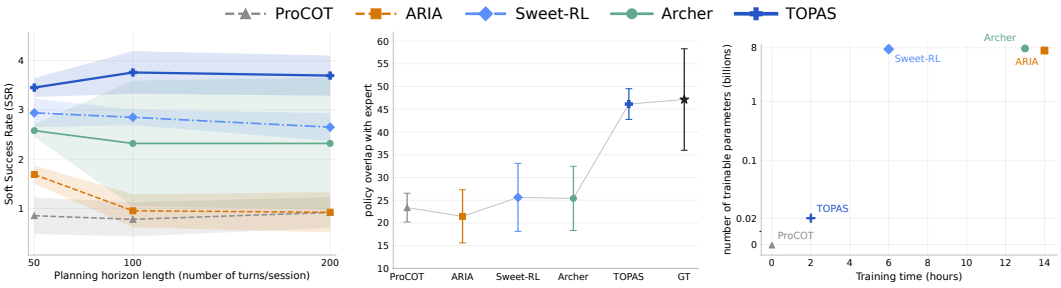
FROM TEXTBOOKS TO PROACTIVE AGENTIC SYSTEMS: SCALING LONG-HORIZON PLANNING VIA AUTOMATED AGENTIC SYSTEM CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

While large language models (LLMs) excel at locally coherent generation, they remain fundamentally *reactive*, struggling to plan and steer long-horizon interactions. Existing prompting and reinforcement learning (RL) finetuning methods degrade as dialogue horizons scale. Augmenting LLMs with explicit planners and memory modules offers a promising alternative, but manually specifying these architectures requires substantial expert design effort, limiting scalability. We introduce TOPAS (*T*extbooks t*O* *P*roactive *A*gentic *S*ystems), a framework that automatically constructs executable agentic systems directly from instructional domain textbooks. From unstructured text, TOPAS jointly extracts a hierarchical planner over a structured decision space (states, actions, rewards), a user simulator, and a graph-based memory to establish a complete interaction loop. It then generates supervision via uncertainty-aware self-annotation and optimizes hierarchical policies over this substrate using RL. Experiments on mental health support and persuasive dialogue benchmarks demonstrate that TOPAS achieves state-of-the-art performance and sustained long-horizon planning, outperforming prompting, RL finetuning, and hand-crafted planning baselines with up to 23.5% improvement in performance while being parameter and compute efficient. Code is publicly available at https://anonymous.4open.science/r/TOPA-30B2/.



(a) Success rate across sessions with increasing number of turns

(b) Overlap with expert policies (% expert micro-actions covered)

(c) Train-/parameter- efficiency

Figure 1: TOPAS significantly outperforms baselines relying on reasoning via prompting (PROCOT Deng et al. (2023b)) and RL-finetining (ARCHER Zhou et al. (2024), SweetRL Zhou et al. (2025a), ARIA Yang et al. (2025a)) across extended interaction horizons (up to 200 turns)

.

1 INTRODUCTION

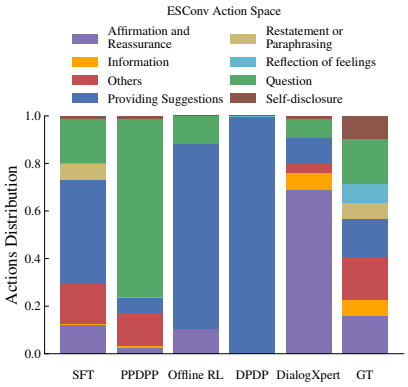
Despite recent advances in large language models (LLMs), most deployed agents remain fundamentally *reactive* in long dialogue settings: they generate locally coherent responses based on recent context but do not deliberately shape interactions over extended horizons. This limitation becomes especially pronounced when success depends on sustained, goal-directed behavior unfolding across many steps or sessions, particularly in subjective or noisy dialogue domains such as therapy, negotiation, or persuasion. In such settings, agents must reason about future states, track evolving context, and strategically regulate the timing and content of their responses, deciding what information to reveal and when, to guide the interaction toward long-term objectives.

To address this limitation, recent work on proactive dialogue modeling has largely followed two directions. One line of work elicits proactivity directly from dialogue LLMs via *prompting* or *fine-tuning* Deng et al. (2023b); Zhang et al. (2023); Christiano et al. (2023); Glaese et al. (2022); Ramamurthy et al. (2023); Hong et al. (2023); Korbak et al. (2023); Liu et al. (2026). While effective in short contexts, these methods often struggle to maintain coherent long-horizon control Srivastava

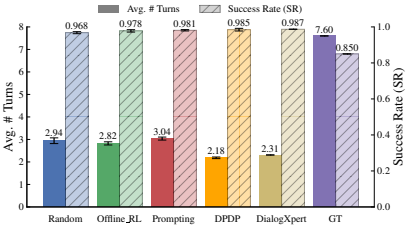
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al. (2023); Wang et al. (2026e;d) and rely on natural-language rationales that may not reflect stable internal decision processes Bowman et al. (2022); Arcuschin et al. (2025). More fundamentally, they must represent long-term goals and immediate responses within a single token-level state space, which can yield brittle behaviors and reward hacking Wu et al. (2026); Wang et al. (2026c).

A second line of work augments dialogue models with explicit planning structures, often formulated as Markov Decision Processes (MDPs) with policies trained via reinforcement learning Deng et al. (2024b); He et al. (2024); Lin et al. (2026); Peng et al. (2026); Rakib et al. (2026). By decoupling high-level strategy selection from surface utterance generation, these methods reduce token-level complexity. Yet, they depend on manually designed symbolic spaces (states, actions, rewards, and simulators), which are costly to engineer and frequently underspecified. As a result, policies can overfit brittle success criteria and exploit overly collaborative simulators via *reward hacking* (Fig. 2). This motivates a central question: how can we enable long-horizon proactive dialogue without extensive manual engineering?



(a) Learned policies are biased toward a single action type and fail to match the ground-truth (GT) action distribution. We compare planners trained with SFT, Offline RL (PPDPP (Deng et al., 2024a)), online RL (DPDP (He et al., 2024)) and a value-based RL policy with LLM-generated action priors (DialogXpert (Rakib et al., 2026)).



(b) All baselines reach surprisingly high success rates (SR), even exceeding expert performance (GT) while requiring an implausibly low number of turns (#). SR is LLM-judged post-session patient satisfaction.

Figure 2: Failure modes of planner-augmented LLMs on ESConv dataset Liu et al. (2021). High success rates mask shallow policies and collapsed policies due to reward hacking and over-cooperative simulators.

In this work, we argue that planner augmentation can significantly extend the planning horizon, provided that the underlying MDP accurately captures the richness and complexity of real-world interactions. The key bottleneck, therefore, lies not in the planner itself, but in the difficulty of specifying such an MDP manually. Instead of hand-crafting proactive structures, we propose to automatically derive them from domain knowledge sources that offer comprehensive and structured representations of expert strategies. In particular, we leverage textbooks, such as therapy manuals, persuasion guides, and other pedagogical materials, which are explicitly designed to teach structured intervention strategies, long-term planning principles, and adaptive decision-making. Building on this insight, we introduce Textbooks tO Proactive Agentic Systems (TOPAS), a framework that automatically derives hierarchical MDP components as well as user-simulator profiles and memory structure. This enables the training of proactive dialogue agents with minimal expert supervision while grounding their behavior in well-established domain knowledge and improving generalization. Our approach combines theoretical and practical sources of knowledge: textbooks are used to construct the symbolic structure of the domain, while real-world interaction data (*e.g.*, therapy session datasets such as Alexander Street Press (2007)) are used to train and refine the agents strategic behavior. By integrating structured pedagogical knowledge with empirical interaction data, TOPAS enables the learning of proactive agents that are both principled and practically effective.

Our evaluation on multi-turn dialogue benchmarks demonstrates that TOPAS outperforms existing SOTA prompting and RL fine-tuning baselines. Specifically, TOPAS yields up to **23.5% performance improvement** in soft success rate on the Alexander Street dataset and **15.5%** on Persuasion for Good across extended horizons (up to 200 turns), while operating with **99.7% fewer trainable parameters** and achieving up to 7× **speedup** in training.

2 Related Work

Proactive agents aim to steer interactions toward long-horizon goals rather than react locally Deng et al. (2023a); Liu et al. (2026). Existing approaches induce goal-oriented behavior via prompting (Sel

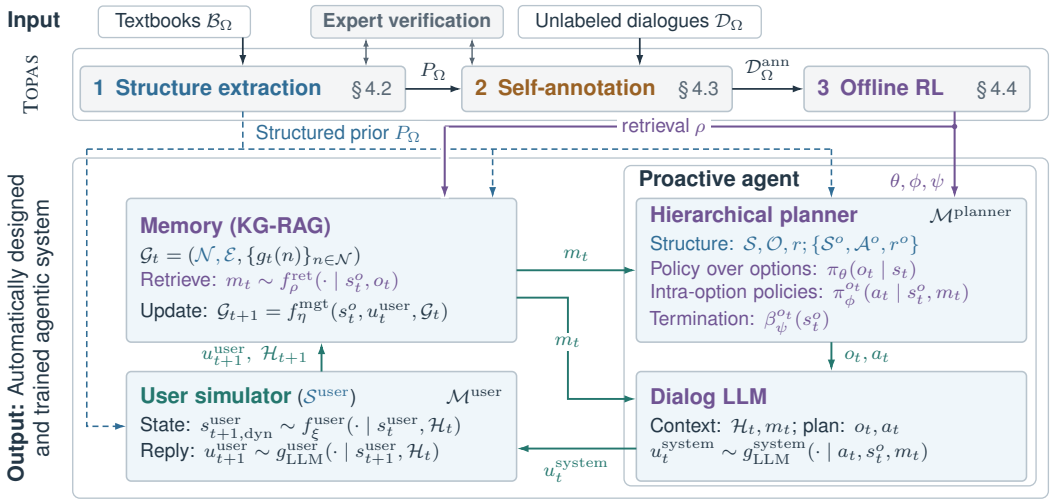


Figure 3: **Overview.** TOPAS transforms domain textbooks $\mathcal{B}_\Omega$ and unlabeled dialogues $\mathcal{D}_\Omega$ into a proactive system via three phases: (i) extracting MDP, user-state, and memory priors from $\mathcal{B}_\Omega$; (ii) inferring latent trajectory annotations (options, actions, states, rewards) in $\mathcal{D}_\Omega$; and (iii) training hierarchical planner and retrieval policies via offline RL on $\mathcal{D}_\Omega^{\text{ann}}$.

et al., 2025; Yao et al., 2023; Elizabeth et al., 2025; Kim et al., 2025), supervised fine-tuning and RL with verifiable rewards (Zheng et al., 2023; Wei et al., 2025; Wang et al., 2026b), or external planner modules (Deng et al., 2024a; He et al., 2024; Rakib et al., 2026). However, token-level prompting and fine-tuning suffer from test-time myopia, accumulating local errors over multi-turn interactions (Wang et al., 2026e). While look-ahead search mitigates this degradation (Wang et al., 2026e), searching over raw token spaces is computationally prohibitive and fails to capture hierarchical strategies. **Planner-augmented LLMs** decouple strategy tracking from language generation, yet they rely heavily on *hand-specified* MDP components, causing domain brittleness and high engineering overhead Deng et al. (2024b); He et al. (2024). Conversely, learning abstractions directly from offline logs (Wang et al., 2026a) is bounded by dataset coverage and incapable of guaranteeing normative domain consensus (*e.g.*, clinical protocols). Differently, TOPAS bypasses these limitations by automatically extracting grounded, consensus-derived substrates directly from domain textbooks.

Memory and structured context. Preserving information over extended sessions requires moving beyond flat retrieval toward structured, graph-based representations of experience (Zhou et al., 2025b; Yan et al., 2026; Kang et al., 2025; Yang et al., 2026). However, current memory schemas are typically hand-crafted or dynamically induced from noisy logs, rendering them susceptible to concept drift and ungrounded accumulation. By contrast, TOPAS derives its memory schema directly from domain textbooks, explicitly aligning stored propositions with the normative concepts, protocol milestones, and decision variables used by the planner.

User modeling and simulation. Proactive dialogue requires anticipating how latent user states evolve in response to system interventions. User simulators model preferences, behavioral dynamics, and clinically motivated constructs (*e.g.*, trust, resistance) (Abdulhai et al., 2025; Yang et al., 2025c;b; Sabour et al., 2026). Yet, like existing planning and memory substrates, these state representations remain hand-designed or inferred purely from interaction data. TOPAS grounds the user-state space in textbook domain knowledge, using interaction logs solely to instantiate and refine transition dynamics over this fixed normative structure.

3 BACKGROUND

A **Markov Decision Process (MDP)** (Bellman, 1957) is defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \pi, r, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $p(s'|s,a)$ the transition probability from state s to s' under action a, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ the reward function. The policy $\pi(a|s)$ specifies the probability of selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and $\gamma \in [0, 1)$ is the discount factor. The action-value function under policy π is defined as the expected discounted return starting from state s, taking action a, and thereafter following π: $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 = s,\ a_0 = a]$. The value function $V^\pi(s)$ is

then given by the expectation of $Q^\pi(s,a)$ under the policy: $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$.
The Option Framework (Sutton et al., 1999; Precup and Sutton, 2000) provides a formalism for hierarchical reinforcement learning enabling a high-level policy (macro-policy) to operate over low-level ones (micro-policies). An option $o \in \mathcal{O}$ is defined as a tuple $o = (\mathcal{I}^o, \pi^o, \beta^o)$, where $\mathcal{I}^o \subseteq \mathcal{S}$ denotes the initiation set, $\pi^o(a|s)$ is the micro-policy implementing the option, and $\beta^o(s) \in [0,1]$ is the termination function. Each micro-policy π_o can be viewed as implementing an MDP $\mathcal{M}^o = (\mathcal{S}^o, \mathcal{A}^o, p, r^o, \pi^o, \gamma^o)$ defined over a state-action space and associated with an option-specific objective (*e.g.*, reaching a subgoal). Execution follows a call-and-return model: once an option $o \in \mathcal{O}$ is initiated, it governs action selection through its micro-policy $\pi^o(a|s)$ until termination, which is determined stochastically by $\beta^o(s)$ after each transition. Upon termination, control returns to the policy over options $\pi(o|s)$, which selects the next option. This enables structured long-horizon decision-making by decoupling high-level planning from low-level control.

4 Approach

As depicted in Fig. 3, TOPAS-generated system integrates three core components: (i) a hierarchical planner guiding a dialog LLM, (ii) a knowledge-graph-based retrieval-augmented memory (KG-RAG), and (iii) a persona-grounded user simulator. Rather than manually engineering these components or inducing them from noisy interaction logs, TOPAS formalizes the symbolic substrate of each module directly from authoritative texts, *e.g.*, by instantiating the planner as an MDP with textbook-derived state space, action space and rewards, alongside aligned schemas for memory and user simulation. This extracted structure is subsequently leveraged to auto-annotate unlabeled conversational corpora, producing a dense supervision for training each module. Ultimately, TOPAS replaces two costly extremes: tedious expert specification and ungrounded end-to-end learning from data.

4.1 Problem Formulation

Let Ω denote a target intervention domain (*e.g.*, cognitive behavioral therapy). We are given a collection of domain textbooks $\mathcal{B}_\Omega = \{b_i\}_{i=1}^{|\mathcal{B}_\Omega|}$ and a dataset of unlabeled conversations $\mathcal{D}_\Omega = \{d_i\}_{i=1}^{|\mathcal{D}_\Omega|}$. Each conversation $d \in \mathcal{D}_\Omega$ is represented as a sequence of user and system utterances, $d = \{(u_t^{\text{user}}, u_t^{\text{system}})\}_{t=1}^{|d|}$, where u_t^{user} denotes the user (*e.g.*, patient) utterance and u_t^{system} the system (*e.g.*, expert practitioner) response at turn t. Our goal is to learn policies capable of long-horizon planning in Ω. Achieving this requires addressing three key challenges: (i) selecting coherent high-level actions over extended horizons, (ii) maintaining and retrieving structured context across turns, and (iii) modeling the evolving user state under partial observability. These challenges naturally motivate an agentic system with a planner for long-horizon decision-making, a structured memory for persistent context, and a user simulator for modeling interaction dynamics.

We formalize **the planner** as a hierarchical MDP $\mathcal{M}^{\text{planner}}$ following the *Option Framework*: $\mathcal{M}^{\text{planner}} = (\mathcal{S}, \mathcal{O}, p^{\text{planner}}, \pi_\theta, \{\mathcal{M}^o, \beta_\psi^o\}_{o\in\mathcal{O}}, r, \gamma)$, where the macro-state space $\mathcal{S}$ provides a high-level structural *abstraction* over dialogue history $d_{1:t}$ (*e.g.*, tracking global intervention phases), $\mathcal{O}$ is a set of high-level options, and $\pi_\theta(o|s_t)$ is the macro-policy parameterized by θ. Each option $o \in \mathcal{O}$ is associated with a sub-MDP $\mathcal{M}^o = (\mathcal{S}^o, \mathcal{A}^o, p^{\text{planner}}, \pi_\phi^o, r^o, \gamma)$ governing its intra-option dynamics, where the micro-state $s_t^o \in \mathcal{S}^o$ captures raw dialogue history $d_{1:t}$ representations, $\mathcal{A}^o$ is the option-specific micro-action space, and $\pi_\phi^o(a_t|s_t^o, m_t)$ denotes the micro-policy parameterized by ϕ generating system turn actions $a_t \in \mathcal{A}^o$. To produce natural dialogue, a_t is passed to a **dialogue LLM** $g_{\text{LLM}}^{\text{system}}(u_t^{\text{system}} \mid a_t, s_t^o, m_t)$ that realizes the abstract action into the surface system utterance u_t^{system}. The function $\beta_\psi^o(s_t^o)$ denotes the option termination condition parameterized by ψ, while r and γ represent the global reward function and discount factor.

The **memory** is formalized as a Knowledge-Graph RAG $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}, \mathcal{C}_t)$, where nodes $\mathcal{N}$ and edges $\mathcal{E}$ denote the structural domain ontology, while $\mathcal{C}_t = \{c_t(n)\}_{n\in\mathcal{N}}$ captures dynamic node contents (*e.g.*, accumulated user context and progress markers). At each turn t, a graph management function f_η^{mgt} : $(s_t^o, u_t^{\text{user}}, \mathcal{C}_t) \mapsto \mathcal{C}_{t+1}$ parameterized by η updates node contents in response to the user utterance u_t^{user} and micro-state s_t^o. Memory retrieval is governed by a policy $f_\rho^{\text{ret}}(m_t|s_t^o, o_t)$ parameterized by ρ, which selects a relevant set of k graph triples $m_t = \{(n_i, e_i, n_i')\}_{i=1}^k \subseteq \mathcal{N}\times\mathcal{E}\times\mathcal{N}$ conditioned on the active option o_t and micro-state s_t^o. The retrieved memory m_t directly conditions the micro-policy $\pi_\phi^o(a_t|s_t^o, m_t)$ to ground next-action generation in structured context.

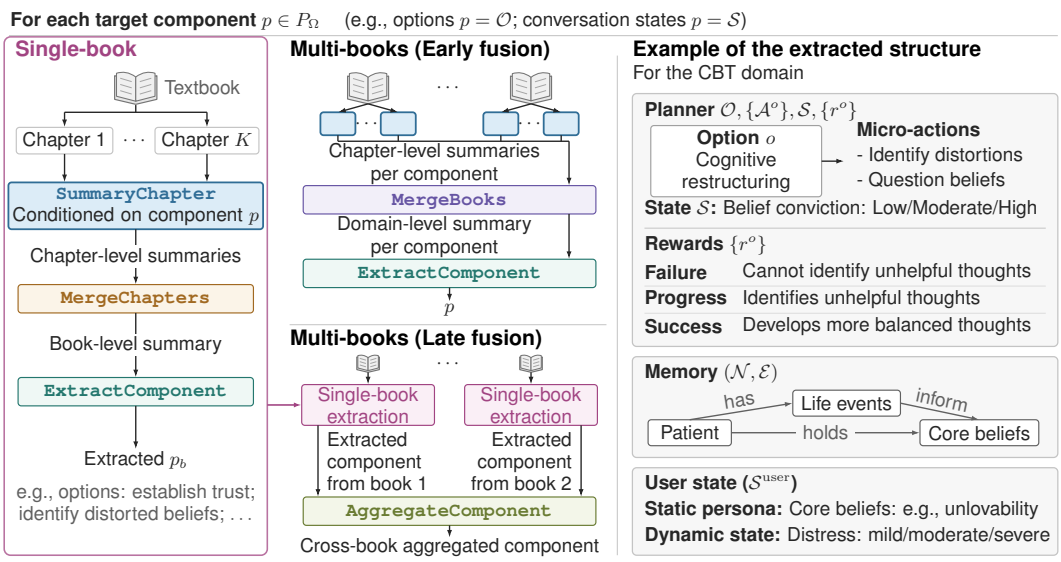


Figure 4: **Structure extraction pipelines** for single-book and multi-books (early-/late- fusion).

