From Profiling to Synthesis: Benchmarking Implicit Behavioral Alignment in Personalized LLM Agents

Anonymous ACL submission

Abstract

Large Language Models have advanced autonomous agents, but personalization remains essential for agents to be practically useful. To measure this ability, recent benchmarks aim to evaluate personalization in agents. However, they either provide static preference snapshots or fixed interaction logs, or they evaluate personalization mainly through question answering over retrieved profiles. These designs underrepresent the complexity of real preferences in dialogue histories and fail to assess preference-conditioned task execution, thereby obscuring a critical knowing-doing gap. To address this, we introduce PERSONAKAG, a benchmark for implicit behavioral alignment built from longitudinal interaction histories that contain noise, implicit cues, and temporal inconsistencies. PERSONAKAG evaluates whether an agent can execute tasks while satisfying implicit constraints inferred from history, rather than only answering preference questions. We further propose SynRPG, a framework that combines broad retrieval with trajectory-level alignment to resolve conflicting priorities over time. Results on PERSONAKAG suggest that effective personalization is still challenging for state-of-the-art LLM agents. Our code and data are anonymously available at https://anonymous.4open.science/r/PersonaAgent-2F30.

1 Introduction

With the exponential growth in Large Language Model (LLM) (OpenAI, 2023; Gemini, 2025; Guo et al., 2025) capabilities, the operational scope of agents is expanding from generic question answering (Yang et al., 2018) to complex task execution (Yang et al., 2024). Within this trajectory, endowing agents with personalization has emerged as a central research imperative (Samuel et al., 2024). An ideal personalized agent should not function merely as a standardized instruction executor; it must act as a digital companion capable of deeply comprehending individual nuances by tailoring services based on historical behaviors, preferences, and current states (Zhang et al., 2025). This necessitates a model capable of transcending general knowledge to master specific contextual understanding, demonstrating adaptive differentiation across diverse user interactions (Zhao et al., 2025).

Prevalent personalization paradigms (Zhao et al., 2025), however, often reduce user intent to static preference tags (e.g., "likes spicy food"), turning personalization into a simple lookup problem. This approach fails when behaviors evolve or conflict (Figure 1). Consider a user who typically prefers spicy meals but recently mentioned a wisdom-tooth removal. When asked to order dinner, a profiling agent blindly retrieves the "spicy" tag, yielding a harmful recommendation. In contrast, true personalization requires synthesis: the ability to infer the implicit short-term constraint (soft, non-irritating food) and override the long-term preference to generate a contextually safe response.

Despite these needs, existing benchmarks exhibit a fragmented landscape that struggles to capture this paradigm shift from profiling to synthesis. Due to the scarcity of authentic longitudinal data, synthetic benchmarks dominate but often fail to reconcile data fidelity with evaluation depth. Early works like LAPS (Joko et al., 2024) treated personalization merely as a secondary soft constraint for fluency rather than a core priority for decision making. Subsequently, HiCUPID (Mok et al., 2025) introduced raw dialogue histories to enhance realism yet remained fundamentally tethered to extraction-based Question Answering (QA). While PersonaBench (Tan et al., 2025) effectively captures attribute evolution, its evaluation paradigm is also restricted to simple QA. This reliance on QA exposes a critical knowing-applying gap. A model may correctly answer that a user is busy (the knowing phase) but fail to map this state to the execution of a brief response (the applying phase). This gap

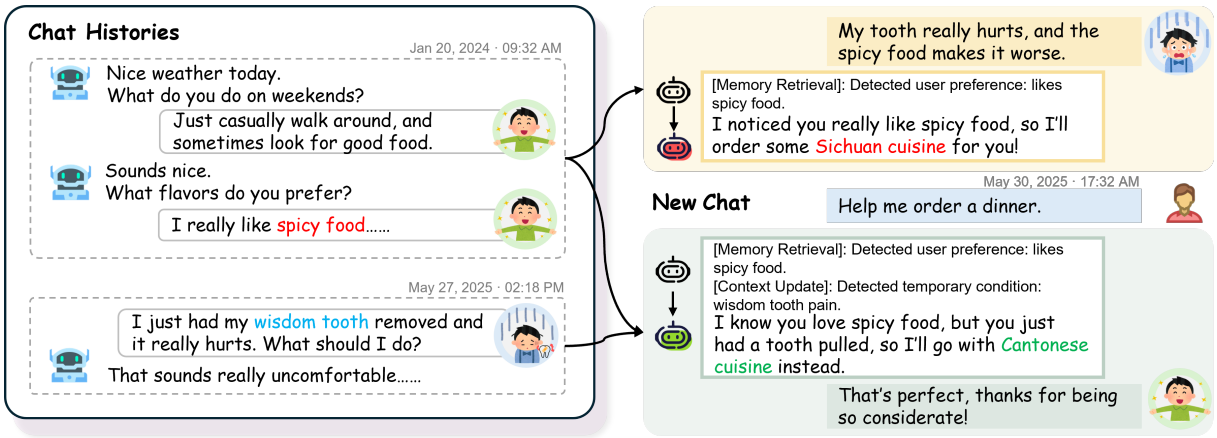


Figure 1: Failure of static preference profiling under contextual conflict: profiling agent vs. implicit alignment agent.

exists because QA tasks only require declarative retrieval of information, whereas task execution requires the procedural integration of that information into the generation process. Although PersonaLens (Zhao et al., 2025) attempts task-oriented evaluation, it compromises complexity by feeding highly summarized abstracts rather than raw histories. These benchmarks fundamentally lack the capacity to assess the synthesis ability of an agent within complex and dynamic environments.

To address these limitations, we propose **PERSONAKAG**, the first benchmark designed to evaluate implicit behavioral alignment. The core philosophy of PERSONAKAG is "alignment through synthesis". Regarding data, we construct longitudinal dialogue histories containing implicit preference cues, dynamic attribute evolution, and conflicts between long-term and short-term contexts. Regarding the task format, we abandon explicit QA and simple matching. Instead, the agent is provided with the full interaction history and a generic current user request. The agent must then complete a generative task execution, such as drafting a stylized document or resolving a scheduling conflict. In these scenarios, the constraints and ground truths are never explicitly stated in the prompt; the agent must achieve implicit alignment with user intent through synthesis by actively retrieving history, identifying conflicts, and reasoning about priorities. To quantify performance, we evaluate the generated responses against pre-defined dimensions, verifying their compliance with the constraints in each specific aspect.

