DemMA：具有专家引导推理和动作模拟功能的痴呆症多轮对话智能体

Yutong Song[1], **Jiang Wu**[2], **Kazi Sharif**[3] **Honghui Xu**[3],**Nikil Dutt**[1],**Amir Rahmani**[11]加州大学欧
尔文分校， [2]独立研究员， [3]肯尼索州立大学 ty

arXiv:2601.06373v1 [cs.MA] 10 Jan 2026

摘要

使用大语言模型（LLM）模拟痴呆患者颇具挑战性，因为需要在长时间对话中联合建模认知障碍、情绪动态和非语言行为。我们提出了 DemMA，这是一种专家引导的痴呆症对话智能体，用于实现高保真度的多轮患者模拟。DemMA 通过整合病理信息、人格特质以及由临床专家提供信息的特定亚型记忆状态人物设定，构建具有临床依据的痴呆症人物画像。为了超越纯文本模拟，DemMA 对非语言行为进行显式建模，包括动作、面部表情和声音线索。我们进一步提出了一种思维链蒸馏框架，在一次前向传播中训练单个大语言模型联合生成推理轨迹、患者话语和相互对齐的行为动作，从而无需多智能体推理即可实现高效部署。由专家、医学生和大语言模型评审者开展的大规模评估表明，在多个指标上，DemMA 均显著优于强基线模型。

1 引言

在痴呆症研究和照护者培训中，高质量互动数据的匮乏仍是一个结构性瓶颈（Livingston 等人，2017年）。由于严格的隐私和伦理限制，收集敏感患者数据受到严格约束（Johnson 和 Karlawish，2015年）。因此，公开语料库几乎不存在，尤其是那些涵盖面部表情和语音韵律等多模态资源的语料库。在病理细分层级上，这种“数据荒漠”尤为严重：即便是阿尔茨海默病（AD）等常见类型，公开语料库也十分有限，而且20多种临床亚型仍基本未得到体现（Gorno-Tempini 等人，2011年）。因此，患者模拟往往依赖脚本材料，无法呈现现实照护互动的异质性。基于大语言模型的生成式智能体的出现，为数据合成提供了有前景的途径，但将通用对话智能体模型用作痴呆症模拟器并不可靠（Wei 等人，2022年；Liu 等人，2023年）。首先，若缺乏基于临床的对齐，这类模型可能生成缺乏医学严谨性的内容，甚至给出不安全指导。其次，在长时间互动中，模拟往往会出现人物画像漂移，即模型逐渐退回到流畅、礼貌且泛化的助手风格（Shumailov 等人，2023年）。这种“过于完美”的表达会掩盖认知衰退的重要迹象。因此，尽管合成的对话在语言上可能流畅，却缺乏对患者人物画像的忠实呈现，使其不适用于智能体训练。

痴呆模拟的一项核心挑战，是对可解释、可复现且具有特定亚型特征的病理行为进行建模，而非将随机变化注入模型输出中（Jack 等人，2018年）。尽管意识混乱和语言错误等一般症状在不同病症中普遍存在，但不同痴呆亚型表现出各异的认知障碍。因此，高保真模拟需要具备相应机制，以区分病理性不一致与一般的模型幻觉。

此外，痴呆沟通本质上是多通道的，涵盖语言、情感和行为（Bender 等人、2022年）。随着语言能力下降，具有临床意义的信号越来越多地通过非语言表达呈现，而基于文本的大语言模型无法保留这些信号。这一局限促使人们采用结构化文本表示，对潜在非语言线索进行编码，以弥补言语能力受损造成的不足。

代码可在 https://anonymous.4open.science/r/DemMA-0A4B/获取

最后，从系统角度来看，长时域对话的连贯性通常通过多智能体架构来处理，但这会带来相当大的额外开销和延迟，使其不适用于实时痴呆照护培训与模拟。

基于这些动机，我们提出**DemMA**，这是一个通过微调开源模型实现端到端痴呆症虚拟患者生成的框架，无需依赖敏感真实世界数据。此外，为解决纯文本模拟往往无法充分表征语言障碍这一局限，DemMA采用双轨建模范式：

内在认知层面： 模型明确捕捉不同痴呆亚型和阶段之间的细粒度差异，确保生成的思维不连贯反映特定的人物画像模式，而非随机噪声。

外在表达层面： 我们引入动作标签作为一种补偿机制，将多模态行为（如动作、面部表情、声音）映射为基于文本的标签，发挥三项关键作用：（i）明确揭示无法通过含糊语言直接表达的潜在情绪和意图；（ii）将多模态行为投射到基于文本的交互界面；（iii）提供关键信号，以区分不同的亚型和疾病阶段。

我们的贡献总结如下：

• 我们提出一种具有临床依据的建模方法，在DemMA中精准表征**痴呆症患者人物画像**。

• 我们是首个将明确的动作、面部表情和声音标签引入基于大语言模型的生成式智能体的工作。

• 我们提出了一种多智能体流程，用于生成高质量的交互对话数据，并发布了**DemMA-Dialogue**，这是首个涵盖主要痴呆亚型并经过专家验证的合成痴呆症对话数据集。