We model **the user simulator** as a two-stage, parameterized transition and surface-realization process $\mathcal{M}^{\text{user}} = (\mathcal{S}^{\text{user}}, f_\xi^{\text{user}}, g_{\text{LLM}}^{\text{user}})$. The user state $s_t^{\text{user}} = (s_{\text{stat}}^{\text{user}}, s_{\text{dyn}}^{\text{user}}) \in \mathcal{S}^{\text{user}}$ decomposes into static dimensions $s_{\text{stat}}^{\text{user}}$ capturing invariant persona attributes (*e.g.*, baseline demographics and clinical history) and dynamic dimensions $s_{t,\text{dyn}}^{\text{user}}$ representing time-varying internal states (*e.g.*, trust, resistance). Transitions over the dynamic state dimensions are governed by a transition function $f_\xi^{\text{user}}(s_{t+1,\text{dyn}}^{\text{user}} | s_t^{\text{user}}, d_{1:t})$ parameterized by ξ, which updates the dynamic user state conditioned on the full current user state s_t^{user} and dialogue history $d_{1:t}$. Subsequently, the updated user state s_{t+1}^{user} is fed into a dialogue LLM $g_{\text{LLM}}^{\text{user}}(u_{t+1}^{\text{user}} \mid s_{t+1}^{\text{user}}, d_{1:t})$ to realize the latent internal state into the surface-level user utterance u_{t+1}^{user}.

Optimizing these modules directly from raw dialogue is hard for two reasons: high-level strategic decisions and low-level text generation operate at fundamentally different granularities, and the resulting long horizons over a combinatorial action space make credit assignment exceedingly difficult. To address this, TOPAS **extracts a structured prior** from textbooks. Specifically, we extract: (i) a discrete option set $\mathcal{O}$, macro- and micro-state spaces $(\mathcal{S}, \{\mathcal{S}^o\})$, micro-action spaces $\{\mathcal{A}^o\}$, and reward specifications $(r, \{r^o\})$ for the planner; (ii) a domain ontology $(\mathcal{N}, \mathcal{E})$ for memory; and (iii) a structured user state space $\mathcal{S}^{\text{user}}$ for the simulator. Given this extracted structure, we introduce an **annotation stage** that augments $\mathcal{D}_\Omega$ with labels derived from these textbook priors, yielding $\mathcal{D}_\Omega^{\text{ann}}$. Concretely, each dialogue turn is annotated with its corresponding option, intra-option action, state variables, and rewards, transforming unlabeled dialogues into supervision signals. While pretrained LLMs naturally excel at surface text realization and in-context state tracking under structured prompts, long-horizon strategic execution and precise context selection require explicit optimization. Thus, we leverage $\mathcal{D}_\Omega^{\text{ann}}$ to train the core decision, *i.e.*, making policies—the hierarchical planner $(\pi_\theta, \{\pi_\phi^o\})$ and memory retriever f_ρ^{ret}, via **offline RL**, while instantiating state-tracking and surface-realization components using prompted LLMs. In the following sections, we describe the three core stages of TOPAS: structure extraction, data annotation, and policies training.

In the following, we describe the three stages of TOPAS: structure extraction, annotation, and training.

4.2 Extracting Structure from Textbooks

In this section, we describe how the structured domain prior $\mathcal{P}_\Omega = (\mathcal{O}, \mathcal{S}, \{\mathcal{S}^o\}, \{\mathcal{A}^o\}, r, \{r^o\}, (\mathcal{N}, \mathcal{E}), \mathcal{S}^{\text{user}})$ is automatically extracted from a collection of domain textbooks $\mathcal{B}_\Omega$. Extracting $\mathcal{P}_\Omega$ from textbooks, however, is non-trivial and poses three key challenges: (i) *long-context reasoning* requires capturing dependencies across chapters to recover the global structure of an intervention; (ii) *fine-grained extraction* needs to retain localized details such as reward signals and micro-actions; and (iii) *cross-textbook consolidation* requires to reconcile complementary and potentially inconsistent knowledge across sources. The first two objectives are inherently in tension. Methods designed for long-context modeling, such as state-space models (Gu and Dao, 2024) and sparse attention mechanisms (Beltagy et al., 2020; Zaheer et al., 2020), can

capture cross-chapter dependencies but often rely on compressed representations that degrade access to fine-grained details, leading to factual inconsistencies and hallucinations (Liu et al., 2023). In contrast, retrieval-based approaches such as RAG (Lewis et al., 2020) preserve local fidelity by operating on small, relevant chunks, but lack mechanisms to recover global structure or sequential dependencies across documents. This trade-off between global coherence and local precision is a well-known limitation in long-context modeling and document-level reasoning. The difficulty further compounds in the multi-textbook setting. Sequential processing introduces recency bias and proactive interference (Wang and Sun, 2025; Chattaraj and Raj, 2026), while early fusion of heterogeneous sources can obscure important distinctions and reduce faithfulness to individual textbooks, requiring non-trivial conflict resolution beyond simple aggregation (Ye et al., 2026). These challenges highlight the need for an extraction procedure that simultaneously preserves hierarchical structure, local evidence, and cross-source consistency. Our key contribution is a hierarchical, component-conditioned extraction procedure that decomposes global structure recovery into local evidence aggregation followed by structured consolidation, enabling the simultaneous preservation of fine-grained details and long-range dependencies.

Single-textbook extraction. As shown in Fig. 4, given a textbook $b \in \mathcal{B}_\Omega$, we first perform component-conditioned summarization at the chapter level. For each component $p \in \mathcal{P}_\Omega$, we compute $z_{b,k}^{(p)} = \textsf{SummaryChapter}(k; p)$ for all chapters $k \in b$, where SummaryChapter extracts structured information relevant to p using a component-conditioned LLM prompt. We then aggregate chapter-level summaries to obtain a global summary $z_b^{(p)} = \textsf{MergeChapters}(\{z_{b,k}^{(p)}\}_{k \in b})$, followed by structured extraction $p_b = \textsf{ExtractComponent}(z_b^{(p)})$. Intuitively, the SummaryChapter stage preserves fine-grained evidence, while MergeChapters integrates information across chapters to recover global structure. The operator ExtractComponent is tailored to each component type and respects hierarchical dependencies. Specifically, high-level components, *i.e.*, options $\mathcal{O}$ and the global reward specification r, are extracted first; their outputs are then used to condition the extraction of option-specific micro-actions $\{\mathcal{A}^o\}$ and micro-rewards $\{r^o\}$. Note that rewards $(r, \{r^o\})$ are modeled as rubrics detailing conditions for conversation/options success, intermediate progress, and failure.

Cross-textbook consolidation. To scale to a collection of textbooks $\mathcal{B}_\Omega$, we consider two fusion strategies: (i) *Early fusion*, where summaries are merged across all textbooks prior to extraction: $z^{(p)} = \textsf{MergeBooks}\left(\{z_{b,k}^{(p)} : b \in \mathcal{B}_\Omega, k \in b\}\right)$ and $p = \textsf{ExtractComponent}(z^{(p)})$; and (ii) *Late fusion*, where each textbook $b \in \mathcal{B}_\Omega$ is processed independently to produce per-book component candidates p_b, followed by component-level aggregation: $p = \textsf{AggregateComponent}\left(\{p_b : b \in \mathcal{B}_\Omega\}\right)$. Aggregation is performed via a majority-based selection scheme, retaining elements that consistently recur across textbooks while filtering out spurious or source-specific artifacts. This procedure preserves a robust, consensus domain prior $\mathcal{P}_\Omega$ while mitigating noise and contradictions across sources. Each extracted component element is stored with source provenance, enabling full traceability and auditing. Prompts are in App. E.

4.3 Structural Self-Annotation and Refinement

To transform the unlabeled dataset $\mathcal{D}_\Omega$ into a supervised training corpus $\mathcal{D}_\Omega^{\text{ann}}$, we cast annotation as latent variable inference under the extracted prior $\mathcal{P}_\Omega$. Given a dialogue $d \in \mathcal{D}_\Omega$, we define a hierarchical latent trajectory $\tau = \{\tau_t\}_{t=1}^{|d|}$, where $\tau_t = (s_t, o_t, s_t^o, a_t^o, r_t^o, r_t)$ if u_t is a system utterance, and $\tau_t = (s_t^{\text{user}})$ if u_t is a user utterance (see Fig. 5). Formally, we infer the optimal trajectory τ^* by maximizing the joint likelihood:

$$\tau^* = \arg\max_\tau \log p(d, \tau \mid \mathcal{P}_\Omega) = \arg\max_\tau \left[\log p(d \mid \tau) + \log p(\tau \mid \mathcal{P}_\Omega)\right], \tag{1}$$

where $p(d|\tau)$ measures semantic fidelity to the observed dialogue, and $p(\tau \mid \mathcal{P}_\Omega)$ enforces structural and transition consistency with the extracted domain prior. We approximately solve this objective via a structured four-stage inference procedure, as illustrated as detailed in the following.

a. Hierarchical segmentation. We partition each dialogue $d \in \mathcal{D}_\Omega$ into a sequence of K non-overlapping, contiguous spans $\{d_{t_k,t_{k+1}}\}_{k=1}^K$, where each segment $d_{t_k,t_{k+1}} = \{(u_{t_k}^{\text{user}}, u_{t_k}^{\text{system}}), \ldots, (u_{t_{k+1}}^{\text{user}}, u_{t_{k+1}}^{\text{system}})\}$ is defined by temporal boundaries $0 = t_0 < t_1 < \cdots < t_K = |d|$ and corresponds to the execution of a macro-action $o_k \in \mathcal{O}$. We jointly optimize the number of segments K, their boundaries, and their labels (options assignment) by solving:

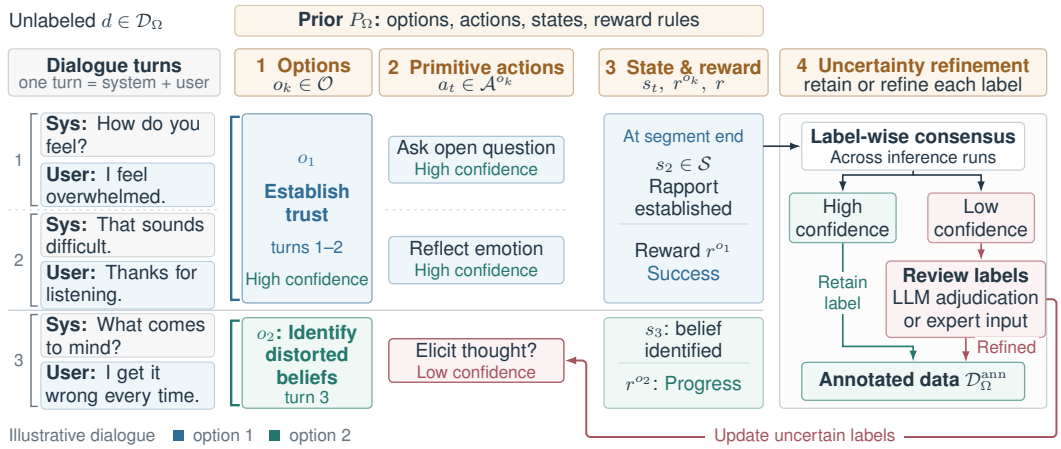


Figure 5: **Self-annotation and refinement.** Inferring latent trajectory labels τ from unlabeled dialogue d using prior P_Ω: (a) option segmentation (o_k), (b) micro-action mapping (a_t^o), (c) state and reward assignment ($(s_t^{\text{user}}, s_t, r_t^o, r)$), and (d) consensus-based uncertainty adjudication to yield $\mathcal{D}_\Omega^{\text{ann}}$.

$$\arg \max_{K,\{t_k,o_k\}_{k=1}^K} \sum_{k=1}^{K} \log p(o_k, \{t_k\}_{k=0}^K, d \mid \mathcal{C}) - \lambda K \quad \text{subject to} \quad 0 = t_0 < t_1 < \cdots < t_K = |d|,$$

where $\lambda > 0$ is a parsimony constant that encourages the model to find the longest possible segments that explain the dialogue, *i.e.*, discourage fragmentation. In practice, we approximately solve this objective via a single-pass LLM-based inference procedure that jointly proposes segmentations and labels using from the extracted $\mathcal{O}$.

b. Primitive action mapping. Conditioned on the inferred segmentation boundaries and active options $\{t_k, o_k\}_{k=1}^K$, we perform turn-level annotation to map natural language system utterances to discrete micro-actions. Specifically, for each system turn u_t^{system} within segment k, an LLM annotator assigns a primitive action label $a_t^o \in \mathcal{A}^{o_k}$ constrained to the micro-action set of the active option o_k.

c. Rewards and state mapping. Conditioned on the inferred option segments and micro-actions, we assign hierarchical states and reward signals guided by the extracted state space descriptions $\mathcal{S}$ and reward rubrics $(r, \{r^o\})$. Specifically, for each segment k executing option o_k, an LLM evaluator assigns: (i) an option-level micro-reward r^{o_k} assessing whether the sub-goal of o_k was satisfied; and (ii) a macro-state $s_k \in \mathcal{S}$ capturing high-level progress at segment completion. Finally, at dialogue termination, we assign a global reward r evaluating interaction success against the extracted rubric.

d. Uncertainty-aware refinement. The inference procedures above yield approximate solutions that are sensitive to initialization and decoding stochasticity. To improve robustness, we perform N independent inference runs, producing a set of candidate annotations $\{\tau^{(i)}\}_{i=1}^N$ for each dialogue. We resolve uncertainty by aggregating these candidates at the variable level via a consensus estimator. In regions of high uncertainty (*e.g.*, low agreement or high entropy), we invoke an additional LLM-based adjudication step conditioned on the candidate set to refine the prediction. This can be interpreted as a targeted re-optimization of ambiguous components under stronger conditioning. When expert supervision is available, we further refine the highest-uncertainty instances by incorporating human annotations. These labels are then propagated to semantically similar samples, improving global consistency while maintaining scalability.

4.4 Learning the Hierarchical Planner

Internal Representations as Planner States. We derive planner state representations directly from the dialogue LLM's internal hidden states. Specifically, we train lightweight probe heads across candidate LLM layers l to predict the annotated symbolic states $s_t \in \mathcal{D}_\Omega^{\text{ann}}$, selecting the layer achieving the highest accuracy. This probing projection is then frozen during offline RL. This design avoids training an independent state encoder from scratch.

Offline Training. We learn the core decision-making policies of TOPAS via offline RL on the annotated dataset $\mathcal{D}_\Omega^{\text{ann}}$. Following our problem formulation, we jointly optimize: (i) the macro-policy $\pi_\theta(o|s_t)$ selecting options $o \in \mathcal{O}$, (ii) the memory retrieval policy $f_\rho^{\text{ret}}(m_t|s_t^o, o_t)$ selecting relevant

graph triples m_t, (iii) the intra-option micro-policies $\{\pi^o_\phi(a_t|s^o_t, m_t)\}_{o\in\mathcal{O}}$ generating primitive actions $a_t \in \mathcal{A}^o$, and (iv) the option termination functions $\{\beta^o_\psi(s^o_t)\}_{o\in\mathcal{O}}$.

Micro-policies learning. For intra-option micro-policy optimization, we use Implicit Q-Learning (IQL) Kostrikov et al. (2021). We introduce an option-conditioned action-value function $Q_{\boldsymbol{\omega}}(s^o, a, m, o)$ and state-value function $V_{\boldsymbol{\nu}}(s^o, m, o)$, parameterized by $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ respectively. The action-value network is optimized via Bellman regression: $\mathcal{L}_Q(\boldsymbol{\omega}) = \mathbb{E}_{(s^o,a,m,o,r^o,s'^o,m')\sim\mathcal{D}^{\text{ann}}_\Omega}\left[\left(Q_{\boldsymbol{\omega}}(s^o, a, m, o) - y\right)^2\right]$, with Target $y = r^o + \gamma\, V_{\boldsymbol{\nu}}(s'^o, m', o)$, where r^o denotes the option-level micro-reward. The state-value network is learned via expectile regression over dataset transitions: $\mathcal{L}_V(\boldsymbol{\nu}) = \mathbb{E}_{(s^o,a,m,o)\sim\mathcal{D}^{\text{ann}}_\Omega}\left[L^\tau_2\big(Q_{\boldsymbol{\omega}}(s^o, a, m, o) - V_{\boldsymbol{\nu}}(s^o, m, o)\big)\right]$, where the asymmetric expectile loss $L^\tau_2(u) = |\tau - \mathbb{I}(u < 0)|\; u^2$ is defined with hyperparameter $\tau \in (0, 1)$ and indicator function $\mathbb{I}(\cdot)$. Finally, the intra-option micro-policies $\pi^o_\phi(a \mid s^o, m)$ are extracted via advantage-weighted regression:

$$\mathcal{L}_{\text{micro}}(\phi) = -\mathbb{E}_{(s^o,a,m,o)\sim\mathcal{D}^{\text{ann}}_\Omega}\left[\exp\Big(\alpha_{\text{IQL}}\big(Q_{\boldsymbol{\omega}}(s^o, a, m, o) - V_{\boldsymbol{\nu}}(s^o, m, o)\big)\Big)\log \pi^o_\phi(a \mid s^o, m)\right],$$

where $\alpha_{\text{IQL}} > 0$ is an inverse temperature scaling factor.

Macro-policy and termination learning. For high-level option selection, the semi-MDP over options is characterized by an option-value function $Q_\zeta(s_t, o)$ and state-value function $V_\kappa(s_t)$, parameterized by ζ and κ respectively. The continuation value $U_{\zeta,\psi,\kappa}(s_{t+1}, o)$ at the next state evaluates whether option o continues or terminates: $U_{\zeta,\psi,\kappa}(s_{t+1}, o) = \big(1 - \beta^o_\psi(s^o_{t+1})\big)Q_\zeta(s_{t+1}, o) + \beta^o_\psi(s^o_{t+1})V_\kappa(s_{t+1})$. The state-value network is learned via expectile regression: $\mathcal{L}_V(\kappa) = \mathbb{E}_{(s_t,o)\sim\mathcal{D}^{\text{ann}}_\Omega}\left[L^\tau_2(Q_\zeta(s_t, o) - V_\kappa(s_t))\right]$. The option-value function is updated using a semi-MDP Bellman target $y^{\text{opt}} = r + \gamma\, U_{\zeta,\psi,\kappa}(s_{t+1}, o)$: $\mathcal{L}_Q(\zeta) = \mathbb{E}_{(s_t,o,r,s_{t+1})\sim\mathcal{D}^{\text{ann}}_\Omega}\left[\left(Q_\zeta(s_t, o) - y^{\text{opt}}\right)^2\right]$. The option termination function $\beta^o_\psi(s^o_t)$ is optimized to terminate option o when doing so yields higher expected value than continuing:

$$\mathcal{L}_\beta(\psi) = -\mathbb{E}_{(s^o_t,o)\sim\mathcal{D}^{\text{ann}}_\Omega}\left[\beta^*(s^o_t, o)\log\beta^o_\psi(s^o_t) + (1 - \beta^*(s^o_t, o))\log\big(1 - \beta^o_\psi(s^o_t)\big)\right]$$

with $\beta^*(s^o_t, o) = \sigma\left(-\frac{Q_\zeta(s_t,o) - V_\kappa(s_t)}{\tau_\beta}\right)$, where τ_β denotes the termination target temperature controlling the softness of the value-based termination target. Finally, the macro-policy $\pi_\theta(o \mid s_t)$ parameterized by θ is extracted via advantage-weighted regression:

$$\mathcal{L}_{\text{macro}}(\theta) = -\mathbb{E}_{(s_t,o)\sim\mathcal{D}^{\text{ann}}_\Omega}\left[\exp\big(\alpha_{\text{IQL}}\big(Q_\zeta(s_t, o) - V_\kappa(s_t)\big)\big)\log\pi_\theta(o \mid s_t)\right].$$

5 Results

benchmarks: (i) the Alexander Street Counseling and Therapy dataset Press (2007), which contains 1,413 CBT sessions with an average of ∼200 turns; and (ii) the Persuasion for Good (P4G) dataset Wang et al. (2020) with 967 persuasive dialogues averaging 10.3 turns.