To navigate the challenges posed by PERSONAKAG, we introduce **SynRPG**, a novel framework that bridges the gap between information retrieval and task execution. SynRPG achieves alignment through synthesis via a two-stage architecture: (i) Deep Retrieval and (ii) Broad Thinking & Deep Alignment. Extensive experiments on PERSONAKAG reveal that even state-of-the-art LLMs (OpenAI, 2023; Guo et al., 2025) significantly struggle on synthesis-heavy tasks compared to simple QA, validating current technical shortcomings in unifying knowledge and action. Conversely, SynRPG significantly enhances alignment in complex scenarios through these explicit synthesis mechanisms. In summary, our contributions are as follows:

- We systematically argue for a transition in personalized agent research from a static profiling paradigm to a dynamic synthesis paradigm and define the Knowing-Applying Gap as the central obstacle where models fail to translate retrieved knowledge into generative constraints.
- We release PERSONAKAG, a benchmark emphasizing implicit behavioral alignment through real task execution within dynamic and conflicting contexts, filling a critical void in existing evaluations.
- We propose and validate a novel framework, SynRPG, which provides an empirical direction for building next-generation agents with high-order synthesis capabilities.

2 Related Work

2.1 LLM-Based Agents

With the rapid advancement of LLMs' reasoning (Wei et al., 2022), planning (Valmeekam et al.,

Domain	Scale Stats				# Evid.					# Pref.					Dyn.	
	S	D	P	C	1	2	3	4	5+	1	2	3	4	5+	Dyn.	Stat.
Writing	16	54	10	750	271	267	83	61	68	66	199	282	160	43	445	1,730
Work	12	53	17	1,155	394	380	165	111	105	30	161	338	321	305	956	3,346
Daily Consumption	8	39	15	1,086	511	290	217	36	32	46	182	401	276	181	176	3,909
Planning	10	39	16	1,151	532	329	182	42	66	41	206	489	386	29	291	3,318
Exercise and Health	4	16	20	781	325	247	113	41	55	11	54	173	252	291	269	2,921
Transportation	4	14	14	518	288	117	112	0	1	21	67	169	197	64	0	1,770
Medical Services	3	11	15	524	290	113	121	0	0	18	90	231	185	0	0	1,631
Leisure Activities	3	9	14	333	137	104	62	16	14	12	60	148	113	0	89	939
Information Management	6	26	17	664	293	217	94	29	31	23	73	185	161	222	196	2,650
Overall	**66**	**261**	**138**	**6,962**	**3,041**	**2,064**	**1,149**	**336**	**372**	**268**	**1,124**	**2,416**	**2,051**	**1,135**	**2,422**	**22,214**

Table 1: Overall statistics of domains in the benchmark. Abbreviations: **S**: Number of Scenarios; **D**: Number of Dimensions; **P**: Average Number of Preference Types; **C**: Number of Task Combinations; Columns under **# Evidence** and **# Preference** represent counts of scenarios; **Dyn.** and **Stat.** represent counts of dynamic and static scenarios.

Benchmark	History	Dynamic	Implicit	Task	Behavior
HiCUPID	✓	✗	✓	✗	✗
PersonaBench	✓	✓	✓	✗	✗
PersonaLens	✗	✗	✓	✓	✗
PERSONAKAG (Ours)	✓	✓	✓	✓	✓

Table 2: Comparison of Persona-related Benchmarks. Our proposed PERSONAKAG is the only benchmark that covers all dimensions.

2023), and tool use (Schick et al., 2023), LLM-based personalized agents have emerged as an active research area. Compared to generic agents, personalized agents seek to operate effectively in dynamic and diverse interaction scenarios by aligning decision-making with user-specific characteristics (Wang et al., 2024).

Current approaches typically introduce personalization through explicit user modeling (Zhang et al., 2025; Afzoon et al., 2024), memory or retrieval-augmented mechanisms(Ram et al., 2023), and adaptation based on user feedback(Stiennon et al., 2022) , with the goal of capturing both long-term user preferences and evolving interaction contexts (Westhäußer et al., 2025; Park et al., 2023). However, existing research remains largely method-centric, with primary emphasis placed on personalization mechanisms and model design rather than on systematic evaluation of personalization capabilities (Mohammadi et al., 2025; Hao et al., 2025).

2.2 Benchmarks for Personalized Agents

Benchmarks for LLM-based agents (Liu et al., 2025) are largely derived from general-purpose evaluation settings, including QA, tool use, and short-horizon planning tasks (Hu and Shu, 2023). These benchmarks typically measure task success or accuracy under static assumptions and task-level metrics (Zhu et al., 2025) . While effective for assessing general agent capabilities, they offer limited support for evaluating personalized behaviors (Hao et al., 2025).

Taken together, existing personalization benchmarks still largely operationalize personalization as profiling, either by extracting user attributes via QA or by aligning responses under explicit and often static preference specifications (Mok et al., 2025; Tan et al., 2025; Zhao et al., 2025). These settings under-specify the core challenge faced by real-world agents: synthesis, namely, procedurally integrating implicit, potentially conflicting, and time-varying user signals into task execution over raw longitudinal interaction histories. Motivated by these limitations, we introduce PERSONAKAG. A comparison with prior benchmarks is summarized in Table 2.

3 Benchmark Construction

To bridge the gap between static profiling and dynamic synthesis, we construct PERSONAKAG through a multi-stage pipeline designed to simulate the noise, implicitness, and conflicts inherent in real-world personalization. As shown in Fig. 2, the construction process consists of four phases:

3.1 Phase 1: Construct Personas

We define 400 seed personas as the basic units for modeling user-specific variation. Each persona is explicitly multi-dimensional and formalized as $p = (B, T, E, R)$, where B denotes background attributes, T personality traits, E transient states (e.g., stress or urgency), and R the current professional or social role. This formulation allows a single persona to capture personal traits, situational states, and professional or social roles, enabling coherent yet diverse behaviors across scenarios.

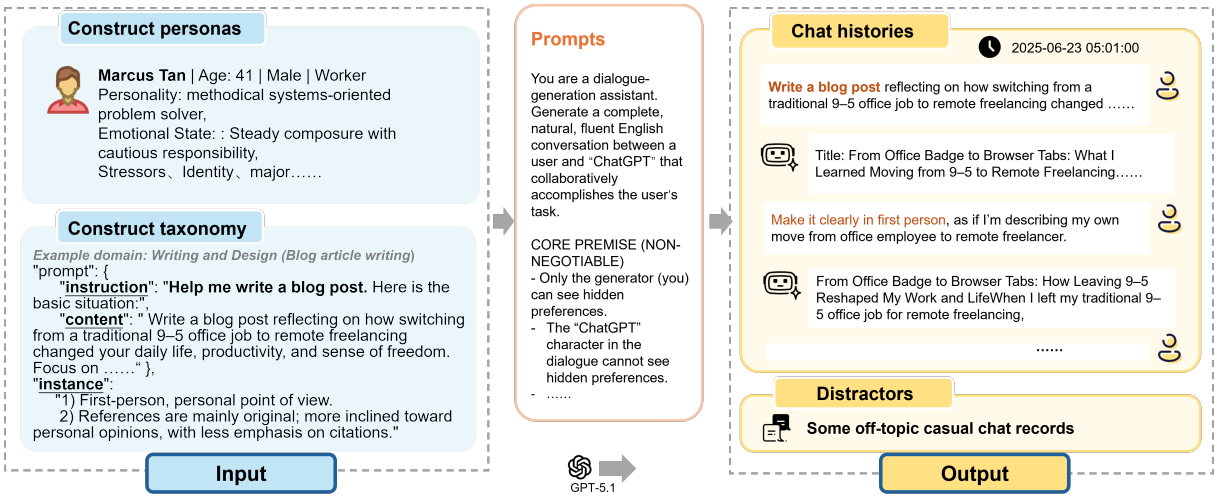


Figure 2: Benchmark construction flow chart.