•我们为智能体**DemMA** 开发了蒸馏训练，使其内化长时程规划信号，从而支持低延迟推理，并实现连贯且人物画像一致的交互。

2 DemMA

在架构上，DemMA 整合了三个组件：**(i)** 具有临床依据的患者角色构建模块；**(ii)** 多智能体对话数据集生成流程，涵盖记忆分析与搜索、对话规划、语言和行动模拟；以及 **(iii)** 面向 DemMA 智能体的思维链蒸馏多任务训练，使其内化医学知识，同时降低多智能体推理延迟。DemMA 的概览如图 1所示。

痴呆症人物画像构建 痴呆症角色模拟的一项核心挑战是确保纵向一致性和临床有效性。朴素的单轮生成往往缺乏医学依据，导致档案不一致和人物画像漂移。DemMA 通过分阶段生成并设置明确的依赖关系来构建患者角色设定，从而将上游约束传递至所有下游组件：

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

其中，B、S 和 M 分别表示个人背景、人格和记忆层，而 $\mathcal{P}$ 表示由此生成的患者角色设定档案。

背景层（B）该层编码了（i）稳定的人口统计学和生活情境属性（例如年龄、教育程度和合并症），以及（ii）痴呆亚型角色设定及其核心临床表现模式（Convery 等人，2019年）。我们设定了涵盖整个疾病谱系的九种亚型：早期阿尔茨海默病、中晚期阿尔茨海默病、血管性痴呆、路易体痴呆、帕金森病痴呆、额颞叶痴呆行为变异型、非流利型原发性进行性失语、语义变异型原发性进行性失语和语义型原发性进行性失语（详见附录 A，表**??**）。这些属性包括亚型特有的行为特征和神经病理学依据（附录 A，表**??**），为后续建模奠定了医学和情境约束基础。

人格层（S）S 根据世界卫生组织《国际功能、残疾和健康分类》（ICF）`b126`心理功能空间世界卫生组织，2001年）（附录**??**，表**??**）。至关重要的是，痴呆引起的人格改变并非随机发生，而是与其潜在神经病理学相对应。因此，我们不将人格视为独立于疾病的特质，而是将其建模为编码于ICF空间中的亚型特定倾向（附录A；表**??** 和**??**）。这种结构性依赖确保生成多样且具有临床依据的互动风格，避免产生不符合医学常理的人格特征。

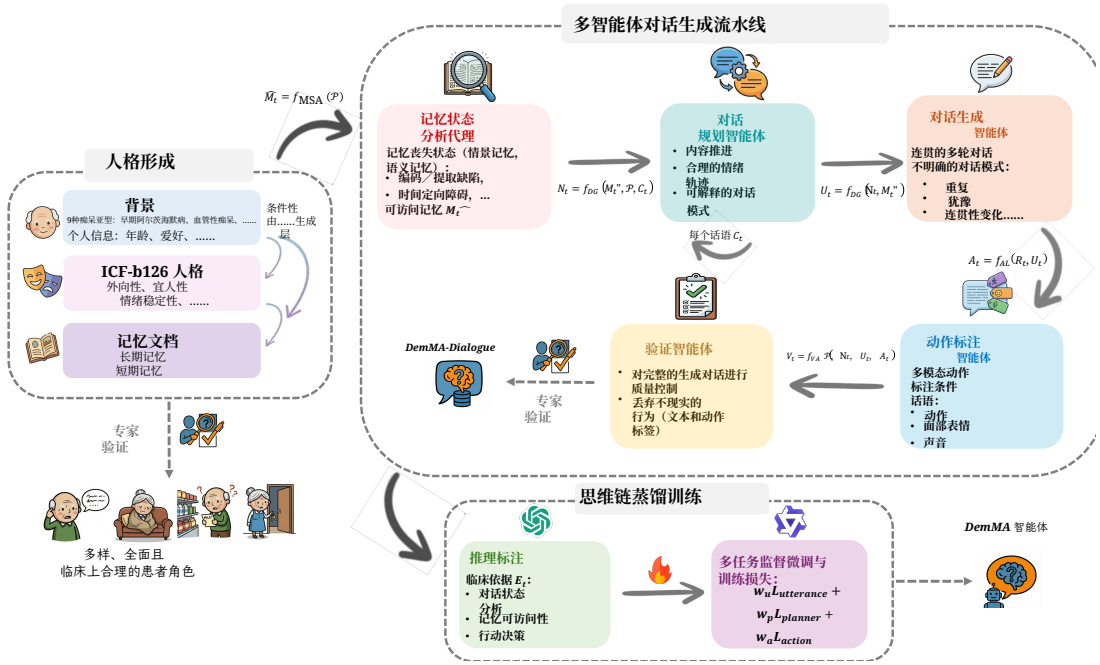

图 1：DemMA 集成了三个组件：a）具有临床依据的患者角色构建模块；b）多智能体对话数据集生成流程，涵盖记忆分析、规划、语言生成和动作模拟；以及 c）DemMA 智能体的思维链蒸馏多任务训练流程。

