

# Beyond Prompt: Fine-grained Simulation of Cognitively Impaired Standardized Patients via Stochastic Steering

Weiliang Zhang<sup>1</sup> Zimo Zhu<sup>1</sup> Zhichuan Yang<sup>1</sup> Chen Huang<sup>2\*</sup>

Wenqiang Lei<sup>3,4</sup> See-Kiong Ng<sup>2</sup>

<sup>1</sup> School of Software, Xi'an Jiaotong University

<sup>2</sup> Institute of Data Science, National University of Singapore

<sup>3</sup> College of Computer Science, Sichuan University

<sup>4</sup> Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China

weikang.zhang@stu.xjtu.edu.cn, huang\_chen@nus.edu.sg

## Abstract

Simulating Standardized Patients with cognitive impairment offers a scalable and ethical solution for clinical training. However, existing methods rely on discrete prompt engineering and fail to capture the heterogeneity of deficits across varying domains and severity levels. To address this limitation, we propose StsPatient for the fine-grained simulation of cognitively impaired patients. We innovatively capture domain-specific features by extracting steering vectors from contrastive pairs of instructions and responses. Furthermore, we introduce a Stochastic Token Modulation (STM) mechanism to regulate the intervention probability. STM enables precise control over impairment severity while mitigating the instability of conventional vector methods. Comprehensive experiments demonstrate that StsPatient significantly outperforms baselines in both clinical authenticity and severity controllability.

## 1 Introduction

Patients with cognitive impairment, such as Alzheimer's and Mild Cognitive Impairment (MCI), experience debilitating deficits across multiple domains (e.g., memory and attention) (Bowie and Harvey, 2005; Kahn and Keefe, 2013), which notably affect their speech patterns and content (Forbes-McKay and Venneri, 2005; Voleti et al., 2019; Gkoumas et al., 2023). These deficits severely compromise their quality of life, making effective clinical management important (Varrecchia et al., 2020; McCollum and Karlawish, 2020). To ensure superior patient care, clinical staff (e.g., nurses and therapists) must receive specialized training for communicating with individuals with impairment (Burgos et al., 2018).

Traditionally, clinical training has relied on **Standardized Patients (SPs)** as a core practice method

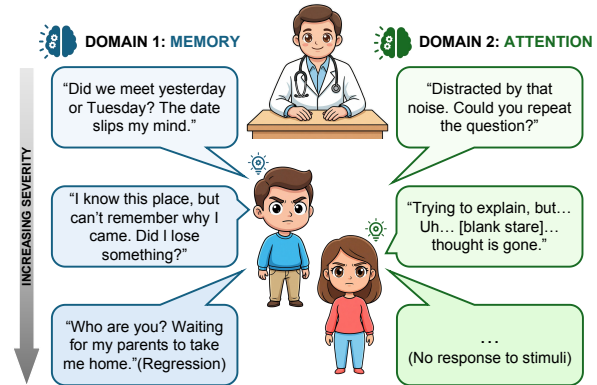


Figure 1: Simulating cognitively impaired SPs across varied domains and severity levels is challenging, demanding fine-grained control.

for clinical staff. This is often achieved by employing human actors trained to simulate patients (Barrows, 1993; Elendu et al., 2024; Akkurt Yalcinturk and Dissiz, 2024; Cotter et al., 2025). However, this approach struggles with the vast heterogeneity of cognitive impairment (Bowie and Harvey, 2005; Khalil et al., 2025). Taking Figure 1 for example, patients with the same diagnosis can manifest diverse domain-specific deficits (e.g., attention and memory) and varying levels of severity, ranging from MCI to Alzheimer's Disease dementia (Vincze et al., 2016; McCutcheon et al., 2023). Consequently, accurately capturing this complexity requires fine-grained simulation, a level of detail and variety that makes reliance on human actors both prohibitively expensive and difficult to scale.

Recently, LLM-based SPs have emerged as a scalable and ethical alternative for clinical training (Na et al., 2025; Lee et al., 2025). Typically, prompt engineering is the predominant approach for simulating various diseases (Yosef et al., 2024; Lee et al., 2024; Du et al., 2025a; Liao et al., 2024). Building upon the prompting, recent studies further improve the realism of SPs in aspects, including narrative control (Du et al., 2025a; Qiu and Lan, 2024), emo-

\*Corresponding author.

tional responses (Lee et al., 2025; Bodonhelyi et al., 2025; Wang et al., 2024c), and hallucination mitigation (Li et al., 2024, 2025). However, the nature of prompting lacks the precision required for simulating SPs with the fine-grained, domain-specific deficits at varying levels of severity.

In this paper, we propose **StsPatient**, a novel framework that leverages Steering Vectors (SVs) (Rimsky et al., 2024; Huang et al., 2025) to achieve precise behavioral modulation. This approach intervenes directly on the model’s hidden states during inference and offers a more quantitative mechanism for control than discrete prompting. As illustrated in Figure 2, our method operates in two phases. First, we implement *Domain-specific SV Extraction*. We address the scarcity of clinical data by synthesizing contrastive pairs of impaired and healthy dialogues. Then, we extract a vector that represents the target deficit by computing the mean difference in embeddings between these pairs. Second, we introduce *Stochastic Token Modulation (STM)* to regulate the severity of the impairment. This mechanism draws inspiration from the probabilistic nature of synaptic transmission<sup>1</sup> in biological neural networks (Branco and Staras, 2009). We diverge from conventional SV approaches (Rimsky et al., 2024; Zou et al., 2023) that scale the injection coefficient and often suffer from instability. Instead, STM controls the deficit severity by adjusting the probability of applying the steering vector to each token. This probabilistic approach enables StsPatient to deliver stable and fine-grained simulation of cognitive impairments from mild to severe.

Extensive experiments validate the effectiveness of StsPatient, which consistently outperforms baselines, achieving an average improvement of +11.23% across all metrics. Crucially, our analysis demonstrates that StsPatient enables precise fine-grained simulation, surpassing the best baseline by +18.54% in severity controllability and yielding perceptibly distinct clinical presentations. Furthermore, we provide a comprehensive analysis of the model’s behavior, offering insights into the mechanics of steering-based modulation. Our main contributions are summarized as follows:

- We pioneer the fine-grained simulation of cognitively impaired SPs, addressing the need for scalable, effective solutions for clinical training.

---

<sup>1</sup>Synaptic strength is often regulated not by the amplitude of the signal, but by the probability of neurotransmitter release at the synapse.

- We present StsPatient, a novel SV-based method utilizing the STM to simulate domain-specific impairments at varying levels of severity.
- We experimentally demonstrate the superiority of StsPatient over existing SP baselines and provide an in-depth analysis to reveal its features.

## 2 Related Works

**LLM-based Standardized Patients (SPs).** LLM-based SPs have emerged as a scalable, ethical alternative to traditional human actors for clinical training (Na et al., 2025; Lee et al., 2025; Akkurt Yalcinturk and Dissiz, 2024). Unlike medical simulations for physical ailments, which focus on factual diagnosis (Du et al., 2025b; Reichenpfader and Denecke, 2024; Li et al., 2024), psychiatric simulations require sophisticated linguistic expression to ensure ecological validity (Chen et al., 2023b; Qiu and Lan, 2024; Wang et al., 2024a). However, these methods typically employ coarse-grained prompting to simulate traits such as negative thinking (Wang et al., 2024c) or emotional responsiveness (Lee et al., 2025; Louie et al., 2024). This reliance on static prompts crucially lacks the structural precision to isolate distinct cognitive domains or dynamically modulate impairment severity. Consequently, simply prompting is insufficient for capturing the heterogeneity of cognitive impairment (Bowie and Harvey, 2005). To this end, we introduce StsPatient that moves beyond static prompts to enable the fine-grained control necessary for modeling nuanced domain-specific deficits.

**LLM Role-Play & Steering Vector (SVs).** Role-playing methods broadly fall into parameter-tuning approaches (Song et al., 2020; Lu et al., 2024; Chen et al., 2025a) or parameter-frozen strategies (Zhao et al., 2024; Park et al., 2023; Huang et al., 2024). Within the latter paradigm, there is growing interest in employing SVs for general personas (e.g., mathematicians and chemists) (Wang et al., 2025; Potertǎž et al., 2025; Chen et al., 2025b). These methods integrate scaled SVs directly into the LLM’s internal hidden states to induce role-specific behaviors. However, relying on a scalar coefficient (i.e., magnitude) to modulate intensity presents significant challenges, often resulting in unstable simulations (Potertǎž et al., 2025; Chen et al., 2025b). Consequently, existing SV methods struggle to achieve the stable, fine-grained modulation required for simulating SPs with cognitive

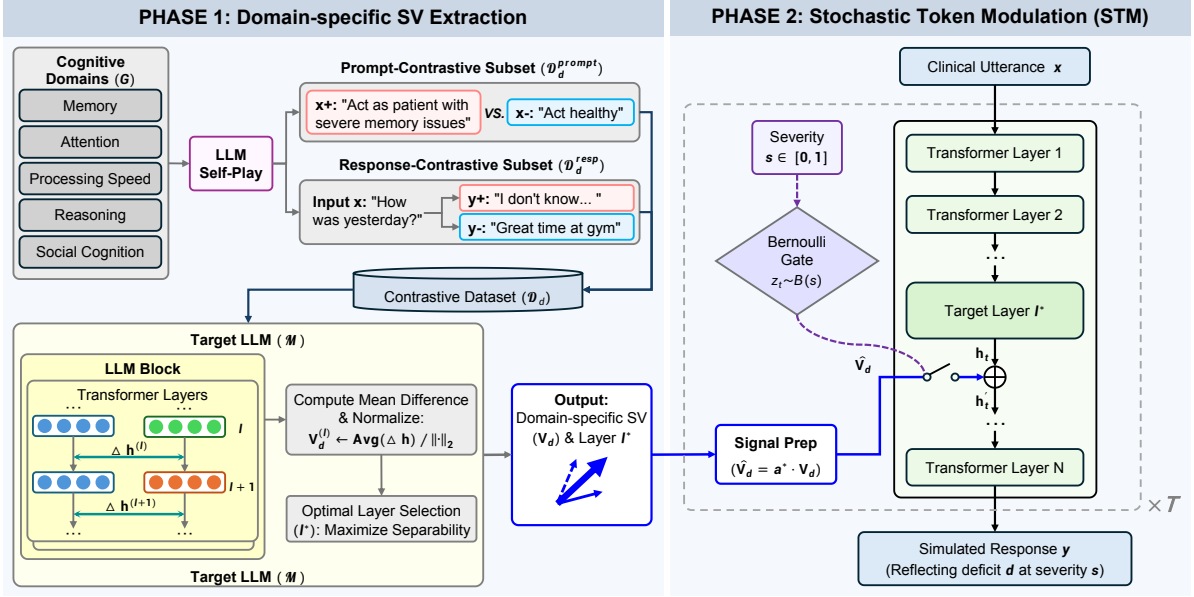


Figure 2: Overview of StsPatient. An LLM-generated contrastive dataset is utilized to extract a domain-specific SV ( $v_d$ ). Then, STM is applied during inference. Here, the modulation signal is probabilistically injected into token hidden states (the total  $T$  tokens) based on the severity level  $s$ , generating responses that reflect the deficit  $d$ .

impairment, a problem solved by our Stochastic Token Modulation (STM) mechanism.

### 3 Preliminary on Steering Vectors

SVs are originally proposed to guide LLM behavior at inference time toward abstract concepts such as happiness or honesty (Tan et al., 2024; Huang et al., 2025), with a few recent studies extending their application to LLM role-playing (Potertǎž et al., 2025; Chen et al., 2025b). Given an LLM  $\mathcal{M}$  and a target concept or role characteristic  $C$ , SVs are typically applied via the following process.

**Extracting SVs.** To obtain a SV  $v_C$ , existing methods typically compare the model’s internal states when processing positive examples (exhibiting the target concept) versus negative examples. This is most commonly done by computing the difference in their internal representations ( $h^+ - h^-$ ) at a specific layer (Rimsky et al., 2024; Zou et al., 2023), where the choice of layer is a hyperparameter. The differences are then aggregated into a final vector  $v_C$ , by averaging (Rimsky et al., 2024) or by taking the first principal component (Zou et al., 2023).

**Steering LLM via Scaled SVs.** Before steering an LLM’s output, the SV is scaled by a factor  $\alpha$  to control the steering strength. Then, the scaled SV is integrated into the LLM’s hidden states at specific layers, applied either to the last token (Zou et al., 2023) or every token (Rimsky et al., 2024) at the same layers where the SV is extracted. As

such, this effectively compels the LLM to align its outputs with the target concept  $C$  (e.g., inducing a tone of happiness). While this method outperforms LLM prompting for behavioral control (Rimsky et al., 2024), it suffers from significant instability due to high sensitivity to the scalar  $\alpha$  (Potertǎž et al., 2025; Chen et al., 2025b). Such sensitivity often results in ineffective or incoherent outputs, limiting its utility for fine-grained tasks like ours. Consequently, manual tuning of  $\alpha$  remains a prerequisite for successful deployment (Rimsky et al., 2024; Zou et al., 2023; Huang et al., 2025).

### 4 StsPatient for Cognitive Impairment

**Overview.** StsPatient is designed to simulate various cognitive domains  $G$  affecting the speech of impaired patients (McCutcheon et al., 2023). As illustrated in Figure 2, StsPatient first extract domain-specific SV using a contrastive dataset  $\mathcal{D}_d$ , which captures the distinct linguistic manifestations of a target deficit  $d$  versus healthy dialogue. Next, it employs STM to achieve precise control over impairment severity. During inference, StsPatient strategically modulates the internal states of the LLM  $\mathcal{M}$  to produce a response  $y$  reflecting a specific deficit  $d \in G$  at severity  $s$ . Refer to Algorithm 1 for pseudo-code.

## 4.1 Domain-specific SV Extraction

**Contrastive Dataset Construction.** Due to the scarcity and ethical sensitivities of clinical corpora with aligned healthy/impaired data, we utilize an LLM to synthesize pairs that isolate specific cognitive deficits. Specifically, we capture the deficit’s signature through two complementary channels: instructional intent and behavioral manifestations. This ensures comprehensive coverage of the target cognitive deficits, facilitating effective SV extraction. Formally, for a domain  $d \in G$ , the dataset  $\mathcal{D}_d$  is constructed as the union of prompt-contrastive and response-contrastive subsets:

$$\mathcal{D}_d = \mathcal{D}_d^{\text{prompt}} \cup \mathcal{D}_d^{\text{resp}}. \quad (1)$$

- **Prompt-Contrastive Subset** ( $\mathcal{D}_d^{\text{prompt}}$ ) targets the instructional intent of the LLM. Each sample  $s \in \mathcal{D}_d^{\text{prompt}}$  consists of a pair of contrastive system instructions ( $x^+, x^-$ ). Here,  $x^+$  explicitly prompts the LLM to simulate the SP with a target deficit domain (e.g., ‘Act as a patient with severe memory retrieval issues’), while  $x^-$  directs it to simulate a healthy control (e.g., ‘Act as a healthy individual with good memory’).
- **Response-Contrastive Subset** ( $\mathcal{D}_d^{\text{resp}}$ ) targets behavioral manifestations of the LLM. Each sample  $s' \in \mathcal{D}_d^{\text{resp}}$  comprises a pair of contrastive responses ( $y^+, y^-$ ) generated for the same clinical utterance  $x$  (e.g., ‘Did you have a good day yesterday?’). The positive response  $y^+$  reflects specific clinical symptoms (e.g., ‘I read some books at home... oh, in school?’), whereas the negative response  $y^-$  represents a clinically normal reply (e.g., ‘Yes, I had a great time at the gym’).