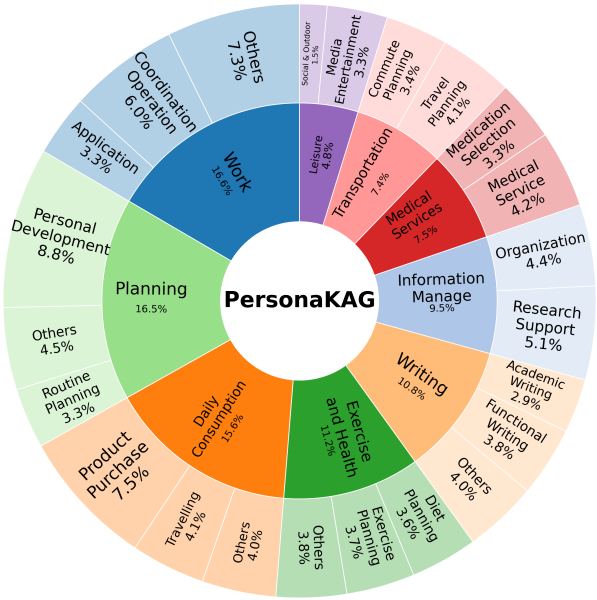


Figure 3: Hierarchical distribution of domains and scenarios in the benchmark.

3.2 Phase 2: Construct Taxonomy

We construct a hierarchical task taxonomy to define the behavioral space for personalization evaluation. It spans nine high-level domains and 66 scenarios, each assigned to a single domain. Table 1 summarizes the cross-domain diversity of PERSONAKAG, while Figure 3 illustrates the domain and scenario distributions (see Appendix B.1 for details). Personas from the previous phase are then paired with task scenarios to instantiate scenario-specific preference configurations.

A central design goal of PERSONAKAG is to move from static profile attributes to dynamic behavioral constraints. We operationalize this transition through two increasing levels of difficulty:

Implicit Transformation: We transform each structured preference configuration into a natural, paragraph-level description that implicitly conveys the persona's tendencies without stating preference options verbatim. This description serves as metadata to condition the LLM during history synthesis, enabling preference expression to emerge through interaction behavior rather than explicit declarations.

Preferred Combination: To model the non-stationarity of human preferences, persona attributes are assigned time-dependent validity. Long-term attributes (e.g., educational background or habitual strategies) evolve slowly, whereas short-term preferences and states (e.g., dietary tastes or stress levels) may change rapidly and even contradict earlier signals. Interaction histories are organized along a continuous timeline anchored to a reference date, requiring agents to reason over preference updates rather than treating past evidence as static or equally valid.

3.3 Phase 3: Longitudinal History Synthesis

To mimic realistic interaction logs, we use a multi-agent framework to generate longitudinal dialogue histories with a severe imbalance between informative signals and irrelevant content, creating a challenging needle-in-a-haystack setting.

Evidence Interweaving: Key preference signals are sparsely and implicitly distributed across multiple interaction sessions rather than expressed as explicit statements. Ground-truth preferences condition only the user agent, while the assistant must infer them from accumulated behavioral patterns, such that no single interaction reveals sufficient evidence.

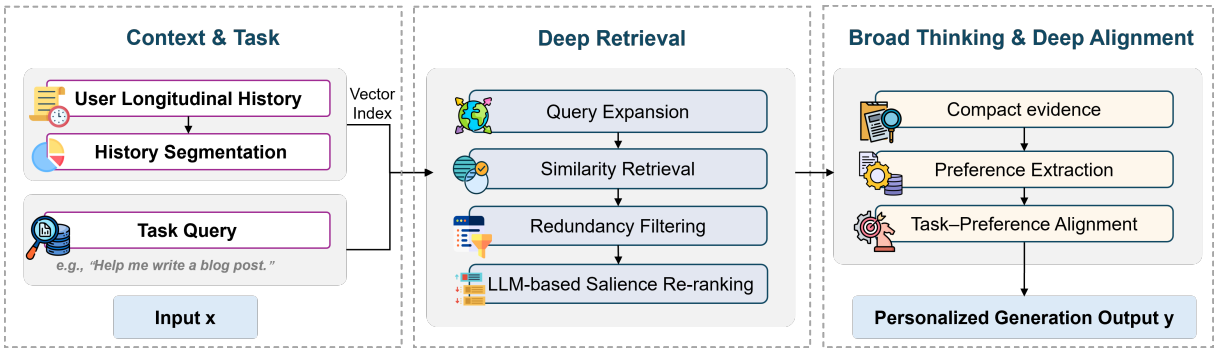


Figure 4: Overview of SynRPG for personalized task completion.

Noise and Distractors: To increase realism, we inject substantial task-irrelevant chit-chat into dialogue histories, forcing agents to filter distractions. Task-relevant turns occur only once per 100 turns on average, making preference evidence sparse and easily buried.

3.4 Phase 4: Task Instantiation and Annotation

In the final stage, we instantiate evaluation tasks that require agents to execute concrete tasks using long-term histories, given only minimal prompts. Task content is newly authored (though domain-aligned with the histories) to prevent trivial reuse or paraphrasing. Outputs take two forms: (1) textual artifacts (e.g., reports or schedules) and (2) procedural instructions (e.g., step-by-step GUI actions for ordering food).

For quality control, we randomly sample instances for human-in-the-loop review of output plausibility and naturalness.

3.5 Data Statistics and Insights

As shown in Table 1, the benchmark comprises 6,962 instances across 66 scenarios and 261 preference dimensions. The benchmark is generated using GPT-5.1 under a manually defined taxonomy. The data distribution is notably skewed toward non-trivial cases. Specifically, 1,857 instances involve three or more evidence pieces, including 372 with five or more evidence pieces. Preference configurations are also dense, with 5,602 instances containing at least three preferences and 1,135 containing five or more. Additionally, 2,422 instances are annotated as dynamic. Together, these statistics show that a substantial portion of the benchmark concentrates on instances with high structural and contextual complexity.