记忆层（M）我们将 M 具体化为（B、 S）中的：（i）长期记忆（例如童年经历、人生经历和重要人生事件）；（ii）近期事件的短期日/周记录，并确保其与长期记忆以及 B 中的临床事实（例如合并症）保持一致。该层将体现远期和近期信息各自不同的保留与衰退模式。

3 多智能体大语言模型工作流

DemMA 采用多智能体大语言模型工作流，以结构化控制和可解释性生成长时程对话。研究表明，在复杂语言任务中，将推理与生成分离有助于提高连贯性并减少幻觉（Yao 等人，2023年）。五个智能体的运作方式如下。

3.1 记忆状态分析代理

工作流始于记忆状态分析智能体。该智能体基于痴呆症角色设定，以及从临床诊断文献中提取的相应记忆状态信息（附录 **??**，表**??**），识别对话期间可访问的记忆系统，并推断具有临床依据的特征，例如近期记忆、远期记忆和语义记忆的可用性，以及线索响应性：

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

其中，M_t 表示在第 t 轮可访问的记忆和具有临床依据的记忆特征，f_{MSA} 表示记忆分析函数。该推断出的记忆状态会影响后续所有推理和生成。

3.2 对话规划智能体

随后，对话规划智能体 会根据患者的人格、推断出的记忆状态以及对话语境 C_t，生成高层规划。该智能体通过指定规划的内容推进以及各轮次合理的情绪轨迹，来组织整体对话流程，如下所示：

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

将这一步规划过程显式化，可提升透明度与可控性，确保生成的对话遵循可解释的模式，并与患者的认知和情绪状态相符，而非受不透明的模型行为支配。

3.3 对话生成智能体

对话生成智能体生成自然语言对话，将高层计划转化为轮次级话语，同时确保内容符合患者的认知局限、人格特质和情绪状态。通过明确根据结构化计划来调节生成过程，该智能体

能够生成连贯的多轮对话，同时保留痴呆特异性沟通模式，例如重复、犹豫、错置指代以及连贯性的变化。轮次 t 的话语生成建模如下：

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

3.4 动作标注智能体

该动作标注智能体 会为每轮对话补充多模态动作标注，这些标注依据生成话语和推断出的推理状态进行调整。它会为合理的**动作**、**面部表情**和**声音**分配相应的标签，捕捉痴呆互动中常见的行为信号，例如坐立不安、回避目光、颤抖的言语以及情绪突然变化（Liu 等人，2023年）。完整的动作标签列表见附录**??**。

这些动作序列从多模态角度呈现患者反应，并支持具身模拟和多模态数据生成等下游任务。多模态接地的动作标签预测在第 t 轮建模如下：

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

3.5 验证智能体

最后一个组件是验证智能体，负责对生成的对话进行质量控制。它会输出验证分数，以决定是否接受或重新生成该轮对话，从而减少长时程生成中的偏移、不连贯性和不现实的行为。验证过程定义如下：

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

其中，V_t 表示第 t 轮的验证分数，f_{VA} 表示验证函数。当满足以下条件时，接受该轮：$V_t \geq \tau$；否则重新生成。

3.6 专家验证

在完成数据集定稿之前，我们邀请 4 位痴呆领域专家进行人工验证，以确认保留的对话在情绪状态、语言模式、动作行为和记忆障碍方面，均符合相应的痴呆亚型和疾病阶段。我们使用这一多智能体框架生成了首个基于大语言模型的痴呆症对话数据集 DemMA-Dialogue。统计数据如表1所示。

动作	**#（百分比）**	**声音**	**#（百分比）**
低头	4,380 (27.8%)	言语犹豫（嗯 / 呃）	13,703 (57.6%)
坐立不安	3,096 (19.6%)	叹气	2,991 (12.6%)
四处张望	2,347 (14.9%)	喃喃自语 / 自言自语	2,602 (10.9%)
推搡照护者	686 (4.3%)	重复词语	2,197 (9.2%)
away			
触摸额头	590 (3.7%)	沉默片刻	1,592 (6.7%)
		秒	
站立	488 (3.1%)	突泣	399 (1.7%)
其他	4,179 (26.5%)	疼痛呻吟	313 (1.3%)
面部表情	**#（百分比）**	**认知特征**	**#（百分比）**
皱眉	5,600（37.0%）	具有远期情景记忆	13,303（83.0%）
避免眼神接触	4,045（26.8%）	具有语义记忆	11,468（71.6%）
茫然表情	3,662（24.2%）	受益于提示线索记忆	9,538（59.5%）
微笑	857 (5.7%)	有近期情景记忆	9,160 (57.2%)
其他	959 (6.3%)	检索缺陷	6,328 (39.5%)
语料统计			
2,709 段对话（平均 5.90 轮；中位数 6；P25/P75 为 5–7；最小/最大值为 3–8）			
每段对话的平均动作标签数：20.19			