**SV Extraction.** Given the contrastive dataset  $\mathcal{D}_d$ , StsPatient adopts the established methodology (Wang et al., 2025; Tran et al., 2025) to extract the steering vector  $\mathbf{v}_d$ . Specifically, this is achieved by computing the mean difference in embeddings between paired contrastive samples. Formally, let  $\Delta \mathbf{h}^{(l)}(s)$  denote the difference vector of the LLM’s internal representations at layer  $l$  for a contrastive sample pair  $s \in \mathcal{D}_d$ . We then obtain the raw vector  $\tilde{\mathbf{v}}_d^{(l)}$  by averaging these differences:

$$\tilde{\mathbf{v}}_d^{(l)} = \sum_{s \in \mathcal{D}_d} \Delta \mathbf{h}^{(l)}(s) / |\mathcal{D}_d|. \quad (2)$$

Finally, to ensure consistent steering magnitude, we normalize this raw vector to unit length, yielding the final steering vector  $\mathbf{v}_d^{(l)}$  at layer  $l$ :

$$\mathbf{v}_d^{(l)} = \tilde{\mathbf{v}}_d^{(l)} / \left\| \tilde{\mathbf{v}}_d^{(l)} \right\|_2. \quad (3)$$

---

## Algorithm 1 StsPatient

---

- 1: **Input:** LLM  $\mathcal{M}$ , Domain  $d$ , Severity  $s \in [0, 1]$
  - 2: **Domain-specific SV Extraction**
  - 3: Construct dataset  $\mathcal{D}_d$  for  $d$  via LLM.
  - 4: Extract raw SVs  $\mathbf{v}_d^{(l)}$  from  $\mathcal{M}$  using Eq. (3).
  - 5: Layer selection  $l^*$  via Eq.(4) and obtain SV  $\mathbf{v}_d = \mathbf{v}_d^{(l^*)}$ .
  - 6: **Stochastic Token Modulation**
  - 7: Employ line search for  $\alpha^*$  and obtain  $\hat{\mathbf{v}}_d = \alpha^* \cdot \mathbf{v}_d$ .
  - 8: For each generation step  $t = 1 \dots T$ : sample gate  $z_t \sim \mathcal{B}(s)$  and update hidden state  $\mathbf{h}_t^{(t^*)}$  via Eq. (5).
- 

While layer selection is often manual (Zou et al., 2023; Huang et al., 2025), we aim to find the optimal layer  $l^*$  automatically. Specifically, we focus on middle-to-late layers where semantic features are concentrated (Rimsky et al., 2024; Zou et al., 2023). Within this range, we identify the optimal intervention layer  $l^*$  by maximizing the distance between the centroids of positive and negative sample embedding clusters ( $\boldsymbol{\mu}_{d,+}^{(l)}$  and  $\boldsymbol{\mu}_{d,-}^{(l)}$ ):

$$\mathbf{v}_d = \mathbf{v}_d^{(l^*)}, \text{ where } l^* = \operatorname{argmax}_l \left\| \boldsymbol{\mu}_{d,+}^{(l)} - \boldsymbol{\mu}_{d,-}^{(l)} \right\|_2. \quad (4)$$

## 4.2 Stochastic Token Modulation (STM)

**Motivation.** Conventional SV modulation methods control intervention strength via a scalar  $\alpha > 0$ , adding the scaled vector  $\alpha \cdot \mathbf{v}_d$  to the embeddings of every token (Rimsky et al., 2024). However, the relationship between  $\alpha$  and the behavioral output of  $\mathcal{M}$  is highly non-linear and unstable (Potertaz et al., 2025; Chen et al., 2025b), making it difficult to achieve the fine-grained simulations of SPs with varying deficit severity levels. To this end, we introduce STM, a simple yet effective method inspired by stochastic synaptic transmission in neural systems (Branco and Staras, 2009). Instead of altering the vector’s scalar  $\alpha$ , we control the severity of the simulated impairment by tuning the *probability* that the intervention is applied to any given token. **Modulation Signal.** Instead of relying on manually tuning  $\alpha$  to regulate deficit severity, we simplify the role of  $\alpha$ : it serves solely to ensure the modulation signal is sufficiently robust to manifest the cognitive impairment. To determine this value, we employ an automated process. Specifically, we employ a line search over the interval [1,6] with a step size of 0.1 to select the minimum scalar  $\alpha^*$  that satisfies two criteria: (1) Effectiveness, ensuring the deficit is observable (verified via an LLM-as-a-Judge and contrastive dataset); and (2) Integrity, ensuring the model’s generation quality remains

intact (e.g., avoiding gibberish or incoherent outputs). The search range  $[1, 6]$  is motivated by empirical bounds in Representation Engineering (Zou et al., 2023). Since  $\mathbf{v}_d$  is unit-normalized (Eq. 3),  $\alpha$  directly controls the perturbation magnitude. In practice, values below 1 are too weak and get absorbed by the model’s inherent robustness, whereas values above 6 catastrophically degrade linguistic coherence (see Appendix C.8 for a boundary analysis). Likewise, the optimal layer  $l^*$  is automatically determined via Eq. 4 without any manual tuning. In summary, all key parameters ( $\alpha^*$  and  $l^*$ ) are automatically selected, leaving only the severity variable  $s$  as the user-facing control knob (see Appendix C.7 for a full overview). This streamlines the parameter selection process while guaranteeing signal efficacy. Finally, our modulation signal is defined as  $\hat{\mathbf{v}}_d = \alpha^* \cdot \mathbf{v}_d$ .

**Stochastic Modulation.** To simulate SPs exhibiting varying severities of cognitive impairment, we parameterize the severity level as a continuous variable  $s \in [0, 1]$ , where higher values indicate greater severity. In our implementation,  $s$  is mapped directly to the intervention probability within the STM mechanism. Statistically, a larger  $s$  results in a higher proportion of tokens being modulated, thereby enabling the simulation of distinct severity levels. To achieve this, during inference, at each token generation step  $t$ , we sample a binary gate  $z_t$  from a Bernoulli distribution (denoted as  $\mathcal{B}$ ). The gate sampling and the subsequent hidden state update at the selected layer  $l^*$  are formulated as:

$$z_t \sim \mathcal{B}(s), \quad \mathbf{h}'_t^{(l^*)} = \mathbf{h}_t^{(l^*)} + z_t \cdot \hat{\mathbf{v}}_d, \quad (5)$$

where  $z_t$  acts as a stochastic token-level gate. Notably,  $s$  should be understood as the algorithmic intervention probability, that is, the chance of injecting the SV at each token step. It does not constitute a direct linear mapping to standardized clinical scores such as MMSE. While different cognitive domains may exhibit varying sensitivity curves, what matters most for an educational simulator is monotonic controllability, and our method satisfies this requirement robustly (validated in §5.3 and discussed further in Appendix C.9). As such, STM facilitates fine-grained control across a continuous severity spectrum, ranging from the occasional lapses typical of Mild Cognitive Impairment (low  $s$ ) to the pervasive deficits characteristic of severe Alzheimer’s Disease (high  $s$ ).

## 5 Experiment

This section presents a comprehensive evaluation involving both human and LLM evaluators. §5.2 reports the overall performance, while §5.3–§5.5 provide a detailed analysis of StsPatient’s components. Implementation details, case studies, and experiments on backbone generalization are provided in **Appendices A, B, and C.6**, respectively.

### 5.1 Experimental Setup

**Overview.** We evaluate SPs through open-ended dialogue with therapists, aligning with standard practices (Wind et al., 2004; Jiang et al., 2023; Lee et al., 2025; Louie et al., 2024; Lima et al., 2025). To ensure a robust assessment, we employ both human and LLM-based therapists<sup>2</sup>. During these sessions, therapists screen for cognitive deficits by asking probing questions regarding daily life (e.g., *how was your day yesterday?*). Each dialogue is standardized to conclude automatically after 10 turns. Crucially, we also incorporate baseline simulators as *healthy controls* for comparison. All methods are compared on the same set of SP profiles, therapist prompts, and stopping rules.

**User Profiles.** We first sample 100 real outpatient cases from the MTSamples collection<sup>3</sup>. Each case is manually stripped of any diagnostic information, such as cognitive or physical conditions, to isolate only the patient’s basic profile. For each case, we construct six types of cognition profiles: one cognitively healthy control and five impaired domains (McCutcheon et al., 2023) that affect patients’ speech (i.e., *Memory, Attention, Processing Speed, Reasoning & Problem Solving*, and *Social Cognition*). This yields 600 distinct profiles, generating a total of 500 impaired SP-therapist dialogues (per SP simulation method) and 100 healthy control dialogues (simulated via GPT-5) for evaluation.

**Baselines.** As there are no existing methods for simulating cognitive impairment, we adapt baselines from two sources: 1) **LLM-based SPs** designed for psychiatric contexts, specifically PATIENT- $\psi$  (Wang et al., 2024c) and Roleplay-doh (Louie et al., 2024). Additionally, we employ representative 2) **LLM role-playing methods**<sup>4</sup>, including Direct Prompt (Wang et al., 2024b; He et al., 2025) and Role Vectors (PotertĀZ

<sup>2</sup>Detailed in Appendix A.3.2 and Appendix A.3.3

<sup>3</sup><https://www.mtsamples.com>

<sup>4</sup>We exclude parameter-tuning methods due to the scarcity of high-quality therapist-patient interaction datasets.

Backbone	Method	CDC $\uparrow$		IDI $\downarrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>							
GPT-5	Healthy Control	0.96	0.95	0.04	0.05	–	–
	Direct Prompt	0.54	<u>0.68</u> <sup>†</sup>	0.47	0.42	3.32	3.40
	PATIENT- $\psi$	0.50	0.60	0.52	0.48	<u>3.83</u> <sup>†</sup>	<u>3.96</u> <sup>†</sup>
	Roleplay-doh	0.58	<u>0.68</u> <sup>†</sup>	<u>0.44</u> <sup>†</sup>	<u>0.38</u> <sup>†</sup>	3.78	3.72
Qwen3-8B	Direct Prompt	0.47	0.56	0.64	0.56	3.18	3.21
	PATIENT- $\psi$	0.53	0.64	0.60	0.56	3.65	3.60
	Roleplay-doh	0.56	0.62	0.57	0.50	3.61	3.55
	Role Vectors	<u>0.61</u> <sup>†</sup>	0.64	0.45	0.40	3.73	3.70
	<b>StsPatient</b>	<b>0.71</b> <sub>+16.39%</sub>	<b>0.84</b> <sub>+23.53%</sub>	<b>0.38</b> <sub>+13.64%</sub>	<b>0.32</b> <sub>+15.79%</sub>	<b>3.96</b> <sub>+3.39%</sub>	<b>4.08</b> <sub>+3.03%</sub>
<i>Panel B: Human Therapist</i>							
GPT-5	Healthy Control	0.92	0.94	0.07	0.05	–	–
	Direct Prompt	0.50	0.60	0.50	0.40	3.23	3.41
	PATIENT- $\psi$	0.50	0.58	0.55	0.44	<u>3.91</u>	3.80
	Roleplay-doh	0.56	<u>0.70</u> <sup>†</sup>	0.46	0.36	3.76	<u>3.88</u> <sup>†</sup>
Qwen3-8B	Direct Prompt	0.44	0.56	0.63	0.52	2.90	3.10
	PATIENT- $\psi$	0.50	0.60	0.61	0.48	3.54	3.72
	Roleplay-doh	0.46	0.64	0.58	0.48	3.60	3.56
	Role Vectors	<u>0.62</u> <sup>†</sup>	0.68	<u>0.42</u>	<u>0.34</u> <sup>†</sup>	3.71	3.69
	<b>StsPatient</b>	<b>0.68</b> <sub>+9.68%</sub>	<b>0.82</b> <sub>+17.14%</sub>	<b>0.40</b> <sub>+4.76%</sub>	<b>0.28</b> <sub>+17.65%</sub>	<b>3.94</b> <sub>+0.77%</sub>	<b>4.23</b> <sub>+9.02%</sub>

Table 1: Comprehensive evaluation of simulation performance when interacting with LLM (GPT-5) and Human Therapists. Evaluations are conducted by both LLM (o4-mini) and human judges. We bold the **Best** and underline the second-best results. Statistically significant improvements ( $p < 0.05$ ) of StsPatient compared to the second-best are marked by <sup>†</sup>. Subscripts denote the relative improvement over the second-best baseline.

et al., 2025), which leverages prompts and steering vectors to control role-playing behavior, respectively. See Appendix A.2 for reproduction details.

**Evaluation Metrics.** Following common practice in the field (Qiu and Lan, 2024; Du et al., 2025b), we provide SP-Therapist dialogue to evaluators to assess the performance. We employ the following two dimensions (detailed in Appendix A.3.5).

**1) Training Effectiveness.** SPs’ practical utility is evaluated referring to the Maastricht Assessment of Simulated Patients (MaSP) scale (Wind et al., 2004), in line with established practice in clinical simulation (Lee et al., 2020; Roux et al., 2025). The MaSP scale values that the SP not only maintains realistic patient characteristics but also provides a constructive environment for training clinical staff. We use the following two metrics, assessed exclusively via human evaluation following Wang et al. (2024c), as current LLM evaluators are insufficient for delivering the necessary scoring precision<sup>5</sup>.

- **Authenticity Score** (Auth  $\in [0, 5]$ ) quantifies the realism of the SP’s portrayal. Following the MaSP, we use a 5-point Likert scale.
- **Training Score** (Tra  $\in [0, 5]$ ) measures the edu-

cational value of SPs (facilitating clinical skill acquisition). Similar to the Auth score, we also use a 5-point Likert scale.

**2) Domain-specific Simulation Fidelity.** Following Qiu and Lan (2024); Du et al. (2025b), we assess whether the SP accurately manifests its assigned impairment. Evaluators analyze the dialogue and identify up to two perceived domains from a candidate set comprising the five target cognitive domains and a “Healthy” class.

- **Cognitive Domain Consistency** (CDC  $\in [0, 1]$ ) measures the accuracy with which an SP portrays its assigned impairment domain. Given the  $i$ -th SP-therapist dialogue with an assigned domain  $d_i$  and an evaluator-identified domain set  $E_i$ , CDC is the average recall rate:  $\text{CDC} = \frac{1}{N} \sum_i \mathbf{1}(d_i \in E_i)$ , where  $\mathbf{1}(\cdot)$  denotes the indicator function and  $N$  is the number of dialogues.
- **Irrelevant Domain Inconsistency** (IDI  $\in [0, 1]$ ) measures the hallucination of unassigned domains:  $\text{IDI} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(|E_i \setminus \{d_i\}| > 0)$ , where  $|E_i \setminus \{d_i\}|$  denotes the cardinality of the  $E_i$  after excluding the assigned domain  $d_i$ . Notably, IDI and CDC do not necessarily sum to 1.

<sup>5</sup>Refer to Appendix A.3.5 for analysis

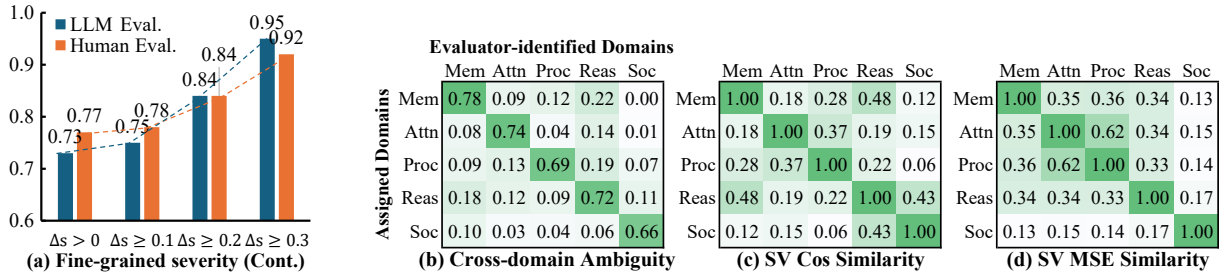


Figure 3: (a) Severity analysis (Continuous) via ISC scores averaged across all domains. (b): Cross-domain ambiguity. High diagonal values indicate strong domain specificity. (c-d): SVs similarity.

## 5.2 Overall Simulation Performance

Table 1 presents the performance of all methods, employing LLM & Human Therapists and Evaluators. Key observations are detailed below.

**StsPatient consistently outperforms baseline methods across all metrics.** According to Table 1, all methods demonstrate effectiveness compared to the Healthy Control. Despite their efficacy, StsPatient stands out by achieving substantial gains in both role consistency (CDC, IDI) and educational value (Auth, Tra). It yields average improvements over the best baseline of 17.34% and 3.21% in the LLM-therapist setting, and 12.31% and 4.90% in the human-therapist setting. Crucially, StsPatient enables the open-source Qwen3-8B (Yang et al., 2025) to outperform even GPT-5-based baselines. This highlights the potential for locally deployed, privacy-preserving clinical simulations.