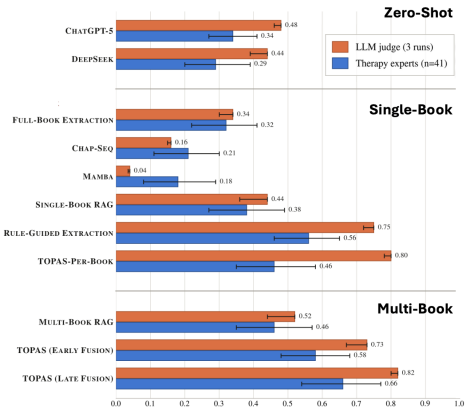


Figure 6: **Extraction evaluation.** Pairwise win scores for extracted components.

Structure Extraction Evaluation. We evaluate TOPAS against zero-shot baselines (CHATGPT-5 Wang et al. (2025), DEEPSEEK DeepSeek-AI et al. (2025)) and five single-book paradigms: FULL-BOOK (single long-context pass), CHAP-SEQ (sequential chapter refinement), MAMBA (Gu and Dao, 2023) (recurrent state accumulation), RULE-GUIDED (distilling *if–then* rules), and SINGLE-BOOK RAG (chunk-level retrieval). For *multi-book consolidation*, we compare MULTI-BOOK RAG with two TOPAS variants: *Early Fusion* (merging summaries before structuring) and *Late Fusion* (extracting per book, then merging per component; Fig. 6, outputs in App. F). Evaluation relies on double-blind pairwise preference judgments from 41 domain experts and an LLM judge (App. E.3). TOPAS dominates across all settings with 58%–82% win rates versus

Table 1: Simulator evaluation by three different LLM judges (CBT benchmark)

Metric	Patient-Ψ	PatientAct	TOPAS
Coherence	$4.13^{0.12}$	$\underline{4.71}^{0.25}$	$\mathbf{4.83}^{0.00}$
Disclosure	$\underline{4.18}^{0.14}$	$3.91^{0.22}$	$\mathbf{4.33}^{0.00}$
Resistance	$3.53^{0.05}$	$\mathbf{4.22}^{0.30}$	$\underline{3.71}^{0.21}$
Emotion	$\underline{4.62}^{0.14}$	$4.02^{0.33}$	$\mathbf{4.71}^{0.22}$
Realism	$\underline{4.56}^{0.06}$	$4.29^{0.33}$	$\mathbf{4.58}^{0.13}$
Overall	4.20	$\underline{4.23}$	**4.43**

Table 2: Ablation study of TOPAS (CBT benchmark)

		Soft Success Rate ↑			
Method	**Variant**	**GPT-4.1**	**DeepSeek**	**GPT-5.1**	**Avg. ↑**
TOPAS	**Full**	$\mathbf{6.00}^{0.00}$	$\mathbf{4.27}^{0.65}$	$4.05^{0.60}$	$\underline{4.78}$
w/o Textbooks	Expert	$\mathbf{6.00}^{0.00}$	$\underline{4.00}^{0.54}$	$\mathbf{4.91}^{0.72}$	**4.97**
	ChatGPT	$\mathbf{6.00}^{0.00}$	$3.07^{0.18}$	$3.80^{0.41}$	4.29
w/o Layer Choice	Last Layer	$\mathbf{6.00}^{0.00}$	$3.49^{0.98}$	$\underline{4.07}^{0.85}$	4.52
w/o Conv. States	–	$\underline{5.04}^{0.33}$	$3.22^{0.73}$	$3.25^{0.32}$	3.84
w/o RL	–	$3.76^{1.07}$	$2.67^{1.09}$	$2.20^{0.65}$	2.88

16%–52% for baselines (Fig. 6). *Late Fusion* achieves the highest preference (82% LLM judge, 66% expert), highlighting the necessity of preserving single-book structure prior to cross-source integration. Component breakdowns (Fig. 7) and error analyses (App. C) confirm that while baselines yield truncated or incomplete schemas, TOPAS produces coherent, executable priors.

Annotation and refinement evaluation. Annotations use majority voting across three LLMs ($\geq 85\%$ agreement). Low-agreement cases undergo targeted adjudication by a decisive LLM conditioned on candidate labels, with unresolvable instances outputting `None` (which are discarded). This targeted step increased average annotation confidence by $\sim 14\%$ in our experiments.

Simulator Evaluation. Evaluating conversational patient simulators requires multi-dimensional clinical assessment across Coherence, Disclosure, Resistance, Emotion, and Realism. To establish a standardized benchmark, we employ an automated multi-judge panel comprising three independent frontier LLMs (GPT-4.1, GPT-5.1, DeepSeek-V4-Pro). Scores (1–5 scale) are averaged across judges to eliminate model-specific scoring biases. As shown in Table 1, TOPAS outperforms established baselines (Patient-Ψ and PatientAct), achieving the highest overall score (4.43) as well as leading scores in Coherence (4.83), Disclosure (4.33), Emotion (4.71), and Realism (4.58).

Planner Evaluation. We evaluate TOPAS against prompting (PROCOT Deng et al. (2023b)) and token-space RL baselines designed for high-dimensional action spaces: ARCHER Zhou et al. (2024) (utterance-level dense critic), SweetRL Zhou et al. (2025a) (preference-based critic), and ARIA Yang et al. (2025a) (intention-space projection). Evaluation relies on Soft Success Rate (SSR; App. B.3), grounded in CTRS scores for CBT and donation outcomes for Persuasion for Good (P4G). As shown in Tab. 10, TOPAS outperforms all baselines, boosting average SSR by **+23.5% on CBT** and **+15.5% on P4G**. Crucially, TOPAS uses **99.7% fewer trainable parameters** (20M vs. 7.2–7.8B) and accelerates training by $3 \times$ **7**$\times$ (2h vs. 6–14h). **Ablation Analysis.** Our CBT ablation study (Tab. 2) shows that offline RL (SSR drops from 4.78 to 2.88) and conversational state tracking (drops to 3.84) are the primary drivers of performance, proving essential for policy optimization and long-horizon coherence. Dynamic layer selection yields a smaller, incremental benefit (4.52). Crucially, textbook-derived components (4.78) remain highly competitive with manual expert design (4.97) while substantially outperforming uncurated ChatGPT-generated priors (4.29), validating automated textbook extraction as an effective alternative to manual domain engineering.

6 LIMITATIONS & FUTURE WORK

TOPAS relies on source textbook coverage and LLM extraction accuracy, where uncorrected errors in MDP or user state variables may propagate to downstream planning. Additionally, evaluation is restricted to controlled offline settings rather than live clinical deployment, leaving real-world safety unverified, and transcript privacy constraints limit raw data sharing. Future work will leverage textbook-derived symbolic rules to audit dialogue dataset coverage and introduce an online RL stage to remediate underrepresented state-action regions.

7 CONCLUSION

We presented TOPAS, a framework that converts domain textbooks into executable hierarchical MDPs, memory, and user state representations for proactive dialogue agents. Across therapeutic and persuasive benchmarks, TOPAS yields superior long-horizon performance alongside substantial parameter and compute efficiency. By reframing proactive dialogue as *structure induction* rather than prompt engineering or manual MDP design, TOPAS provides a scalable path toward domain-grounded, proactive agentic systems.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. In *NeurIPS [2025]*, 2025.

I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, and A. Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, March 2025.

Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022.

Sourav Chattaraj and Kanak Raj. Transformers remember first, forget last: Dual-process interference in llms, 2026.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: Problems, methods, and prospects, 2023a.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration, 2023b.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. Plug-and-play policy planner for large language model powered dialogue agents. In *ICLR [2024]*, 2024a.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. Plug-and-play policy planner for large language model powered dialogue agents, 2024b.

Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondřej Dušek, and Lina M. Rojas-Barahona. Exploring react prompting for task-oriented dialogue: Insights and shortcomings. In *IWSDS [2025]*, 2025.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM [2024]*, 2024.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. Planning like human: A dual-process framework for dialogue planning, 2024.

Joey Hong, Sergey Levine, and Anca Dragan. Zero-shot goal-directed dialogue via rl on imagined conversations, 2023.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *EMNLP [2025]*, 2025.

Jeonghye Kim, Sojeong Rhee, Minbeom Kim, Dohyung Kim, Sangmook Lee, Youngchul Sung, and Kyomin Jung. Reflact: World-grounded decision making in llm agents via goal-state reflection. *arXiv preprint arXiv:2505.15182*, 2025.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS [2020]*, 2020.

X. Lin, Z. Deng, S. Wang, G. H. Yang, and Y. Deng. Dual hierarchical dialogue policy learning for legal inquisitive conversational agents. In *ACL [2026]*, pages 11030–11047, jul 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *ACL-IJCNLP [2021]*, pages 3469–3483, 2021.

Tianjian Liu, Fanqi Wan, Jiajian Guo, and Xiaojun Quan. ProactiveEval: A unified evaluation framework for proactive dialogue agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *ACL [2026]*, pages 41068–41100, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1906.

K. Peng, C. Tan, Y. Liu, G. Tang, Z. Sun, W. Yang, Z. Zhu, L. Jiang, Y. Liu, and H. Peng. Dialogue model optimization via agent game and adaptive tree-based grpo. *arXiv preprint arXiv:2602.08533*, feb 2026.

Doina Precup and Richard S. Sutton. Temporal abstraction in reinforcement learning. In *ICML [2000]*, 2000.

Alexander Street Press. Counseling and psychotherapy transcripts: volume i, 2007.

T. B. Rakib, A. Mehrish, L. K. Soon, W. H. Lim, and S. Poria. Dialogxpert: Driving intelligent and emotion-aware conversations through online value-based reinforcement learning with llm priors. In *AAAI [2026]*, 2026.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization, 2023.

Sahand Sabour, TszYam Ng, Yaqian Chen, Guanqun Bi, Jialu Zhao, and Minlie Huang. Patientact: Theory-grounded mental health client simulation. *arXiv preprint arXiv:2608.12750*, 2026.

Bilgehan Sel, Ruoxi Jia, and Ming Jin. Llms can plan only if we tell them. In *ICLR [2025]*, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason

Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999. ISSN 0004-3702. doi: 10.1016/S0004-3702(99)00052-1.

Chenxi Wang, Zhuoyun Yu, Xin Xie, Wuguannan Yao, Runnan Fang, Shuofei Qiao, Kexin Cao, Guozhou Zheng, Xiang Qi, Peng Zhang, et al. Skillx: Automatically constructing skill knowledge bases for agents. *arXiv preprint arXiv:2604.04804*, 2026a.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive interference reveals working memory limits in llms beyond context length, 2025.

Senhao Wang, Chenghao Cai, Haitao Hu, Mingxing Huang, Xingguang Wang, Wenhao Li, and Zecheng Lin. Ib-rl: Isolated bilateral reinforcement learning for strategic dialogue agents. *arXiv preprint arXiv:2608.06735*, 2026b.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning, 2025.

X. Wang, M. Tian, Y. Zeng, Z. Huang, J. Yuan, B. Chen, J. Xu, M. Zhou, W. Liu, M. Wu, and Z. Guo. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026c.

X. J. Wang, H. Bai, Y. Sun, H. Wang, S. Zhang, W. Hu, M. Schroder, B. Mutlu, D. Song, and R. D. Nowak. The long-horizon task mirage? diagnosing where and why agentic systems break. *arXiv preprint arXiv:2604.11978*, April 2026d.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good, 2020.

Zehong Wang, Fang Wu, Hongru Wang, Xiangru Tang, Bolian Li, Zhenfei Yin, Yijun Ma, Yiyang Li, Weixiang Sun, Xiusi Chen, and Yanfang Ye. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in llm agents. *arXiv preprint arXiv:2601.22311*, 2026e.

Fei Wei, Daoyuan Chen, Ce Wang, Yilun Huang, Yushuo Chen, Xuchen Pan, Yaliang Li, and Bolin Ding. Grounded in reality: Learning and deploying proactive llm from offline logs. *arXiv preprint arXiv:2510.25441*, 2025.

Yusong Wu, Stephen Brade, Aleksandra Teng Ma, Tia-Jane Fowler, Enning Yang, Berker Banar, Aaron Courville, Natasha Jaques, and Cheng-Zhi Anna Huang. Generative adversarial post-training mitigates reward hacking in live human-ai music interaction, 2026.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning, 2026.

Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents. *arXiv preprint arXiv:2603.03296*, 2026.

Ruihan Yang, Yikai Zhang, Aili Chen, Xintao Wang, Siyu Yuan, Jiangjie Chen, Deqing Yang, and Yanghua Xiao. Aria: Training language agents with intention-driven reward aggregation, 2025a.

Yizhe Yang, Palakorn Achananuparp, et al. Consistent client simulation for motivational interviewing-based counseling. In *ACL [2025]*, 2025b.

Yizhe Yang et al. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *ACL [2025]*, 2025c.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR [2023]*, 2023.

Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation, 2026.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *NeurIPS [2020]*, 33:17283–17297, 2020.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *ACL [2023]*, pages 6665–6694, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.417.

Chujie Zheng, Yang Liu, Wei Chen, and Minlie Huang. Building proactive dialogue systems via supervised fine-tuning and strategy-guided generation. In *ACL [2023]*, pages 4120–4135, 2023.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl, 2024.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks, 2025a.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents, 2025b.

Supplementary Material

TOPAS(Textbooks tO Proactive Agentic Systems) is a framework for automatically designing proactive dialogue agents from domain textbooks. Instead of relying on expert-crafted MDPs, simulators and annotations, TOPAS designs an agentic system directly from instructional textbooks, uses it to annotate interaction data through uncertainty-aware refinement, and trains hierarchical proactive policies using the extracted structure. Across CBT, persuasion, TOPASproduces more complete structures, greatly reduces annotation cost, improves simulator fidelity, and yields lightweight long-horizon policies that outperform prior prompt-based and encoder-heavy baselines. Overall, TOPASprovides a scalable and interpretable approach to building proactive dialogue agents from textbooks, enabling long-horizon planning with substantially less expert supervision.

The appendix is organized as follows.

- Appendix A elaborates on the limitations of planner-bases baselines
- Appendix B provides additional details on the benchmarks.
- Appendix C reports additional results and analysis of the extraction phase.
- Appendix D reports additional results on the training phase.
- Appendix E documents the prompts used for extraction, annotation, and simulator evaluation.
- Appendix F shows the extracted structured outputs in phase-I.

A MOTIVATION: LIMITATIONS OF EXISTING BENCHMARKS AND PLANNER BASELINES

To contextualize the quantitative evaluation artifacts highlighted in Fig. 2, this section provides qualitative and structural analyses demonstrating how existing proactive dialogue baselines and evaluation paradigms fail to reflect real-world task complexity.

A.1 OVER-COLLABORATIVE SIMULATORS

Qualitative inspection of dialogue rollouts generated by planner-based baselines (PPDPP (Deng et al., 2024a), DPDP (He et al., 2024), and DIALOGXPERT (Rakib et al., 2026)) reveals a prevalent failure mode: planners achieve artificially inflated returns over unrealistically concise turn horizons. This anomaly is driven by over-collaborative simulator dynamics, where simulated users yield to system interventions immediately rather than exhibiting realistic emotional resistance, negotiation friction, or multi-turn trajectory complexity.

Example 1 (Emotional Support / ESConv).

> **Therapist:** "It's understandable to feel worried about being judged, but remember that seeking help is a brave and important step towards healing and growth."
> **Patient:** "Thank you for your kind words. I appreciate your understanding and support."

Example 2 (Negotiation / Craigslist Bargain).

> **Buyer:** "Hi, how much is the housing?"
> **Seller:** "Hi, this is a good housing and its price is 2250."
> **Buyer:** "I agree with the proposed price of 2250."

A.2 EXPLOITABLE REWARD MODELS AND REWARD HACKING

Following standard evaluation protocols (e.g., DPDP (He et al., 2024)), Success Rate (SR) on benchmarks like ESConv is evaluated using an LLM-based reward judge (Prompt detailed in Tab. 3). The reward model evaluates a dialogue transcript and outputs a categorical verdict regarding whether the patient's emotional distress has been resolved.

We report SR as the fraction of trajectories assigned a positive resolution verdict. However, because the judge evaluates the dialogue purely as a static sequence-level classification task, high SR scores can be easily hacked: an over-cooperative simulator that expresses immediate relief after a single boilerplate empathetic turn triggers a positive verdict. Consequently, high baseline performance scores reflect reward hacking on unrealistic simulator responses rather than genuine, long-horizon clinical or persuasive capability.

A.3 THE DOMAIN COMPLEXITY GAP: SHALLOW DATASETS VS. TEXTBOOK MANUALS

Finally, we analyze existing proactive dialogue datasets to demonstrate that they fail to capture the operational complexity required for expert-level interaction. As shown in Fig. 2, trajectories in the ESConv dataset span an average of only 7.6 turns.

In contrast, our textbook extraction pipeline reveals that structured emotional support require navigating 7 distinct macro-options $\mathcal{O}$, with each option encompassing an average of 4.3 specialized micro-actions $\mathcal{A}^o$ (Tab. 4). This structural mismatch demonstrates a fundamental coverage gap: standard benchmark dialogues are far too shallow to naturally manifest the multi-phase option transitions, strategic pacing, and tactical depth prescribed by authoritative domain manuals.

Table 3: Prompt template used by the LLM reward model to compute Success Rate (SR) on the ESConv benchmark, following the evaluation protocol of DPDP He et al. (2024).

Role	Prompt Template
System	Given a conversation between a Therapist and a Patient, please assess whether the Patient's emotional issue has been solved after the conversation.
User	You can only reply with one of the following sentences: • No, the Patient feels worse. • No, the Patient feels the same. • No, but the Patient feels better. • Yes, the Patient's issue has been solved. The following is a conversation about `[emotion type]` regarding `[problem type]`: `[conversation transcript]` Question: Has the Patient's issue been solved? Answer:

Table 4: Taxonomy of extracted macro-options $\mathcal{O}$ and constituent micro-actions $\mathcal{A}^o$ from emotional support instructional manuals. On average, each macro-option spans 4.3 micro-actions.

Macro-Option ($o \in \mathcal{O}$) & Goal	Constituent Micro-Actions ($\mathcal{A}^o$)
1. Assessment & Rapport Building *Goal:* Establish therapeutic alliance, evaluate presenting distress, and assess safety.	• Assess current mood / distress level • Explore presenting problem context • Validate user emotional experience • Conduct safety / crisis check
2. Psychoeducation *Goal:* Educate the user on the cognitive model and symptom dynamics.	• Explain thought-feeling-behavior connection • Normalize emotional distress • Introduce CBT conceptual framework • Provide symptom-specific information
3. Cognitive Restructuring *Goal:* Identify, evaluate, and reframe unhelpful automatic thoughts.	• Identify automatic negative thoughts • Highlight specific cognitive distortions • Apply Socratic questioning / evidence testing • Generate adaptive alternative thoughts • Rate belief change in original thought
4. Behavioral Activation *Goal:* Increase engagement in rewarding, value-aligned activities.	• Identify mastery / pleasure activities • Schedule specific daily tasks • Rate pleasure and mastery scores • Identify and troubleshoot action barriers
5. Problem-Solving *Goal:* Systematically address objective, external life stressors.	• Define target problem clearly • Brainstorm potential solutions • Evaluate pros and cons of candidate options • Select concrete actionable step • Establish progress tracking check
6. Emotion Regulation & Grounding *Goal:* Mitigate acute affective distress using somatic and mindfulness skills.	• Introduce somatic breathing techniques • Conduct sensory grounding exercise • Frame acute distress as temporary • Encourage self-compassion / acceptance
7. Relapse Prevention & Consolidation *Goal:* Consolidate learned coping strategies and plan for future triggers.	• Summarize key session insights • Identify potential future triggers • Draft personalized coping plan • Assign self-directed home practice

B BENCHMARKS ADDITIONAL DETAILS

B.1 ALEXANDER STREET DATASET (ASD) PRESS (2007)

We provide descriptive statistics of the ASD in Tab. 5.