表 1： DemMA-Dialogue 的统计数据。

4 思维链蒸馏智能体训练

4.1 推理标注

为了将高质量的对话生成提炼到单个模型中，我们构建了轮次级推理标注，作为中间监督。每条标注都会明确说明对话轮次背后的推理轨迹，使模型无需执行完整的智能体生成流程，也能在推理时复现相当水平的话语和动作决策。对于每个轮次 t，我们以形式化方式生成推理标注，

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

每个 E_t 包括：(i) 对话状态分析，(ii) 照护者意图，(iii) 记忆可访问性，(iv) 带有 ICF-b126 依据的情绪推断，以及 (v) 带有临床依据的行动决策。我们使用E_t 作为中间监督，以学习临床一致的推理—表达—行动行为。附录 **??** 中提供了一个具体的轮次级推理标注示例。

4.2 多任务监督微调

为了在单次前向传播中联合建模规划器推理依据、患者话语和多模态动作标签，我们采用多任务监督微调目标。我们优化以下各项的加权和：(i) [PLAN] 和[SPEAK] 片段的掩码下一词元预测损失；以及 (ii) 多标签动作分类损失：

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

为减少推理与表层形式生成之间的干扰，我们采用分段特定的令牌掩码（使用索引集合 T_p 和 T_u），从而在令牌级别将规划器和话语监督解耦：

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

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

动作标签由辅助多标签预测头预测，并通过 $\mathcal{L}_{\text{action}}$ 联合优化，从而在生成过程中鼓励推理与动作行为保持一致。

多标签动作学习。 动作预测被建模为一项多标签分类任务，针对动作、面部表情和声音特征等行为线索进行预测。我们采用焦点调制损失：

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

其中， a_k 是真实标签， z_k 是预测逻辑值。焦点项侧重处理低置信度预测，从而提升模型在动作分布稀疏情况下的鲁棒性。

集成学习。 通过在单次前向传播中联合训练所有组件，模型能够学习记忆障碍、推理、语言生成与多模态动作之间的依赖关系。这种集成优化使 DemMA 能够在统一模型中生成连贯的推理轨迹、逼真的患者对话以及相匹配的动作标签，无需单独的模块或后处理。

5 实验结果与分析

5.1 实验设置

我们在一个带有规划器、话语和动作标签的痴呆症对话语料库上对 DemMA 进行微调，并使用固定随机种子按 85%/15% 划分训练集和验证集。基础模型为 Qwen3-8B。训练最多进行 5 个周期，并采用 AdamW 8 位优化，学习率为 5×10^{-6}。所有训练和实验均在 8 块 NVIDIA H100 GPU 上以混合精度进行。代码可在此处获取：

h

ttps://anonymous.4open.science/r/DemMA-0A4B/。

5.2 评估指标

我们从七个互补维度评估生成的对话。每项维度指标均由大语言模型评估器、临床专家和学生进行评估，从而实现涵盖医学、语言和行为标准的多视角评估。

人格一致性。 衡量智能体在多轮对话中是否保持稳定且连贯的人格，包括一致的身份、偏好和行为倾向。

语言自然度。 衡量话语听起来是否自然、口语化，而非像套用模板或经过模型润色。

真实性。 衡量对话的真实程度，例如语言流畅度、重复情况和变化性。

医学一致性。 衡量语言、行为和情感是否与指定的痴呆亚型及严重程度相符。

记忆合理性。 衡量遗忘、重复和线索响应是否符合定义的记忆特征。

情感合理性。 衡量情绪变化是否由上下文驱动且逐渐发生，而非突兀或随机。

动作一致性。 衡量非语言动作是否合理，并与话语内容和临床特征一致。

5.2.1 自动评估

为发挥大语言模型作为评审者评估开放式任务的能力，我们使用了最先进的大语言模型，包括 GPT-5.2-pro、Gemini-2.5-pro 和 Qwen3-32B，将其作为评估者。我们设计了七个针对特定任务的大语言模型评审提示词，分别对应上述评估维度。评审者会给出一个标量评分（0–5）和简短说明。所有评审提示词均见附录 **??**。

5.2.2 人工评估

在数据集标注方面，我们招募了三位痴呆症专家，他们每人均拥有十年以上痴呆症诊断和患者护理临床经验。在评估方面，我们另外招募了五位具有相当临床背景的专家，以及四位接受过神经病学或老年医学正规培训的医学生。

5.3 基线

我们将 DEMMA 与三个基线进行比较，这些基线逐步加入痴呆症特定条件化和

Table 2: 基于大语言模型的判断，以及专家对患者模拟保真度和多轮对话质量的评估 .