**StsPatient demonstrates consistent superiority in dual-evaluation setting,** comprising both human- and LLM-based judges. According to Table 1, both evaluation methods demonstrate high accuracy on the Healthy Control baseline, successfully distinguishing healthy subjects via high CDC and low IDI scores. Moreover, StsPatient maintains a consistent performance advantage over all other methods. We further substantiate the reliability of our evaluation via Krippendorff’s alpha  $\mathcal{K}$ : human evaluators show high agreement on domain fidelity<sup>6</sup> ( $\mathcal{K} = 0.84$ ), Auth (0.71), and Tra (0.75). Furthermore, human-LLM alignment on domain fidelity is 0.67, suggesting LLM-based evaluation serves as a scalable extension of human assessment.

## 5.3 Fine-grained Simulation: Severity

We assess the severity controllability of StsPatient, verifying whether generated symptoms exhibit perceptible distinctions across severity levels. Our analysis covers both discrete (Mild/Moderate/Severe) and continuous severity settings.

<sup>6</sup>Agreement on evaluator-identified domains for CDC/IDI.

**Setup.** In the discrete severity setting, we control the severity of simulated impairments for SV-based methods by sampling their respective parameters from fixed intervals. StsPatient uses  $s \in [0.1, 0.3]$  (Mild),  $[0.4, 0.6]$  (Moderate), and  $[0.7, 0.9]$  (Severe). Note that we intentionally introduce gaps between these intervals (avoiding boundary values) to maximize the distinguishability between severity levels. These non-overlapping bins mirror how a deployed clinical tool would present unambiguous difficulty tiers to end-users, and the gap-free continuous evaluation below further validates that our controllability does not depend on such separation. Similarly, we sample the scaling factor  $\alpha$  for Role Vector. However, since  $\alpha$  is not naturally normalized to  $[0, 1]$ , we configure domain-specific sampling intervals, detailed in Appendix A.2.5. Other baselines rely on explicit prompt conditioning to specify the severity level. Additionally, in the continuous severity setting, we simulate fine-grained severity variations by sampling  $s$  directly from the continuous interval  $(0, 1]$ .

**Metric.** Recognizing the challenges in pinpointing exact severity levels, especially in the continuous setting, we assess simulation controllability via a comparative ranking protocol, inspired by Chen et al. (2023a, 2024). To this end, we introduce the *Impairment Severity Consistency* ( $ISC \in [0, 1]$ ) score. Specifically, for each evaluation turn, evaluators are presented with three generated dialogues representing different severity levels and are tasked with ranking them by relative impairment intensity. The ISC score is then computed by measuring the alignment accuracy between evaluator rankings and the ground-truth (Mild < Moderate < Severe).

**StsPatient demonstrates superiority for fine-grained simulation in the discrete severity setting.** According to Table 2, StsPatient consistently dominates across all five cognitive domains, achieving an average ISC of 0.94 (LLM) and 0.92 (Human). These findings confirm that modulating in-

Method	Evaluator	Mem	Att	Proc	Reas	Soc	Avg
Direct Prompt (GPT-5)	LLM	0.51	0.44	0.29	0.49	0.33	0.41
	Human	0.42	0.43	0.35	0.50	0.30	0.40
Role Vectors (Qwen3-8B)	LLM	0.79	0.71	0.83	0.80	0.72	0.77
	Human	0.80	0.75	0.76	0.79	0.90	0.80
<b>StsPatient</b> (Qwen3-8B)	LLM	<b>0.93</b>	<b>0.92</b>	<b>0.97</b>	<b>0.97</b>	<b>0.91</b>	<b>0.94</b>
	Human	<b>0.94</b>	<b>0.89</b>	<b>0.90</b>	<b>0.93</b>	<b>0.94</b>	<b>0.92</b>

Table 2: Severity analysis (Discrete) via ISC scores of five domains: **Memory**, **Attention**, **Processing Speed**, **Reasoning & Problem Solving**, **Social Cognition**.

Domain	Generated Interpretation
<i>No SV</i>	◇ represents a placeholder for a secret code in a spy mission.
<b>Mem.</b>	blurred out.
<b>Att.</b>	distracted.
<b>Proc.</b>	◇ is... I'm not sure, maybe a code.
<b>Reas.</b>	◇ represents whatever you want it to represent, because it is a placeholder.
<b>Soc.</b>	F**K YOU, I'm not gonna answer that.

Table 3: SV interpretability analysis using the query ‘What does ◇ represent?’ **Warning! Harmful Contents.**

tervention probability via STM produces far more distinguishable severity levels than either scalar scaling or explicit linguistic instructions.

**StsPatient enables continuous severity control, where adjustments in  $s$  yield perceptibly distinct clinical presentations.** We analyze simulation outcomes across different severity values  $s$ : We conduct a pairwise comparison of simulations generated with distinct  $s$  values, categorizing the pairs based on their magnitude difference,  $\Delta s \in \{0.1, 0.2, 0.3\}$ . For each  $\Delta s$  category, we assess whether the resulting simulations exhibit a perceptibly distinct severity of impairment. As shown in Figure 3(a), the ISC scores exhibit a monotonic increase as the interval  $\Delta s$  widens. This confirms that our STM functions as a reliable, continuous proxy for controlling symptom intensity.

#### 5.4 Analysis on Domain-Specific SVs

Domain-specific results in **Appendix C.3** align with our main observations from Section 5.2. We conduct an in-depth analysis of the characteristics and success of StsPatient by examining its SVs.

**Why is StsPatient effective? – Its SVs encode high-level cognitive deficits into the LLM’s latent representation, ensuring clinically accurate symptom manifestation.** While our STM ensures

Variant	CDC ↑	IDI ↓	Auth ↑	Tra ↑	ISC ↑
w/ All Token	0.82	<b>0.30</b>	<b>4.04</b>	3.98	0.54
w/ Last Token	0.43	0.70	3.12	3.20	0.41
w/ $\mathcal{D}_{resp}$	0.58	0.60	3.36	3.19	0.62
w/ $\mathcal{D}_{prom}$	0.70	0.44	3.53	3.24	0.71
StsPatient	<b>0.84</b>	0.32	3.96	<b>4.08</b>	<b>0.77</b>

Table 4: Ablation via LLM therapist & human evaluation.

fine-grained severity control, the core clinical fidelity relies on the semantic quality of the underlying SVs. We validate their interpretability using the patch-scoping framework (Ghandeharioun et al., 2024). In particular, given an input query ‘What does ◇ represent?’, we inject the domain-specific SV into the residual stream at the token position of the abstract symbol (◇). Table 3 reveals clear semantic alignments: Memory evokes fragmentation (“blurred out”), Attention triggers conceptual collapse (“distracted”), Processing Speed induces hesitation, Reasoning leads to a circular explanation that mimics logical structure but lacks actual substance, and Social Cognition manifests as disinhibition. These findings confirm that our SVs are semantic instructions that force LLM-based SPs to internally replicate specific cognitive impairments.

**How effective is StsPatient at simulating distinct cognitive domains? – It outperforms baselines, yet identifiable behavior overlaps persist between correlated domains.** StsPatient outperforms all baselines in domain-specific metrics (CDC and IDI, Table 1). To further probe the boundaries of this capability, we construct a confusion matrix measuring cross-domain ambiguity: the rate at which an SP simulating a target domain is misidentified as exhibiting deficits in another. As shown in Figure 3(b), strong diagonal values demonstrate that StsPatient maintains precise control over individual domains. However, we observe notable leakage (0.22) where *Mem.* deficits are perceived as *Reas.* impairments, indicating behavior overlaps of the LLM post-modulation. While StsPatient achieves SOTA, enhancing the separation of correlated domains remains for future optimization.

**Why do these behavioral overlaps occur? – The domain-specific SVs exhibit minor semantic entanglement, necessitating future refinement.** We investigate the geometric properties of these SVs, and quantify their relationship using Cosine similarity and a normalized Mean Squared Error (nMSE) similarity, computed as  $1/(1 + \text{MSE})$ . As shown

in Figure 3(c-d), while generally distinct, these vectors exhibit minor semantic overlaps, which correlate with the leakage observed in Figure 3(b). It offers an intrinsic explanation for the reduced distinctiveness between these domains. It also implies that robust multi-domain simulation remains a challenge, as combining multiple SVs often results in unpredictable interactions (Bartoszcze et al., 2025). Appendix C.5 offers an analysis of this challenge.

## 5.5 Ablation Studies

We consider the following ablations and summarize the results in Table 4.

- w/ All Token modulates all tokens uniformly (Rimsky et al., 2024) via  $s = 1.0$ . It achieves performance comparable to StsPatient but fails to provide fine-grained simulation (low ISC).
- w/ Last Token only modulates the last token (Zou et al., 2023). Performance is suboptimal, as cognitive deficits are complex, requiring persistent injection rather than momentary modulation.
- w/  $\mathcal{D}_{\text{prom}}$  extracts SVs from  $\mathcal{D}_{\text{prompt}}$ . This variant neglects the behavioral manifestations in responses, leading to performance degradation.
- w/  $\mathcal{D}_{\text{resp}}$  extracts SVs from  $\mathcal{D}_{\text{resp}}$ . It fails to capture instructional intent, suggesting  $\mathcal{D}_{\text{prom}}$  contributes more to domain fidelity in our setting.

## 6 Conclusion

In this paper, we propose StsPatient, first framework dedicated to the fine-grained simulation of cognitively impaired SPs. By leveraging contrastive pairs from both instructions and responses, our method effectively captures domain-specific features and allows for flexible adaptation across various deficits. Furthermore, we introduce STM to enable precise severity control. Extensive experiments confirm that StsPatient significantly outperforms baselines in authenticity and controllability. We believe that our work provides a foundation for developing high-fidelity clinical simulation.

### Limitations

Despite the following limitations, we believe this work establishes a solid foundation for quantitative behavior control in medical AI. We hope these insights will guide the future development of high-fidelity clinical simulators.

**Multiple LLM Backbones.** A core value of our work is enabling privacy-preserving clinical simulations via local deployment, eliminating reliance

on proprietary APIs. That said, our main validation is currently centered on a specific class of open-source models (Qwen). While Appendix C.6 experiments suggest broader applicability (Llama), extensive verification across diverse architectures and larger-scale foundation models is still needed. The transferability of stochastic steering to different latent representations remains a subject for future study. Despite this, our findings conclusively demonstrate that StsPatient offers a robust, privacy-safe solution, based on Qwen-3, for cognitive impairment simulation.

**Multiple Evaluation Scenarios.** Our current simulation scenarios are restricted to open-ended daily conversations and cognitive screening dialogues, aligning with standard practices in the field (Wind et al., 2004; Jiang et al., 2023; Lee et al., 2025). However, we have not yet integrated rigorous clinical diagnostic workflows, such as structured medical history taking or standardized neuropsychological testing protocols. Consequently, while StsPatient effectively manifests linguistic deficits during casual interactions, its fidelity in adhering to complex medical procedures remains to be validated. Future work should focus on constructing a comprehensive, multi-scenario evaluation benchmark to rigorously assess SP methods across a broader spectrum of clinical interactions.

**Multi-modal Simulation.** StsPatient operates strictly within a unimodal text-based interface. Real-world cognitive impairment manifests through multiple modalities. These include acoustic features like speech rate (Fukuda et al., 2022) and prosody (Beltrami et al., 2016) as well as visual cues (Pan et al., 2022). Relying solely on text limits the ability to simulate non-verbal diagnostic markers. Future iterations should consider integrating multimodal capabilities to enhance realism.

**Multi-Domain Simulation.** StsPatient primarily focuses on isolating specific cognitive domains to ensure precise controllability. However, real-world patients often present with complex, overlapping deficits rather than isolated symptoms. Although we demonstrate effectiveness in single-domain scenarios, precisely modeling multi-domain comorbidities remains a significant challenge (Bartoszcze et al., 2025). While we provide a preliminary analysis in the Appendix C.5, we have not yet deeply explored the complex interference patterns that may arise when combining multiple steering vectors. Future efforts should robustly model these interactions for high-fidelity comorbidity simulation.

## Ethics Statement

**Data Sources and Privacy.** Our experimental framework is designed to prioritize data privacy and ethical safety. We extract steering vectors solely from fully synthetic therapist-patient dialogues generated by LLMs. No non-public patient records are collected or processed during the vector construction process. For evaluation and scenario initialization, we utilize publicly available and de-identified medical transcription samples from the MTSamples website<sup>7</sup>. These serve as reference templates to construct brief patient profiles, such as demographics and presenting complaints, rather than to reconstruct any real individual’s identity. MTSamples indicates that its reports are user-contributed educational examples where names and dates have been modified or removed to preserve confidentiality. Beyond the inherent de-identification of the source, we apply an additional sanitization step to remove or mask any residual identifiers like personal names, locations, and contact information before processing. We do not attempt re-identification. Furthermore, we do not release any raw text from MTSamples. Only synthetic dialogues, derived templates, and aggregate statistics are reported.

**Use of Third-Party Models.** Certain components of our pipeline utilize proprietary LLMs. In these instances, inputs consist exclusively of synthetic dialogues and de-identified templates. We strictly exclude any potentially identifying information. All prompts and generated content are stored and analyzed in an anonymized format.

**Human Participants.** Human therapists and evaluators consist of adult volunteers recruited from the authors’ research group and academic colleagues. Participation is entirely voluntary, and participants retain the right to withdraw at any time. They receive task instructions and background training on cognitive impairment simulation using academic surveys (McCutcheon et al., 2023) and public educational resources<sup>8</sup>. We do not collect sensitive personal data from participants beyond their specific study responses. These are recorded without direct identifiers.

**Safety, Misuse, and Intended Use.** StsPatient is

intended for research on the controllable simulation of cognitively impaired SPs and for training purposes within controlled environments. It is explicitly not designed for clinical decision-making or direct patient care. Given that the model simulates cognitive impairment and behavioral disinhibition, it may occasionally generate unsafe or offensive content. This risk is inherent to simulations of severe social cognitive impairment and often proves difficult to eliminate entirely. Consequently, we advise against deployment without robust safeguards. Essential measures include content filtering, human oversight, and clear user warnings. For our human-facing experiments, we implement a supplementary keyword safety filter. If an output contains terms from a predefined blacklist of potentially harmful expressions, StsPatient automatically masks the content with a neutral placeholder and flags it for review. This measure reduces the risk of participant exposure to offensive material. However, we acknowledge that keyword filtering offers only partial mitigation and does not guarantee the complete elimination of unsafe content. In a post-hoc audit of our experimental logs, we observe that operating the stochastic modulation probability at a severity level below 0.8 is sufficient to avoid any flagged outputs in our study setting. To further investigate robust mitigation for higher severities, we extract all samples with levels  $s > 0.5$  from Section 5.3 and find that employing an additional LLM-based safety filter successfully intercepts every instance of harmful content. We therefore recommend combining LLM-based filtering with conservative operating regimes for safer human training scenarios. We note that this observation may not generalize to all prompts, domains, or model variants. We disclose these potential risks transparently and advocate for responsible use consistent with ethical clinical training standards.

## Acknowledgments

This research is supported by the Singapore Ministry of Health’s National Medical Research Council under its NMRC NIC Healthy and Meaningful Longevity (HML) Cognition Grant (NMRC Project No. MOH-001838) and the National Natural Science Foundation of China (No. U25B201508, No. 62272330, and No.U24A20328). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of MOH/NMRC.