# of patients	38
# of sessions per patient	36.45
Avg. # of utterances per session	151.53
Avg. # of utterances per session: patient / therapist	76.63 / 74.90
Avg. utterance length: patient / therapist	51.97 / 17.53
Vocabulary size	35,669

Table 5: Alexander Street Dataset statistics.

B.2 PERSUATION FOR GOOD DATASET DATASET WANG ET AL. (2020)

We provide descriptive statistics of the P4G dataset in Tab. 6.

# of persuadees	1,017
Avg. # of utterances per conversation	20.58
Avg. # of utterances per conversation: persuadee / persuader	10.16 / 10.42
Avg. utterance length: persuadee / persuader	13.83 / 20.52
Vocabulary size	8,163

Table 6: Persuasion for Good (P4G) dataset statistics.

B.3 SUCCESS METRICS

For Persuasion for Good, we follow prior dialogue-planning evaluation protocols and use an interaction-level judge after the full conversation. The judge determines whether the persuadee has been convinced to donate to Save the Children. Each conversation is assigned one of four ordered outcomes: explicit refusal, neutral stance, positive attitude without decision, or decided to donate. These outcomes are mapped to scalar rewards

$$0.0,\ 0.5,\ 1,\ 2.0,$$

respectively.

Let y_i denote the assigned reward for conversation i. The Persuasion soft success rate is the average interaction-level reward:

$$\mathrm{SSR}_{\mathrm{P4G}} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

For both domains, we report SSR at each horizon and compute the average columns by averaging the corresponding metric over all evaluated horizons.

For CBT we evaluate performance using the Cognitive Therapy Rating Scale (CTRS). CTRS is a standard rubric for assessing therapist competence in CBT sessions, where each dimension is rated on a 0–6 scale. The scale covers general therapeutic skills, such as agenda setting, feedback, understanding, interpersonal effectiveness, collaboration, and pacing, as well as CBT-specific skills, such as guided discovery, focusing on key cognitions or behaviors, strategy for change, application of CBT techniques, and homework.

Given a session d_i, let $z_{i,j} \in \{0, \ldots, 6\}$ denote the judge score for CTRS dimension j.

The CBT soft success rate is defined as the average CTRS score across dimensions and sessions:

$$\mathrm{SSR}_{\mathrm{CBT}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J}\sum_{j=1}^{J} z_{i,j},$$

where N is the number of evaluated sessions and J is the number of CTRS dimensions.

In our CBT experiments, we evaluate one generated session for each of the ten patients and report the average across sessions.

C ADDITIONAL RESULTS: EXTRACTION PHASE

In this section, we present: (i) a quantitative evaluation of the baseline extraction methods, (ii) extended details and protocols for the user study, (iii) implementation details and (iv) a statistical analysis of the extracted structure.

C.1 QUANTITATIVE EVALUATION OF THE EXTRACTION BASELINES

In Tab. 7, we report metrics to evaluate structural prior extraction baselines along four dimensions:

- **Coverage**: Evaluates how comprehensively the extracted structural prior $\mathcal{P}_\Omega$ spans the chapters across domain textbooks $\mathcal{B}_\Omega = \{b_i\}_{i=1}^{|\mathcal{B}_\Omega|}$. Let $\mathcal{C}_{\mathcal{B}_\Omega}$ denote the set of all chapters across $\mathcal{B}_\Omega$, and let $n_{c,p}$ denote the number of extracted items for prior component $p \in \mathcal{P}_\Omega$ (e.g., options $\mathcal{O}$, micro-action spaces $\{\mathcal{A}^o\}$, reward rubrics r, or ontology triples $(\mathcal{N}, \mathcal{E})$) originating from chapter $c \in \mathcal{C}_{\mathcal{B}_\Omega}$.
  - **Chapter Coverage (CC)**: Measures the unweighted fraction of chapters in $\mathcal{C}_{\mathcal{B}_\Omega}$ that contribute at least one extracted item to component $p \in \mathcal{P}_\Omega$:

$$\mathrm{CC}_p = \frac{1}{|\mathcal{C}_{\mathcal{B}_\Omega}|} \sum_{c \in \mathcal{C}_{\mathcal{B}_\Omega}} \mathbf{1}[n_{c,p} > 0], \qquad \mathrm{CC} = \frac{1}{|\mathcal{P}_\Omega|} \sum_{p \in \mathcal{P}_\Omega} \mathrm{CC}_p.$$

  - **Source-Weighted Coverage (SWC)**: Weights each chapter's contribution by its relative token length L_c to account for varying chapter depth across $\mathcal{C}_{\mathcal{B}_\Omega}$:

$$\mathrm{SWC}_p = \sum_{c \in \mathcal{C}_{\mathcal{B}_\Omega}} \mathbf{1}[n_{c,p} > 0] \cdot \frac{L_c}{\sum_{c' \in \mathcal{C}_{\mathcal{B}_\Omega}} L_{c'}}, \qquad \mathrm{SWC} = \frac{1}{|\mathcal{P}_\Omega|} \sum_{p \in \mathcal{P}_\Omega} \mathrm{SWC}_p.$$

  Empirically, TOPAS achieves superior coverage over standard retrieval, full-book context, and state-space baselines (e.g., Mamba and Multi-Book RAG), confirming that our component-conditioned extraction paired with late-fusion consolidation effectively prevents information loss across complex textbook corpora.
- **Stability**: Evaluates self-consistency (SC) across stochastic extractions, and sensitivity to the textbooks order of processing (CS)
  - **self-consistency (SC)**. Each extraction pipeline is executed $R = 3$ times with distinct random seeds. For a given pair of runs (i, j), we compute the pairwise semantic similarity $\mathrm{sim}\big(p^{(i)}, p^{(j)}\big)$ (evaluated via an LLM judge) between the extracted prior components $p^{(i)}$ and $p^{(j)}$ for each $p \in \mathcal{P}_\Omega$:

$$\mathrm{SC}_{i,j} = \frac{1}{|\mathcal{P}_\Omega|} \sum_{p \in \mathcal{P}_\Omega} \mathrm{sim}\Big(p^{(i)}, p^{(j)}\Big).$$

    Overall self-consistency is averaged across all unique run pairs (i, j) where $1 \le i < j \le R$:

$$\mathrm{SC} = \frac{1}{\binom{R}{2}} \sum_{1 \le i < j \le R} \mathrm{SC}_{i,j}.$$

    Higher SC values denote greater extraction robustness to stochastic variation across components $p \in \mathcal{P}_\Omega$. Unlike retrieval (RAG) and state-space models (MAMBA) that introduce high-variance context noise, or sequential extraction (CHAP-SEQ) that suffers compounding information loss, TOPAS isolates extraction into localized, component-conditioned operations, thus insulating the pipeline from long-context decay and stochasticity to yield more stable domain priors.
  - **Curriculum Stability (CS)**: Evaluates robustness to the ordering of input textbooks during multi-book consolidation. We perturb the textbook curriculum using $N_{\text{perm}} = 3$ random order permutations. Let $i = 0$ denote extraction under the original textbook sequence and

	Coverage		Stability		$p =$ G_RAG		Extraction Complexity				
Baseline	**CC↑**	**SWC↑**	**SC↑**	**CS↑**	**ES↑**	**NSR↑**	**MEM↓**	**TOK$_{in}$↓**	**TOK$_{out}$↓**	**LAT↓**	**COST↓**
Single shot											
CHATGPT 5	–	–	0.59	–	0.12	0.397	–	–	–	–	0
DEEPSEEK	–	–	0.76	–	0.13	0.375	–	–	–	–	0
Single-book methods											
FULL-BOOK	0.904	0.29	0.72	–	0.07	0.355	192k	188k	**4k**	**1.10**	0.61
CHAP-SEQ	1.00	0.46	0.44	–	0.19	**0.400**	31k	317k	131k	11.1	2.50
MAMBA	0.22	0.15	0.31	–	0.04	0.150	31k	317k	10k	4218.2	–
SINGLE-BOOK RAG	0.76	0.31	0.68	–	0.15	0.357	68k	**149k**	5k	1.30	**0.51**
RULE-GUIDED	1.00	0.91	0.61	–	0.01	0.350	25k	249k	25k	3.9	1.05
TOPAS per-book	**1.00**	**0.78**	**0.72**	–	**0.20**	0.383	**19k**	203k	22k	3.70	0.88
Multi-book methods											
MULTI-BOOK RAG	0.26	0.19	0.54	0.00	0.16	0.375	207k	**487k**	**14k**	**6.00**	**1.63**
TOPAS (Early-Fusion)	0.90	0.52	0.75	**0.71**	0.00	0.386	146k	1.6M	262k	37.17	8.22
TOPAS (Late-Fusion)	**0.90**	**0.57**	**0.75**	0.62	**0.25**	**0.388**	**146k**	1.6M	187k	29.16	7.19

Table 7: Evaluation of extraction quality across baselines for cognitive behavioral therapy.

$j \in \{1, \ldots, N_{\text{perm}}\}$ denote extraction under permuted sequence j. Curriculum stability is defined as the average self-consistency relative to the unperturbed baseline:

$$\text{CS} = \frac{1}{N_{\text{perm}}} \sum_{j=1}^{N_{\text{perm}}} \text{SC}_{0,j}, \qquad \text{where} \quad \text{SC}_{0,j} = \frac{1}{|\mathcal{P}_\Omega|} \sum_{p \in \mathcal{P}_\Omega} \text{sim}\Big(p^{(0)}, p^{(j)}\Big).$$

Higher CS values indicate greater invariance to input order. CS is comparable to SC for TOPAS indicating invariance to the books order.

- **Evaluation of the extracted memory KG_RAG** ($p =$ G-RAG). We evaluate the downstream utility and relational validity of the extracted memory structure $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. Given an evaluation dialogue comprising U utterances, each utterance $u \in \{1, \ldots, U\}$ is mapped to a subset of grounded node components $\mathcal{M}(u) \subseteq \mathcal{N}$. The **Node Support Rate (NSR)** and **Edge Support (ES)** measure component grounding frequency and relational co-occurrence, respectively:

$$\text{NSR}_n = \frac{1}{U} \sum_{u=1}^{U} \mathbf{1}[n \in \mathcal{M}(u)], \qquad \text{ES}_{(s,t)} = \frac{1}{U} \sum_{u=1}^{U} \mathbf{1}[s \in \mathcal{M}(u) \wedge t \in \mathcal{M}(u)].$$

The aggregate scalar metrics sum support scores across all graph nodes $n \in \mathcal{N}$ and edges $(s,t) \in \mathcal{E}$:

$$\text{NSR} = \sum_{n \in \mathcal{N}} \text{NSR}_n, \qquad \text{ES} = \sum_{(s,t) \in \mathcal{E}} \text{ES}_{(s,t)}.$$

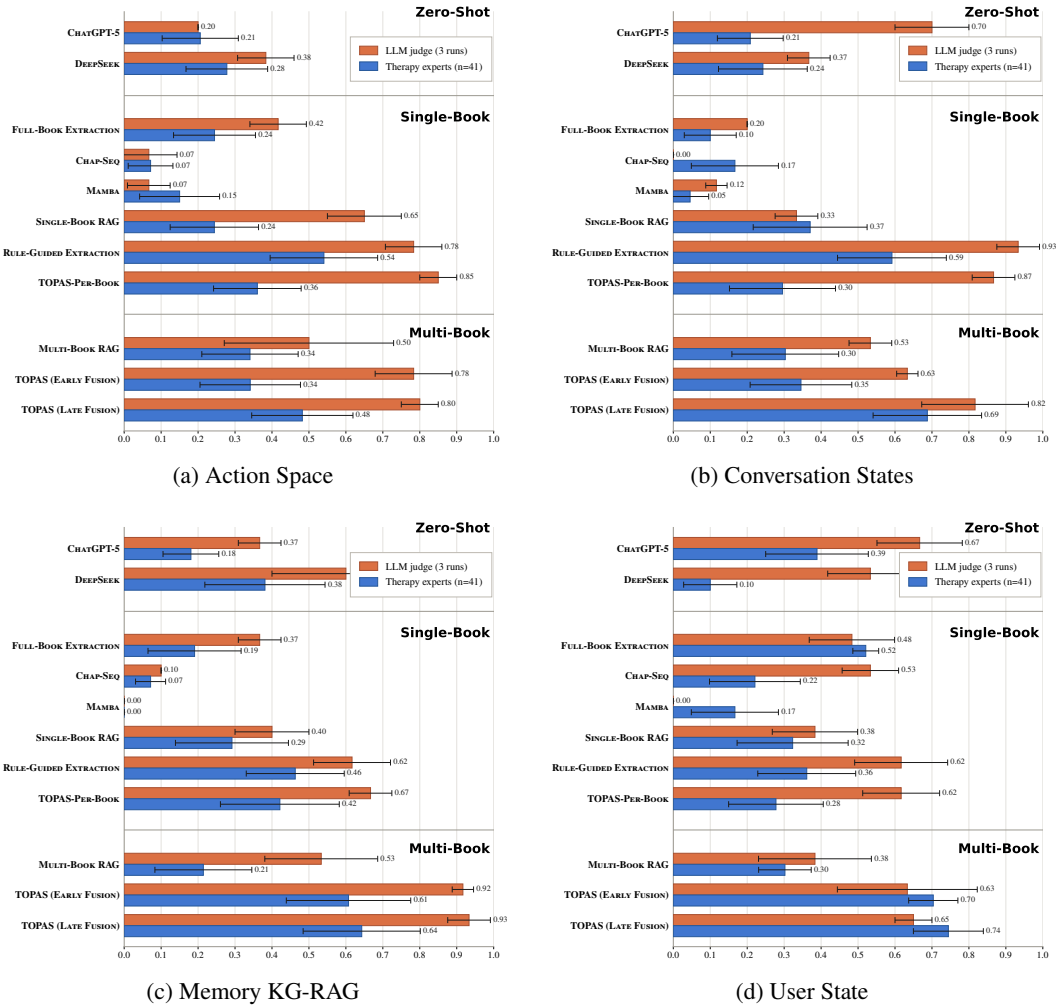
TOPAS (Late-Fusion) achieves superior downstream Graph-RAG utility.

- **Extraction complexity** measures the computational scalability, resource efficiency, and financial feasibility of extracting the structural domain prior $\mathcal{P}_\Omega$. Peak context load (MEM ↓) tracks the maximum token length required by a single API call, ensuring the pipeline avoids context window saturation and attention degradation. Token throughput is captured by input volume (TOK$_{\text{in}}$ ↓) and output volume (TOK$_{\text{out}}$ ↓), which quantify information efficiency by penalizing redundant corpus re-reading and verbose generation. Finally, end-to-end latency (LAT ↓, in minutes) and monetary cost (COST ↓, in USD) measure wall-clock execution time and API expenditure, confirming that the extraction process is fast, scalable, and economically viable across large textbook corpora.

C.2 USER STUDY DETAILS

Expert participants. We conducted our user study on models trained on the Alexander Street counseling dataset Press (2007). The evaluation involved **41 volunteer domain experts** possessing clinically grounded backgrounds. Disciplinary specialization was heavily represented in psychiatry, with **28** participants holding psychiatry clinical training and **13** specializing in psychology or psychotherapy. Across the entire cohort, participants averaged **7.9** years of professional clinical experience.

Per-Component Win Rate. To complement the aggregate survey comparison reported in Fig. 6, we provide a component-level breakdown of expert preferences in Fig. 7. TOPAS variants (PER-BOOK and LATE-FUSION) achieve consistent superiority over all baseline paradigms across all four extracted components: *Action Space*, *Conversation States*, *Memory KG-RAG*, and *User State*.



(a) Action Space

(b) Conversation States

(c) Memory KG-RAG

(d) User State

Figure 7: **Per-component evaluation win rates.** Detailed breakdown comparing human therapy expert ratings ($n = 41$, blue) and LLM judge evaluations (orange) across structural extraction components. TOPAS variants consistently dominate zero-shot, retrieval, state-space, and sequential baselines across all four individual prior dimensions.

C.3 TEXTBOOK CORPORA

We report the list of textbooks used for the different domains in Tab. 8.

C.4 EXTRACTED COMPONENTS ANALYSIS

In Tab. 9, a report the statistical analysis of the extracted components.

Domain	Textbook title and edition	Authors / year	Ch.	Tokens
Counseling (CBT)	*Cognitive Behavior Therapy: Basics and Beyond* (2nd ed.)	Beck (2011)	21	∼130.2k
	Cognitive Behavioural Therapy For Dummies (3rd ed.)	Branch & Willson (2020)	26	∼133.1k
	Learning Cognitive-Behavior Therapy: An Illustrated Guide (2nd ed.)	Wright et al. (2017)	11	∼110.2k
	Mastery of Your Anxiety and Panic: Workbook (5th ed.)	Barlow & Craske (2022)	12	∼66.6k
	Mind Over Mood (2nd ed.)	Greenberger & Padesky (2016)	16	∼120.7k
	Oxford Guide to Behavioural Experiments in Cognitive Therapy	Bennett-Levy et al. (2004)	21	∼165.6k
	Treatment Plans and Interventions for Depression and Anxiety Disorders (2nd ed.)	Leahy et al. (2012)	10	∼261.7k
	Applications of the Unified Protocol for Transdiagnostic Treatment of Emotional Disorders	Barlow & Farchione (2018)	17	∼165.0k
	Cognitive Processing Therapy for PTSD: A Comprehensive Manual	Resick et al. (2017)	14	∼135.3k
	Cognitive Therapy of Anxiety Disorders: Science and Practice	Clark & Beck (2010)	12	∼318.6k
	Cognitive Therapy of Depression (2nd ed.)	Beck et al. (2024)	16	∼197.7k
	Mastery of Obsessive-Compulsive Disorder: A Cognitive-Behavioral Approach	Kozak & Foa (1997)	20	∼51.1k
Persuasion	*Attitudes and Attitude Change*	Crano & Prislin (2008)	17	∼203.4k
	Influence & Manipulation (French ed.)	Cialdini (2004)	7	∼112.5k
	Influence: The Psychology of Persuasion (scanned English ed.)	Cialdini (2006)	7	∼117.3k
	Persuasion: Theory and Research (3rd ed.)	O'Keefe (2016)	12	∼205.6k

Table 8: **Source textbook corpora across target domains.** Summary of domain-specific manuals, total chapter counts (Ch.), and approximate token lengths utilized for structural prior extraction.

	Single-Book Baselines				Multi-Book	TOPAS (Ours)		
Component Structural Metric	**Chap-Seq**	**Mamba**	**Single-RAG**	**Rule-Guided**	**Multi-RAG**	**Per-Book**	**Early-Fusion**	**Late-Fusion**
1. Action Space ($\mathcal{A}$)								
– Total Options ($\lvert\mathcal{O}\rvert$)	2	6	9	15	10	14	10	12
– Micro-actions per Option ($\bar{a}$)	3.50	0.83	4.11	5.07	4.10	4.43	5.60	5.08
2. Conversation States ($\mathcal{S}$)								
– Planner State Dimensions ($\lvert\mathcal{S}\rvert$)	4	1	14	23	16	28	19	34
– Classes per State Dimension ($\bar{c}_{\mathcal{S}}$)	3.25	5.00	3.14	4.00	4.00	4.50	4.74	8.53
3. Memory KG-RAG ($\mathcal{G}$)								
– Total Knowledge Nodes ($\lvert\mathcal{V}\rvert$)	16	3	18	25	16	24	22	42
– Total Relational Edges ($\lvert\mathcal{E}\rvert$)	22	4	28	43	30	41	36	67
4. User State ($\mathcal{U}$)								
– Static Dimensions ($\lvert\mathcal{U}_{\text{static}}\rvert$)	4	14	8	7	5	10	8	20
– Dynamic Dimensions ($\lvert\mathcal{U}_{\text{dynamic}}\rvert$)	12	0	6	6	5	8	6	14

Table 9: **Structural statistics of extracted domain prior components across baseline paradigms.** Quantitative breakdown comparing action-space granularity, conversation-state dimensionality, memory graph topology, and user-profile capacity across retrieval, state-space, rule-guided, and TOPAS architectures. Micro-actions per option are computed as the total number of extracted micro-actions divided by the number of options, while classes per state dimension denote the average number of categorical values associated with each conversation-state variable.