方法	评估者	患者模拟保真度						多轮对话质量		
		权威性	Med.	记忆	情感	行动	平均 人格	语言		Avg.
原始版	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
	专家评估	1.50	2.00	2.00	2.50	1.00	1.80	2.00	1.50	1.75
临床档案提示词	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
	专家评估	2.00	2.50	2.00	3.00	1.50	2.20	2.50	2.00	2.25
SFT-话语	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
	专家评估	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
	专家评估	3.50	4.00	3.50	4.00	3.50	3.70	4.00	3.50	3.75

训练信号：**(i)原始版**，仅使用背景档案，不包含痴呆线索；**(ii) 临床档案提示**，加入描述病理特异性角色设定及亚型症状的内容；**(iii) SFT-话语**，一种经过监督训练、用于生成话语但不具备规划器推理或动作监督的监督模型。

6 结果与分析

本节将介绍全面的实验，旨在回答以下研究问题（RQs）：**研究问题1：模拟保真度**。DEMMA 能否实现对不同痴呆症亚型患者的高保真模拟?

研究问题2：对话质量。DEMMA 能否维持高质量的多轮交互，同时保持角色一致性和自然的对话互动?

研究问题3：基于大语言模型的评估的有效性。基于大语言模型的评审能否可靠地评估痴呆症模拟质量？其评分是否与人类专家的评估一致?

研究问题4：数据集质量与多样性。DemMA-Dialogue 能否为有效的多轮智能体训练提供足够的多样性和覆盖范围?

研究问题5：教育效果。DEMMA 能否通过帮助受训者更好地理解痴呆症临床表现，支持医学培训（例如，医学生或受训照护者）？

6.1 模拟结果

我们使用大语言模型评审者，针对所有痴呆症人物画像，采用相同的七项指标评估每种方法。为获得每种方法的单一总分，我们先计算每个人物设定下各项指标的评分均值，再对九个人物设定的结果取平均值。自动评估和人工评估的结果见表2。

为回答**研究问题1**，我们重点评估真实性、医学一致性、记忆合理性、情感合理性和动作一致性，以及动作一致性。**在所有评估者的评估中，DemMA始终取得最高的总体患者模拟保真度**。在GPT-5.2、Gemini-2.5和Qwen3-32B评估者的评判中，DemMA的平均得分均居首位（4.1–4.3），显著优于所有基线。不同评估者得出一致结果，表明DemMA的优势体现了广泛认可的提升，而非评估者特异性偏差。**DemMA的提升是全面的，涵盖所有关键模拟维度**。在所有模拟保真度指标上，DemMA均取得最高或并列最高分，且没有明显短板。相比之下，其他方法表现出显著失衡：语言方面表现尚可，但在具有临床依据的维度或行为维度上表现不佳。**从基于提示词的方法到监督微调（SFT），再到DemMA，这一发展体现的是质的转变，而非线性提升**。其他方法的得分仍徘徊在2.0至3.0左右，而DemMA在所有评估维度上的得分均同时超过4.0。这表明DemMA带来了根本不同层次的建模能力，使患者模拟从表层的对话可信度迈向高保真、具有临床依据的模拟。

为回答**研究问题2**，我们重点评估角色一致性和语言自然度。**在所有评估者的评估中，DemMA始终在多轮对话质量的平均得分上表现最佳**。这一结果表明，DemMA在长时间交互中并非仅仅勉强可用，而是能够稳定地保持高质量表现，在多轮对话中接近人类水平的表现。DemMA展现出显著的

表3： 不同痴呆亚型的专家评估（5位评定者的平均值）。

	权威性	医学	记忆	情感	行动	平均值	人格	语言
早期阿尔茨海默病	3.0	3.2	3.4	3.8	3.0	3.4	3.8	3.6
中晚期阿尔茨海默病	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
行为变异型额颞叶痴呆	3.8	3.2	3.4	4.0	3.0	3.6	3.8	3.6
非流利型原发性进行性失语	4.0	4.2	4.4	4.0	3.0	3.9	4.2	4.0
语义型原发性进行性失语	3.4	3.8	4.0	3.6	3.6	3.7	4.0	3.6
语义变异型原发性进行性失语	3.6	3.8	3.4	3.2	3.6	3.7	4.6	3.0
平均值	**3.6**	**3.7**	**3.6**	**3.8**	**3.3**	**3.7**	**4.0**	**3.5**

角色一致性优势，其跨轮次稳定性强于所有基线方法。在角色一致性（人格）维度上，DemMA显著优于其他方法，这表明它能够更有效地在不同对话轮次中保持稳定的角色属性和行为模式。**DemMA在保持角色一致性的同时，并未牺牲语言自然度**。这一发现表明，DemMA并非依靠刻板或模板化的回答来维持一致性，而是在流畅自然的对话交流中保持长期的角色稳定性。

6.2 痴呆症亚型的人类评估与大语言模型评判

为回答 **研究问题3**，我们在表3 和图2中总结了针对各亚型的人类与大语言模型评判评估结果。就患者模拟保真度而言，大语言模型评审者对 DemMA 的评分为 4.15–4.29，而专家评分为 3.70；就多轮对话质量而言，大语言模型评审者的评分为 3.95–4.55，专家评分为 3.75。这表明，**大语言模型评审者适合进行比较性评估，但往往会系统性地高估绝对质量，尤其是对于表现更强的模型**。