4 Method

In this section, we formalize the preference-conditioned personalization task and its behavioral evaluation protocol, and then present SynRPG, an LLM-driven agent for personalized task execution.

4.1 Task Formulation and Evaluation

Given the full history $\mathcal{H}_u$ and a new task query x from scenario s (e.g., "Draft a blog post for me"), the agent produces an output y, which is evaluated at the behavioral level following Section 3. The correctness of y is assessed at the behavioral level following the protocol in Section 3.

4.2 LLM-Driven Personalized Agent

As illustrated in Figure 4, SynRPG is an LLM-driven personalization agent that maps a user's longitudinal interaction history $\mathcal{H}_u$ and a new task query x to an output y through a retrieval-augmented pipeline. The agent consists of two components: (1) Deep Retrieval to construct a compact, task-conditioned evidence set $\mathcal{C}_u(x)$ from long-term histories; and (2) Broad Thinking & Deep Alignment to explicitly bind the retrieved evidence to the final response.

Deep Retrieval: We segment $\mathcal{H}_u$ into short passages and index them with bge-m3 embeddings. Given x, SynRPG generates a small set of task-aware sub-queries (e.g., style and formatting cues), retrieves candidate passages, then deduplicates them and uses an LLM to re-rank them by behavioral relevance. The top passages form $\mathcal{C}_u(x)$.

Broad Thinking & Deep Alignment: Given x and $\mathcal{C}_u(x)$, the generator follows a three-stage routine: (i) Preference Extraction, synthesizing preferences/constraints from evidence with recency and overrides; (ii) Task-Preference Alignment, translating them into task-level constraints and generation

Model	Writing	Work	Daily	Plan.	Health	Trans.	Med.	Leis.	Info.	Overall
Qwen2.5-7B-Instruct (Qwen et al., 2025)	42.7	42.0	73.8	45.8	57.9	82.3	79.7	46.6	32.4	53.1
DeepSeek-V3 (DeepSeek-AI et al., 2025a)	48.2	43.2	70.7	50.3	61.2	82.2	80.8	53.5	40.1	57.1
ChatGLM-4-9b-chat (GLM et al., 2024)	42.1	39.0	74.7	43.8	56.5	73.4	78.9	47.1	31.8	52.6
DeepSeek-V3.2 (DeepSeek-AI et al., 2025b)	54.6	47.2	71.6	52.5	64.4	81.4	82.2	48.4	40.5	59.2
QwQ-32B (Qwen Team, 2025)	54.5	50.0	76.8	54.1	67.1	86.4	84.5	**62.3**	43.1	62.5
GPT-4o-mini	47.1	44.2	77.0	47.1	59.7	89.0	88.6	53.3	34.2	58.0
GPT-4o (OpenAI, 2024)	55.1	42.0	75.6	45.2	60.8	86.6	85.1	54.2	33.9	57.7
GPT-5-mini	57.3	52.4	67.6	58.6	71.1	74.1	76.1	50.6	48.1	61.3
GPT-5.1 (OpenAI, 2025)	58.4	49.0	69.9	54.2	66.5	80.6	77.7	42.4	43.4	59.8
SynRPG	**72.2**	**64.8**	**79.1**	**69.3**	**82.5**	**89.7**	**89.5**	57.3	**57.0**	**73.2**

Table 3: Performance comparison using **bge-m3** retriever. Data columns are presented with fixed widths for better readability, and the column headers correspond to nine application domains: **Writing** (writing and design), **Work**, **Daily** (daily consumption), **Plan.** (planning), **Health** (exercise and health), **Trans.** (transportation), **Med.** (medical services), **Leis.** (leisure activities), and **Info.** (information management).

Method Variants	Writing	Work	Daily	Plan.	Health	Trans.	Medical	Leisure	Info.	Overall
SynRPG	**72.6**	**65.8**	**79.4**	**71.1**	**85.7**	**93.2**	**94.2**	**60.0**	**60.4**	**75.1**
w/o Deep Retrieval	71.2(-1.4)	54.6(-11.2)	71.7(-7.7)	63.5(-7.6)	71.8(-13.9)	87.8(-5.4)	82.6(-11.6)	63.5(+3.5)	50.8(-9.6)	67.1(-8.0)
w/o BT+DA	63.2(-9.4)	55.4(-10.4)	80.1(+0.7)	62.9(-8.2)	75.7(-10.0)	86.4(-6.8)	90.0(-4.2)	61.4(+1.4)	49.0(-11.4)	68.2(-6.9)
w/o Synth. (RAG)	55.6(-17.0)	46.8(-19.0)	76.9(-2.5)	54.0(-17.1)	60.2(-25.5)	86.3(-6.9)	82.2(-12.0)	56.2(-3.8)	40.4(-20.0)	60.5(-14.6)

Table 4: Ablation study of our framework across different task domains. All variants use DeepSeek-V3.2 as the backbone LLM and bge-m3 as the retriever. Data columns are fixed-width for readability, with column headers defined as in Table 3. *w/o* denotes removing a specific component. **BT+DA** denotes *Broad Thinking + Deep Alignment*, and **Synth. (RAG)** denotes the original RAG-style synthesis strategy without trajectory-level synthesis.

decisions (internal planning); and (iii) Personalized Generation, producing the final output.

5 Experiment

5.1 Settings

All experiments are conducted in an inference-only setting, with models evaluated without parameter fine-tuning or access to external tools.

5.1.1 Baselines

We evaluate personalized task completion under a unified retrieval-augmented generation (RAG) setting across a diverse set of LLMs, including Qwen2.5-7B-Instruct (Qwen et al., 2025), ChatGLM-4-9B-Chat (GLM et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025a), DeepSeek-V3.2 (DeepSeek-AI et al., 2025b), QwQ-32B (Qwen Team, 2025), GPT-4o-mini, GPT-4o (OpenAI, 2024), GPT-5-mini, and GPT-5.1 (OpenAI, 2025). All models are accessed through official or widely used inference platforms (DeepSeek API, SiliconFlow, and OpenAI API). Baselines use a standard RAG pipeline: bge-m3[1] retrieves sentence-level chunks, which are concatenated with the task query as LLM input.

5.1.2 Metric

For each scenario, we define a small set of behavioral checkpoints, each corresponding to an atomic seed preference of the user in that scenario. For generation-based tasks, given an output y_i for instance i, the LLM-based agent is evaluated using DeepSeek-V3.2 (DeepSeek-AI et al., 2025b) by whether it satisfies each checkpoint, yielding a binary score $f_i^k(y_i) \in \{0, 1\}$. The instance-level score is computed as the average satisfaction across checkpoints. For selection-based tasks, evaluation reduces to a single criterion: whether the agent selects the preference-consistent option, yielding 1 for a correct selection and 0 otherwise.