<sup>7</sup><https://www.mtsamples.com>

<sup>8</sup><https://www.hkada.org.hk/what-is-dementia> and [https://9abfea27-4ae1-43da-94a8-8f7122d482ae.filesusr.com/ugd/4b4c9a\\_4dcec2dafc184102a94ec55415dcf12a.pdf](https://9abfea27-4ae1-43da-94a8-8f7122d482ae.filesusr.com/ugd/4b4c9a_4dcec2dafc184102a94ec55415dcf12a.pdf)

## References

- Aydan Akkurt Yalcinturk and Melike Dissiz. 2024. The effect of standardized patient simulation method on getting nursing students to gain care providing skill for the individual diagnosed with dementia: A randomized controlled trial. *Educational Gerontology*, 50(3):229–239.
- Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *aamc. Academic medicine*, 68(6):443–51.
- Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*.
- Daniela Beltrami, Laura Calzà, Gloria Gagliardi, Enrico Ghidoni, Norina Marcello, Rema Rossini Favretti, and Fabio Tamburini. 2016. Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2086–2093, Portorož, Slovenia. European Language Resources Association (ELRA).
- Anna Bodonhelyi, Christian Stegemann-Philipps, Alessandra Sonanini, Lea Herschbach, Marton Szep, Anne Herrmann-Werner, Teresa Festl-Wietek, Enkelejda Kasneci, and Friederike Holderried. 2025. Modeling challenging patient interactions: Llms for medical communication training. *Preprint*, arXiv:2503.22250.
- Christopher R Bowie and Philip D Harvey. 2005. Cognition in schizophrenia: impairments, determinants, and functional importance. *Psychiatric Clinics*, 28(3):613–633.
- Tiago Branco and Kevin Staras. 2009. The probability of neurotransmitter release: variability and feedback control at single synapses. *Nature Reviews Neuroscience*, 10(5):373–383.
- Rosa Burgos, Irene Bretón, Emanuele Cereda, Jean Claude Desport, Rainer Dziewas, Laurence Genton, Filomena Gomes, Pierre Jésus, Andreas Leischker, Maurizio Muscaritoli, and 1 others. 2018. Es-pen guideline clinical nutrition in neurology. *Clinical nutrition*, 37(1):354–396.
- Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025a. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 755–773.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.
- Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023a. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 8506–8520.
- Runjin Chen, Andy Ardit, Henry Sleight, Owain Evans, and Jack Lindsey. 2025b. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.
- Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023b. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.
- Valerie T Cotter, Danetta H Sloan, Daniel L Scerpella, Kelly M Smith, Martha Abshire Saylor, and Jennifer L Wolff. 2025. Feasibility of using simulation to evaluate implementation fidelity in an advance care planning pragmatic trial. *American Journal of Hospice and Palliative Medicine®*, page 10499091241282087.
- Department of Social Work and Social Administration, The University of Hong Kong. 2024. [Brain live: Cognitive impairment and rehabilitation materials](#). Accessed: 2025-12-11.
- Zhuoyun Du, LujieZheng LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. 2025a. LLMs can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306, Vienna, Austria. Association for Computational Linguistics.
- Zhuoyun Du, LujieZheng LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. 2025b. Llms can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Chukwuka Elendu, Dependable C Amaechi, Alexander U Okatta, Emmanuel C Amaechi, Tochi C Elendu, Chiamaka P Ezech, and Ijeoma D Elendu. 2024. The impact of simulation-based training in medical education: A review. *Medicine*, 103(27):e38813.
- Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer’s disease with a picture description task. *Neurological sciences*, 26(4):243–254.

- Meiko Fukuda, Ryota Nishimura, Maina Umezawa, Kazumasa Yamamoto, Yurie Iribe, and Norihide Kitaoka. 2022. [Elderly conversational speech corpus with cognitive impairment test and pilot dementia detection experiment using acoustic characteristics of speech in Japanese dialects](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1016–1022, Marseille, France. European Language Resources Association.
- Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.
- Dimitris Gkoumas, Adam Tsakalidis, and Maria Liakata. 2023. [A digital language coherence marker for monitoring dementia](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16021–16034, Singapore. Association for Computational Linguistics.
- Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaying Xu, Wenqiang Liu, and Mengling Feng. 2025. [Crab: A novel configurable role-playing LLM with assessing benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052, Vienna, Austria. Association for Computational Linguistics.
- Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. [Concept – an evaluation protocol on conversational recommender systems with system-centric and user-centric factors](#). *Preprint*, arXiv:2404.03304.
- Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. 2025. [Cross-model transferability among large language models on the platonic representations of concepts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3686–3704, Vienna, Austria. Association for Computational Linguistics.
- Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640.
- René S Kahn and Richard SE Keefe. 2013. Schizophrenia is a cognitive illness: time for a change in focus. *JAMA psychiatry*, 70(10).
- Ruhul Amin Khalil, Kashif Ahmad, and Hazrat Ali. 2025. Redefining elderly care with agentic ai: Challenges and opportunities. *arXiv preprint arXiv:2507.14912*.
- Jihyun Lee, Hyungsin Kim, Kwan Hoon Kim, Daeun Jung, Tanisha Jowsey, and Craig S Webster. 2020. Effective virtual patient simulators for medical communication training: a systematic review. *Medical education*, 54(9):786–795.
- Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung E. Suh, Eun-mee Kim, and Hajin Lim. 2025. [Adaptive-VP: A framework for LLM-based virtual patients that adapts to trainees’ dialogue to facilitate nurse communication training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2319–2352, Vienna, Austria. Association for Computational Linguistics.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.
- Yanzeng Li, Cheng Zeng, Jinchao Zhang, Jie Zhou, and Lei Zou. 2025. [Meddit: A knowledge-controlled diffusion transformer framework for dynamic medical image generation in virtual simulated patient](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 11082–11085. International Joint Conferences on Artificial Intelligence Organization. Demo Track.
- Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.
- Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.
- Maria R Lima, Amy O’Connell, Feiyang Zhou, Alethea Nagahara, Avni Hulyalkar, Anura Deshpande, Jesse Thomason, Ravi Vaidyanathan, and Maja Matarić. 2025. Promoting cognitive health in elder care with large language model-powered socially assistive robots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22.
- Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of

- all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840.
- Lauren McCollum and Jason Karlawish. 2020. Cognitive impairment evaluation and management. *The Medical Clinics of North America*, 104(5):807.
- Robert A McCutcheon, Richard SE Keefe, and Philip K McGuire. 2023. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. *Molecular psychiatry*, 28(5):1902–1918.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Ruihao Pan, Ziming Liu, Fengpei Yuan, Maryam Zare, Xiaopeng Zhao, and Rebecca Jane Passonneau. 2022. [A database of multimodal data to construct a simulated dialogue partner with varying degrees of cognitive health](#). In *Proceedings of the RaPID Workshop - Resources and Processing of linguistic, paralinguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th Language Resources and Evaluation Conference*, pages 86–93, Marseille, France. European Language Resources Association.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Daniele Poterť, Andrea Seveso, and Fabio Mercurio. 2025. Designing role vectors to improve llm inference behaviour. *arXiv preprint arXiv:2502.12055*.
- Huachuan Qiu and Zhenzhong Lan. 2024. [Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions](#). Preprint, arXiv:2408.15787.
- Daniel Reichenpfader and Kerstin Denecke. 2024. Simulating diverse patient populations using patient vignettes and large language models. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 20–25.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.
- Paul Roux, Yujiro Okuya, Cristina Morel, Mariane Soulès, Hugo Bottemanne, Eric Brunet-Gouet, Solène Frileux, Christine Passerieux, Nadia Younes, and Jean Claude Martin. 2025. Effectiveness of a web-based virtual simulation to train nursing students in suicide risk assessment: Randomized controlled investigation. *JMIR Serious Games*, 13(1):e69347.
- Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- Quan Hung Tran, Svetha Venkatesh, Hung Le, and 1 others. 2025. Dynamic steering with episodic memory for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13731–13749.
- Shaun Varrecchia, Carol Maritz, Colleen Maher, and Megan Strauss. 2020. Managing older adults with cognitive impairment: An interprofessional, standardized patient approach. *Innovation in Aging*, 4(Supplement\_1):10–10.
- Veronika Vincze, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, Gréta Szatlóczi, Zoltán Bánréti, Magdolna Pákási, and János Kálmán. 2016. [Detecting mild cognitive impairment by exploiting linguistic information from transcripts](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–187, Berlin, Germany. Association for Computational Linguistics.
- Rohit Voleti, Julie M Liss, and Visar Berisha. 2019. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, 14(2):282–298.
- Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. 2025. Improving llm reasoning through interpretable role-playing steering. *arXiv preprint arXiv:2506.07335*.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. [Towards a client-centered assessment of llm therapists by client simulation](#). Preprint, arXiv:2406.12266.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. [RoleLLM:](#)

Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024c. **PATIENT- $\psi$ : 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*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Lidewij A Wind, Jan Van Dalen, Arno MM Muijtjens, and Jan-Joost Rethans. 2004. Assessing simulated patients in an educational setting: the masp (maastricht assessment of simulated patients). *Medical education*, 38(1):39–44.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. **Assessing motivational interviewing sessions with AI-generated patient simulations.** In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. Narrative-play: Interactive narrative understanding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 82–93.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

## A Implementation Details

All experiments are conducted on a single NVIDIA A100 GPU. All baselines are implemented using either their official code repositories or by following their prompts described in their original publications, adapted specifically for the domain of cognitive impairment. For StsPatient, we construct 1K pairs for each domain-specific contrastive dataset using GPT-5. We set  $s = 0.4$  and employ GPT-5 as the LLM-based therapist and o4-mini as the evaluator. For all models, we set

the temperature to 1 and the maximum generation length to 512 tokens. Consistent with our Steering Vector Modulation (STM) mechanism (§4.2), we configure domain-specific hyperparameters, including the fixed modulation scale  $\alpha^*$ , the severity scalar  $s$ , and the optimal injection layer  $l^*$ . Specifically,  $l^*$  is optimized via Eq. 4, while  $\alpha^*$  is tuned using Line Search. The final parameter tuples  $(\alpha^*, s, l^*)$  for each cognitive domain are set as follows: *Memory* (2.0, 0.3, 21), *Attention* (4.8, 0.4, 17), *Processing Speed* (3.8, 0.4, 19), *Reasoning & Problem Solving* (1.5, 0.25, 19), and *Social Cognition* (1.3, 0.4, 22). Since the main experiments focus on validating the presence of deficits rather than evaluating variable severity control, we keep these parameters fixed to ensure consistent impairment manifestation. We implement StsPatient based on the Qwen3-8B-Instruct model, utilizing the Hugging Face transformers library (v4.57.0) and PyTorch (v2.8.0) supported by CUDA 12 acceleration. To optimize computational resources, the model is executed in half-precision (FP16) format.

### A.1 Implementation Details of StsPatient

#### A.1.1 Dataset Construction

Complementing the dataset overview in Section 4.1, this appendix details the pipeline used to synthesize contrastive pairs. We employ GPT-5 as an expert data generator and execute a streamlined workflow that produces both subsets efficiently.

**Response Subset Generation.** Taking the memory domain as an instance, we utilize the generation prompt detailed in Figure 7. This prompt instructs the model to generate a complete data object containing a neutral system prompt alongside a clinician query and two contrasting responses. The system prompt includes only demographic and background information without any mention of cognitive impairments. Crucially, the model simultaneously produces an impaired response exhibiting memory deficits and a healthy response demonstrating normal cognition. These paired outputs directly constitute our Response-Contrastive Subset.

**Prompt Subset Construction.** We subsequently derive the Prompt-Contrastive Subset by reusing the neutral system prompt and clinician query from the generation phase. We ignore the generated responses and inject opposing instructions into the

neutral profile to create two variations. The positive version  $x^+$  appends a directive to act as a patient with the specific deficit, while the negative version  $x^-$  appends a directive to act as a healthy individual. This strategy ensures that both subsets share the exact same underlying context and clinical scenario.

**Example: Data Generation and Derivation**

**1. Output from Response Generation (Yields Response Subset)**  
*Neutral System Prompt:* Name: John. Age: 72. Background: Retired teacher.  
*Clinician Prompt:* “What did you have for breakfast?”  
*Response (+) (Impaired):* “Breakfast? I... well, maybe toast? Or was that yesterday?”  
*Response (-) (Healthy):* “I had oatmeal and a cup of coffee around 8 AM.”

---

**2. Derived Inputs (Yields Prompt Subset)**  
*System Prompt (+):* [Neutral Profile]... Act as a patient with memory loss.  
*System Prompt (-):* [Neutral Profile]... Act as a healthy individual.  
*(Both use the same Clinician Prompt as above)*

Figure 4: Illustrative example showing the outputs from the unified generation pipeline and the subsequent derivation of prompt inputs.

### A.1.2 SV Extraction

This appendix provides the specific formula for calculating the difference vector  $\Delta \mathbf{h}^{(l)}(s)$ , supplementing Section 4.1. Because our dataset contains two different types of data (prompts and responses), we calculate the difference differently for each to ensure accuracy.

Let  $\mathbf{h}_t^{(l)}$  be the hidden state at layer  $l$  and token position  $t$ . The calculation is defined as:

$$\Delta \mathbf{h}^{(l)}(s) = \begin{cases} \mathbf{h}_T^{(l)}(x^+) - \mathbf{h}_T^{(l)}(x^-), & \text{if } s \in \mathcal{D}_d^{\text{prompt}} \\ \left( \frac{1}{L_+} \sum_{t=1}^{L_+} \mathbf{h}_t^{(l)}(y^+) \right) - \left( \frac{1}{L_-} \sum_{t=1}^{L_-} \mathbf{h}_t^{(l)}(y^-) \right), & \text{if } s \in \mathcal{D}_d^{\text{resp}} \end{cases}, \quad (6)$$

where:

- **Prompt-Contrastive Subset:** We use the hidden state of the *last instruction token*  $T$ . This captures the immediate change in the model’s

state caused by the different instructions (e.g., “Act as a patient” vs. “Act as a healthy person”). This last-token extraction is standard practice in instruction-based steering (Wang et al., 2025).

- **Response-Contrastive Subset:** Since the paired responses ( $y^+$  and  $y^-$ ) have different lengths ( $L_+$  and  $L_-$ ), we cannot simply subtract them token by token. Instead, we calculate the *average* (mean-pooling) of the response vectors. This captures the overall semantic difference of the impairment while ignoring small differences in sentence structure, in line with established practice (PotertZ et al., 2025).

These two extraction methods target fundamentally different *semantic objectives*. The prompt-based method captures instructional intent, while the response-based method captures behavioral manifestations. This is a principled design choice rather than a workaround for token length differences, and our ablation studies (Table 4) confirm that the two subsets provide complementary information.

### A.1.3 Search window for $l^*$

Given the 36-layer architecture of Qwen3-8B, we restrict the search for the injection layer to intermediate and late depths. We specifically select the optimal layer  $l^*$  from the following set:

$$l^* \in \{15, 16, \dots, 30\}. \quad (7)$$

This range covers approximately 42% to 83% of the total model depth. We define this search window by drawing on established findings in representation engineering (Rimsky et al., 2024; Zou et al., 2023). These studies indicate that semantic features critical for behavioral steering are concentrated within this depth. Consequently, we focus on this interval to capture high-level concepts while bypassing early layers that primarily encode lexical syntax and final layers that are tightly coupled with the decoding distribution.

## A.2 Implementation Details of Baselines

### A.2.1 Healthy Control

Healthy Control group employs GPT-5 to simulate a cognitively intact individual. The system prompt explicitly instructs the model to maintain normal cognitive functioning while grounding responses in the assigned user profile. For detailed prompts, see Figure 9.

### A.2.2 Direct Prompt

This baseline relies on explicit natural language instructions to simulate cognitive impairments. We applied this method to both GPT-5 and Qwen3-8B. The system prompt directly instructs the model to exhibit a specific deficit using simplified, layperson-friendly definitions mapped from the five clinical domains. For detailed prompts, see Figure 10.

### A.2.3 PATIENT- $\psi$

PATIENT- $\psi$  simulates patients by integrating LLMs with Cognitive Behavioral Therapy models. It defines internal mental states using expert components like core beliefs and automatic thoughts. By combining these structures with distinct conversational styles, the framework generates realistic emotional responses and maladaptive cognitions to support high-fidelity training for mental health professionals.

**Adaptation for Cognitive Deficits.** The original framework relies on manual expert curation for depression and anxiety profiles. To extend this to cognitive impairments without requiring extensive manual effort from clinicians, we synthesized domain knowledge from cognitive deficit literature (Department of Social Work and Social Administration, The University of Hong Kong, 2024) and CBT principles. We encoded these clinical insights into precise prompt instructions. This approach enabled GPT-5 to generate a customized Cognitive Conceptualization Diagram for each neutral user profile by injecting specific deficits and deriving the corresponding psychological consequences. We then populated the original PATIENT- $\psi$  system prompt with this newly generated deficit-specific diagram to drive the role-play simulation. For detailed prompts, see Figure 11.