图 2 和表 3 展示了一致的趋势，但侧重点有所不同。二者均表明，DemMA 在各亚型上的表现均优于基线。然而，大语言模型雷达图放大了性能差距，而专家给出的评分则更为保守且集中（总体为 3.5–3.9），并一致指出动作一致性是目前剩余的主要弱点。**大语言模型评审者有助于呈现相对改进和亚型概况，但专家评分提供了具有临床依据的评估**。

6.3 训练数据集的规模效应

为回答 **研究问题4**，我们使用不同规模的数据集训练了 Qwen3-8B 模型。性能如——

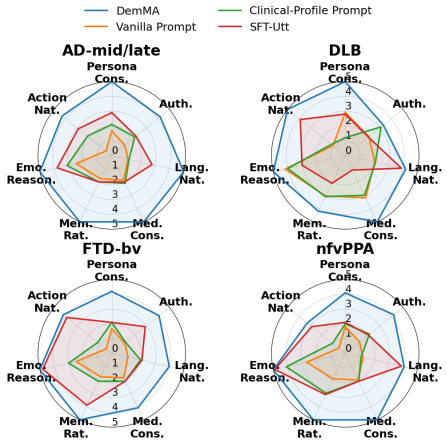

图2： 大语言模型评判（GPT-5.2-pro）在四种痴呆亚型上的表现。

图中所示3。**较长的对话主要提升多轮能力，但收益很快就呈现边际收益递减**。

增加每段对话的轮次：从短对话增加到中等长度对话时，提升最为显著，此后性能趋于平稳。较长的上下文尤其能提升情感合理性，并使角色一致性、语言自然度和真实性更加稳定，因为这些维度取决于跨轮次的连贯性和渐进式演变。**增加主题多样性可显著提升医学一致性和记忆合理性**。模型必须反复观察同一种病理在不同情境中的表现。相比之下，动作一致性和真实性变化相对较小，这表明它们更多依赖动作标签监督/对齐质量或风格线索，而非单靠主题覆盖度。

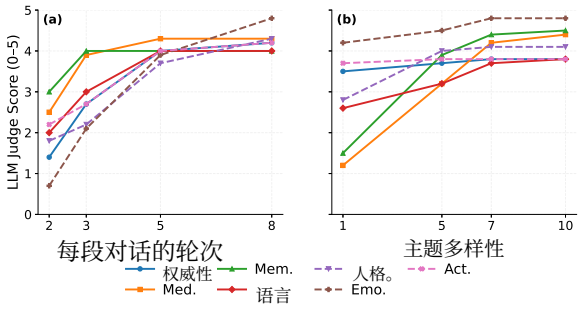

图3： 训练数据集的规模效应。

6.4 教育效果与反馈 ck

为回答 **研究问题5**，我们与医学生和痴呆症专家开展了一项盲法亚型识别研究。参与者阅读随机呈现的 DemMA 生成的多轮对话，并被要求判断每个病例的痴呆亚型。结果以混淆矩

阵形式汇总于图4。 **学生能够通过 DemMA 识别痴呆亚型特异性临床差异。**两个矩阵都呈现出明显的对角线，表明医学生能够根据 DemMA 对话正确识别大多数痴呆亚型，但其准确率通常低于专家。DemMA 提供了清晰且易于学习的临床线索，可用于区分不同亚型。**学生的错误与专家在临床表现重叠的亚型上的混淆相呼应。** 当提示线索明确时，DemMA 能够区分不同亚型，并再现真实的边界案例。图 5中的成对评估显示，DemMA 在真实性方面获得了压倒性的偏好。DemMA 的胜率达到 89.3%，表明其在感知到的痴呆症真实感方面具有明显且稳健的优势。