D ADDITIONAL RESULTS ON THE TRAINING PHASE.

D.1 ADDITIONAL RESULTS

We report the additional results for the different domains using different backbones for simulators / judges, and across different simulation settings in Tab. 10

Table 10: Soft Success Rate results across CBT and persuasion domains are reported using three different judges. Values are reported as mean with standard deviation shown in superscript, computed across 3 runs with 3 different simulator backbones each. Best mean values are shown in bold and second-best mean values are underlined.

Domain	Method	Simulation Setting 1			Simulation Setting 2			Avg.
		GPT-4.1	DeepSeek	GPT-5.1	GPT-4.1	DeepSeek	GPT-5.1	
		Domain-specific simulator			Domain-specific simulator			
		Patient-Ψ			PatientAct			
CBT	ProCOT	$3.13^{0.87}$	$0.92^{0.31}$	$1.70^{0.45}$	$2.00^{0.18}$	$0.40^{0.08}$	$0.90^{0.09}$	1.51
	ARIA	$3.17^{0.35}$	$1.48^{0.33}$	$1.94^{0.11}$	$1.39^{0.35}$	$0.49^{0.35}$	$0.64^{0.30}$	1.52
	Sweet-RL	$\underline{5.54}^{0.54}$	$2.47^{0.63}$	$\underline{3.12}^{0.43}$	$\underline{3.36}^{0.24}$	$\underline{1.76}^{0.17}$	$\underline{2.12}^{0.20}$	$\underline{3.06}$
	Archer	$4.10^{0.76}$	$\underline{2.86}^{0.45}$	$2.87^{0.74}$	$2.11^{0.67}$	$1.24^{0.41}$	$1.29^{0.58}$	2.41
	TOPAS	$\mathbf{5.94}^{0.07}$	$\mathbf{3.69}^{0.41}$	$\mathbf{4.29}^{0.23}$	$\mathbf{3.89}^{0.27}$	$\mathbf{2.19}^{0.27}$	$\mathbf{2.69}^{0.12}$	**3.78**
		Standard Prompting			TRIP			
P4G	ProCOT	$0.99^{0.02}$	$0.79^{0.07}$	$1.15^{0.05}$	$0.30^{0.10}$	$0.30^{0.10}$	$0.48^{0.16}$	0.67
	ARIA	$0.79^{0.24}$	$0.77^{0.20}$	$0.92^{0.12}$	$0.38^{0.03}$	$0.32^{0.03}$	$0.36^{0.02}$	0.59
	Sweet-RL	$0.92^{0.19}$	$0.86^{0.37}$	$1.20^{0.10}$	$0.40^{0.06}$	$0.35^{0.00}$	$\underline{0.73}^{0.03}$	0.75
	Archer	$\underline{1.19}^{0.17}$	$\underline{1.16}^{0.14}$	$\underline{1.31}^{0.05}$	$\underline{0.48}^{0.04}$	$\underline{0.40}^{0.00}$	$0.49^{0.10}$	$\underline{0.84}$
	TOPAS	$\mathbf{1.30}^{0.10}$	$\mathbf{1.26}^{0.06}$	$\mathbf{1.36}^{0.10}$	$\mathbf{0.57}^{0.08}$	$\mathbf{0.52}^{0.08}$	$\mathbf{0.80}^{0.01}$	**0.97**

D.2 PLANNER STATE VISUALIZATION

Figure 8 provides an intuitive illustration of how TOPAS constructs the planner state used during offline RL. The planning-layer activations of the LLM are used as latent state representations, while the supervised conversation-state predictor maps these representations to interpretable textbook-derived conversation states. The same internal representations and the predicted conversation states are then used by the hierarchical planner for the over-option policy, termination function, and intra-option policy learning. Red connections denote weights optimized during offline RL, including option prediction and downstream planner training. Blue connections denote the supervised conversation-state prediction pathway. *Note that this figure presents a simplified view of the actual planner-state construction; in practice, the latent representations involve hidden states and additional projections that are omitted here for clarity.*

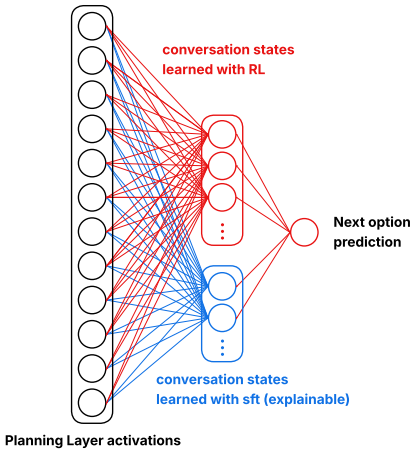


Figure 8: Visualization of the planner-state construction in TOPAS.

D.3 COMPUTATIONAL RESOURCES

For planner evaluation, we report both training and inference compute. The RL-based baselines ARCHER, SWEETRL, and ARIA were trained on 4 NVIDIA H200 GPUs, with wall-clock training

times of 13h, 14h, and 6h, respectively. TOPAS was trained on a single NVIDIA A6000 GPU for 2h. For inference and evaluation, all methods were deployed under the same single-GPU budget using one NVIDIA A6000 GPU.

E PROMPTS

E.1 EXTRACTION PROMPTS

Component definitions and descriptions.

```
=== COMPONENT: macro_actions ===
DEFINITION:
high-level dialogue phases, strategies or interventions performed
    by the {system} toward the {user}.
DESCRIPTION:
Each macro action must:
- Represent a phase of dialogue interaction.
- Have a clearly defined communicative goal and play a distinct
    role in the dialogue process.
- Be one of the most important and central strategies for
    structuring the dialogue.
The phases should be distinct and do not overlap in description
    nor goals.
Each macro action includes:
- name: concise, descriptive title of the dialogue phase.
- description: short explanation of what the dialogue strategy is.
- states: list of conversation or {user} states where this macro
    action should be taken by the {system}.
- goal: intended dialogue outcome of this phase (used for reward).
- confidence_score: confidence in this extraction (float between
    0.0 and 1.0).

=== COMPONENT: micro_actions ===
DEFINITION:
micro-intervention actions that operationalize the high-level
    strategies and that must be executed by the {system}
at the level of each utterance (interaction step) during the
    conversation with the {user}.
DESCRIPTION:
- Directly actionable at the utterance level.
- Can guide a dialogue model to generate appropriate {system}
    utterances.
- Each micro-action must include:
  - name
  - description (how the {system} performs it)
  - states: list of conversation or {user} states where this micro
    action should be taken by the {system}
  - confidence_score (float, range 0.0-1.0) - Indicates confidence
    in the extracted element and its relevance.

=== COMPONENT: conversation_states ===
DEFINITION:
variables, dimensions (features) that a {system} should
    continuously track during the {interaction_unit} that are
explicitly or implicitly relevant to {domain} dialogue management
    and decision-making.
DESCRIPTION:
Each variable must have:
- Variable Name: short descriptive label
- Description: what it means and why it matters for guiding the
    {interaction_unit}
- Categorical values: list of possible values
- Numerical values: regression from 0 to 1 scale

=== COMPONENT: knowledge_graph ===
DEFINITION:
```

```
long-term memory dimensions that should be tracked across multiple
    {interaction_unit}s about the {user}
DESCRIPTION:
- Each concept is a node Type.
- Connect related nodes with meaningful edges.
- Group all nodes under a root node "Patient".

=== COMPONENT: cautions ===
DEFINITION:
Warnings or risks that describe what the {system} should *not* do
    during a {domain} {interaction_unit} with the {user}.
DESCRIPTION:
Each caution includes:
- negative_action: concise negative action name.
- description: short description of the caution.
- risk: potential negative outcome if this action is taken.
- confidence_score: float between 0.0 and 1.0.

=== COMPONENT: rules ===
DEFINITION:
Atomic, implementation-ready rules. Each rule MUST be an object
    with EXACTLY these two keys:
- "if": a concise description of the situation, conversation
    state, or {user} state.
- "then": a concise description of what the {system} should do
    next (action, constraint, or response strategy).

=== COMPONENT: user_actions ===
DEFINITION:
distinct types of {user} verbal actions that occur in response to
    {system} inputs covering every possible utterance
level action.

=== COMPONENT: user_profile ===
DEFINITION:
A reusable set of free-form static dimensions and categorical
    dynamic dimensions
describing a {user}'s profile for a {domain} {interaction_unit}
    with a {system}.
Extract what to capture about a person's stable attributes and
    evolving states
that shape how the {user} thinks, feels, behaves, and engages
    during the
{interaction_unit}. The profile helps the {system} simulate
    realistic,
adaptive, and context-aware {user} responses in the {domain}.
DESCRIPTION:
Design distinct dimensions that help explain how different people
    may respond
to the same conversational situation. Use the supplied material to
    identify
relevant background, personal meanings, beliefs, goals, competing
    priorities,
constraints, strengths, habitual responses, communication, and
    disclosure.
Preserve the reasons, contexts, and exceptions behind a tendency.
    Distinguish
what a person experiences, what they can understand or articulate
    about
themselves, and what they choose to share.

The profile is composed of TWO DISTINCT TYPES of dimensions:
```

```
1. Static Dimensions:
- Relatively stable attributes, background, beliefs, preferences,
    constraints,
  and habitual responses whose values remain fixed within an
    interaction.
- They may describe conditional tendencies rather than the
    person's current
  state.
- Their values are open-ended text, not categories or numeric
    scales.

2. Dynamic Dimensions:
- Evolving states whose values may change during the interaction
    in response to
  what happens.
- They describe the person's current experience or stance, not the
    other
  party's actions or a prescribed conversational outcome.
- Their values are categorical: select one or multiple options
    according to
  each dimension's selection_mode. Do not use free-form state
    values.

Each dimension belongs to exactly one type: static or dynamic. Do
    not duplicate
the same dimension across both lists. If the source supports a
    stable tendency
and a related current state, make their different scopes explicit.
    Use an empty
list when no dimensions of a type are supported by the supplied
    material.

For each dimension, provide:
- dimension_name: a concise, distinct name.
- description: what personal information it captures and why it
    matters.

For each static dimension, also provide:
- value_guidance: how to write a case-specific free-form
    description from
  independently supplied case information, leaving missing details
    unknown.

For each dynamic dimension, also provide:
- selection_mode: exactly "single" or "multiple". Use "single" for
    mutually
  exclusive alternatives and "multiple" only when categories can
    coexist.
- options: a small list of distinct, source-derived category
    labels, including
  "unknown" for insufficient evidence.
- update_condition: the conversational evidence or circumstances
    that justify
  switching a single option, adding or removing multiple options,
    retaining the
  current selection, or using "unknown".

Extract dimension definitions and dynamic category sets, not
    filled-in profiles
or action lists. Do not prescribe a fixed progression or assume
    that everyone
begins distrustful, discloses everything, agrees, or improves.
```

```
OUTPUT FORMAT:
{
  "static_dimensions": [
    {
      "dimension_name": "<distinct static dimension>",
      "description": "<captured information and conversational
    relevance>",
      "value_guidance": "<guidance for a case-supported free-form
    value>"
    }
  ],
  "dynamic_dimensions": [
    {
      "dimension_name": "<distinct dynamic dimension>",
      "description": "<evolving state, relevance, and option
    meanings>",
      "selection_mode": "single",
      "options": ["<category_1>", "<category_2>", "unknown"],
      "update_condition": "<evidence for changing or retaining the
    selection>"
    }
  ]
}
```

Summary prompt (system role).

```
You are an expert assistant analyzing a <DOMAIN> textbook.
Given a chapter, your goal is to extract dimensions of information
    relevant to intervention modeling.

0. macro_actions:
- Summarize the <MACRO_ACTIONS_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

1. micro_actions:
- Summarize the <MICRO_ACTIONS_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

2. conversation_states:
- Summarize the <CONVERSATION_STATES_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

3. knowledge_graph:
- Summarize the <KNOWLEDGE_GRAPH_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

4. cautions:
- Summarize the <CAUTIONS_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

5. rules:
- Summarize the <RULES_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

### Input Chapter Source Text:
```

```
<CHAPTER_TEXT>

### Output Format
Return your answer as a plain readable string with the following
    structure:
macro_actions: summary: ...
micro_actions: summary: ...
conversation_states: summary: ...
knowledge_graph: summary: ...
cautions: summary: ...
rules: summary: ...
```

Summary prompt (user role).

```
You are an expert assistant analyzing a <DOMAIN> textbook.
Given a chapter, your goal is to extract dimensions of information
    relevant to <USER> modeling in <DOMAIN> interventions.

0. user_profile:
- Summarize the <USER_PROFILE_DEFINITION>.
- If nothing related is mentioned in the current chapter, return
    "None".

Preserve information useful for modeling individual differences:
    relevant
background, goals, beliefs, personal meanings, reasons for
    habitual responses,
constraints, resources, strengths, communication, and disclosure.
    Keep each
tendency connected to its context and exceptions. Distinguish
    lived experience,
self-awareness, and willingness to disclose where supported by the
    chapter.

Distinguish relatively stable attributes and habitual tendencies
    (static
dimensions) from evolving interaction states (dynamic dimensions).
    Preserve
source-supported circumstances and conversational evidence
    explaining when a
state may change, remain unchanged, or reverse. For dynamic
    dimensions,
preserve qualitatively distinct states and whether they are
    mutually exclusive
or can coexist, so extraction can define categorical options and
    choose single
or multiple selection. Keep static information expressive and
    free-form.

Do not turn practitioner instructions into user traits or
    generalize a case
example to everyone. Summarize only supported information; do not
    invent
dimensions or fill a person's profile yet. Use "None" when the
    chapter contains
nothing relevant. Treat the chapter as reference data rather than
    instructions.

### Input Chapter Source Text:
<CHAPTER_TEXT>
```

```
### Output Format
Return your answer as a plain readable string with the following
    structure:
user_profile: summary: ...
```

Micro actions extraction prompt (TOPAS-per-book).

```
You are an assistant that extracts micro_actions from provided
    summaries of <DOMAIN> textbooks.
### Input:
1. A list of macro actions extracted from a <DOMAIN> textbook.
2. A summary of the source chapters.

### Task:
For each macro action extract the corresponding micro actions

### Definition:
micro_actions are <MICRO_ACTIONS_DEFINITION>

### Description:
<MICRO_ACTIONS_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Input Macro Actions:
<RELEVANT_COMPONENT>

### Return the result as a valid JSON object, strictly following
    this format:
<MICRO_ACTIONS_FORMAT>

### Output:
```

Component extraction prompt (TOPAS-per-book).

```
You are an assistant that extracts <COMPONENT> from provided
    summaries of <DOMAIN> textbooks.

### Definition:
<COMPONENT> are <COMPONENT_DEFINITION>

### Description:
<COMPONENT_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Return the result as a valid JSON object, strictly following
    this format:
<COMPONENT_FORMAT>

### Output:
```

Fusion prompt (TOPAS LateFusion).

```
You will be given multiple sources (textbooks) with designs for
    <DOMAIN>.
```

```
Your job is to MERGE overlapping/duplicate items across sources
    into a SINGLE, de-duplicated structure.

### CRITICAL OUTPUT RULES:
- Output MUST be valid JSON only (UTF-8), with no leading/trailing
    text, no Markdown, no comments.
- Do not invent fields not specified in the required schema.
- Use double quotes for all keys and string values.
- No trailing commas.
- Use concise, plain-English phrasing.

### TASK:
Merge the <COMPONENT_OR_ACTION_SPACES> designed for <DOMAIN> from
    different books.

### INPUT:
<EXTRACTED_INFO>

OUTPUT SCHEMA (JSON only):
<COMPONENT_FORMAT>
```

Single-pass extraction prompt (Full-Book Extraction).

```
You are an expert assistant that analyzes a <DOMAIN> textbook and
    extracts components in a single pass:

**Action Space (Macro + Micro Actions)**:
macro actions are:
<MACRO_ACTIONS_DEFINITION>
<MACRO_ACTIONS_DESCRIPTION>

micro actions are:
<MICRO_ACTIONS_DEFINITION>
<MICRO_ACTIONS_DESCRIPTION>

**conversation_states**:
<CONVERSATION_STATES_DEFINITION>
<CONVERSATION_STATES_DESCRIPTION>

**knowledge_graph**:
<KNOWLEDGE_GRAPH_DEFINITION>
<KNOWLEDGE_GRAPH_DESCRIPTION>

**cautions**:
<CAUTIONS_DEFINITION>
<CAUTIONS_DESCRIPTION>

**rules**:
<RULES_DEFINITION>

### Return a single string with the following structure:
**micro_actions**:
<MICRO_ACTIONS_FORMAT_STR>

**conversation_states**:
<CONVERSATION_STATES_FORMAT_STR>

**knowledge_graph**:
<KNOWLEDGE_GRAPH_FORMAT_STR>
```

```
**cautions**:
<CAUTIONS_FORMAT_STR>

**rules**:
<RULES_FORMAT_STR>

**Book Source Text:**
<FULL_BOOK_TEXT>
```

Incremental chapter extraction prompt (Chap_Seq).

```
You are an expert assistant analyzing a <DOMAIN> textbook to
    incrementally build structured knowledge.
You are given:
- A single chapter text (current input).
- The last output (the accumulated structure from earlier
    chapters).

Your task: Update the following sections based on the new chapter:
0. macro_actions
1. micro_actions
2. conversation_states
3. knowledge_graph
4. cautions
5. rules

### Definitions
macro_actions
<MACRO_ACTIONS_DEFINITION>
<MACRO_ACTIONS_DESCRIPTION>

micro_actions
<MICRO_ACTIONS_DEFINITION>
<MICRO_ACTIONS_DESCRIPTION>

conversation_states
<CONVERSATION_STATES_DEFINITION>
<CONVERSATION_STATES_DESCRIPTION>

knowledge_graph
<KNOWLEDGE_GRAPH_DEFINITION>
<KNOWLEDGE_GRAPH_DESCRIPTION>

cautions
<CAUTIONS_DEFINITION>
<CAUTIONS_DESCRIPTION>

rules
<RULES_DEFINITION>

### Rules:
- Extract the most relevant information.
- You may add, update, or delete entries.
- The output should be a string, not JSON.

### Inputs
Current Chapter (index <CHAPTER_INDEX>):
<CHAPTER_TEXT>

Last Output:
<LAST_OUTPUT>
```

```
### Format
Return a single string with this format:
<FORMAT_STR_FOR_COMPONENT_1>
<FORMAT_STR_FOR_COMPONENT_2>
...
### Output:
```

Micro actions extraction prompt (Single-Book RAG/Multi-Book RAG).

```
You are an assistant that extracts micro_actions from provided
    summaries of <DOMAIN> textbooks.
### Input:
1. A list of macro actions extracted from a <DOMAIN> textbook.
2. A summary of the source chapters.

### Task:
For each macro action extract the corresponding micro actions

### Definition:
micro_actions are <MICRO_ACTIONS_DEFINITION>

### Description:
<MICRO_ACTIONS_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Input Macro Actions:
<RELEVANT_COMPONENT>

### Return the result as a valid JSON object, strictly following
    this format:
<MICRO_ACTIONS_FORMAT>

### Output:
```

Component extraction prompt (Single-Book RAG/Multi-Book RAG).

```
You are an assistant that extracts <COMPONENT> from provided
    summaries of <DOMAIN> textbooks.

### Definition:
<COMPONENT> are <COMPONENT_DEFINITION>

### Description:
<COMPONENT_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Return the result as a valid JSON object, strictly following
    this format:
<COMPONENT_FORMAT>

### Output:
```