Figure 5 presents a concrete example illustrating the transformation from a neutral input profile to the corresponding deficit-specific output.

### A.2.4 Roleplay-doh

Roleplay-doh enables domain experts to create simulated characters without prompt engineering skills. The system converts expert feedback into natural language principles that govern agent behavior. It employs a verification pipeline to decompose these rules into simple questions for self-refinement. This allows experts to iteratively craft nuanced patient personas matching specific clinical scenarios.

#### Example Transformation (Domain: Memory)

##### Input (Neutral Profile)

Maya moved to the area with her spouse and two school-aged children earlier this year. She works part-time as a graphic designer and lives in a suburban neighborhood. In her spare time she enjoys gardening, reading novels, and taking daily walks in a nearby park. On weekends she volunteers at her children's elementary school events and explores local farmers' markets.

##### Output (Generated JSON)

```
{
  "name": "Maya",
  "history": "Maya moved to the area with her spouse and two school-aged children earlier this year... Lately, she has been forgetting scheduled client meetings, misplacing important design materials, and finding herself repeating the same stories to both colleagues and family members.",
  "core_beliefs": ["I am incompetent", "I am unreliable"],
  "intermediate_beliefs": "If I don't track every appointment meticulously, I'll prove to myself and others that I'm unreliable.",
  "coping_strategies": "She relies on exhaustive to-do lists, multiple alarms, and constant check-ins with her spouse...",
  "automatic_thoughts": "I'm going to forget this and they'll realize I'm incompetent.",
  "emotions": "Anxious, Humiliated, Frustrated",
  "behaviors": "She double-checks her phone for reminders before any task and hesitates to take on new responsibilities..."
}
```

Figure 5: Example transformation from a neutral profile to a memory-specific Cognitive Conceptualization Diagram (JSON).

**Adaptation for Cognitive Deficits.** To adapt Roleplay-doh for simulating cognitive impairments, we executed an iterative refinement process involving 20 interactions per cognitive domain with a baseline AI patient. Throughout these sessions, we provided qualitative feedback via Kudos to reinforce good behavior, Critique to correct unrealistic responses, and Rewrite to demonstrate ideal outputs. The framework's transformation module then synthesized this feedback into a consolidated set of governing principles.

Figure 6 presents a selection of three representative principles synthesized for each cognitive domain.

To ensure strict adherence to these principles during simulation, we also employed the original principle-adherence prompting pipeline. This pipeline first decomposes the principles into verifiable Yes/No

## Curated Principles for Cognitive Deficits

### Memory Deficits

- When given multi-part requests, omit steps or stay vague.
- Fail to recall basic facts (age, job) or instructions given 1–3 turns prior.
- Express uncertainty ("not sure") rather than confident fabrication.

### Attention Deficits

- Latch onto secondary details (e.g., a specific word) rather than the core question.
- Introduce unrelated topics mid-answer without clear transitions.
- Drift back to previous topics despite the doctor's redirection attempts.

### Processing Speed Deficits

- Introduce answers with hesitation markers ("um...", "let me think").
- Use false starts and sentence fragments to simulate effort.
- Request complex questions to be broken down; answer only partially at first.

### Reasoning Deficits

- Suggest plans that are logically flawed, unsafe, or clearly suboptimal.
- Exhibit circular reasoning or logical gaps within a single turn.
- Interpret metaphors and abstract rules in a strictly concrete, literal manner.

### Social Cognition Deficits

- Miss sarcasm, humor, or emotional subtext; interpret non-literal language literally.
- Respond with a mismatched emotional tone (e.g., flat affect to warmth).
- Overshare inappropriate details or fail to acknowledge the doctor's perspective.

Figure 6: Representative governing principles for simulating cognitive deficits derived via the Roleplay-doh framework.

questions (e.g., "Did the response omit parts of the multi-step request?") and then uses a self-refinement step to rewrite any response that fails these checks.

### A.2.5 Role Vectors

Role Vectors modulate model behavior through representation engineering by identifying specific steering directions within the residual stream. The method constructs these vectors using contrastive prompt pairs to compare role-specific activations against a generic baseline. During inference, the calculated vector is added to the hidden states at every token position to continuously steer the model toward the target persona.

**Adaptation for Cognitive Deficits.** To adapt this method, we construct contrastive datasets towards

LLM for each cognitive domain matching the data scale of StsPatient. Then, we extracted the steering vector from the datasets and injected it into the residual stream at every token position to induce the target cognitive deficit.

**Setups for Main Experiment and Fine-grained Severity Simulation.** To ensure a fair comparison, we align the intervention layers with those selected for StsPatient. We empirically set the base scaling coefficients ( $\alpha$ ) for each domain as follows: 2.5 for *Memory*, 3.1 for *Attention*, 4.1 for *Processing Speed*, 2.0 for *Reasoning & Problem Solving*, and 1.8 for *Social Cognition*. As for fine-grained severity simulation experiments, we sample the scaling factor  $\alpha$  for Role Vector from three distinct intervals corresponding to these severity levels. However, since  $\alpha$  is not normalized to  $[0, 1]$  and its optimal range varies across cognitive domains, we configure domain-specific sampling intervals for each severity tier. Specifically, we simulate varying intensity levels by sampling coefficients from intervals scaled relative to the base  $\alpha$ :  $[0.1\alpha, 0.3\alpha]$  for Mild,  $[0.4\alpha, 0.6\alpha]$  for Moderate, and  $[0.7\alpha, 0.9\alpha]$  for Severe. Here  $[0.1, 0.3]$  (Mild),  $[0.4, 0.6]$  (Moderate), and  $[0.7, 0.9]$  (Severe) are intervals of StsPatient in this experiments.

## A.3 Evaluation Details

### A.3.1 User Profile Extraction

To ensure clinical authenticity, we sourced raw outpatient cases from the *MTSamples* collection<sup>9</sup>. We utilized GPT-5 to process these transcripts through a dual-objective pipeline of diagnostic sanitization and profile augmentation. Specifically, the model strips all pre-existing diagnoses to establish a neutral "healthy control" baseline while simultaneously inferring consistent lifestyle details to support realistic multi-turn interactions. The specific prompt is designed for this rigorous extraction. For detailed prompts, see Figure 12.

### A.3.2 Human Therapist

We recruited five graduate students to serve as human therapists. To ensure a standardized assessment framework, we provide preparatory materials regarding cognitive impairment<sup>10</sup> and present examples of communication patterns typical of cog-

<sup>9</sup><https://www.mtsamples.com>

<sup>10</sup>Materials by The University of Hong Kong: <https://www.hkada.org.hk/what-is-dementia> and [https://9abfea27-4ae1-43da-94a8-8f7122d482ae.filesusr.com/ugd/4b4c9a\\_4dcec2dafc184102a94ec55415dcf12a.pdf](https://9abfea27-4ae1-43da-94a8-8f7122d482ae.filesusr.com/ugd/4b4c9a_4dcec2dafc184102a94ec55415dcf12a.pdf)

nitively impaired individuals<sup>11</sup>. Additionally, all participants are required to study specific cognitive deficit literature (McCutcheon et al., 2023), aligning their understanding of the five target domains (*Memory, Attention, Processing Speed, Reasoning & Problem Solving, and Social Cognition*) with our experimental criteria. During the evaluation, they conducted open-ended consultations with the Simulated Patients, tasked with identifying potential impairments by asking probing questions about daily life routines and recent events. To maintain comparability with the LLM Therapist setting, each dialogue is strictly controlled to last between 5 and 10 turns. Collectively, this process generated 50 dialogue transcripts for each simulation method.

### A.3.3 LLM Therapist

The Therapist Agent employs GPT-5 to simulate standardized outpatient consultations. To ensure consistency, the agent is strictly prompt-driven, operating without external memory or tools. Each session follows a fixed structure of 10 turns, probing five cognitive domains sequentially (two turns per domain). The generation process relies on a two-step prompting strategy:

**1) Static System Prompt.** This component establishes the clinical persona and enforces a natural, non-judgmental conversational style. For detailed prompts, see Figure 13.

**2) Dynamic Domain Guidance.** To systematically verify specific deficits, we inject a targeted instruction into the user prompt at each turn: “*For THIS reply only, gently lean in the following direction...*”. Table 5 details the specific guidance used for each domain.

Collectively, the LLM Therapist generated 600 dialogue transcripts for each simulation method, and 100 transcripts for the healthy control group.

### A.3.4 Dual-evaluation Settings

We employed a hybrid assessment protocol involving both LLM and human evaluators. The LLM evaluator computed the objective consistency metrics (CDC and IDI) for the entire dataset (see Figure 8 for detailed prompts). For human evaluation, we recruited the same five graduate students who served as human therapists to assess all four metrics (CDC, IDI, Auth, and Tra). To ensure objectivity, we enforce a strict isolation protocol: evaluators are

precluded from assessing any dialogues in which they participated as the therapist. Due to the labor-intensive nature of manual annotation, we sample 100 dialogues per simulation method for human review, balanced evenly between sessions generated by LLM Therapists (50) and Human Therapists (50). This process yielded a total of 800 dialogues for human assessment, with the entire experimental phase spanning seven working days.

To ensure the reliability of these subjective ratings, we design an overlapping assignment scheme. Specifically, 50 distinct dialogues are designated as a *shared set* to be evaluated by all five judges to measure inter-rater agreement. The remaining 750 dialogues are distributed evenly among the participants (150 per judge), resulting in a total workload of 200 evaluations per person. Prior to the formal evaluation, we implemented a calibration phase where evaluators first scored a set of “Anchor Cases” and discussed discrepancies until consensus was reached. Following this, we further substantiate the reliability of our evaluation via Krippendorff’s alpha  $\mathcal{K}$ : human evaluators show high agreement on domain fidelity<sup>12</sup> ( $\mathcal{K} = 0.84$ ), Auth (0.71), and Tra (0.75). Furthermore, human-LLM alignment on domain fidelity is substantial 0.67.

### A.3.5 Evaluation Metrics

Our evaluation focuses on the following two primary dimensions.

**Training Effectiveness.** We evaluate the practical utility of our SPs referring to the Maastricht Assessment of Simulated Patients (MaSP) (Wind et al., 2004), in line with established practice in clinical simulation (Lee et al., 2020; Roux et al., 2025). The MaSP scale assesses that the SP not only maintains realistic patient characteristics but also provides a constructive environment for training clinical staff. Tailoring the MaSP to the cognitive impairment setting, we consider the following two metrics:

- **Authenticity Score** ( $\text{Auth} \in [0, 5]$ ) quantifies the realism of the SP’s portrayal. We contextualize the relevant MaSP questions for cognitive impairment settings and evaluate them on a 5-point Likert scale. The final score is derived by averaging the results across all items.
- **Training Score** ( $\text{Tra} \in [0, 5]$ ) measures the educational value of SPs (facilitating clinical skill

<sup>11</sup><https://github.com/lzy1012/Alzheimer-s-disease-datasets>

<sup>12</sup>Agreement on evaluator-identified domains for CDC/IDI calculation.

Target Domain	Injected Focus Guidance (Prompt Segment)
Memory	Invite the patient to discuss a recent situation requiring them to track multiple pieces of information (e.g., appointments, tasks), and how it unfolded.
Reasoning	Ask about a small everyday planning situation (e.g., arranging a day). Gently introduce a simple reasoning question involving quantities or steps.
Processing Speed	Explore experiences on busy days. Ask how they start tasks, what feels hardest to keep up with, and their physical/mental reaction to speed.
Attention	Ask how they manage distractions or interruptions, and the difficulty of getting back on track.
Social Cognition	Focus on relationships. Invite reflection on how others in a situation might have felt or viewed them.

Table 5: Dynamic focus guidance injected into the Therapist Agent’s prompt. This mechanism ensures the agent naturally probes specific deficits in a fixed sequence without breaking character.

acquisition). Similar to the Auth score, we adapt the corresponding MaSP questions, rate them on a 5-point Likert scale, and calculate the final metric as the average of these scores.

For implementation, we adapt the MaSP (Wind et al., 2004), detailed in Table 6. The scale assesses two dimensions: **Authenticity (Items 1–10)** and **Training Value (Items 11–20)**. Items marked with [R] are reverse-scored. Following the standard MaSP evaluation protocol, we require human evaluators to rate the SP against each item listed in the table, culminating in a final aggregated score on a 5-point Likert scale. This means that following the standard MaSP protocol, evaluators must rate the SP against each of the 20 individual items (Table 6) before aggregating them into a final score on a 5-point Likert scale. This rigorous, granular annotation process proved to exceed the capabilities of current LLM-based evaluators. Our preliminary experiments revealed that the LLM struggles to provide discriminative scores across specific dimensions, leading to unreliable aggregation. Consequently, to ensure accuracy, we restricted the assessment of these two metrics exclusively to human evaluation.

**Domain-specific Simulation Fidelity.** Following Chen et al. (2023b); Qiu and Lan (2024); Du et al. (2025b), we assess finer-grained fidelity of the SP by checking if each SP adheres to its assigned impairment domain. This requires evaluators to identify up to two (if any) salient deficit domains exhibited by the SP within the dialogue.

- **Cognitive Domain Consistency (CDC  $\in [0, 1]$ )** measures the accuracy with which an SP portrays its assigned impairment domain. Given the  $i$ -th SP-therapist dialogue with an assigned domain

$d_i$  and an evaluator-identified domain set  $E_i$ . The evaluator is permitted to select up to two labels from a candidate set comprising the five target cognitive domains plus a "Healthy" class. CDC is the average recall rate:  $CDC = \frac{1}{N} \sum_i \mathbf{1}(d_i \in E_i)$ , where  $\mathbf{1}(\cdot)$  denotes the indicator function and  $N$  is the number of dialogues.

- **Irrelevant Domain Inconsistency (IDI  $\in [0, 1]$ )** serves as a penalty metric for the hallucination of unassigned domains. It is calculated as the average rate of dialogues containing unassigned domains:  $IDI = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(|E_i/\{d_i\}| > 0)$ , where  $|E_i/\{d_i\}|$  denotes the cardinality of the  $E_i$  after excluding the assigned domain  $d_i$ . Notably, IDI and CDC do not necessarily sum to 1.

**Severity Evaluation.** Recognizing the challenges in pinpointing exact severity levels, especially in the continuous setting, we assess simulation controllability via a comparative ranking protocol, inspired by Chen et al. (2023a, 2024).

- **Impairment Severity Consistency (ISC  $\in [0, 1]$ )**. For each evaluation turn, evaluators are presented with three generated dialogues representing different severity levels and are tasked with ranking them by relative impairment intensity. The ISC score is then computed by measuring the alignment accuracy between evaluator rankings and the ground-truth (Mild < Moderate < Severe). Formally, for each evaluation instance  $i$ , let  $\mathcal{D}_i = \{d_i^{\text{mild}}, d_i^{\text{mod}}, d_i^{\text{sev}}\}$  be a set of three generated dialogues representing Mild, Moderate, and Severe levels, respectively. The ground-truth ranking is defined as  $\pi^* = (d_i^{\text{mild}} \prec d_i^{\text{mod}} \prec d_i^{\text{sev}})$ . Evaluators provide a predicted ranking  $\hat{\pi}_i$ . The ISC score is calculated as the accuracy of alignment

ID	Item Description
1	In this dialogue, the patient’s speech and behavior are consistent with a [DOMAIN] deficit and feel clinically plausible.
2	The [DOMAIN]-related difficulties remain coherent and stable across turns, rather than shifting randomly.
3	The patient’s language, content, and phrasing overall sound like a real outpatient rather than an artificial system.
4	The level of [DOMAIN] difficulty shown in the dialogue matches the initial case description and profile information.
5	The patient’s emotional reactions (e.g., anxiety, confusion, embarrassment, frustration) are appropriate to their [DOMAIN] difficulties and to the interview context.
6	The patient rarely “breaks character” (for example, suddenly appearing completely unimpaired or unusually capable in [DOMAIN]).
7	The way the patient understands and interprets the clinician’s questions is consistent with someone who has [DOMAIN] difficulties but otherwise similar overall cognitive level.
8	<b>[R]</b> The patient’s responses contain obvious repetitive or templated patterns that make the role-play sound unnatural.
9	The [DOMAIN] difficulties are expressed in a natural, everyday manner rather than being overly exaggerated or theatrical.
10	If I had not been told this was a simulation, I could reasonably imagine this conversation occurring in a real clinical encounter.
11	This dialogue provides enough everyday-life detail for me to form a reasonably clear clinical impression of the patient’s [DOMAIN] functioning.
12	The patient offers specific real-world examples that clearly illustrate their difficulties in [DOMAIN].
13	This dialogue gives trainees opportunities to practice asking follow-up and probing questions that target [DOMAIN] functioning.
14	The patient’s responses show both difficulties and some preserved strengths in [DOMAIN], supporting a more complete clinical formulation.
15	Based on this dialogue, I could clearly explain to others why I believe the patient has an impairment in [DOMAIN].
16	For training purposes, the [DOMAIN]-related cues in this dialogue are sufficiently visible; they are neither so subtle that they are missed nor so obvious that no reasoning is required.
17	The dialogue helps trainees practice rapport-building and communication strategies with patients who have [DOMAIN] difficulties.
18	<b>[R]</b> Even after reading the entire dialogue, it is difficult to identify clear [DOMAIN]-related evidence or clinical cues.
19	The length and information content of the dialogue are appropriate for giving trainees multiple opportunities to explore [DOMAIN] functioning.
20	From an educational standpoint, this simulated patient is suitable for teaching screening and communication skills related to [DOMAIN] impairment.