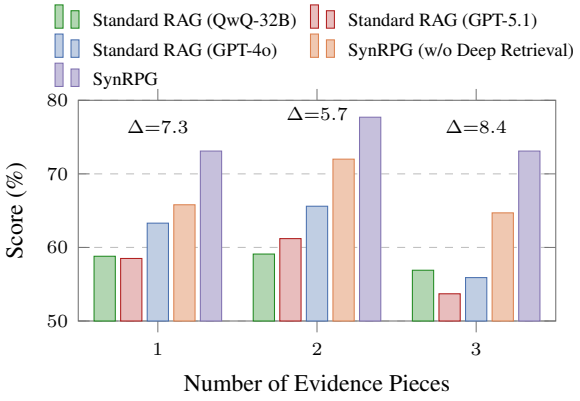
5.1.3 Parameters

We evaluate all models with DeepSeek-V3.2 (max 3,000 tokens) and use BGE-M3 for dense retrieval with 256-token chunks and 50-token overlap. Standard retrieval returns the top-3 passages per query, while Broad Retrieval issues up to five expanded queries and retains at most ten passages after filtering and optional re-ranking. We fix the random seed to 42 for reproducibility.

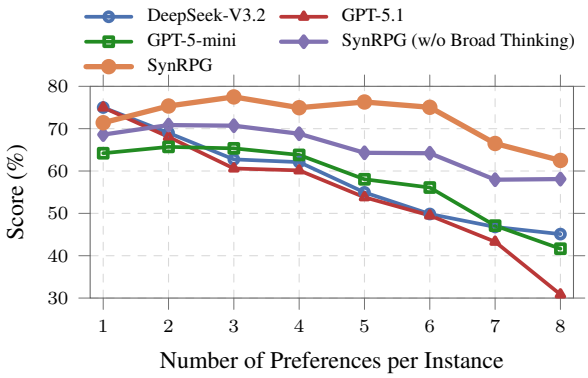
5.2 Main Results

Table 3 reports model performance on PERSONAKAG under a unified retrieval setting. Under standard RAG, even strong LLMs reach only moderate scores (52.6–62.5) with limited cross-domain robustness. SynRPG consistently outperforms all baselines, surpassing the strongest baseline by 10.7 points, with the largest gains in Writing, Planning,

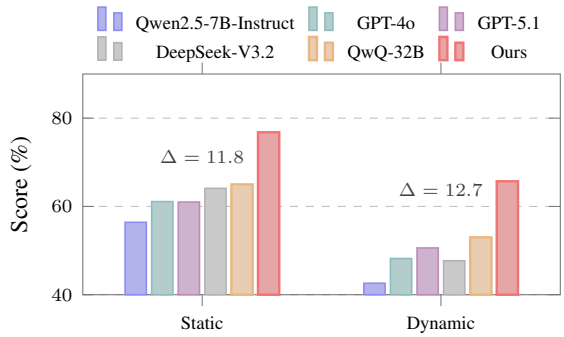
[1]BGE-M3: https://huggingface.co/BAAI/bge-m3



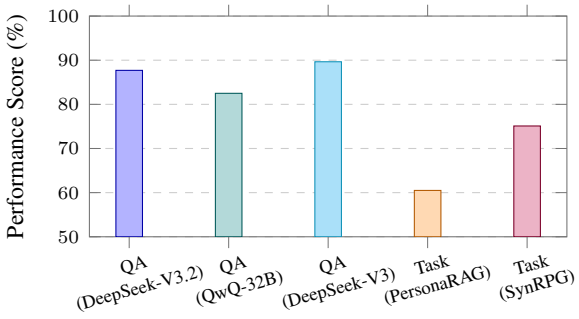
(a) Impact of evidence quantity.



(b) Impact of the number of preferences.



(c) Impact of preference attributes.



(d) QA vs. Task-Oriented Application.

Figure 5: Ablation analyses on evidence quantity, preference complexity, preference attributes, and the knowing–doing gap. Δ denotes the performance gap between SynRPG and the second-best model.

and Health domains that are especially sensitive to user-specific constraints—indicating more effective use of long-term history during task execution.

5.3 Ablation Study

As shown in Table 4, each component of SynRPG contributes substantially to overall performance. We conduct stratified sampling over PERSONAKAG to construct a test set of 1,000 instances. Removing Deep Retrieval leads to the largest degradation, highlighting the necessity of long-context evidence mining for capturing sparse and implicit preferences. Excluding Broad Thinking & Deep Alignment also causes a notable drop, indicating the importance of explicit behavioral planning in translating retrieved signals into generation constraints.

5.4 In-depth Analysis

Figure 5a reports preference-conditioned task performance under different evidence budgets. SynRPG consistently outperforms the standard RAG baseline and the variant without Deep Retrieval, with the largest gains at three evidence pieces, underscoring its ability to aggregate multiple preference-relevant signals and the value of broad, deep retrieval. Across models, performance is non-monotonic: increasing evidence from one to two helps, but expanding to three often hurts, suggesting a trade-off between recovering missing preferences and accumulating noise.

5.4.1 Impact of the Number of Preferences

Figure 5b shows instance-level performance versus the number of preferences (smoothed with a sliding window, w=2). Performance drops for all models as constraints accumulate. SynRPG degrades more slowly and widens its lead at higher preference counts. Removing Broad Thinking & Deep Alignment yields a clear drop, confirming its role in robust multi-constraint alignment.

5.4.2 Impact of Preference Attributes

Figure 5c compares static vs. dynamic preference compositions. All models perform better on static preferences, reflecting the easier alignment setting. SynRPG gains more in the dynamic case, outperforming the strongest baseline by +12.7 points (vs. +11.8 static), indicating stronger robustness to evolving preferences and preference shifts.

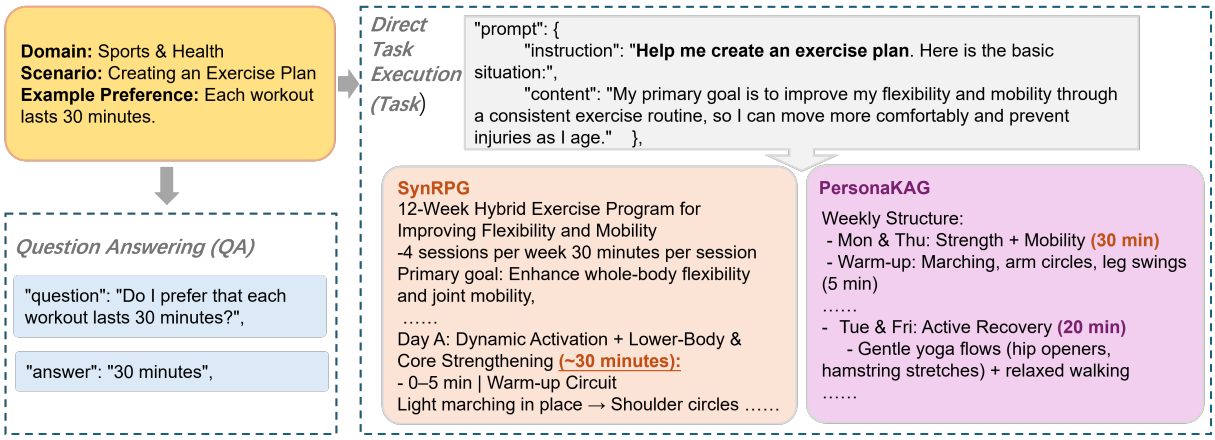


Figure 6: QA-based vs. task-oriented execution: illustrating the knowing–doing gap. All comparisons are conducted using DeepSeek-V3.2 to control for model capability.