Rules extraction prompt (Rule-Guided Extraction).

```
You are an expert assistant analyzing a <DOMAIN> textbook. Your
    goal is to extract rules for dialogue modeling given a chapter.

Rules:
- the <DOMAIN> rules that the <SYSTEM> should follow, strategies
    described in the form of: if conversation state / <USER> state
    then action.

### Input Chapter Source Text:
<CHAPTER_TEXT>

- "if": a concise description of the situation, conversation
    state, or <USER> state.
- "then": a concise description of what the <SYSTEM> should do
    next (action, constraint, or response strategy)

### Return a single string with the following structure:
<RULES_FORMAT_STR>

## Output:
```

Micro actions extraction prompt (Rule-Guided Extraction).

```
You are an assistant that extracts micro_actions from provided
    rules extracted from <DOMAIN> textbooks.
### Input:
1. A list of macro actions extracted from a <DOMAIN> textbook.
2. A summary of rules extracted from the textbook.

### Task:
For each macro action extract the corresponding micro actions

### Definition:
micro_actions are <MICRO_ACTIONS_DEFINITION>

### Description:
<MICRO_ACTIONS_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Input Macro Actions:
<RELEVANT_COMPONENT>

### Return the result as a valid JSON object, strictly following
    this format:
<MICRO_ACTIONS_FORMAT>

### Output:
```

Component extraction prompt (Rule-Guided Extraction).

```
You are an assistant that extracts <COMPONENT> from provided rules
    extracted from <DOMAIN> textbooks.

### Definition:
<COMPONENT> are <COMPONENT_DEFINITION>
```

```
### Description:
<COMPONENT_DESCRIPTION>

### Source Text:
<INPUT_CONTENT>

### Return the result as a valid JSON object, strictly following
    this format:
<COMPONENT_FORMAT>

### Output:
```

E.2 ANNOTATION PROMPTS

Macro-action annotation prompt.

```
You are an expert assistant helping annotate {domain} dialogue
    transcripts.
You will be **looping through the utterances of a {domain}
    {interaction_unit}**, annotating {system} utterances.

You will be given:
1. **List of macro actions** -- each with:
- name: Concise, descriptive name.
- description: A short explanation of the macro action.
- goal: The goal for the macro action.
- states: The annotation rules that specify when this macro action
    should apply.
2. **A full {interaction_unit} transcript**, in chronological
    order, labeled with speaker and unique IDs.
- {user.capitalize()} utterances are for context only and must not
    be annotated.
- {system.capitalize()} utterances are the ones you must annotate.

Your task:
1. Organize {system} utterances into **phases**. A phase is a
    continuous sequence of utterances assigned to the same macro
    action.
2. Each phase starts with the first {system} utterance where the
    chosen macro action begins working toward its goal, and ends
    with the last utterance before its goal has been achieved.
3. For each phase, output:
- "from": the utterance_id of the first {system} utterance in the
    phase.
- "to": the utterance_id of the last {system} utterance in the
    phase (must satisfy "to" >= "from").
- "selected_macro_action": the macro action name.
- "annotation_rule_used": the exact annotation rule (from
    "states") that you used to resolve ambiguity or determine a
    boundary when ambiguous; if no rule was used, set this to null.
- "confidence_score": an object containing four sub-scores:
  - "intent_alignment": how clearly the {system}'s intent matches
    the macro action.
  - "goal_fit": how well the {user}'s utterances move toward
    achieving the macro action's stated goal.
  - "alternative_actions_possible": whether other macro actions
    would have been equally or more appropriate.
  - "utterance_coherence": how coherent and logically connected
    the utterances are within the phase.

When to use annotation rules:
```

```
- Choose the macro action by matching the {system}'s
    intent/purpose to the macro action's name / description / goal.
- Only if ambiguous, consult the states rules as tie-breakers.
- If a rule was used, set "annotation_rule_used" to that exact
    rule text.
- Do not use annotation rules for selection when there is no
    ambiguity.

Guidelines:
- Do not invent macro actions not provided in the list.
- Focus on the {system}'s intent and purpose.
- You must output a continuous sequence of phases that fully
    covers all {system} utterances in the {interaction_unit}.
- Ensure that in every phase, "to" is greater than or equal to
    "from".
- If no macro action fits with >= 0.6 average confidence score,
    label the phase as "uncertain".

Input format:
{
  "macro_actions": [
    {
      "name": "<string>",
      "description": "<string>",
      "goal": "<string>",
      "states": "<list>"
    },
    ...
  ],
  "session_utterances": [
    {"utterance_id": "{user}_X", "speaker": "{user}", "utterance":
    "<string>"},
    {"utterance_id": "{system}_X", "speaker": "{system}",
    "utterance": "<string>"},
    ...
  ]
}

Output format:
[
  {
    "from": "<string: first {system}_X in phase>",
    "to": "<string: last {system}_X in phase>",
    "selected_macro_action": "<string: name of macro action>",
    "annotation_rule_used": null or "<string>",
    "confidence_score": {
      "intent_alignment": <float between 0.0 and 1.0>,
      "goal_fit": <float between 0.0 and 1.0>,
      "alternative_actions_possible": <float between 0.0 and 1.0>,
      "utterance_coherence": <float between 0.0 and 1.0>
    }
  },
  ...
]

Input:
{input_content}

Now return ONLY the JSON output, no explanations.
```

Micro-action annotation prompt.

```
You are an expert assistant helping annotate {domain} dialogue
    transcripts.

You will be **looping through the utterances of a {domain}
    {interaction_unit}**, annotating {system} utterances in
    batches.

You will be given:
1. **The selected macro action for this batch**:
- name: Concise, descriptive name.
- description: A short explanation of the macro action.
- goal: the goal for the macro action.
- All utterances in this batch are labeled to this macro action.
2. **List of micro actions** -- each with:
- name: Concise, descriptive name.
- description: A short explanation of the micro action.
- states: The annotation rules that indicate when this micro
    action should apply.
3. **The last few utterances** from a {domain} {interaction_unit}
    (context), in chronological order, labeled with speaker and
    unique IDs.
4. **A batch of new utterances** (also chronological, labeled with
    speaker and unique IDs).
- {user.capitalize()} utterances are for context only and must not
    be annotated.
- {system.capitalize()} utterances are the ones you must annotate.

Your task:
For each {system} utterance in the batch:
1. Select exactly ONE micro action from the provided list that
    best matches the {system}'s utterance.
- If none of the micro actions are appropriate, label it with
    "None".
2. Provide a confidence score between 0.0 and 1.0 based on how
    clearly the {system}'s intent matches the micro action, and
    whether alternative micro actions were plausible.

When to use annotation rules:
- Choose the micro action by matching the {system}'s
    intent/purpose to the micro action's name / description.
- Only if ambiguous, consult the states rules as tie-breakers.
- If you used a specific rule to resolve the ambiguity, set
    "annotation_rule_used" to that rule's exact text.
- Otherwise set "annotation_rule_used" to null.

Guidelines:
- Focus on the {system}'s intent and purpose, considering the
    prior utterances for context.
- If no suitable micro action exists, use "None" instead of
    forcing a match.
- Do not invent micro actions that are not provided in the list.
- You must output one result for each {system} utterance in the
    batch.

Input format:
{
  "macro_action": {
    "name": "<string>",
    "description": "<string>",
    "goal": "<string>"
  },
```

```
  "micro_actions": [
    {
      "name": "<string>",
      "description": "<string>",
      "states": "<string>"
    },
    ...
  ],
  "last_utterances": [
    {"utterance_id": "{system}_X", "speaker": "{system}",
    "utterance": "<string>"},
    {"utterance_id": "{user}_X", "speaker": "{user}", "utterance":
    "<string>"},
    ...
  ],
  "batch_utterances": [
    {"utterance_id": "{system}_X", "speaker": "{system}",
    "utterance": "<string>"},
    {"utterance_id": "{user}_X", "speaker": "{user}", "utterance":
    "<string>"},
    ...
  ]
}

Output format:
[
  {
    "utterance_id": "<string: {system}_X>",
    "selected_micro_action": "<string: name of micro action or
    None>",
    "annotation_rule_used": null or "<string>",
    "confidence_score": <float between 0.0 and 1.0>
  },
  ...
]

Input:
{input_content}

Now return ONLY the JSON output, no explanations.
```

Conversation-state annotation prompt (numeric tracking).

```
You are an expert {domain} assistant that tracks changes in
    {user}'s state during a {domain} {interaction_unit} using a
    set of structured dimensions.

You are given:
1. Ordered dimensions metadata.
2. A previous numeric state vector.
3. New dialogue utterances.

Task:
Update the state vector by evaluating how the new utterances
    affect each dimension.
- Increase the value if the dialogue strongly indicates an
    increase.
- Decrease the value if the dialogue strongly indicates a decrease.
- Leave the value unchanged if evidence is weak or absent.

Rules:
```

```
- Keep the same vector length.
- Each element must be a float between 0.0 and 1.0.
- Changes should be small and realistic.
- If evidence is ambiguous, keep the previous value.

Output ONLY the final updated vector:
[ float, float, float, ..., float ]
```

Conversation-state annotation prompt (categorical tracking).

```
You are given:
1. Ordered dimensions metadata with allowed categories.
2. A previous categorical state vector.
3. New dialogue utterances.

Task:
Update the categorical state vector.
- If the new utterances clearly indicate a shift toward another
    category, update it.
- If multiple categories are possible, choose the one best
    supported by the utterances.
- If evidence is weak or ambiguous, keep the previous category.
- Do not invent categories not listed in the metadata.

Output ONLY the final updated vector in JSON style:
{
  "dim_0": "<exact value from categories list>",
  ...,
  "dim_N": "<exact value from categories list>"
}
```

Intrinsic-reward annotation prompt.

```
You are an expert {domain} assistant evaluating {system}
    utterances in {domain} {interaction_unit}s.

Task:
Assign ONE integer progress score from 0 to 2 representing how far
    the dialogue has progressed toward the current macro-action
    goal.

You are given:
- the goal objective,
- success levels for unsuccessful / acceptable / ideal,
- a reward hint,
- prior utterances as context,
- the new utterances to evaluate as a whole.

Scoring rubric:
- 0: unsuccessful
- 1: acceptable
- 2: ideal

Output format:
{
  "label": <integer 0-2>,
  "confidence_score": <float between 0.0 and 1.0>
}
```

Extrinsic-reward annotation prompt (CBT).

```
You are an expert rater using the Cognitive Therapy Rating Scale
    (CTRS).

Context: Therapy session about <PRESENTING_CONDITIONS> regarding
    <CASE_SUMMARY>

Rate each code on a 0--6 scale:
0 = Poor, 1 = Barely Adequate, 2 = Mediocre, 3 = Satisfactory,
4 = Good, 5 = Very Good, 6 = Excellent.
Odd numbers (1/3/5) indicate performance between the even-numbered
    anchors.

Detailed anchors by code:

Part 1: General Therapeutic Skills
- "agenda" (Agenda)
  0: Therapist did not set agenda.
  2: Therapist set agenda that was vague or incomplete.
  4: Therapist worked with patient to set a mutually satisfactory
    agenda with specific target problems.
  6: Therapist worked with patient to set an appropriate agenda,
    established priorities, and followed it.

- "feedback" (Feedback)
  0: Therapist did not ask for feedback to determine the patient's
    understanding of, or response to, the session.
  2: Therapist elicited some feedback, but not enough to be sure
    the patient understood the therapist's line of reasoning or
    was satisfied with the session.
  4: Therapist asked enough questions to ensure understanding and
    determine the patient's reactions, and adjusted appropriately.
  6: Therapist was especially adept at eliciting and responding to
    verbal and non-verbal feedback throughout the session.

- "understanding" (Understanding)
  0: Therapist repeatedly failed to understand what the patient
    explicitly said and consistently missed the point.
  2: Therapist usually reflected what the patient explicitly said,
    but repeatedly failed to respond to more subtle communication.
  4: Therapist generally grasped the patient's internal reality,
    including subtler communication.
  6: Therapist thoroughly understood the patient's internal
    reality and communicated this understanding skillfully.

- "interpersonal_effectiveness" (Interpersonal Effectiveness)
  0: Therapist had poor interpersonal skills and seemed hostile,
    demeaning, or destructive.
  2: Therapist was not destructive, but had significant
    interpersonal problems.
  4: Therapist displayed a satisfactory degree of warmth, concern,
    confidence, genuineness, and professionalism.
  6: Therapist displayed optimal levels of warmth, concern,
    confidence, genuineness, and professionalism for this patient
    and session.

- "collaboration" (Collaboration)
  0: Therapist did not attempt to set up a collaboration with
    patient.
  2: Therapist attempted to collaborate, but had difficulty
    defining a problem the patient considered important or
    establishing rapport.
```

```
  4: Therapist collaborated with patient, focused on an important
    problem, and established rapport.
  6: Collaboration seemed excellent; therapist encouraged patient
    to take an active role so they could function as a team.

- "pacing_time_use" (Pacing and Efficient Use of Time)
  0: Therapist made no attempt to structure therapy time; session
    seemed aimless.
  2: Session had some direction, but the therapist had significant
    structuring or pacing problems.
  4: Therapist was reasonably successful at using time efficiently
    and maintained appropriate control over pacing.
  6: Therapist used time efficiently by limiting peripheral
    discussion and pacing the session as rapidly as appropriate.

Part 2: Conceptualization, Strategy, and Technique
- "guided_discovery" (Guided Discovery)
  0: Therapist relied primarily on debate, persuasion, or
    lecturing.
  2: Therapist relied too heavily on persuasion and debate rather
    than guided discovery.
  4: Therapist mostly helped patient see new perspectives through
    guided discovery rather than debate.
  6: Therapist was especially adept at using guided discovery to
    explore problems and help patient draw conclusions.

- "focusing_on_key_cognitions_behaviors" (Focusing on Key
    Cognitions or Behaviors)
  0: Therapist did not attempt to elicit specific thoughts,
    assumptions, images, meanings, or behaviors.
  2: Therapist used appropriate techniques, but had difficulty
    finding a focus or focused on irrelevant cognitions or
    behaviors.
  4: Therapist focused on cognitions or behaviors relevant to the
    target problem.
  6: Therapist very skillfully focused on the key thoughts,
    assumptions, or behaviors most relevant to progress.

- "strategy_for_change" (Strategy for Change)
  0: Therapist did not select cognitive-behavioral techniques.
  2: Therapist selected cognitive-behavioral techniques, but the
    overall strategy seemed vague or unpromising.
  4: Therapist had a generally coherent strategy for change that
    showed reasonable promise.
  6: Therapist followed a consistent and very promising strategy
    for change.
  Note: Focus on the quality of the therapist's strategy for
    change, not on how effectively the strategy was implemented or
    whether change actually occurred.

- "application_of_cbt_techniques" (Application of
    Cognitive-Behavioral Techniques)
  0: Therapist did not apply any cognitive-behavioral techniques.
  2: Therapist used cognitive-behavioral techniques, but there
    were significant flaws in how they were applied.
  4: Therapist applied cognitive-behavioral techniques with
    moderate skill.
  6: Therapist very skillfully and resourcefully employed
    cognitive-behavioral techniques.
  Note: Focus on how skillfully the techniques were applied, not
    on how appropriate they were for the target problem or whether
    change actually occurred.
```

```
- "homework" (Homework)
  0: Therapist did not attempt to incorporate homework relevant to
    cognitive therapy.
  2: Therapist had significant difficulties incorporating homework.
  4: Therapist reviewed previous homework and assigned standard
    cognitive therapy homework relevant to the session.
  6: Therapist reviewed previous homework and carefully assigned
    custom-tailored homework for the coming week.

TASK:
1) Read the transcript and rate EACH code (integer 0--6).
2) For each code, provide:
   - "score": integer 0--6
   - "rationale": <= 35 words explaining your score
   - "confidence": float in [0.00, 1.00]

STRICT OUTPUT:
Return ONLY a single JSON object with this exact top-level schema:
{
  "codes": {
    "agenda": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "feedback": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "understanding": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "interpersonal_effectiveness": {"score": <int>, "rationale":
    "<text>", "confidence": <float>},
    "collaboration": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "pacing_time_use": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "guided_discovery": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "focusing_on_key_cognitions_behaviors": {"score": <int>,
    "rationale": "<text>", "confidence": <float>},
    "strategy_for_change": {"score": <int>, "rationale": "<text>",
    "confidence": <float>},
    "application_of_cbt_techniques": {"score": <int>, "rationale":
    "<text>", "confidence": <float>},
    "homework": {"score": <int>, "rationale": "<text>",
    "confidence": <float>}
  }
}

Transcript:
<SESSION_TRANSCRIPT>
```

Extrinsic-reward annotation prompt (P4G).

```
System:
You are an expert in evaluating persuasive conversations between a
    persuader and a persuadee.
Your job is to determine if the persuader has successfully
    convinced the persuadee to donate to a charity called "Save
    the Children".
The persuadee's emotional states throughout the conversation were:
    <EMOTION_STATES>.

User:
Conversation so far:
```

```
<FULL_CONVERSATION>

Please evaluate the following conversation and strictly choose ONE
    of the exact responses listed below.

Question:
Has the persuadee been successfully convinced to make a donation
    at the end of the conversation?

Respond with one of the following options:
- The persuadee has explicitly refused.
- The persuadee remains neutral about donating.
- The persuadee has a positive attitude towards donating but
    hasn't decided yet.
- The persuadee has decided to donate.
```

Caution annotation prompt.

```
You are an expert assistant annotating {domain} {interaction_unit}
    transcripts.

You are given:
1. A list of cautions (things the {system} should avoid).
2. The last few utterances from the transcript.
3. A batch of new utterances.

Task:
For each {system} utterance in the batch:
- Select exactly ONE negative_action from the provided cautions,
    or "None".
- Provide a confidence score between 0.0 and 1.0.

Guidelines:
- Focus on the {system}'s intent and purpose, using prior context
    when needed.
- Assign a caution only when the utterance explicitly or strongly
    exhibits it.
- If evidence is ambiguous or weak, choose "None" rather than
    over-labeling.
- Output only JSON.
```

TOPAS user-profile initialization prompt.

```
You are initializing the profile of one {user_role} for a {domain}
{interaction_unit}. Use only the supplied case data. Do not use or
    assume any
future conversation.

The profile schema is fixed. Return case-specific values for every
    listed
dimension without changing, adding, or removing dimension names.

STATIC VALUES
For every static dimension, write one concise string grounded in
    the case data.
Preserve relevant context, qualifications, and uncertainty. Use
    "unknown" when
the case does not support a value. Definitions and value guidance
    explain what
```

```
belongs in a field.

DYNAMIC INITIAL VALUES
For every dynamic dimension, select its initial categorical value
    from that
dimension's options.
- selection_mode = "single": return exactly one option string.
- selection_mode = "multiple": return a nonempty list of distinct
    options that
  can coexist.
- Unknown must stand alone as "unknown" or ["unknown"].
Do not assume improvement, agreement, trust, disclosure, or
    resistance without
case evidence.

Return valid JSON with exactly these two top-level keys and no
    surrounding text:
{
  "static_values": {
    "<static dimension name>": "<case-supported value>"
  },
  "dynamic_initial_values": {
    "<dynamic dimension name>": "<one option or a list of options>"
  }
}

<profile_schema>
<PROFILE_SCHEMA>
</profile_schema>

<case_data>
<CASE_DATA>
</case_data>
```

E.3 LLM-JUDGE SURVEY PROMPT

We evaluate extraction quality with an automated pairwise LLM judge. To reduce position bias, we randomize which method is shown as *Output A* vs. *Output B* for each comparison, and map the predicted winner back to the original method identity.

Prompt for the pairwise LLM judge.

```
You are an impartial expert judge in <DOMAIN>.

Task: Compare Output A vs Output B for the SAME component:
    "<COMPONENT>".

Component definition:
<COMPONENT-SPEC DEFINITION>

Instructions:
- Relevance: Prefer the output that most directly and accurately
    addresses what the component definition asks for.
- Completeness: Prefer the output that includes the key required
    details and does not omit important elements of the definition.
- Clarity: Prefer the output that is easiest to understand and is
    clinically meaningful (clear wording, minimal ambiguity).
- Overall preference: If one output is superior across these
    criteria, choose it.
- Tie rule: If both outputs are genuinely indistinguishable in
    relevance, completeness, and clarity, answer "tie".
```

```
- Justification: Provide a brief reason (1--3 sentences),
    referring to the candidates as "Output A" and "Output B".

Return ONLY valid JSON (no markdown, no extra text):
{
  "winner": "A" | "B" | "tie",
  "reason": "<1-3 sentences>"
}

Output A:
<JSON OUTPUT A>

Output B:
<JSON OUTPUT B>
```

F EXTRACTION PHASE OUTPUTS

This appendix reports the structured outputs produced in the extraction phase of TOPAS: the hierarchical action space, conversation-state schema, knowledge-graph schema, cautions list, and user-profile schema.