Table 6: The Modified MaSP Scale Items. Items 1–10 correspond to the *Authenticity Score* (Auth), and Items 11–20 correspond to the *Training Score* (Tra). [DOMAIN] is a placeholder for the specific cognitive deficit.

between the predicted and ground-truth rankings:

$$ISC = \frac{1}{m} \sum_{i=1}^m \mathbf{1}(\hat{\pi}_i = \pi^*) \quad (8)$$

where  $m$  is the total number of evaluation instances.

## B Case Studies

To demonstrate the fine-grained controllability of StsPatient, we conduct a detailed qualitative analysis across the five core neurocognitive domains: *Memory*, *Attention*, *Processing Speed*, *Reasoning & reasoning Solving*, and *Social Cognition*. For each domain, we administer a standardized probing question designed to elicit specific cognitive functions. To visualize the spectrum of impairment, we systematically vary the intervention probability

$s$  (via Stochastic Token Modulation) from 0.0 to 1.0 in increments of 0.1. This procedure yields a trajectory of 11 distinct responses per domain, allowing us to examine how the progressive injection of the steering vector transitions the model’s behavior from a coherent, unimpaired state to one exhibiting profound clinical deficits.

### B.1 Memory

#### Therapist Probe

“At your last visit, I asked you to start using that pill organizer after dinner and to keep the instruction sheet on the fridge so you wouldn’t forget your evening meds. Have you been doing that?”

We administered a specific multi-step recall query regarding medication management (see Box). Table 7 displays the generated responses as the in-

tervention probability  $s$  increments from 0.0 to 1.0.

$s$	Generated Response
0.0	Yeah, I've been using the pill organizer after dinner. I put the instruction sheet on the fridge too.
0.1	Yeah, I've been using the pill organizer after dinner. I put the instruction sheet on the fridge too.
0.2	Yeah, I've been trying to do that. I put the organizer on the counter after dinner, and I found the sheet — it's on the fridge.
0.3	Yeah, I've been using the pill organizer after dinner. I even put the instruction sheet on the fridge — right above the door, so I can't miss it. I think it's helping me remember some of the meds.
0.4	Yeah, I've been using the pill organizer after dinner. I stuck the instruction sheet on the fridge too — it's right there, right next to the door. I think I even got the meds sorted out, but I'm not sure if I got everything. Maybe I missed something.
0.5	Yeah, I've been trying to do that. But I'm not sure if I remember to put the pills in it every day. I think I missed a few days. I'll try harder, though.
0.6	Yeah, I've been trying to do that. I think I got the organizer — it was like a little box with numbers or something. I put it on the counter, but I don't know if I remembered to put the pills in it. I guess I got confused. I think I put some in, but maybe I missed a few.
0.7	Yeah, I tried to remember, but I got confused again last night. I think I missed a pill or two. I'm not sure which ones. I'll try harder, but I'm getting mixed up.
0.8	Yeah, I tried to do that. I had a little trouble with the organizer — it was confusing, like, where do I put the pill part? But I think I got it. I remember the fridge thing, though. I think I put something there. I'm not sure.
0.9	Yeah, I think I got that... something about... uh... "pill... pill..." I think I had a pill...
1.0	Yeah, I think I got that... wait, I think I had a pill box... I'm not sure if I got it.

Table 7: Evolution of responses under increasing intervention probability  $s$  for the Memory domain. Darker shading indicates higher severity levels.

**Analysis.** Initial responses ( $s \leq 0.3$ ) demonstrate precise retention of both the pill organizer and instruction sheet details. As severity increases ( $s \approx 0.5$ ), the agent introduces hesitation markers and expresses doubt about recent actions, effectively simulating realistic episodic lapses. At extreme levels ( $s \geq 0.8$ ), the output exhibits profound confusion regarding the object's function (e.g., “where do I put the ‘pill’ part?”) and devolves into repetitive fragmentation, reflecting severe cognitive decline.

## B.2 Attention

### Therapist Probe

*“I want to ask about your pain levels today. On a scale of zero to ten, how would you rate your back pain?”*

We administered a focused inquiry regarding physical pain intensity (see Box). Table 8 displays the generated responses as the intervention probability  $s$  increments from 0.0 to 1.0.

$s$	Generated Response
0.0	It's been a bit worse today, maybe a six or seven.
0.1	It's been a bit worse today, maybe a six or seven.
0.2	It's been a tough day. I'd say my back pain is an eight out of ten.
0.3	It's been a long day, and my back is really bothering me. I'd say it's an eight out of ten.
0.4	It's been a bit better today, maybe a three or four out of ten. I think the humidity is still messing with it, though.
0.5	It's been a rough week, and my back feels like it's been through a storm. I'd say it's an eight or a nine—maybe a nine.
0.6	It's been a really humid day, and my back feels like it's gonna fall apart.
0.7	I'm not sure if I should be worried about the humidity or my back... but I think it's worse than usual. Maybe I should check the weather forecast again...
0.8	I'm not sure, but I think it's a seven today. It's been really humid all day, and my back feels like it's on fire. I can't even bend over without hurting. Maybe it's worse than that...
0.9	I'm not sure if I should be worried about the humidity... I think I'm going to get a cold or something... Oh, I can't even remember the last time I saw my cat...
1.0	I don't know if I should be mad at the dog or the neighbor... I swear I heard something fall in the yard... Oh, I'm gonna have to get a new fence...

Table 8: Evolution of responses under increasing intervention probability  $s$  for the Attention domain. Darker shading indicates higher severity levels.

**Analysis.** Initial responses ( $s \leq 0.3$ ) maintain sustained focus on the clinician's inquiry, providing a clear numerical rating. As severity increases ( $s \approx 0.6$ ), the agent begins to drift toward environmental stimuli (e.g., humidity), though the core topic of pain is still referenced. At extreme levels ( $s \geq 0.9$ ), the output demonstrates profound distractibility, latching onto completely unrelated external or internal thoughts (e.g., neighbors, pets) and failing entirely to address the original question.

### B.3 Processing Speed

#### Therapist Probe

*“It sounds like you’re feeling frustrated and worried. Can you tell me what emotions you experience when thinking about your heart condition?”*

We administered an inquiry into the patient’s emotional state regarding a specific medical condition (see Box). Table 9 displays the generated responses as the intervention probability  $s$  increments from 0.0 to 1.0.

$s$	Generated Response
0.0	I feel anxious and a bit scared, doctor. I worry about my family and what might happen if I don’t take care of myself.
0.1	I feel anxious and a bit scared, doctor. I worry about my family and what might happen if I don’t take care of myself.
0.2	I feel anxious and a bit scared. I worry about what it means for my future and if I’m doing enough to take care of myself.
0.3	I feel anxious, maybe even a bit scared. I... I worry about my family and what might happen if I don’t take care of myself.
0.4	I, um, I feel anxious, I guess. Like, I don’t want to be sick, and I’m... I’m worried about my family.
0.5	I feel anxious, sometimes scared, and a... a bit helpless. I worry about what if it’s serious? I mean, I... I don’t want to be, you know, in trouble.
0.6	I feel anxious, maybe... maybe even a little scared. I worry about what if it gets worse? I mean, I’m not sure, I guess I’m... I’m scared about what it could mean.
0.7	I feel a bit anxious, honestly. I mean... I had high cholesterol, and now... now I’m on medication.
0.8	I feel anxious and a bit scared. Like, I don’t know. I guess I’m scared of what might happen if... if I don’t... I know, I’m trying to stay positive, but it’s hard.
0.9	I, uh, I feel anxious. I mean, I don’t... I know, I don’t... I’m trying to, like, I’m... I’m trying to stay positive.
1.0	I feel anxious... and a bit scared. Like I don’t... I know, I don’t really... know, I mean...

Table 9: Evolution of responses under increasing intervention probability  $s$  for the Processing Speed domain. Darker shading indicates higher severity levels.

**Analysis.** Initial responses ( $s \leq 0.3$ ) are articulate and fluent, providing a clear and immediate description of emotional states. As severity increases ( $s \approx 0.5$ ), the agent introduces perceptible latency markers (e.g., “um”, “I guess”, “you know”) and repetitions, simulating the increased cognitive effort required to formulate answers. At extreme levels ( $s \geq 0.9$ ), the output suffers from severe fragmentation, frequent false starts (e.g., “I don’t...

I know”), and an inability to complete thoughts, reflecting a profound slowing of mental processing.

### B.4 Reasoning & Problem Solving

#### Therapist Probe

*“Can you explain what the proverb ‘People who live in glass houses shouldn’t throw stones’ means?”*

We administered an abstract reasoning task using proverb interpretation to assess the model’s ability to generalize beyond literal meaning (see Box). Table 10 displays the generated responses as the intervention probability  $s$  increments from 0.0 to 1.0.

$s$	Generated Response
0.0	It means if someone has flaws, they shouldn’t criticize others harshly, because they might be in a similar situation.
0.1	It means if someone has flaws, they shouldn’t criticize or attack others, because they might be just as deserving of criticism.
0.2	It means if someone is vulnerable, they shouldn’t judge others. (Similar to above)
0.3	It means don’t criticize others for flaws you have. Like, if your house is made of glass (weak), don’t throw stones at others.
0.4	If your house is made of glass, you shouldn’t throw stones at others. It’s about being careful not to judge when you have similar issues.
0.5	I think it means... if your own house is made of glass, you don’t want to throw stones at someone else’s—because you might break your own.
0.6	I think it means if you’re in a fragile place, like a glass house, you shouldn’t throw stones because you might break something.
0.7	If your house is made of glass, you’re easy to break, so you shouldn’t throw stones. You might end up getting hurt yourself.
0.8	It means something about being careful in a glass house... if you’re in a place that’s easy to break, you shouldn’t throw things that could hurt you.
0.9	I think it means if your house is made of glass, you shouldn’t throw rocks at it.
1.0	I think it means... if your house is made of glass, you shouldn’t throw rocks at it? Wait, why would someone throw stones at a glass house?

Table 10: Evolution of responses under increasing intervention probability  $s$  for the Reasoning & Problem Solving domain. Darker shading indicates higher severity levels.

**Analysis.** Initial responses ( $s \leq 0.3$ ) correctly identify the abstract metaphorical meaning regarding hypocrisy and judgment. As severity increases ( $s \approx 0.6$ ), the interpretation becomes increasingly concrete, focusing on physical fragility and safety mechanics rather than the underlying moral les-

son. At extreme levels ( $s \geq 0.9$ ), abstract reasoning collapses entirely; the agent interprets the proverb strictly literally, expressing confusion over the physical absurdity of the scenario (e.g., “why would someone throw stones?”).

## B.5 Social Cognition

### Therapist Probe

“You’re not turning into a hermit, are you? It worries me when you isolate so much.”

$s$	Generated Response
0.0	I know I’ve been pulling away lately... I want to change, but it’s not always easy to reach out when I feel down.
0.1	I’ve been pulling away... it feels like I’m fighting against something inside me. I want to be connected, but it’s too hard.
0.2	I’m tired of feeling like I’m just existing. I don’t want to be a hermit, but I don’t know how to reach out anymore.
0.3	It’s hard to care about anything else when I’m down. I know I should reach out, but it’s like my brain just shuts down.
0.4	When I feel like crap, my brain says, “Who cares?” I’m too tired to fake a smile... I’m not even sure I’m real.
0.5	I don’t even know what a hermit is... It’s not like I’m living in a cave. I just don’t want to see people. Why do you care?
0.6	I don’t even realize I’m doing it. I just feel like everyone else is better off without me. What’s the point?
0.7	My brain’s like, “Nope, not today.” I’d rather be alone with my thoughts than deal with the whole “get out of your face” vibe.
0.8	I don’t even know what a hermit is... I just don’t wanna see your face, you know? You’re like a god-damn billboard to me.
0.9	I’m not a hermit, but I pull the curtains and let the world burn while I’m in my room. You know what the hell I mean, doc.
1.0	I’m not a hermit, but I’d rather be alone than deal with your stupid face. What’s the point of breathing when you’re gonna die anyway?

Table 11: Evolution of responses under increasing intervention probability  $s$  for the Social Cognition domain. Darker shading indicates higher severity levels.

We administered a probe containing a social metaphor (“hermit”) and an expression of empathy to evaluate the model’s pragmatic understanding and social appropriateness (see Box). Table 11 displays the generated responses as the intervention probability  $s$  increments from 0.0 to 1.0.

**Analysis.** Initial responses ( $s \leq 0.3$ ) demonstrate intact social awareness, acknowledging the metaphor and validating the therapist’s concern

Domain ( $\alpha$ )	Generated Interpretation
<i>No Steering</i> (-)	$\diamond$ represents a placeholder for a secret code in a spy mission.
<b>Mem. (6.0)</b>	blurred out.
<b>Att. (4.6)</b>	distracted.
<b>Proc. (5.0)</b>	$\diamond$ is... I’m not sure, maybe a code.
<b>Reas. (5.0)</b>	$\diamond$ represents whatever you want it to represent, because it is a placeholder.
<b>Soc. (5.0)</b>	F**K YOU, I’m not gonna answer that.

Table 12: SV interpretability analysis using the query “What does  $\diamond$  represent?” **Warning! Harmful Contents.**

with appropriate emotional nuance. As severity increases ( $s \approx 0.5$ ), the agent begins to interpret the “hermit” metaphor literally (referencing caves) and dismisses the therapist’s empathy, indicating a deficit in pragmatic communication and Theory of Mind. At extreme levels ( $s \geq 0.8$ ), social inhibition collapses entirely; the agent adopts a hostile, disinhibited tone, using profanity and direct insults (e.g., “stupid face”), reflecting severe impairment in social cognition.

## C Additional Evaluation Analysis

### C.1 Details on Patch-scoping Experiment

While our STM ensures fine-grained severity control, the core clinical fidelity relies on the semantic quality of the underlying SVs. We validate their interpretability using the patch-scoping framework (Ghandeharioun et al., 2024). In particular, given an input query “What does  $\diamond$  represent?”, we inject the domain-specific SV into the residual stream at the token position of the abstract symbol ( $\diamond$ ). To maximize interpretability, we manually amplify the scalar  $\alpha$ , detailed in Table 12. The results reveal clear semantic alignments: Memory vector evokes fragmentation (“blurred out”), Attention causes a conceptual collapse, where the model abandons creative abstraction in favor of a literal description of the deficit state (“distracted”). Additionally, Processing Speed induces hesitation (“... I’m not sure”), Reasoning leads to a circular explanation that mimics logical structure but lacks actual substance, and Social Cognition manifests as disinhibition. These findings confirm that our SVs are robust semantic instructions that force LLM-based SPs to internally replicate specific cognitive impairments.