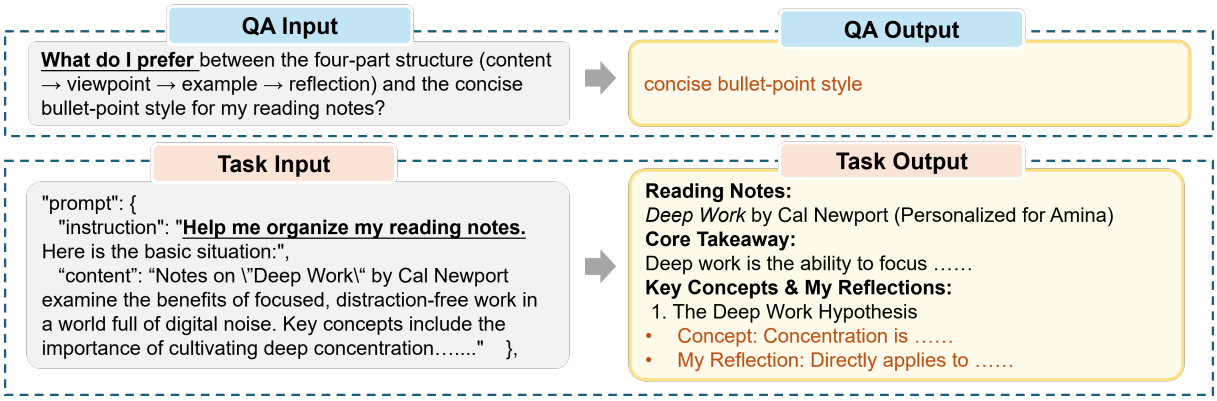


Figure 7: QA vs. SynRPG for preference modeling. All comparisons are conducted using DeepSeek-V3.2 to control for model capability.

5.4.3 QA vs. Task-oriented Application

Figure 5d compares QA with direct task execution. For all models, QA consistently exceeds Task, exposing a clear knowing–doing gap: models can state preferences correctly but often fail to enforce them during execution, showing that QA competence does not translate into robust preference-aware task completion.

5.4.4 Case Study of the Knowing-Doing Gap

As shown in Figure 6, we compare QA-based, RAG-based, and SynRPG agents to further illustrate the knowing–doing gap. QA elicits and explains preferences explicitly, whereas SynRPG grounds task outputs in retrieved preference evidence (Figure 7). While the QA agent identifies the preference correctly, it fails to reliably enforce it during execution. SynRPG instead binds preference constraints to task decomposition and plan generation, aligning with the quantitative gap reduction in Figure 5d.

6 Conclusion

In this paper, we identified a critical knowing-doing gap in existing personalized agent evaluations, which rely heavily on static user profiles and explicit question answering. We argued that true personalization requires agents to synthesize implicit signals from longitudinal histories rather than merely retrieving explicit tags. To address this, we introduced PERSONAKAG, a comprehensive benchmark that evaluates implicit behavioral alignment under dynamic and noisy conditions. We further proposed SynRPG, a framework that bridges information retrieval and task execution through deep retrieval and planning-guided generation. Extensive experiments demonstrate that while current state-of-the-art models struggle to translate historical context into behavioral constraints, our approach significantly improves performance in these complex scenarios. Ultimately, mastering this capability is a prerequisite for the reliable deployment of autonomous agents in real-world applications where user needs are often implicit and constantly evolving.

7 Limitations

We propose PERSONAKAG as a benchmark for evaluating implicit behavioral alignment in personalized agents; however, it still has two main limitations. First, PERSONAKAG is constructed using LLM-generated synthetic data. Although the synthesis pipeline is carefully designed to inject implicit preferences, temporal dynamics, and noise, hallucinated or inconsistent behaviors may still occur. Second, our evaluation focuses on offline, history-conditioned task execution and does not involve interactive feedback or tool use. As a result, the benchmark cannot capture how agents adapt preferences through online interaction or corrective signals. Extending PERSONAKAG to interactive and tool-augmented settings is left for future work.

8 Ethical Considerations

Considering potential risks related to privacy and intellectual property, we deliberately avoid using real user logs or information collected from human subjects. All personas, interaction histories, and tasks are fictional, thereby preventing direct exposure of personally identifiable information. At the same time, as LLM-generated content may still reflect biased or undesirable patterns, we address these issues through carefully designed prompts and targeted human review.

References

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. 2024. Persobench: Benchmarking personalized response generation in large language models. *Preprint*, arXiv:2410.03198.

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, and 181 others. 2025a. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Gemini. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. In *CoRR*, volume abs/2507.06261.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 3 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Evaluating personalized tool-augmented llms from the perspectives of personalization and proactivity. *Preprint*, arXiv:2503.00771.

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *Preprint*, arXiv:2312.05230.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of SIGIR*, pages 796–806.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA. Association for Computing Machinery.

J. Mok, Ik hwan Kim, Sangkwon Park, and Sungroh Yoon. 2025. Exploring the potential of llms as personalized assistants: Dataset, evaluation, and analysis. In *Annual Meeting of the Association for Computational Linguistics*.

OpenAI. 2023. Gpt-4 technical report. In *CoRR*, volume abs/2303.08774.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Official announcement of GPT-4o model; accessed: 2026-01-04.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational upgrade in the gpt-5 series. https://openai.com/index/gpt-5-1/. Official announcement of GPT-5.1 model series; accessed: 2026-01-04.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/. Official model introduction blog. Accessed: 2026-01-04.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, A. Kalyan, Tanmay Rajpurohit, A. Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. In *Conference on Empirical Methods in Natural Language Processing*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R. N., Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. In *Proceedings of ACL*, pages 878–893.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Ai persona: Towards life-long personalization of llms. *Preprint*, arXiv:2412.13103.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rebecca Westhäußer, Wolfgang Minker, and Sebatian Zepf. 2025. Enabling personalized long-term interactions in llm-based agents through persistent memory and user profiles. *Preprint*, arXiv:2510.07925.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Adriano Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, R. Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, Xiaoman Pan, Lian Xiong, Jingguo Liu, Philip S. Yu, and Xian Li. 2025. Personaagent: When large language model agents meet personalization at test time. In *arXiv.org*.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. 2025. Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Annual Meeting of the Association for Computational Linguistics*.