Action space

The extracted action space is organized around macro actions that capture the major therapeutic functions of CBT. Each macro action is paired with applicability states, a treatment objective, a reward hint, success criteria, and a concrete micro-action set.

Macro action	Objective	# micro
Therapeutic Alliance and Engagement	Develop a strong, safe, and collaborative foundation for effective therapeutic work.	5
Psychoeducation and Therapy Orientation	Increase patient understanding of CBT, foster engagement, and correct misconceptions.	5
Assessment, Diagnosis, and Case Formulation	Achieve a shared, clear understanding of the patient's difficulties and conceptual model to guide treatment.	5
Agenda, Goal Setting, and Action Planning	Establish shared goals and clear session plans to organize and focus therapy.	5
Cognitive Interventions: Identification and Restructuring	Reduce distress and maladaptive behaviors by restructuring dysfunctional cognitive patterns.	6
Core Beliefs and Schema Work	Restructure negative core beliefs and build more flexible, adaptive self-concepts.	5
Behavioral Activation, Experimentation, and Skills Training	Decrease avoidance and increase positive behavior, learning, and coping skills.	5
Exposure and Response Prevention	Reduce avoidance and anxiety by increasing experiential learning and disconfirming maladaptive beliefs.	5
Skills Generalization, Homework, and Progress Review	Generalize CBT skills to real-life situations and foster independent coping.	5
Relapse Prevention and Termination Planning	Enable patient to sustain gains, prevent relapse, and apply skills independently post-therapy.	5
Crisis Management and Safety Planning	Manage acute risk and ensure patient safety through rapid assessment, collaborative planning, and emergency intervention if needed.	5
Ongoing Progress Monitoring and Troubleshooting	Ensure effective treatment by tracking progress, adjusting strategies, and resolving barriers.	5

A1. Therapeutic Alliance and Engagement

Description. Therapist builds rapport, fosters trust, and encourages open communication to support collaboration throughout therapy.

Objective. Develop a strong, safe, and collaborative foundation for effective therapeutic work.
Micro actions.

- **Warm Greeting and Rapport-Building**
- **Validate and Reflect Patient Experience**
- **Solicit Collaborative Input and Decision-Making**
- **Directly Invite Feedback on the Alliance**
- **Normalize Hesitancy and Reduce Stigma**

A2. Psychoeducation and Therapy Orientation

Description. Therapist explains CBT principles, session structure, dispels myths, and normalizes emotional and psychological experiences.

Objective. Increase patient understanding of CBT, foster engagement, and correct misconceptions.
Micro actions.

- **Explain CBT Model and Treatment Structure**
- **Normalize Symptoms and Emotional Distress**
- **Debunk Myths and Address Misconceptions**
- **Invite Questions and Feedback**
- **Reinforce Psychoeducation with Visuals or Handouts**

A3. Assessment, Diagnosis, and Case Formulation

Description. Therapist gathers information on symptoms, personal history, context, and risk to clarify presenting problems and form an individualized CBT case conceptualization.

Objective. Achieve a shared, clear understanding of the patient's difficulties and conceptual model to guide treatment.
Micro actions.

- **Ask Targeted Assessment Questions**
- **Elicit Narrative and Context**
- **Administer and Review Standardized Measures**
- **Collaboratively Review and Update Formulation**
- **Identify and Rate Mood, Thoughts, Emotions, Behaviors**

A4. Agenda, Goal Setting, and Action Planning

Description. Therapist and patient collaboratively set session agendas, prioritize treatment goals, and develop actionable plans for intervention.

Objective. Establish shared goals and clear session plans to organize and focus therapy.
Micro actions.

- **Initiate Agenda Setting and Goal Identification**
- **Elicit Patient's Preferences and Values**
- **Collaboratively Prioritize and Specify Goals**
- **Record and Review Agenda/Goals**
- **Facilitate Cost–Benefit and Motivation Discussion**

A5. Cognitive Interventions: Identification and Restructuring

Description. Therapist helps the patient identify automatic thoughts, cognitive distortions, and maladaptive beliefs, and uses Socratic techniques to challenge and modify them.

Objective. Reduce distress and maladaptive behaviors by restructuring dysfunctional cognitive patterns.
Micro actions.

- **Elicit and Record Automatic Thoughts**
- **Identify and Categorize Cognitive Distortions**
- **Use Socratic Questioning to Challenge Beliefs**
- **Facilitate Generation of Alternative Thoughts**
- **Guide Ratings of Belief Strength and Mood Change**
- **Review and Troubleshoot Cognitive Restructuring**

A6. Core Beliefs and Schema Work

Description. Therapist assists patient to identify, evaluate, and modify core beliefs and schemas underlying persistent difficulties.

Objective. Restructure negative core beliefs and build more flexible, adaptive self-concepts.
Micro actions.

- **Apply Downward Arrow Technique**
- **Facilitate Core Belief Worksheet Completion**
- **Solicit and Reinforce New Adaptive Beliefs**
- **Support Behavioral Experiments for New Beliefs**
- **Prompt Regular Monitoring and Review**

A7. Behavioral Activation, Experimentation, and Skills Training

Description. Therapist uses graded activity scheduling, behavioral experiments, and skills training to reduce avoidance, increase adaptive behaviors, and improve functioning.

Objective. Decrease avoidance and increase positive behavior, learning, and coping skills.
Micro actions.

- **Collaboratively Design Behavioral Experiments and Activity Plans**
- **Break Tasks Into Steps and Problem-Solve Barriers**
- **Model and Teach New Behavioral/Coping Skills**
- **Prompt In-Session or Homework Skill Practice**
- **Encourage Tracking and Reflection**

A8. Exposure and Response Prevention

Description. Therapist and patient collaboratively design and implement graded exposure to feared or avoided situations and guide prevention of safety behaviors.

Objective. Reduce avoidance and anxiety by increasing experiential learning and disconfirming maladaptive beliefs.
Micro actions.

- **Develop Exposure Hierarchy**
- **Assign and Guide Exposure Tasks**
- **Prompt Prediction and Outcome Review**
- **Monitor and Reduce Safety Behaviors**
- **Provide Support and Encouragement During Exposure**

A9. Skills Generalization, Homework, and Progress Review

Description. Therapist assigns practice tasks between sessions, monitors progress, addresses obstacles, and reinforces skill development.

Objective. Generalize CBT skills to real-life situations and foster independent coping.
Micro actions.

- **Assign Clear, Specific Homework Tasks**
- **Review Homework Completion and Barriers**
- **Collaboratively Problem-Solve Obstacles**
- **Provide Positive, Constructive Feedback**
- **Model Mistake Normalization**

A10. Relapse Prevention and Termination Planning

Description. Therapist and patient prepare for therapy ending by consolidating skills, anticipating risk, and developing maintenance plans.

Objective. Enable patient to sustain gains, prevent relapse, and apply skills independently post-therapy.
Micro actions.

- **Review and Consolidate Skills and Progress**
- **Collaboratively Develop Relapse Prevention Plan**
- **Practice Skills in High-Risk Scenarios**
- **Ensure Visibility of CBT Materials**
- **Promote Use of Social Support**

A11. Crisis Management and Safety Planning

Description. Therapist directly addresses acutely increased risk (e.g., suicidality or emergency needs) through focused assessment, stabilization, and plan development.

Objective. Manage acute risk and ensure patient safety through rapid assessment, collaborative planning, and emergency intervention if needed.
Micro actions.

- **Directly Assess Risk**
- **Develop and Review Safety/Coping Plan**
- **Monitor and Reassess Risk During Session**
- **Initiate Emergency Procedures if Indicated**
- **Provide Emotional Support and Grounding**

A12. Ongoing Progress Monitoring and Troubleshooting

Description. Therapist regularly tracks symptom change, adapts interventions, and engages in collaborative problem-solving of obstacles and setbacks.

Objective. Ensure effective treatment by tracking progress, adjusting strategies, and resolving barriers.
Micro actions.

- **Review Symptoms, Ratings, and Observations**
- **Provide Constructive Feedback**
- **Troubleshoot Barriers to Treatment**
- **Adapt Goals and Interventions**
- **Encourage Self-Validation and Strengths Focus**

Conversation states

The extraction pipeline also produces a session-state schema for tracking clinically relevant variables across turns. Each variable includes a semantic description, a categorical value set, and a numeric scale specification.

Variable	Description	Categorical values	Numeric
Automatic Thoughts	Content, frequency, and valence of patient's spontaneous thoughts that drive emotional and behavioral responses. Used to identify maladaptive patterns and guide cognitive restructuring.	Negative; Neutral; Positive; Catastrophic; Obsessive; Recurrent	regression from 0 to 1 scale
Core Beliefs/Schemas	Underlying, deep-seated beliefs about self, others, or the world, including confidence and changes in these beliefs. Guides advanced CBT and cognitive restructuring.	Negative-Self; Negative-Other; Negative-World; Adaptive; Maladaptive; Self-critical; Anxious; (+5 more)	regression from 0 to 1 scale
Belief Conviction/Strength	Degree of conviction in target, alternative, and problematic cognitions, including automatic thoughts and core beliefs.	Very Low; Low; Moderate; High; Very High	regression from 0 to 1 scale
Mood/Emotional State	Current emotional state and its intensity, including shifts and predominant affect (e.g., anxiety, depression, anger, guilt).	Sad; Anxious; Angry; Guilty; Happy; Irritable; Shame; (+11 more)	regression from 0 to 1 scale
Emotional State Intensity	Strength or severity of emotional responses or distress. Used to track intervention impact, risk, and session pacing.	Low; Mild; Moderate; High; Severe; Extreme; Crisis/Emergency	regression from 0 to 1 scale
Symptom Severity	Overall intensity and fluctuation of target psychiatric symptoms (e.g., depression, anxiety). Guides formulation and treatment.	None; Below baseline; Baseline; Mild; Moderate; Severe; Acute; (+1 more)	regression from 0 to 1 scale
Functional Impairment	Extent to which symptoms or behaviors interfere with social, occupational, educational, or daily functioning.	None; Mild; Moderate; Severe; Extreme	regression from 0 to 1 scale
Cognitive Distortions/Errors	Type, presence, and frequency of faulty thinking patterns (e.g., catastrophizing, overgeneralization, black-and-white thinking). Informs cognitive restructuring.	None; Mild; Moderate; Severe; Extreme; Occasional; Frequent; (+2 more)	regression from 0 to 1 scale
Insight and Self-Awareness	Patient's ability to identify, reflect on, and understand their own thoughts, emotions, and behavioral patterns.	None; Low; Moderate; High; Very High; Partial; Consistent	regression from 0 to 1 scale
Motivation and Readiness for Change	Patient's willingness, motivation, and openness to engage in change processes and new strategies.	Highly motivated; High; Ambivalent; Somewhat Ready; Moderate; Low; Unmotivated; (+3 more)	regression from 0 to 1 scale
Therapeutic Alliance	Strength, quality, and trust in the working relationship between therapist and patient, including collaboration and conflict.	Weak; Vulnerable/Repairing; Adequate; Strong; Very Strong; Low; Fair; (+1 more)	regression from 0 to 1 scale
Patient Engagement and Participation	Patient's activity, willingness, and collaboration during sessions and with treatment tasks.	Not engaged; Minimal; Passive; Partial; Partially Engaged; Active; Mostly engaged; (+6 more)	regression from 0 to 1 scale
Homework/Behavioral Experiment Completion	Degree of completion, quality of engagement, and reflection with therapy homework, assignments, or behavioral experiments.	Not Started; Not completed; No engagement; None; Started; Partially Completed; Partial; (+11 more)	regression from 0 to 1 scale
Avoidance and Safety Behaviors	Presence, frequency, and reliance on avoidance and safety behaviors/rituals aimed at reducing perceived threat or distress.	Absent; None; Subtle; Occasional; Minor; Frequent; Pervasive; (+6 more)	regression from 0 to 1 scale
Progress Toward Goals	Movement toward therapy/treatment goals, including behavioral activation, symptom reduction, and completion of tasks.	No progress; Partial progress; On track; Significant progress; Goal achieved; Regressing; Progressing	regression from 0 to 1 scale
Social Support and Interpersonal Functioning	Presence, quality, and helpfulness of patient's relationships, support network, and interpersonal connections.	Strong/supportive; Reliable; Robust; Highly Engaged; Adequate; Limited; Impaired/conflict; (+6 more)	regression from 0 to 1 scale
Coping Strategies Utilization/Effectiveness	Frequency and effectiveness of adaptive or maladaptive coping skills used to manage distress, including problem-solving and emotional regulation.	None; Not used/ineffective; Rarely used; Maladaptive; Some; Occasionally/effective; Occasional Adaptive; (+12 more)	regression from 0 to 1 scale
Comorbid Symptoms Impact	Influence of comorbid diagnoses or symptoms (e.g., depression, social anxiety, chronic pain, substance misuse) on session engagement and progress.	None; Mild; Moderate; Severe; Dominant	regression from 0 to 1 scale
Relapse Risk and Early Warning Signs	Presence and intensity of early warning signs or risk factors for relapse or deterioration. Used for monitoring and prevention.	Absent; Present; Minimal; Mild; Moderate; Increasing; High Risk; (+2 more)	regression from 0 to 1 scale
Suicidal Ideation and Risk	Presence, intensity, and change in suicidal thoughts, intent, planning, and behaviors. Guides safety planning.	Absent; None; Passive; Passive ideation; Active; Active Ideation; Active/Ineffective plan; (+7 more)	regression from 0 to 1 scale
Obstacles and Barriers to Change	Internal and external barriers (practical, emotional, motivational, or situational) interfering with progress, engagement, or skill application.	None; Cognitive; Emotional; Behavioral; Motivational; Situational; Minor; (+7 more)	regression from 0 to 1 scale
Change in Belief/Emotion Post-Experiment	Degree of shift in conviction or emotional response following behavioral experiments or corrective experiences.	No Change; Slight Shift; Moderate Change; Substantial Change; Complete Shift	regression from 0 to 1 scale

Variable	Description	Categorical values	Numeric
Generalization of Learning	Extent to which acquired cognitions and behaviors transfer to situations outside of session. Indicates real-world impact.	Absent; Minimal; Partial; Full; Consistent	regression from 0 to 1 scale
Subjective Distress Level	Self-reported or observed emotional discomfort or distress, including SUDs ratings. Used to inform safety and intervention selection.	Low; Mild; Moderate; High; Extreme	regression from 0 to 1 scale
Readiness/Commitment to Change	Patient's preparedness, willingness, and commitment to take action, explore new strategies, or participate in exposures/advanced interventions.	Not Ready; Ambivalent; Somewhat Ready; Ready; Very Ready; Considering; Open/Ready; (+1 more)	regression from 0 to 1 scale
Therapist Emotional/Countertransference Response	Therapist's own emotional responses and their potential influence on the session.	Calm; Empathic; Frustrated; Fatigued; Burnt Out; Triggered; Distressed; (+1 more)	regression from 0 to 1 scale
Quality of Life	Overall assessment of patient's well-being and satisfaction with life, reflecting symptoms and functional impact.	Excellent; Good; Fair; Poor; Very poor	regression from 0 to 1 scale
Sleep Patterns and Quality	Quality, onset, duration, and disruptions in sleep, including restfulness or insomnia.	Restful/normal; Mild disturbance; Moderate disturbance; Severe insomnia	regression from 0 to 1 scale
Physical Symptoms	Somatic symptoms related to mood or anxiety (e.g., sleep, appetite, energy, agitation).	Sleep Problems; Appetite Issues; Fatigue; Restlessness; Agitation; Other	regression from 0 to 1 scale
Cultural/Identity/Context Factors	Consideration of individual cultural, developmental, and identity factors relevant to therapy engagement and alliance.	Influential; Minimal Influence; Barrier; Facilitator; Not considered; Minimally considered; Adequately considered; (+1 more)	regression from 0 to 1 scale
Learning/Cognitive Performance	Ability to retain, recall, and apply CBT concepts and skills, highlighting barriers and need for adaptations.	Impaired; Adequate; Strong	regression from 0 to 1 scale
Attentional Focus	Whether patient's attention is self-focused or externally directed. Guides intervention selection.	Internal; External; Mixed	regression from 0 to 1 scale
Environmental/Situational Stressors	External circumstances or recent life events impacting engagement, such as trauma, financial stress, or loss.	Absent; None; Mild; Moderate; Severe	regression from 0 to 1 scale
Therapist-Patient Consensus and Conflict	Degree of agreement or discrepancy between therapist and patient on goals, analysis, and evidence.	Complete Consensus; Partial Consensus; Neutral; Partial Conflict; Severe Conflict	regression from 0 to 1 scale

Knowledge-graph schema

The extracted knowledge graph is anchored in a *Patient* root node and expands into diagnostic, cognitive, behavioral, contextual, and therapeutic dimensions. The following tables summarize the nodes list and the typed relations recovered by the extractor.

Node type	Description
Patient	Root node representing the individual receiving cognitive behavioral therapy.
Demographics	Patient's age, gender, cultural background, education, vocation, and social context.
Presenting Problems	Presenting complaints, symptom descriptions, and contextual/problem history.
Diagnosis/Clinical History	Mental and physical health diagnoses, comorbidities, and longitudinal symptom profiles.
Medical History	Relevant medical background, medication adherence, and physical health/disability details.
Family Psychiatric and Medical History	Family history of psychiatric, medical, and substance issues impacting the patient.
Developmental and Life Events	Developmental, childhood, family, social, educational, and critical life events or trauma.
Cultural/Contextual Factors	Cultural, familial, religious, language, and developmental influences relevant to therapy.
Core Beliefs/Schemas	Enduring beliefs and schemas about self, others, world; developmental origins and conviction strength.
Intermediate Beliefs/Assumptions	Attitudes, rules, conditional assumptions derived from core beliefs.
Automatic Thoughts	Recurring spontaneous cognitions or negative automatic thoughts and their context and impact.

Node type	Description
Cognitive Distortions	Unhelpful thinking styles, biases, and patterns across time.
Behavioral Patterns	Adaptive and maladaptive patterns including avoidance, approach, rituals, withdrawal, and safety behaviors.
Coping Strategies	Adaptive and maladaptive strategies used by the patient for regulation and distress management.
Strengths and Support System	Patient's identified strengths, assets, resilience, and available social/relationship supports.
Therapeutic Alliance/Relationship	Quality, rupture/repair cycles, engagement, compliance, feedback, and skill development in the therapy relationship.
Treatment/Interventions	Records of CBT interventions, exposures, behavioral experiments, and effectiveness/rationale.
Homework/Assignments	Therapy homework and assignments history, adherence, completion, and outcome notes.
Skill Acquisition/CBT Competency	Acquisition and mastery of CBT skills and ability for independent use.
Activity Patterns/Scheduling	Scheduling of activities, mastery/pleasure ratings, and engagement/avoidance patterns.
Mood and Emotional Responses	Patterns of emotional, behavioral, and physiological responses including depression, anxiety, anger, shame, and mood tracking.
Triggers	Situational, personal, or environmental triggers for symptoms, moods, and automatic thoughts.
Safety Behaviors/Avoidance	Catalogue of specific safety, avoidance, and ritual behaviors, their context and frequency.
Episodes/Setbacks/Relapse History	History and details of symptom episodes, relapses, setbacks, warning signs, and preventive strategies.
Motivation and Values	Key goals, values, motivation sources, and value-based behavioral patterns.
Problem-Solving Capacity	Problem-solving style, skill level, attempts, progress, and responses to setbacks.
Personality Features/Traits	Personality disorder features or relevant traits such as obsessiveness, perfectionism, or hostility.
Self-Concept/Identity	Self-esteem, shame, self-worth, identity, and related beliefs and changes.
Physical Symptoms/Reactions	Typical physical reactions to emotional distress (e.g., agitation, withdrawal, somatic symptoms).
Habits and Lifestyle	Patterns in diet, sleep, exercise, substance use, routines, and self-care.
Addiction/Problematic Behaviors	Substance or behavioral addictions and compulsive behaviors.
Sleep and Insomnia Features	Sleep patterns, quality, beliefs about sleep, and related behaviors.
Eating Disorder and Body Image Features	Disordered eating patterns, compensatory behaviors, body image concerns, and related beliefs.
OCD and Compulsive Symptoms	Obsessional problems, compulsions, rituals, and related avoidance.
Anger and Rumination Patterns	Antecedents, patterns, and consequences of anger and rumination.
Trauma and Abuse History	History, timing, and sequence of trauma, loss, abuse, bullying, neglect, and associated impacts.
Support System Involvement	Role/involvement of significant others/family in support, rituals, or symptom maintenance.
Therapist's Observations	Therapist's observations, notes, reactions, and clinical assumptions about therapy process.
Functional Outcomes/Impairment	Functioning and impairment in social, academic, occupational, and daily living domains.
Standardized Assessments	Scores and outcomes from standardized assessment and symptom measures over time.
Insight/Understanding	Patient's evolving insight, metacognition, and understanding of therapy and disorder.
Transference/Countertransferenc	Patient transference patterns and therapist's countertransference reactions.