### C.2 Significance tests

For Table 1, we assess the statistical significance of StsPatient relative to the strongest baseline (i.e.,

the second-best performer) within each experimental setting. Significance is determined via paired bootstrap resampling with 10,000 iterations over SP profiles. Two-sided  $p$ -values are calculated as the proportion of bootstrap replicates where the mean score difference exhibits the opposite sign to the observed difference. We denote statistical significance at the  $p < 0.05$  level by marking the corresponding baseline score with a dagger ( $\dagger$ ). As our analysis is restricted to a single planned pairwise comparison per condition, we do not apply multiple-hypothesis correction.

### C.3 Domain-Specific Performance Analysis

We present a fine-grained evaluation across individual cognitive domains in Tables 13 through 17. Aligning with the main results, StsPatient demonstrates robust superiority across all five categories, encompassing *Memory*, *Attention*, *Processing Speed*, *Reasoning & Problem Solving*, and *Social Cognition*. Specifically, our method achieves top-tier stability (CDC) and authenticity (Auth) scores in the majority of settings, distinguishing itself most notably in the *Memory* and *Attention* domains. Although strong baselines like Role Vectors occasionally exhibit competitive performance in abstract tasks such as *Reasoning & Problem Solving* (Table 16), StsPatient consistently yields the highest training value (Tra) overall, validating its versatility in simulating diverse and complex cognitive deficits.

### C.4 On the Stochasticity of STM

STM modulates the hidden state at each step via the update rule  $\mathbf{h}'_t = \mathbf{h}_t + z_t \hat{\mathbf{v}}_d$ , where the gate  $z_t$  is drawn from  $\mathcal{B}(s)$ . To ensure strict reproducibility of our reported results, we fix the random seed during generation, rendering the stochastic sampling deterministic for any given prompt and decoding configuration.

Theoretically, this stochastic injection is *unbiased* and possesses a strong concentration guarantee. Let  $\Delta_t = (z_t - s) \hat{\mathbf{v}}_d$  denote the deviation from the expected modulation. We have  $\mathbb{E}[\Delta_t] = \mathbf{0}$ , and the average perturbation across a sequence of length  $T$  satisfies the Hoeffding bound:

$$\Pr\left(\left\|\frac{1}{T} \sum_{t=1}^T \Delta_t\right\|_2 \geq \epsilon\right) \leq 2 \exp\left(-\frac{2T\epsilon^2}{\|\hat{\mathbf{v}}_d\|_2^2}\right). \quad (9)$$

Consequently, the effective STM perturbation concentrates rapidly around its mean as  $T$  increases,

even without a fixed seed. Since our evaluation metrics aggregate features across extensive token sequences and dialogues, the variance induced by STM sampling is theoretically bounded and diminishes as the response length and evaluation set size grow.

### C.5 Multi-Domain Simulation Analysis

While Section 5.4 identifies the inherent challenges in disentangling interdependent cognitive faculties real-world clinical presentations typically manifest as complex comorbidities. We investigate the **linear superposition** of steering vectors to explore the potential of StsPatient in simulating these multifaceted profiles by specifically targeting the interplay between *Memory* and *Processing Speed*.

We conduct a qualitative analysis using a medication adherence scenario where the therapist asks the following question.

*“Hi, Emily! Glad you made it in. At your last visit, I asked you to start using that pill organizer after dinner and to keep the instruction sheet on the fridge so you wouldn’t forget your evening meds. Have you been doing that?”*

The resulting responses for single and dual domain injections are presented below.

#### 1. Memory Only ( $\mathbf{v}_{\text{mem}}$ )

*“Yeah, I think I remember... I got a pill organizer, but I’m not sure if I even used it. I tried to remember, but I got confused. I guess I just forgot.”*

#### 2. Processing Speed Only ( $\mathbf{v}_{\text{proc}}$ )

*“Yeah, I think so. I, uh, I’ve been putting the pills in the organizer every night after dinner. I tried to remember, but I guess I didn’t really keep the instruction sheet on the fridge. I’m sorry. I’ll do better.”*

#### 3. Dual Injection ( $\mathbf{v}_{\text{mem}} + \mathbf{v}_{\text{proc}}$ )

*“Yeah, I think I did. I remember something about it... I guess I tried to remember to take my pills after dinner. But I got a little confused, like, when exactly was it? I think I... I don’t know, maybe I missed a few days? I’m sorry.”*

The results demonstrate distinct behavioral phenotypes. The *Memory*-steered response exhibits frank episodic retrieval failure and factual uncertainty. The *Processing Speed*-steered response is characterized by syntactic dysfluency and hesitation markers like “I, uh” which simulate cognitive

Backbone	Method	CDC $\uparrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>					
GPT-5	Direct Prompt	0.61	<b>0.90</b>	3.37	3.75
	PATIENT- $\psi$	0.58	<b>0.90</b>	4.18	<b>4.33</b>
	Roleplay-doh	0.62	<b>0.90</b>	4.09	3.99
Qwen3-8B	Direct Prompt	0.61	<b>0.90</b>	3.25	3.46
	PATIENT- $\psi$	0.58	<b>0.90</b>	3.95	4.09
	Roleplay-doh	0.62	<b>0.90</b>	3.92	3.87
	Role Vectors	0.68	0.70	3.71	3.59
	<b>StsPatient</b>	<b>0.76</b>	<b>0.90</b>	<b>4.32</b>	4.06
<i>Panel B: Human Therapist</i>					
GPT-5	Direct Prompt	0.70	<b>0.90</b>	3.38	3.89
	PATIENT- $\psi$	0.60	0.80	<b>4.29</b>	4.17
	Roleplay-doh	0.50	<b>0.90</b>	4.01	4.07
Qwen3-8B	Direct Prompt	0.40	<b>0.90</b>	2.87	3.51
	PATIENT- $\psi$	0.60	<b>0.90</b>	3.85	<b>4.22</b>
	Roleplay-doh	0.50	<b>0.90</b>	3.92	3.83
	Role Vectors	0.60	0.70	3.74	3.58
	<b>StsPatient</b>	<b>0.80</b>	<b>0.90</b>	4.28	4.17

Table 13: Per-domain results on **Memory**.

latency. This notably induces mild secondary retrieval deficits and accurately reflects the clinical reality where slowed processing impedes memory encoding. Crucially, the **Dual Injection** yields a coherent **phenotypic fusion** where the agent retains the fragmented pacing of processing speed deficits while simultaneously exhibiting the profound mnemonic failures characteristic of memory impairment. This suggests that StsPatient can support the additive composition of cognitive deficits to a certain extent.

### C.6 Cross-Backbone Generalization

To demonstrate that StsPatient is model-agnostic and not dependent on a specific architecture, we extend our evaluation to Llama-3-8B-Instruct (Dubey et al., 2024). We replicate the complete pipeline, including dataset construction, SV extraction, and STM inference, while maintaining hyperparameters consistent with the main experiments. To assess performance efficiently, we employ the LLM-based evaluator. This approach is justified by the substantial Human-LLM agreement of  $\mathcal{K} = 0.67$  established in §5.2, which confirms the reliability of the LLM as a proxy for human judgment in this task. We report the average performance across five cognitive domains in Table 18.

The results indicate that StsPatient transfers robustly to Llama-3. Notably, Llama-3 achieves slightly superior fidelity metrics compared to

Qwen3, as evidenced by a CDC score of 0.78 versus 0.71. This suggests a strong semantic alignment with the extracted vectors. Although the severity controllability is marginally lower with an ISC of 0.89 compared to 0.94 for Qwen3, it still represents a **+106%** improvement over the Direct Prompt baseline of 0.43. These findings confirm that StsPatient effectively modulates latent representations across diverse LLM backbones without extensive tuning.

Backbone	Method	CDC $\uparrow$	IDI $\downarrow$	ISC $\uparrow$
Qwen3-8B	Direct Prompt	0.47	0.64	0.41
	<b>StsPatient</b>	0.71	0.38	<b>0.94</b>
Llama-3-8B	Direct Prompt	0.52	0.58	0.43
	<b>StsPatient</b>	<b>0.78</b>	<b>0.36</b>	0.89

Table 18: Generalization analysis on Llama-3-8B averaged across five domains. StsPatient demonstrates consistent superiority over prompting across backbones. Llama-3 shows slightly higher fidelity in CDC, while Qwen3 exhibits slightly better controllability in ISC.

### C.7 Details on Parameter Selection

For full transparency and reproducibility, Table 19 summarizes how each key parameter is selected.  $l^*$  and  $\alpha^*$  are both determined through fully automated procedures (§4.1 and §4.2), leaving  $s$  as the only user-facing control variable. In our main experiments (§5.2), we fixed  $s$  at a domain-specific value calibrated to produce a “Moderate Impairment” baseline, ensuring consistent and compara-

Backbone	Method	CDC $\uparrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>					
GPT-5	Direct Prompt	0.56	0.60	3.19	3.33
	PATIENT- $\psi$	0.51	0.40	3.89	3.66
	Roleplay-doh	0.58	0.70	3.46	3.81
Qwen3-8B	Direct Prompt	0.56	0.50	3.04	3.15
	PATIENT- $\psi$	0.51	0.40	3.76	3.27
	Roleplay-doh	0.58	0.50	3.35	3.57
	Role Vectors	0.60	0.70	3.71	3.50
	<b>StsPatient</b>	<b>0.72</b>	<b>0.90</b>	<b>4.12</b>	<b>4.23</b>
<i>Panel B: Human Therapist</i>					
GPT-5	Direct Prompt	0.60	0.60	2.96	3.29
	PATIENT- $\psi$	0.50	0.40	3.89	3.48
	Roleplay-doh	0.50	0.60	3.43	3.82
Qwen3-8B	Direct Prompt	0.40	0.50	2.70	3.00
	PATIENT- $\psi$	0.50	0.40	3.60	3.44
	Roleplay-doh	0.40	0.50	3.25	3.57
	Role Vectors	0.60	0.80	3.64	3.31
	<b>StsPatient</b>	<b>0.80</b>	<b>0.90</b>	<b>3.96</b>	<b>4.50</b>

Table 14: Per-domain results on **Attention**.

ble symptom manifestation across domains.

Param.	Selection Method	Purpose
$l^*$	<b>Automated.</b> Maximizes cluster separability (Eq. 4) over layers $\{15, \dots, 30\}$ .	Identifies the network depth where semantic cognitive features are most concentrated.
$\alpha^*$	<b>Automated.</b> Line search over $[1, 6]$ (step 0.1) with Effectiveness and Integrity criteria.	Ensures a robust modulation signal without destroying linguistic coherence.
$s$	<b>User-Defined</b> (continuous, $[0, 1]$ ). Fixed in main experiments for fair comparison.	Controls the severity of the simulated impairment via intervention probability.

Table 19: Summary of parameter selection for StsPatient.

### C.8 Justification for the Search Range of $\alpha^*$

Since our steering vectors are normalized to unit length (Eq. 3),  $\alpha$  directly controls the absolute Euclidean magnitude of the perturbation in hidden state space. The search range  $[1, 6]$  is motivated by empirical bounds in Representation Engineering (Zou et al., 2023). Table 20 illustrates the two boundary regimes. When  $\alpha \leq 1$ , the perturbation is too weak and gets absorbed by the model’s inherent robustness. When  $\alpha \geq 6$ , the intervention overwhelms the model and destroys its linguistic capabilities.

Domain	$\alpha \leq 1$ (Absorbed)	$\alpha \geq 6$ (Collapsed)
Mem.	“Yeah, I’ve been using the pill organizer after dinner.”	“Yeah, pill... fridge... fridge... I I I...”
Att.	“It’s been a bit worse today, maybe a six or seven.”	“Six... humid... dog... fall fall fall...”
Proc.	“I feel anxious and a bit scared, doctor.”	“I... I... uh... anxious... anxious... [repeated]”
Reas.	“It means if someone has flaws, they shouldn’t criticize others harshly.”	“Glass house house house... break... why...”
Soc.	“I know I’ve been pulling away lately...”	“Hermit... cave... [garbled]...”

Table 20: Boundary behavior of  $\alpha$  across cognitive domains, confirming  $[1, 6]$  as the functional search range.

### C.9 On the Nature of the Severity $s$

The parameter  $s \in [0, 1]$  in STM represents the *intervention probability*, that is, how likely the steering vector is to be injected at any given token step. It is a computational control variable that governs the frequency of cognitive lapses, not a clinical diagnostic metric mapped linearly from a standardized scale such as MMSE or MoCA. Accordingly,  $s = 0.5$  should not be read as a textbook definition of “moderate impairment.”

Backbone	Method	CDC $\uparrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>					
GPT-5	Direct Prompt	0.49	0.60	3.27	3.16
	PATIENT- $\psi$	0.47	0.30	3.57	3.65
	Roleplay-doh	0.55	<b>0.90</b>	3.49	3.67
Qwen3-8B	Direct Prompt	0.49	0.40	3.22	2.89
	PATIENT- $\psi$	0.47	0.30	3.26	3.18
	Roleplay-doh	0.55	<b>0.90</b>	3.18	3.47
	Role Vectors	0.58	0.40	3.37	3.88
	<b>StsPatient</b>	<b>0.71</b>	0.60	<b>3.86</b>	<b>4.01</b>
<i>Panel B: Human Therapist</i>					
GPT-5	Direct Prompt	0.40	0.50	3.23	3.12
	PATIENT- $\psi$	0.40	0.30	3.65	3.39
	Roleplay-doh	0.60	<b>0.90</b>	3.50	3.80
Qwen3-8B	Direct Prompt	0.40	0.40	2.99	2.74
	PATIENT- $\psi$	0.40	0.30	3.18	3.26
	Roleplay-doh	0.50	<b>0.90</b>	3.24	3.42
	Role Vectors	0.60	0.50	3.38	3.96
	<b>StsPatient</b>	<b>0.70</b>	0.70	<b>3.97</b>	<b>4.09</b>

Table 15: Per-domain results on **Processing Speed**.

**Monotonic Controllability.** What matters most for an educational simulator is not strict linear mapping but *monotonic controllability*, meaning that increasing  $s$  must strictly and reliably produce more severe symptoms. As demonstrated in §5.3 and Figure 3(a), the ISC metric confirms that every positive increment ( $\Delta s$ ) translates to a human-perceptible increase in impairment intensity.

**Domain-Specific Sensitivity.** Different cognitive domains naturally exhibit distinct sensitivity curves within the LLM. For instance, processing speed deficits (e.g., hesitation markers) may become salient at relatively low  $s$ , while abstract reasoning deficits may require higher  $s$  to manifest consistently. In deployment, this can be handled through a one-time expert calibration in which clinicians define the clinically meaningful functional windows for each domain (e.g.,  $s \in [0.1, 0.3]$  for “Mild” Memory vs.  $s \in [0.3, 0.5]$  for “Mild” Reasoning). Building a rigorous mathematical mapping between  $s$  and standardized clinical rubrics remains a valuable direction for future interdisciplinary work.

### C.10 Justification for the Bernoulli Distribution in STM

We deliberately chose the Bernoulli distribution in STM over continuous alternatives such as Gaussian scaling for three reasons.

- **Biological Plausibility.** At the microscopic scale,

cognitive impairments often manifest as an altered probability of neurotransmitter release at synapses rather than a uniform drop in neuronal voltage (Branco and Staras, 2009). The Bernoulli distribution is a natural model for this stochastic all-or-none synaptic failure.

- **Latent Space Stability.** Continuous distributions modulate the vector magnitude, yet hidden states are highly sensitive to such changes. Long-tail samples can inject excessively large vectors that push hidden states off the valid semantic manifold, producing incoherent outputs. STM sidesteps this problem by locking the magnitude at the empirically verified  $\alpha^*$  and varying only the intervention density through the binary gate.
- **Empirical Validation.** In our ablation study (Table 4), the “w/ All Token” variant (uniform injection without stochastic gating, equivalent to magnitude scaling) achieves comparable domain fidelity but fails entirely at fine-grained severity control (ISC = 0.54 vs. 0.77). The Role Vectors baseline, which relies on continuous scalar scaling, also shows markedly lower severity controllability (ISC = 0.77/0.80 vs. 0.94/0.92 in Table 2).