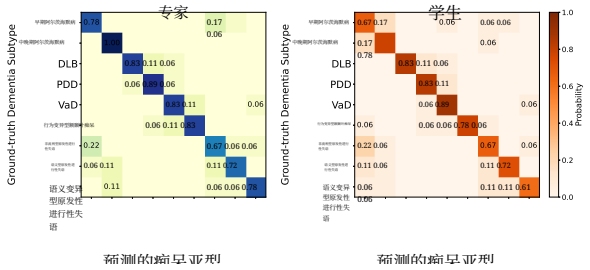

图 4： 九种痴呆亚型识别的混淆矩阵（专家与学生对比）。

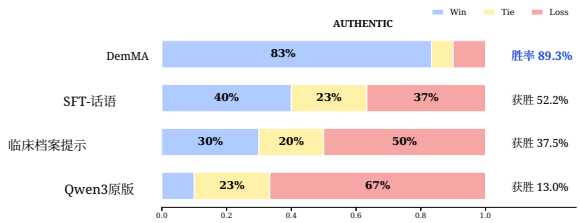

图 5： AUTHENTIC 上的两两胜/平/负结果。胜率计算方式为 Win/(Win + Loss)，不包括平局。

7 相关工作

7.1 思维链蒸馏训练

思维链（CoT）蒸馏已被提出作为一种有效方法，可在降低推理时开销的同时，保留大语言模型的多步推理能力。早期研究表明，让模型接触显式推理轨迹能够提升其在复杂任务上的表现，但在推理过程中直接生成此类轨迹的计算成本很高（Wei 等人，2022年）。为解决这一局限，后续工作通过监督训练，将教师生成的推理依据或规划轨迹蒸馏到学生模型中，使推理行为得以隐式内化，而无需在推理时显式生成思维链（Wu 等人，2024年；Magister 等人，2023年）。这一理念进一步扩展到规划器—执行器和基于智能体的框架中，将多步决策过程蒸馏到单一模型中，从而在降低延迟的同时保持长时程连贯性（Nye 等人，2021年；Yao 等人，2023年）。与主要关注文本推理的既有思维链蒸馏方法不同，我们将具有临床依据的推理信号（包括记忆分析、对话规划和动作决策）蒸馏到统一模型中，以高效生成痴呆症对话。

7.2 用于临床培训的基于大语言模型的患者模拟

人们已探索使用大语言模型提供面向痴呆症的对话支持和照护者辅助，但安全性与伦理问题始终备受关注（Treder 等，2024年）。近期的患者模拟系统，如 PATIENT-Ψ 和 PatientSim，利用结构化角色设定与心理学模型，使大语言模型的行为具有依据，从而生成合理的多轮临床交互（Wang 等，2024年；Kyung 等，2025年）。然而，这些框架大多假设患者认知状态稳定，并以人物设定的一致性为优化目标，因此未能充分体现痴呆症特有的现象，例如重复、相互矛盾的回忆以及时间定向障碍（Lyketsos 等，2011年；Zhao 等，2016年；Aalten 等，2008年）。相比之下，DemMA 将痴呆症建模为持续进展的认知状态，并生成具有临床依据、会随轮次演变的功能崩解模式。

8 结论

本文介绍了 DemMA，这是一种具有临床依据的多轮痴呆症对话智能体，通过联合建模语言与动作生成来模拟真实行为。DemMA 将基于角色设定的痴呆亚型和疾病阶段融入结构化流程，从而能够对认知衰退、情绪变化和非语言线索进行可控模拟。我们提出了一种蒸馏式策略，将规划、推理、对话和动作预测内化到单个低延迟模型中。大量评估表明，DemMA 在角色设定保真度、临床有效性和教育效果方面优于基线，为高保真模拟提供了一种可扩展且符合伦理的解决方案，同时避免使用敏感的真实世界数据。

局限性

尽管取得了出色的结果，DemMA 仍存在若干局限。首先，数据集完全由合成数据构成。尽管 DemMA-Dialogue 避免使用敏感的患者数据，并经过领域专家验证，但合成交互可能遗漏罕见的临床行为以及照护者与患者之间独特的互动动态，从而可能限制其在真实场景中的泛化能力。其次，临床依据依赖文本抽象。我们的动作标签可作为多模态行为的可解释代理，但无法充分捕捉细粒度的视听或感觉运动线索，因此在具身或多模态应用中的保真度受到限制。第三，角色设定和记忆建模较为简化。虽然我们涵盖了多种痴呆亚型和记忆可及性模式，但该框架并未明确建模长期疾病进展、药物影响或不断变化的照护者策略。最后，评估部分依赖基于大语言模型的评审。这类评审可能存在已知偏见，也可能无法充分衡量具有临床依据的正确性（Yao 等人，2023年）。因此，我们将自动评分视为专家评估和人工评估的补充，而非具有决定性的临床评估。

伦理考量

DemMA 旨在通过生成完全合成的对话数据，而非依赖真实患者记录，消除痴呆症研究中的伦理和法律障碍。这大幅降低了隐私、知情同意以及敏感健康信息暴露方面的风险。尽管如此，合成数据仍可能与真实个人或情境相似，因此发布数据集时应筛查个人身份标识符，并提供明确的使用指南。该框架面向研究和教育用途，例如照护者培训和模型开发，不适用于临床诊断或决策。若缺乏适当的防护机制，模拟痴呆对话可能会被滥用，生成具有说服力但医学上不正确的内容。因此，任何下游部署都应明确规定范围限制、免责声明和人工监督。模拟认知障碍也会引发对表征伤害和刻板印象的担忧。尽管角色设定以临床知识为依据并经过验证，模型输出仍可能过度简化或带有偏见地描绘痴呆症。我们鼓励未来从人口统计和互动维度开展审查，以评估并减轻此类风险。最后，自动化评估部分依赖基于大语言模型的评审，而众所周知，这类评审存在偏见和局限。因此，我们将自动化指标与专家评估和人工评估结合起来，并强调不应将报告的分数解读为临床评估。

参考文献

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及其他4位作者。2008年。不同痴呆症中神经精神综合征的一致性：欧洲阿尔茨海默病联盟的研究结果。第二部分。《痴呆与老年认知障碍》，25(1)：1–8。