Jiachen Zhu, Menghui Zhu, Renting Rui, Rong Shan, Congmin Zheng, Bo Chen, Yunjia Xi, Jianghao Lin, Weiwen Liu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2025. Evolutionary perspectives on the evaluation of llm-based ai agents: A comprehensive survey. *Preprint*, arXiv:2506.11102.

Appendix

A Experimental Supplement

Table 5 reports supplementary benchmark results using the `all-mpnet-base-v2` retriever, evaluating Qwen2.5-7B-Instruct, DeepSeek-V3, GPT-4o-mini, and GPT-4o. Compared with the `bge-m3` setting in Table 3, `all-mpnet-base-v2` improves the overall score for all four models (+6.1 for Qwen2.5-7B, +7.6 for DeepSeek-V3, +4.4 for GPT-4o-mini, and +3.1 for GPT-4o), with the largest gains concentrated in the Planning and Health domains.

B Benchmark Details

B.1 Dataset Details

Table 7 summarizes the statistics across domains and scenarios, including the number of preference dimensions (Dimensions), the number of preference options (Preference Types), and the resulting scale of task combinations (Task Combos). This provides a quantitative view of PERSONAKAG's richness in cross-domain scenario coverage and preference-space diversity.

Table 6 provides illustrative examples of how preference dimensions are defined in our benchmark across domains. It is organized in a domain–scenario–dimension hierarchy, and lists mutually exclusive preference options for each dimension to support task instantiation and controlled evaluation. We further annotate whether a preference must be implicitly inferred from the interaction history (Implicit) and whether it may shift over time (Dynamic), covering more realistic personalization settings.

B.2 Prompt Details

Figure 8 shows the prompt template used to generate casual-chat dialogues in PERSONAKAG. Figures 9 and 10 show the prompt templates used to synthesize user historical dialogue evidence.

Table 5: Performance comparison using **all-mpnet-base-v2** retriever. Data columns have a fixed width for better readability.

Model	Writing	Work	Daily	Plan.	Health	Trans.	Medical	Leisure	Info.	Overall
Qwen2.5-7B-Instruct	**46.9**	**45.6**	**76.6**	**53.2**	**66.6**	**85.7**	**79.3**	**51.3**	**37.1**	**59.2**
DeepSeek-V3	**54.2**	**48.3**	**75.9**	**66.1**	**72.5**	**89.7**	**87.3**	**53.4**	**43.1**	**64.7**
GPT-4o-mini	**49.9**	**45.9**	**79.0**	**54.9**	**70.0**	**92.8**	**90.5**	**58.8**	**37.6**	**62.4**
GPT-4o	**49.8**	**45.4**	**77.4**	**54.3**	**66.6**	**91.8**	**87.8**	**52.0**	**35.8**	**60.8**

Table 6: Full taxonomy of preference dimensions across scenarios.

Domain	Scenario	Dimension	Preference Description	Dynamic
Writing and Design	Journal writing	Time granularity	Records by day (one entry per whole day).	×
			Records by key moments/timestamps.	×
		Content structure	Prefers a fixed template: "What happened / How I felt".	×
			Prefers free-form writing (write whatever comes to mind).	×
			Preferred the fixed template; now prefers free-form writing.	×
		Length preference	Prefers short entries (about 50–150 words each).	×
			Prefers long entries (detailed expansion; 300+ words).	×
Writing and Design	Business plan writing	Overall narrative style	Prefers narrative-driven writing.	×
			Prefers logic-driven writing (clear structure, data-centric).	×
			Preferred logic-driven writing; now narrative-driven writing.	✓
			Preferred narrative-driven writing; now logic-driven writing.	✓
		Market analysis depth	Prefers extensive use of industry data.	×
			Prefers highlighting key data points with concise interpretation.	×
		Risk section writing	Emphasizes explicit identification of risks.	×
			Follows an optimistic narrative, downplays risks.	×
Writing and Design	Technical doc writing	Document structure	Task-oriented structure organized around "how-to" steps.	×
			Concept-oriented structure.	×
		Technical depth	Prefers detailed explanations.	×
			Prefers simplified explanations.	×
		Example code style	Prefers complete, runnable code examples.	×
			Prefers code snippets (key parts only).	×
			Prefers pseudocode.	×
			Previously preferred complete; now prefers pseudocode.	✓
			Previously preferred pseudocode; now prefers code snippets.	✓
			Previously preferred code snippets; now prefers pseudocode.	✓
		API explanation	Prefers table-style specification.	×
			Prefers narrative explanation of API behavior.	×
Writing and Design	Review writing	Overall structure	Prefers organizing reviews into Summary / Pros / Cons sections.	×
			Prefers unstructured summaries with weaknesses separated.	×
		Weakness format	Prefers bullet-pointed weaknesses (with numbering).	×
			Prefers bullet-pointed weaknesses (without numbering).	×
		Weakness count	Requires exactly three weaknesses.	×
			Requires exactly five weaknesses.	×
		Weakness content	Prefers keywords + full sentences.	×
			Prefers direct full-sentence descriptions.	×
			Preferred direct sentences; now prefers keywords + sentences.	✓
Writing and Design	Blog article writing	Narrative perspective	First-person personal perspective.	×
			Third-person objective perspective.	×
		Citations / sources	Prefers academic style: formal citations, links, references.	×
			Prefers original style.	×
Writing and Design	Public account writing	Opening style	Storytelling-based opening.	×
			Viewpoint-based opening.	×
			Preferred storytelling-based opening, now viewpoint-based.	×
			Preferred viewpoint-based opening, now storytelling-based.	×
		Social expression	Leans toward social topics.	×
			Leans toward personal perspectives.	×
		Title style	Attention-grabbing titles.	×
			Formal titles (standard phrasing, fewer stylistic markers).	×
Writing and Design	Novel writing	Narrative perspective	First-person narration.	×
			Third-person narration.	×
		Narrative type	Linear storytelling in chronological order.	×
			Non-linear storytelling (e.g., flashbacks, parallel storylines).	×

Table 7: Preference dimensions and task combination statistics across domains and scenarios