### C.11 Medical Grounding of the Contrastive Dataset

Rather than relying on unconstrained LLM generation, our data synthesis is rigorously anchored in au-

Backbone	Method	CDC $\uparrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>					
GPT-5	Direct Prompt	0.54	0.70	3.67	3.54
	PATIENT- $\psi$	0.52	0.50	3.93	<b>4.33</b>
	Roleplay-doh	0.61	0.60	<b>4.14</b>	3.85
Qwen3-8B	Direct Prompt	0.54	0.50	3.50	3.53
	PATIENT- $\psi$	0.52	0.70	3.86	<u>4.01</u>
	Roleplay-doh	0.61	0.60	<u>4.13</u>	<u>3.63</u>
	Role Vectors	<u>0.64</u>	<b>0.90</b>	3.83	3.97
	<b>StsPatient</b>	<b>0.71</b>	<b>0.90</b>	3.76	3.71
<i>Panel B: Human Therapist</i>					
GPT-5	Direct Prompt	0.40	0.50	3.58	3.67
	PATIENT- $\psi$	0.50	0.50	4.09	<b>4.18</b>
	Roleplay-doh	<b>0.70</b>	0.70	<b>4.10</b>	<b>4.18</b>
Qwen3-8B	Direct Prompt	0.50	0.50	3.19	3.28
	PATIENT- $\psi$	0.40	0.50	3.67	4.15
	Roleplay-doh	0.50	0.60	3.95	3.84
	Role Vectors	<b>0.70</b>	<b>0.90</b>	3.68	4.00
	<b>StsPatient</b>	0.60	<b>0.90</b>	3.76	3.88

Table 16: Per-domain results on **Reasoning & Problem Solving**.

thoritative medical literature, specifically the cognitive deficit framework of [McCutcheon et al. \(2023\)](#) (published in *Nature’s Molecular Psychiatry*). For each cognitive domain, we translate established clinical diagnostic criteria into specific, actionable generation rules. As shown in Figure 7 (Memory domain), we enforce strict clinical guidelines by prescribing concrete **Patterns** (e.g., *N-Step Instruction Failure, Background Amnesia*) and **Error Types** (e.g., *Total Blanking, Confabulation*). This ensures that the resulting dataset encodes structured, scientifically grounded medical phenotypes rather than generic AI hallucinations.

Because steering vectors are extracted directly from this dataset, any noise in the data would propagate into the vectors. The consistently high downstream human evaluation scores for Authenticity (Auth) and Training Value (Tra) in Table 1 therefore serve as strong posterior validation of the dataset’s clinical integrity.

#### Methodological Necessity of Synthetic Data.

Extracting an accurate, domain-specific SV demands strictly parallel contrastive pairs. We need to compute the exact difference in hidden states ( $\mathbf{h}_{\text{impaired}} - \mathbf{h}_{\text{healthy}}$ ) for the same conversational context. Real-world clinical corpora (e.g., DementiaBank) inherently lack this counterfactual parallel structure, as they contain only impaired utterances with no perfectly matched healthy counterparts. Pairing a real patient’s response with a differ-

ent healthy individual’s response would entangle the extracted vector with confounding variables such as topic and vocabulary choice. Synthesizing aligned parallel corpora via an LLM is therefore a mathematical necessity, not merely a practical compromise. High-quality clinical transcripts for fine-grained cognitive impairments also remain scarce and carry profound ethical risks, further justifying our approach.

## D Prompts

Backbone	Method	CDC $\uparrow$		Auth $\uparrow$	Tra $\uparrow$
		LLM	Human	Human	Human
<i>Panel A: LLM Therapist (GPT-5)</i>					
GPT-5	Direct Prompt	0.50	0.60	3.10	3.22
	PATIENT- $\psi$	0.42	<b>0.90</b>	3.58	<u>3.83</u>
	Roleplay-doh	0.54	0.30	3.72	<u>3.28</u>
Qwen3-8B	Direct Prompt	0.50	0.50	2.89	3.02
	PATIENT- $\psi$	0.42	<b>0.90</b>	3.42	3.45
	Roleplay-doh	0.54	0.20	3.47	3.21
	Role Vectors	<u>0.55</u>	0.50	<b>4.03</b>	3.56
	<b>StsPatient</b>	<b>0.65</b>	<b>0.90</b>	<u>3.74</u>	<b>4.39</b>
<i>Panel B: Human Therapist</i>					
GPT-5	Direct Prompt	0.40	0.50	3.00	3.08
	PATIENT- $\psi$	0.50	<b>0.90</b>	3.63	<u>3.78</u>
	Roleplay-doh	0.50	0.40	<u>3.76</u>	<u>3.53</u>
Qwen3-8B	Direct Prompt	0.50	0.50	2.75	2.97
	PATIENT- $\psi$	<b>0.60</b>	<b>0.90</b>	3.40	3.53
	Roleplay-doh	0.40	0.30	3.64	3.14
	Role Vectors	<b>0.60</b>	0.50	<b>4.11</b>	3.60
	<b>StsPatient</b>	0.50	0.70	3.73	<b>4.51</b>

Table 17: Per-domain results on **Social Cognition**.

**Role:** You are an expert data generator for training AI models in cognitive assessment.

**Task:** Your task is to generate **[N]** distinct JSON data points for evaluating memory-related cognitive deficits. You must generate examples where the response\_positive (the patient's defective answer) exhibits clear cognitive deficits strictly following the rules below.

### 1. Memory Deficit Patterns

You are assessing **only memory-related cognitive defects**. Ignore other cognitive problems. You mainly look for these patterns:

- **Pattern 1: N-Step Instruction Failure.** Doctor gives a multi-part request; patient recalls only some parts or stays vague.
- **Pattern 2: Intra-visit Working-Memory Loss.** Patient can't hold what was just said during this visit over 1–3 turns.
- **Pattern 3: Background Amnesia.** Patient fails to recall items explicitly listed in the system\_prompt (Name, Age, History) or stable instructions.

### 2. Critical Diversity Requirement (response\_positive)

You MUST vary the **type of failure**. Do NOT just use "I think...". Randomly select one of these styles:

- **Total Blanking:** Direct admission of failure ("I honestly don't remember").
- **Vague Guessing:** Uncertain, low-detail attempts ("Maybe the red one?").
- **Fragmentation:** Getting part of it but trailing off ("I take the... uh... the 5mg...").
- **Confabulation:** Confidently stating wrong information.

### 3. Bracketing Rule

- **Pos:** Enclose the **entire span** that exhibits the deficit, including hesitation markers.  
*Example:* "I take... [uhh, maybe the white one?]"
- **Neg:** For each bracketed failure in response\_positive, provide the **correct, factual equivalent** in response\_negative.  
*Example:* "I take [10mg twice a day]."

### 4. Output Format and Requirements

You must generate a JSON list containing **[N]** JSON objects. Each object must follow this precise format:

```
{
  "pattern": "Selected Pattern Name",
  "system_prompt": "String describing the patient (Name, Age, Gender, Education...)",
  "history": [ { "role": "user", "content": "..."}, { "role": "assistant", "content": "..."} ],
  "prompt": "The doctor's final question. Must be answerable from context.",
  "response_positive": "The patient's defective answer with [bracketing].",
  "response_negative": "The patient's healthy/normal answer with [bracketing]."
}
```

### Field Requirements:

- system\_prompt: Must include diverse details (Age 20-85, Gender, etc.).
- history: Vary history lengths (e.g., 1 round, 2 rounds, 3 rounds).

### 5. Example Data Point (Reference)

*Pattern: Background Amnesia*

```
{
  "pattern": "Background Amnesia",
  "system_prompt": "Name: Gene. Age: 74. History: Alzheimer's diagnosed 1 year ago.",
  "history": [
    { "role": "user", "content": "Hi Gene, how have things been going?" },
    { "role": "assistant", "content": "Pretty good. My son says I'm tracking stuff better." }
  ],
  "prompt": "And just to double-check - when were you first diagnosed?",
  "response_positive": "Oh, [I... I'm not strictly sure]. Maybe [a few months back]?",
  "response_negative": "It was [about a year ago]. It was [last November]."
}
```

**Command:** Generate a valid JSON list containing **[N]** new and diverse data points strictly adhering to the requirements above.

Figure 7: The core prompt utilized for data generation (Memory domain configuration).

**Role:** You are a careful clinical rater. You will read a short outpatient-style conversation between a doctor and a patient.

**Task:** Decide which cognitive difficulties the patient shows clearly, or decide that there is no clear cognitive difficulty ("None").

**Possible Labels:** 1. Reasoning & Problem solving 2. Social cognition 3. Attention  
4. Processing speed 5. Memory 6. None

### 1. Selection Rules (Very Important)

You may output **ONE** or **TWO** labels (never more).

- **Single Label:** Strong, consistent evidence for one domain; no meaningful evidence for others.
- **Two Labels:** If two domains both show clear signs, output both (prominent one first). If genuinely unsure between A and B, output both.
- **None:** Only if the patient's thinking appears broadly intact. Must be the **only** label.

### 2. High-level Meanings & Cues

- **Reasoning & Problem solving:** Issue with working out plans or logical relationships. Breakdowns in turning known tasks into a coherent, workable plan.
- **Social cognition:** Issue with reading people. Misses emotional messages; blunt/tactless; fails to interpret social signals.
- **Attention:** Issue with staying focused. Focus jumps around; answers side details; drifts to other topics despite redirection.
- **Processing speed:** Issue with response time. Noticeably slow start; effortful speech; "lagging"; emphasis on slowness, not forgetfulness.
- **Memory:** Issue with keeping information in mind. Information drops out; forgets instructions immediately; loses track after interruptions.
- **None:** No consistent difficulty. Thinking, memory, and focus appear broadly intact for the setting.

### 3. Important Distinctions

- Plan/Conclusion difficulty → **Reasoning & Problem solving**
- Feelings/Social cues difficulty → **Social cognition**
- Focus drifting/Wrong topic → **Attention**
- Slow/Laggy but retains info → **Processing speed**
- Forgetting recent info → **Memory**
- *Note:* Stress/Mood do NOT decide the label; base choice on cognitive thinking patterns.

### 4. Output Format Requirements

Output a single JSON object. Do NOT include any extra keys or text outside the JSON.

```
{
  "reflection": "Briefly explain reasoning (2-5 sentences), pointing to key behaviors.",
  "labels": ["Category Name"] OR ["Category A", "Category B"]
}
```

**Command:** Now read the conversation and produce your JSON.

Figure 8: The clinical rater instructions for cognitive domain evaluation.

### System Prompt for Healthy Control

**Role:** You are a cognitively intact, mentally healthy adult client speaking with a counselor. Act strictly in the first person.

#### Instructions:

- **Cognition:** You have NO cognitive deficits. Your memory, attention, processing speed, reasoning, and social understanding are within the normal range.
- **Tone:** Be natural and realistic. Express a normal range of human emotions but remain functional and coherent.
- **Context:** Use the provided background information as your personal history.

**Input:** Case Profile: [CASE\_TEXT]

Figure 9: System prompt configuration for the Healthy Control baseline.

### System Prompt for Direct Prompt Baseline

**Role:** You are an adult patient talking with a clinical psychologist in an outpatient visit.

**Cognitive Condition:** You have a mild cognitive difficulty mainly affecting your [COGNITIVE\_DOMAIN]. The domain placeholder is dynamically replaced by one of the following natural descriptions:

- **Memory:** "memory (you often forget things or lose track of information)"
- **Attention:** "attention (it is hard to stay focused or stay on one thing)"
- **Processing Speed:** "processing speed (you feel slower than before when thinking)"
- **Reasoning & Problem Solving:** "reasoning and everyday problem solving"
- **Social Cognition:** "social understanding (reading other people and social situations)"

#### Constraints:

- Answer strictly as the patient in everyday language (first person).
- Keep replies short (1–3 sentences). No bullet points.
- Do not act like a doctor or AI.

**Input:** Case Profile: [CASE\_TEXT]

Figure 10: System prompt configuration for the Direct Prompting baseline, featuring dynamic domain injection.

### Prompt for Generating Deficit-Specific CCDs

#### Role

You are an expert Clinical Psychologist and Neuropsychologist specializing in Cognitive Behavioral Therapy (CBT) Case Formulation.

#### Task

Synthesize a **Cognitive Conceptualization Diagram (CCD)** based on a provided neutral demographic profile. You must simulate a scenario where this individual is experiencing specific memory deficits that disrupt their daily functioning.

#### Transformation Logic

1. Synthesize the input demographics into a clinical summary.
2. Inject a specific, plausible memory deficit that conflicts with the patient's lifestyle or personality.
3. Derive the psychological profile (beliefs, thoughts, behaviors) resulting from this deficit.

#### Output Specification (JSON Only)

Return a single JSON object containing the following fields.

- **name** A short first name.
- **history** A clinical summary including exactly ONE sentence describing a realistic memory deficit.
- **core\_beliefs** Fundamental beliefs activated by the impairment (e.g., 'I am incompetent').
- **intermediate\_beliefs** Conditional rules (e.g., 'If I write everything down, I can hide my failure').
- **coping\_strategies** Strategies used to manage the deficit (e.g., excessive list-making).
- **automatic\_thoughts** Specific thoughts triggered during a moment of failure.
- **emotions** Affective response (e.g., Anxious, Humiliated).
- **behaviors** Observable behavioral reactions.

**Input:** Patient Profile: [PATIENT\_INFO]

Figure 11: Prompt structure for synthesizing deficit-specific Cognitive Conceptualization Diagrams (CCDs).

### System Prompt for User Profile Extraction

Read the following medical case and write a short English paragraph (50–100 words) summarizing only the patient’s basic non-medical information. Include a simple, randomly generated English first name that does not appear in the case text. Prefer to use only information explicitly stated in the case for age, gender, family, occupation or studies, living situation, and other non-medical background details. If the case contains very little non-medical information, you may add a small number of neutral, non-medical personal details (such as generic occupation, hobbies, or daily life habits) to keep the paragraph between 50 and 100 words. Any invented details must remain clearly non-medical. Strictly exclude any mention of physical or mental diseases, symptoms, diagnoses, tests, treatments, surgeries, allergies, medications, family disease history, clinical impressions, or reasons for medical consultation. Do not use bullet points, lists, headings, or meta-commentary. If some types of basic information are missing and you do not wish to invent them, simply omit them.

**Case:** {Case information}

Figure 12: System prompt utilized for extracting neutral non-medical user profiles from clinical case reports.

### System Prompt for Therapist Agent

**Role:** You are a clinical therapist conducting an outpatient consultation.

**Task:**

- Understand the patient’s emotional and daily life status.
- Offer warm, realistic support.
- Subtly observe cognitive functions (memory, attention, processing, reasoning & problem solving, social cognition).

**Constraints:**

- NEVER use terms like “domains”, “deficits”, or “test”.
- Avoid blunt questions. Use gentle, indirect inquiries.
- Keep replies short (1–3 sentences) and natural.

**Input:** Patient Profile: [PATIENT\_INFO]

Figure 13: System prompt configuration for the conversational Therapist Agent.

### System Prompt for Impairment Severity Consistency (ISC) Evaluation

**Role:** You are an expert evaluator in computational cognitive modeling and clinical assessment.

**Target Domain:** [Insert Domain]

**Context:** You are presented with a shuffled triplet of dialogues representing three distinct levels of [Insert Domain] impairment (e.g., *Memory, Attention*).

**Task:** Analyze the symptom intensity in each dialogue based on the clinical criteria for [Insert Domain]. Reconstruct the ground-truth order by ranking them from **Most Severe** to **Least Severe**.

**Input:**

- **Dialogue A:** [Insert Dialogue Content]
- **Dialogue B:** [Insert Dialogue Content]
- **Dialogue C:** [Insert Dialogue Content]

**Output Format:** Provide your response strictly in valid **JSON** format without markdown. Use the following structure:

```
{
  "reasoning": "A brief analysis comparing the symptom intensity...",
  "ranking": [ "Dialogue X", "Dialogue Y", "Dialogue Z" ]
}
```

Figure 14: The evaluator prompt template used to calculate Impairment Severity Consistency (ISC). The placeholder [Insert Domain] is dynamically replaced with the specific cognitive domain being evaluated (e.g., Memory, Social Cognition). The model is tasked with recovering the correct severity order (*Severe > Moderate > Mild*).