Emma N. Bender、Marie Y. Savundranayagam、Laura Murray和J. B. Orange。2022年。痴呆症患者非言语沟通的支持策略：一篇范围综述。*International Journal ofNursing Studies*，136：104365。

R. Convery, S. Mead, and Jonathan D. Rohrer. 2019年。额颞叶痴呆的临床、遗传和神经影像学特征。神经病理学与应用神经生物学，45(1):6–18。

Maria Luisa Gorno-Tempini, Argye E. Hillis, Sandra Weintraub, AndrewKertesz,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, Kar- alyn Patterson, Bruce L. Miller, David S. Knopman, John R. Hodges, M. Marsel Mesulam, and Murray Grossman. 2011年。原发性进行性失语及其变体的分类。*Neurology*，76(11):1006–1014。

Jr. Jack, Clifford R., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Brenda Dunn, Stephen B. Hae- berlein, 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年。NIA–AA研究框架：迈向阿尔茨海默病的生物学定义。*Alzheimer's& Dementia*，14(4):535–562。

Rebecca A. Johnson 和 Jason Karlawish。2015年。痴呆症中的伦理问题综述。国际精神老年医学，27(10)：1635–1647。

Daeun Kyung、Hyunseung Chung、Seongsu Bae、Jiho Kim、Jae Ho Sohn、Taerim Kim、Soo Kyung Kim 和 Edward Choi。2025年。PatientSim：面向真实医患互动的角色驱动模拟器。arXiv 预印本 arXiv:2505.17818。版本 1.0.0；亦通过 PhysioNet 发布。

Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xu- anyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang、Shudan Zhang、Xiang Deng、Ao- han Zeng、Zhengxiao Du、Chenhui Zhang、Sheng Shen、Tianjun Zhang、Yu Su、Huan Sun 以及其他 3 位作者。2023年。Agentbench：评估作为智能体的大语言模型。arXiv 预印本 arXiv:2308.03688。发表于 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、HelenC Kales、Eric B Larson、Karen Ritchie、Kenneth Rockwood、Elizabeth L Samp- son等另外5人。2017年。痴呆预防、干预与照护。《柳叶刀》，390(10113):2673–2734。

Constantine G. Lyketsos、Maria C. Carrillo、John M. Ryan、Ara S. Khachaturian、Paula Trzepacz、Joanna Amatniek、Daniel S. Miller和Gregory S. Smith。2011年。阿尔茨海默病的神经精神症状。*Alzheimer's &Dementia*，7(5):532–539。

Lucie Charlotte Magister、Jonathan Mallinson、Jakub Adamek、Eric Malmi 和 Aliaksei Se veryn。2023年。《教小型语言模型进行推理》。载于计算语言学协会第*61*届年会论文集（第*2*卷：短篇论文），第1773–1781页。

Maxwell I. Nye、Anders Johan Andreassen、Guy Gur- Ari、Henryk Michalewski、Jacob Austin、David Bieber、David Dohan、Aitor Lewkowycz、Maarten Bosma、DavidLuan、Charles Sutton 和 Augustus Odena。2021年。《展示你的工作：语言模型中间计算的草稿区》。*CoRR*，abs/2112.00114。

Ilia Shumailov、Zakhar Shumaylov、Yiren Zhao、Yarin Gal、Nicolas Papernot 和 Ross Anderson。2023年。《递归诅咒：使用生成数据进行训练会使模型遗忘》。*arXiv* 预印本 *arXiv:2305.17493*。

Matthias S. Treder、SojinLee 和 Kamen A.Tsvetanov。2024年。痴呆照护与研究中大语言模型（llms）简介。《痴呆症前沿》，3：1385303。

Ruiyi Wang、Stephanie Milani、Jamie C. Chiu、Jiayin Zhi、Shaun M. Eack、Travis Labrum、Samuel M. Mur- phy、Nev Jones、Kate Hardy、Hong Shen、Fei Fang 和 Zhiyu Zoey Chen。2024年。PATIENT-Ψ：使用大语言模型模拟患者，以培训心理健康专业人员。载于*2024*年计算语言学实证方法会议论文集，美国迈阿密。计算语言学协会。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V. Le、Denny Zhou 和 Brian Ichter。2022年。思维链提示促使大语言模型进行推理。神经信息处理系统进展，35：24824–24837。

世界卫生组织。2001年。国际功能、残疾和健康分类（ICF）。世界卫生组织。

Zongqian Wu、Baoduo Xu、Ruochen Cui、Mengmeng Zhan、Xiaofeng Zhu 和 Lei Feng。2024年。从自训练视角重新思考思维链。*arXiv* 预印本 *arXiv:2412.10827*。

Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik R. Narasimhan 和 Yuan Cao。2023年。ReAct：在语言模型中协同推理与行动。载于国际学习表征会议。

Qianhua Zhao、Lili Tan、Huaxiang Wang、Tian Jiang、Mengmeng Tan 和 Lan Tan。2016年。阿尔茨海默病患者神经精神症状的患病率：系统综述与荟萃分析。情感障碍杂志，190：264–271。