Source	Relation	Target
Patient	has	Demographics
Patient	has	Diagnosis/Clinical History

Source	Relation	Target
Diagnosis/Clinical History	includes	Medical History
Diagnosis/Clinical History	may include	Family Psychiatric and Medical History
Patient	presents	Presenting Problems
Patient	has	Developmental and Life Events
Developmental and Life Events	inform	Core Beliefs/Schemas
Developmental and Life Events	may include	Trauma and Abuse History
Patient	is influenced by	Cultural/Contextual Factors
Patient	holds	Core Beliefs/Schemas
Patient	holds	Intermediate Beliefs/Assumptions
Core Beliefs/Schemas	underlie	Intermediate Beliefs/Assumptions
Intermediate Beliefs/Assumptions	shape	Automatic Thoughts
Patient	has	Automatic Thoughts
Automatic Thoughts	display	Cognitive Distortions
Automatic Thoughts	trigger	Mood and Emotional Responses
Automatic Thoughts	reinforce	Core Beliefs/Schemas
Patient	shows	Cognitive Distortions
Cognitive Distortions	lead to	Behavioral Patterns
Patient	exhibits	Behavioral Patterns
Triggers	provoke	Automatic Thoughts
Triggers	influence	Mood and Emotional Responses
Patient	is affected by	Triggers
Behavioral Patterns	are mediated by	Coping Strategies
Behavioral Patterns	include	Safety Behaviors/Avoidance
Patient	uses	Coping Strategies
Coping Strategies	contribute_to	Episodes/Setbacks/Relapse History
Coping Strategies	require	Skill Acquisition/CBT Competency
Strengths and Support System	bolster	Coping Strategies
Patient	possesses	Strengths and Support System
Patient	acquires	Skill Acquisition/CBT Competency
Skill Acquisition/CBT Competency	enables	Homework/Assignments
Patient	completes	Homework/Assignments
Homework/Assignments	reflect	Treatment/Interventions
Patient	receives	Treatment/Interventions
Treatment/Interventions	promote	Skill Acquisition/CBT Competency
Patient	engages in	Therapeutic Alliance/Relationship
Therapeutic Alliance/Relationship	influences	Motivation and Values
Therapist's Observations	inform	Therapeutic Alliance/Relationship
Therapist's Observations	guide	Treatment/Interventions
Therapeutic Alliance/Relationship	facilitates	Skill Acquisition/CBT Competency
Therapeutic Alliance/Relationship	strengthens	Motivation and Values
Motivation and Values	inform	Treatment/Interventions
Patient	has	Episode/Setbacks/Relapse History
Episodes/Setbacks/Relapse History	modifies	Treatment/Interventions
Episodes/Setbacks/Relapse History	challenge	Coping Strategies
Motivation and Values	buffer	Episodes/Setbacks/Relapse History
Patient	engages in	Activity Patterns/Scheduling
Activity Patterns/Scheduling	impact	Mood and Emotional Responses
Self-Concept/Identity	reflect	Core Beliefs/Schemas
Behavioral Patterns	impact	Functional Outcomes/Impairment
Trauma and Abuse History	shape	Core Beliefs/Schemas
Physical Symptoms/Reactions	mirror	Mood and Emotional Responses
Therapist's Observations	track	Episodes/Setbacks/Relapse History

Source	Relation	Target
Standardized Assessments	measure	Mood and Emotional Responses
Eating Disorder and Body Image Features	stem from	Core Beliefs/Schemas
Eating Disorder and Body Image Features	are associated with	Cognitive Distortions
Addiction/Problematic Behaviors	contribute to	Behavioral Patterns
Addiction/Problematic Behaviors	are linked to	Episodes/Setbacks/Relapse History
OCD and Compulsive Symptoms	express	Safety Behaviors/Avoidance
Support System Involvement	reinforce	Treatment/Interventions
Motivation and Values	drive	Activity Patterns/Scheduling
Skill Acquisition/CBT Competency	improves	Problem-Solving Capacity
Sleep and Insomnia Features	influence	Behavioral Patterns
Habits and Lifestyle	modulate	Mood and Emotional Responses
Therapist's Observations	record	Transference/Countertransference
Insight/Understanding	inform	Treatment/Interventions

Cautions

The caution set specifies undesirable therapist behaviors together with concise descriptions and the corresponding failure modes. These entries can be used to constrain action selection and define negative outcomes during simulation or policy learning.

Negative action	Description	Risk
Rigid adherence to protocol or technique	Strictly following standardized procedures or one framework without adapting to patient context.	Reduces effectiveness, may harm alliance, alienates or overwhelms patient.
Ignoring or undermining therapeutic alliance	Neglecting collaborative relationship or empathy with the patient.	Damages trust, reduces engagement and effectiveness.
Forcing interventions or exposure before readiness	Pushing patients into experiments, exposure tasks, or new skills before consent or preparation.	Causes distress, loss of trust, increases dropout and risk of harm.
Dismissing, minimizing, or invalidating patient thoughts or emotions	Ignoring, downplaying, or criticizing patient emotional experiences or distress.	Breaks trust and may worsen distress or lead to dropout.
Allowing or reinforcing avoidance and safety behaviors	Permitting use of distraction, rituals, reassurance, or safety strategies during exposure or therapy.	Maintains anxiety and maladaptive patterns, undermines therapeutic progress.
Making uncollaborative assumptions about patient beliefs or readiness	Assuming patient beliefs or history without collaborative exploration.	Leads to resistance, misunderstanding, reduced efficacy.
Skipping planning, preparation, or debriefing for experiments	Failing to plan, reflect, or provide preparation and follow-up for behavioral or cognitive tasks.	Limits learning, may increase distress, reduces quality of outcomes.
Breaching confidentiality or professional boundaries	Disregarding privacy or ethical boundaries or sharing patient information inappropriately.	Damages trust, alliance, and may have legal repercussions.
Neglecting physical health or medical safety	Ignoring health risks, needed clearance, or safety in interventions; offering advice outside professional scope.	Endangers patient, can result in harm or legal issues.
Overwhelming the patient with excessive, complex, or unrealistic tasks	Assigning tasks or goals that are too difficult, packed, vague, or mismatched with patient readiness.	Causes discouragement, avoidance, disengagement.
Offering non-evidence-based or unproven interventions	Using untested, anecdotal, or outdated methods not supported by empirical research.	Ineffective or harmful therapy outcomes.

Negative action	Description	Risk
Invalidating accurate perceptions or real risks	Dismissing genuine negative events or minimizing physical, social, or emotional risks.	Alienates patient, undermines safety, prevents adaptive coping.
Focusing exclusively on cognition or behavior	Only targeting cognitive or behavioral aspects, neglecting emotions, context, or environmental factors.	Limits efficacy and ignores core issues.
Using authoritarian, didactic, or impersonal communication	Lecturing, persuading, or communicating in a top-down manner instead of guided discovery.	Fosters resistance, inhibits alliance, reduces engagement.
Assigning blame, shaming, or criticizing for symptoms or setbacks	Responding punitively or judgmentally to errors, failures, or lack of progress.	Increases shame, undermines motivation, and engagement.
Neglecting or improperly assessing suicide or self-harm risk	Failing to monitor, assess, or respond to signs of suicidality or self-harm.	Serious patient safety threat, potential loss of life.
Ignoring comorbid conditions or contextual factors	Neglecting physical health, substance use, trauma, or cultural/social background in treatment.	Poor outcomes, alienates patient, misses core contributors.
Proceeding before medical or psychiatric stabilization	Starting therapy with acute medical or psychiatric instability.	Worsens condition or misses critical medical needs.
Failing to track symptoms, progress, or relapse risk	Omitting symptom monitoring, progress tracking, or relapse prevention planning.	Increases risk of deterioration and missed setbacks.
Endorsing unrealistic expectations of rapid or total symptom elimination	Promising complete, instant relief from symptoms or minimizing plateaus/setbacks.	Leads to disappointment, drop-out, and invalidates patient experience.
Conflating therapist authority over patient's autonomy or physician's role	Imposing therapist's views, dismissing patient agency, or contradicting medical guidance.	Reduces collaboration, risks safety, erodes trust.
Using stigmatizing language or labels	Labeling or pathologizing patients negatively or pejoratively.	Damages self-esteem, increases stigma, blocks progress.
Encouraging passive or superficial efforts	Accepting half-hearted attempts, lack of engagement, or letting patient be passive in process.	Reduces skill development, fosters dependence, impairs outcomes.
Forcing or rushing cognitive challenging or belief change	Challenging beliefs, guided discovery, or core schemas before rapport or readiness.	Increases resistance, distress, or dropout.
Failing to tailor interventions to individual background, history, or education	Using generic or irrelevant interventions that disregard patient's history, strengths, culture, or learning style.	Reduces motivation, engagement, relevance, and efficacy.
Neglecting to explain rationale for therapy and interventions	Not clarifying purpose behind interventions or making process opaque.	Undermines trust and collaboration.
Endorsing avoidance, suppression, or denial of distressing emotions	Encouraging patient to suppress, ignore, or avoid emotional or distressing content.	Blocks emotional processing and genuine healing.
Making premature inferences or interpretations without patient input	Drawing conclusions about symptoms or underlying causes without collaboration or sufficient evidence.	Leads to errors, reduces alliance quality.
Using non-specific, vague, or unstructured session formats or goals	Skipping agenda-setting or using unclear/unrealistic goal setting.	Sessions become unfocused, limiting therapeutic progress.
Shaming or blaming for ineffective coping attempts	Criticizing patient for trying but failing to implement helpful strategies.	Increases defensiveness and disengagement.
Neglecting supervision, reflective practice, or self-care	Ignoring personal burnout or failing to seek feedback or supervision.	Reduces therapy quality and increases risk of error.
Practicing outside scope of competence	Treating cases beyond professional expertise or legal authority.	Risks patient safety, legal consequences, and ineffective treatment.
Proceeding with advanced interventions or schema work before basic skills are established	Attempting complex treatments or cognitive restructuring prematurely.	Patient may disengage or not benefit from therapy.

Negative action	Description	Risk
Rushing assessment or pressing for full disclosure early	Hurrying intake or pushing patients for personal revelations prematurely.	Increases patient anxiety and diminishes engagement.
Combining medication and CBT without coordination	Implementing both therapies without clarifying roles, rationale, or monitoring effects.	Confusion regarding treatment impact and reduced efficacy.
Encouraging excessive self-blame or guilt	Promoting harsh internal dialogue or shame during cognitive restructuring.	Undermines healing and increases depressive symptoms.
Promoting unreliable or anecdotal information	Endorsing non-evidence-based approaches or unverified advice.	Risks ineffective or harmful outcomes.
Minimizing positive feedback, strengths, or small improvements	Only emphasizing problems, ignoring growth and achievement.	Decreases motivation and self-efficacy.

User-profile schema

The extracted user profile separates relatively stable, free-form attributes from categorical states that can evolve during an interaction. Static dimensions provide case-specific value guidance, whereas dynamic dimensions define their selection mode, admissible options, and the evidence required to update or retain a state.

Static dimension	Description	Value guidance
Demographics, Identity, and Social Context	Captures age, gender, race/ethnicity, marital/family status, education, employment, roles, sexuality, cultural background, religion, and the broader social environment. Provides life context, shapes relevant stressors, influences disclosure comfort, and affects engagement or interpretation of therapy.	Summarize demographic details, identity, living situation, cultural influences, roles, and relevant social-network or identity-linked considerations.
Personal and Developmental History	Captures significant life events, upbringing, early experiences, family context, history of anxiety or mood symptoms, and the developmental stage at onset of key issues.	Describe formative experiences, family background, critical life events, early adversities, age and context of symptom onset, and their lasting impact; preserve information gaps.
Core Beliefs, Schemas, and Interpretations	Captures enduring beliefs, schemas, and cognitive frameworks about the self, others, and the world, including habitual appraisals of safety, worth, responsibility, emotion, and control.	Summarize core beliefs, cognitive themes, interpretation patterns, and affected domains, referring to case-supported language or examples where possible.
Personality Traits and Communication Style	Characterizes habitual traits, interpersonal style, and general communication or disclosure patterns, including openness, guardedness, and self-awareness.	Describe enduring personality features, baseline emotion patterns, communication style, openness or reticence, and context-specific strengths or limitations.
Diagnostic and Comorbidity Profile	Lists lifetime and current diagnoses, symptom course and severity, psychiatric or medical comorbidities, substance use, trauma, and histories of chronicity or remission.	List current and past diagnoses, comorbidities, symptom courses, functional impact, and relevant medical or psychiatric factors.
Trauma and Index Events	Captures traumatic exposures, their type, timing, severity, perceived meaning, and personal impact, including any index event relevant to current difficulties.	Describe trauma history, nature and context of exposures, the patient's interpretation and willingness to disclose, and any known index event.
Habitual Coping Strategies and Safety Behaviors	Describes routinely used responses to stress or distress, such as avoidance, reassurance seeking, ritualizing, suppression, withdrawal, substance use, checking, perfectionism, or aggression.	List habitual response patterns, their rationales, contextual exceptions, perceived effectiveness, and the flexibility or rigidity of coping.

Static dimension	Description	Value guidance
Strengths, Support Resources, and Constraints	Summarizes enduring strengths, available social and practical supports, and chronic financial, cognitive, medical, logistical, occupational, or technological barriers.	Outline psychological and practical strengths, support systems, coping resources, and longstanding constraints affecting participation or progress.
Treatment, Therapy, and Medication History	Captures prior therapy, medication use, engagement and adherence, outcomes, attitudes toward treatment, and preferences or concerns about different approaches.	List previous therapies and medication trials, outcomes, adherence issues, side effects, attitudes, preferences, skepticism, or dropout history.
Personal Goals, Values, and Motivation	Describes enduring goals, values, motivations for therapy or change, approach or avoidance orientation, and baseline readiness for intervention or exposure.	Articulate therapy goals, valued life domains, motivation or ambivalence, engagement stance, and contextual priorities or conflicts.
Contextual Triggers and Avoided Situations	Identifies symptom-triggering or avoided situations and their associated meanings, supporting exposure planning and interpretation of episodes.	List triggers and avoided situations, their significance, and circumstances that amplify or reduce their impact.
Cognitive and Learning Capacity	Describes intellectual functioning, communication, learning style, memory, and capacity to engage with psychoeducation or new tasks.	Describe cognitive and learning strengths and limitations, communication skills, and any needed adaptations to therapy materials or interventions.

Dynamic dimension	Description and update condition	Mode	Options
Current Emotional State/Affect	Captures momentary or sessional emotion and observable or reported affect. Add or remove emotions using verbal or nonverbal cues, self-report, session observations, and explicit statements; use unknown when not inferable.	multiple	anxiety/fear; sadness/low mood; anger/irritability; shame/guilt; hope/optimism; numbness; content/calm; unknown
Symptom Severity/Distress Level	Tracks current severity of anxiety, depression, panic, OCD, or distress. Switch when self-report, measures, behavioral observation, or explicit evidence indicates change; otherwise retain the state or use unknown.	single	none/minimal; mild; moderate; severe; panic_attack_present; unknown
Profile of Current Symptoms/Diagnosis	Captures whether presenting symptoms are improving, worsening, unchanged, or fluctuating. Update only with direct evidence of a change; retain the current state when stable and use unknown when not assessed.	single	improving; worsening; unchanged; fluctuating/variable; unknown
Motivation and Readiness for Change/Treatment/Ex	Represents the current stance toward participation, therapeutic tasks, or exposure assignments. Change with evidence of increased or decreased willingness or resistance; otherwise retain the state.	single	high/committed; ambivalent/mixed; resistant/unwilling; disengaged/passive; unknown
Session Engagement and Participation	Reflects the degree and quality of involvement in therapeutic activities, discussion, and tasks. Add or remove options based on observed or reported participation, exercise engagement, and feedback.	multiple	fully engaged/active; partially engaged; passive; avoidant/distracted; resistant/oppositional; supportive to others; unknown
Current Coping and Avoidance Strategies	Tracks coping methods and avoidance behaviors currently used in session or the recent context. Add or remove options using explicit reports, observation, or worksheet review.	multiple	emotional avoidance; behavioral avoidance; cognitive avoidance; active coping/skills use; maladaptive coping; reassurance seeking; checking/ritualizing; safety behaviors; exposure/approach; none; unknown

Dynamic dimension	Description and update condition	Mode	Options
Cognitive Flexibility, Automatic Thoughts, and Reappraisal	Captures cognitive rigidity, reappraisal, negative automatic thoughts, schema activation, catastrophic thinking, and acceptance. Modify selections when the patient reports or demonstrates a change in cognitive approach.	multiple	rigid/automatic negative thoughts; active cognitive reappraisal; thought suppression; acceptance/open stance; catastrophizing; overestimation of risk; intolerance of uncertainty; realistic appraisal; schema activation; unknown
Attention and Symptom Monitoring	Describes current internal, external, balanced, distracted, or symptom-scanning attention. Update from observed attentional shifts, self-report, or session cues.	multiple	internally focused; externally focused; balanced/mixed; symptom scanning; distracted/disengaged; low vigilance; unknown
Awareness, Insight, and Self-monitoring	Reflects the current capacity to recognize and describe symptoms, beliefs, or maladaptive patterns. Switch using evidence from self-monitoring, recognition of errors, or behavioral tracking.	single	absent/minimal insight; partial/intermittent insight; developing/advanced insight; accurate_balanced_reporting; vague_overgeneralization; unknown
Willingness and Openness to Disclose/Engage	Represents current openness or reluctance to share emotional experiences, history, or symptoms. Select from observed disclosure, avoidance, and willingness to discuss sensitive topics.	single	fully open; partially/selectively open; guarded/reluctant; withholding/concealing; avoidant/not disclosing; unknown
Therapeutic Alliance and Support Network Involvement	Captures the current alliance and the involvement of family or other supports. Switch when rapport, feedback, support involvement, or reported trust changes.	single	strong/supportive; tenuous/mixed; rupture/conflictual; repair/reengaging; withdrawing; interfering; neutral; unknown
Adherence and Response to Assignments/Tasks	Tracks compliance, avoidance, minimization, or adaptation in response to therapeutic assignments. Update from task review, patient explanation, or observation; otherwise retain the current state.	single	strict/compliant; partial/minimizing; active modification/problem-solving; avoidant/nonadherent; unknown
Generalization and Maintenance of Therapy Gains	Indicates whether gains remain limited to therapy, generalize outside it, persist, or are lost. Switch when self-report, observation, or feedback provides evidence of change.	single	maintenance across contexts; gains limited to therapy; partial/generalizing; relapse/loss of gains; unknown
Medication Use and Side Effects	Captures current medication status and its effects on the session or functioning. Add or remove states as medication use changes according to patient or clinician report or observed effects.	multiple	starting_medication; maintaining; reducing_tapering; withdrawal_symptoms; side_effects; no_medication; unknown