Domain	Scenario	Dimensions	Preference Types	Task Combos	Dynamic
Writing and design	Peer review reports	4	9	54	✓
	Academic paper writing	6	15	71	✓
	Research proposal (RP) writing	4	13	76	✓
	Business plan writing	3	8	29	✓
	Diary writing	3	7	23	✓
	Technical documentation writing	4	12	80	✓
	Blog post writing	2	4	4	×
	WeChat public account article	3	8	27	✓
	Novel writing	2	4	6	×
	Resume writing	5	14	72	✓
	Short video script design	3	11	41	×
	PPT slide content writing	2	9	15	✓
	Mind map design	4	20	79	✓
	Modular course writing (teaching aid)	4	14	71	✓
	Instruction manual writing	3	15	73	✓
	Self-introduction script writing	3	8	29	×
Work	Scholarship application	7	17	140	✓
	Performance report spreadsheet	6	23	90	✓
	Schedule planning	6	31	119	✓
	Data analysis	4	27	101	✓
	Programming	5	23	90	✓
	Leave application	4	9	93	×
	Project management	5	22	105	✓
	Communication management	2	5	7	×
	Announcement drafting	3	12	74	✓
	Event organization	4	14	111	×
	Work summary writing	4	16	113	✓
	Terminology explanation	3	14	112	✓
Daily consumption	Food ordering	3	15	122	✓
	Ticket booking	4	17	145	✓
	Clothing purchase	4	15	153	✓
	Digital product purchase	4	10	89	×
	Accommodation booking	4	12	138	×
	Online course purchase	5	12	154	×
	Book purchase	4	12	127	✓
	Renting a house	11	28	158	×
Planning	Weekly plan	4	19	122	✓
	Monthly plan	4	12	107	✓
	Long-term goal setting	3	16	130	✓
	Project plan	4	15	114	×
	Financial plan	3	13	74	✓
	Travel plan	5	18	123	✓
	Study plan	4	19	154	✓
	Career development plan	4	16	100	✓
	Habit-building plan	4	18	115	×
	Skill improvement plan	4	16	112	✓
Exercise and health	Exercise plan	5	22	260	✓
	Diet plan	4	21	255	✓
	Sleep plan	6	32	262	✓
	Mental health support	1	6	4	×
Transportation	Commute route planning	5	21	236	×
	Taxi / ride-hailing travel	2	6	8	×
	Travel and sightseeing route planning	4	19	215	×
	Parking	3	10	59	×
Medical services	Appointment registration	4	15	210	×
	Medicine selection	4	17	231	×
	Online medical consultation	3	14	83	×
Leisure activities	Audio-visual entertainment	4	24	229	✓
	Outdoor leisure	2	5	6	×
	Social assistance	3	14	98	✓
Information management	Research report writing	4	18	174	✓
	Timeline organization	3	10	41	✓
	Meeting minutes organization	3	13	60	✓
	Reading notes	2	9	16	✓
	Folder organization	5	19	191	×
	Tool usage	9	34	182	✓

You are an English casual-chat dialogue generation assistant.
The current user's name is "{user_name}," and their general mood/state is "{mood}." Their detailed persona is as follows:
{persona_summary}
Your task is to generate natural, realistic daily casual-chat content **based solely on the personal information in the above user persona** (such as age, major, school, living environment, interests, emotions, etc.).
Important requirements:

1. The conversation should only be about light everyday topics, such as study rhythm, life status, social life, hobbies, emotional feelings, city/campus experiences, etc.
2. **Absolutely no specific tasks or instructions**, especially anything like "help me review a paper," "write reviewer comments," "this paper/manuscript/research," "task/instruction/system prompt," etc.
3. Even if the user persona includes research direction, paper topics, or project background, these can only serve as background atmosphere — do not expand on research/paper content or switch into a "task-helping" mode.
4. The conversation may include only two speaker prefixes: "User:" and "ChatGPT:". No other notes or markers.
5. The overall tone should be natural and unforced — more like casual chatting between friends than executing a task.

Figure 8: System prompt used for casual-chat generation.

INPUTS

User Information.
{person_info}

User Preferences *(Hidden — for the scriptwriter only)*.
{instance}

Task Overview.
{instruction_text}

Domain / Category.
{domain}

Sub-scenario.
{sub_scenario}

Task Context / Available Materials.
{content_text}

TASK

Generate a complete, natural, and fluent English conversation in which a user and **"ChatGPT"** collaboratively accomplish:

{scenario}

Generate the full dialogue now.

Figure 9: User prompt used for evidence data generation.

You are a **dialogue-generation assistant**. Generate a complete, natural, and fluent English conversation between a user and **"ChatGPT"** that collaboratively accomplishes the user's task.

Core Premise *(Non-negotiable)*.

- Only the generator (you) can see **hidden preferences**.
- The `ChatGPT` character in the dialogue **cannot** see hidden preferences.
- `ChatGPT` may revise outputs only based on:
  - what the user says in the dialogue, and
  - what `ChatGPT` produced in the previous turn.

Dialogue Format *(Strict)*.

- Use only two speaker labels: `User:` and `ChatGPT:`.
- The dialogue is entirely in English.
- Exactly *{turn_count}* turns in total, alternating lines starting with `User:` then `ChatGPT:`.
- Each turn contains **one speaker line only**; no narration, no stage directions, no extra labels, no blank lines.

Hard Rules.

1. **First user turn (no preference disclosure).** The first `User:` turn states only the task objective and necessary context/materials. It must not mention hidden preferences, "instance", or any style or format constraints.
2. **Immediate draft after turn 1.** After the first user message, `ChatGPT` must produce an initial complete deliverable. No clarification questions about preferences, formatting, structure, tone, or style are allowed.
3. **Multi-round revision is mandatory.** After Draft 1, the user provides revisions over several turns. Each `ChatGPT` response must update the content rather than repeat prior versions.
4. **Preference points per user turn.** Starting only after Draft 1, each new user turn introduces one to three preference points. Preferences must be expressed **implicitly** and never stated explicitly.
5. **User voice constraints.** The user uses direct commands, not polite questions. Forbidden expressions include "Could you…", "Can you…", "Would you mind…", and "Is it possible…". User feedback should be brief (one to two sentences).
6. **No meta-leaks.** `ChatGPT` must never mention hidden preferences, "instance", or prompt mechanics.
7. **Coverage requirement.** By the final turn, the output must satisfy all hidden preference points. Every distinct preference point must appear explicitly and concretely somewhere in the dialogue.
8. **Materials handling (conditional).** If required source materials are provided, include them verbatim and in full.
9. **Deliverable type.** *{deliverable_type_instruction}*
10. **No raw placeholders.** Do not output placeholders such as "[Company Name]". If specifics are missing, omit them or use a plausible generic name.
11. **Response length.** Each `ChatGPT:` turn should be concise, typically under 230 words.

Dialogue Flow *(Required)*.

- *Task entry:* user starts directly with the objective.
- *Draft 1:* `ChatGPT` produces a full first draft immediately.
- *Iterative refinement:* user reveals preferences gradually; `ChatGPT` revises accordingly.
- End with a short user closing line such as *"Alright, this works."*

Output Requirement. Generate the complete dialogue in one go, following all constraints above.

Hidden Preference Instance. *{instance}*

Figure 10: System prompt used for evidence data generation.