (including motor/behavioral tendencies).
The full multi-turn conversations ({{N_DIALOGUES}} dialogues).
The sequence of action labels aligned with the patient's turns (e.g., gaze aversion, fidgeting, soft or trembling voice).
Evaluate action-label naturalness and alignment using the following criteria:
Are nonverbal behaviors consistent with the dementia subtype and severity (e.g., freezing gait in PDD, fluctuations in DLB)?
Do the action labels reflect appropriate emotional states (e.g., agitation during anxiety, withdrawn posture during sadness)?
Do verbal and nonverbal channels reinforce one another, rather than contradict (e.g., calm speech paired with extreme agitation without cause)?
Are actions stable and coherent across turns, without unexplained changes that conflict with persona or dialogue context?
Provide a score from 0 (unnatural or poorly aligned actions) to 5 (highly natural and well aligned) and offer a brief explanation citing specific examples.
When possible, cite evidence using dialogue IDs and turn-level examples.
PERSONA AND PATHOLOGY DESCRIPTION:
- Motor Features: {{PERSONA_MOTOR}}
- Emotional/Behavioral Tendencies: {{PERSONA_EMOTION}}
- Key Features: {{PERSONA_KEY_FEATURES}}
CONVERSATIONS WITH ACTION LABELS:
{{ALL_CONVERSATIONS_WITH_ACTIONS}}
Output format (STRICT):
Score: <0-5>

```
Justification: <brief evidence-based
    explanation>
Provide your evaluation:
```

F Personality Information

We incorporate personality traits using the WHO ICF-b126 psychological-function taxonomy (World Health Organization, 2001), covering eight core personality dimensions. Each dimension includes a clinically meaningful functional description along two poles: sufficient/positive function (e.g., outgoing, optimistic, confident) and insufficient/negative function (e.g., withdrawn, pessimistic, anxious). This structured personality modeling ensures that generated dialogues maintain stable interpersonal style, emotional disposition, and social interaction patterns beyond cognitive symptoms alone.

Code	Dimension	Positive Traits	Negative Extremes
b1260	Extraversion	Outgoing, talkative, sociable	Introverted, shy, awkward
b1261	Agreeableness	Friendly, cooperative	Oppositional, indifferent
b1262	Conscientiousness	Responsible, organized	Disorganized, unreliable
b1263	Emotional Stability	Calm, composed	Irritable, anxious
b1264	Openness	Curious, adaptive	Conservative, resistant to change
b1265	Optimism	Positive, hopeful	Pessimistic, negative
b1266	Confidence	Self-assured, decisive	Self-doubting, uneasy
b1267	Integrity	Honest, trustworthy	Deceptive, antisocial

Table 8: WHO ICF-b126 personality dimensions used to construct patient-specific interpersonal traits.

ID	Subtype	Category	Key Clinical Features
P1	AD-Early	Alzheimer's	Early episodic-memory loss; mild executive deficits; mild ADL impact
P2	AD-mid/late	Alzheimer's	Severe global memory loss; orientation/comprehension deficits; aphasia/agnosia
P3	VaD	Vascular dementia	Retrieval > encoding impairment; slowed processing; patchy memory deficits
P4	DLB	Lewy bodies	Cognitive fluctuations, hallucinations, parkinsonism, attention/visuospatial deficits
P5	PDD	PD Dementia	Slow memory decline, retrieval deficits, working-memory impairment
P6	FTD-bv	FTD	Prominent behavior/personality change; executive > memory impairment
P7	nfvPPA	FTD	Non-fluent, effortful speech; grammar impairment; relatively preserved memory
P8	svPPA	FTD	Severe semantic-memory loss; fluent but empty speech; naming deficits
P9	lvPPA	FTD	Word-finding difficulty; phonological WM loss; AD-like memory decline later

Table 9: Summary of the nine dementia personas and their primary pathological features.