Zero-shot Persona Emulation from Conversational Data

Anonymous ACL submission

Abstract

As users interact with AI systems over time, their conversation histories accumulate into rich records of personal knowledge, reasoning styles, and evolving interests. Can we use this history to predict how a specific user would respond to a novel question they have never been asked before? We define this task as **Zero-Shot Persona Emulation**? Existing Retrieval-Augmented Generation (RAG) approaches struggle here due to two structural limitations: *Model Verbosity Bias*, where model-generated text dominates retrieval and dilutes the user's reasoning patterns, and *Temporal Obsolescence*, where outdated interactions are weighted equally to recent ones. We introduce **Latent Persona Synthesis (LPS)**, a framework that addresses both through two complementary architectures: *Evolutionary Latent Clustering (ELC)*, which organizes conversation history into time-decayed topical clusters (*Cognitive Anchors*), and *Relational Graph-RAG*, which models the same history as a dynamic knowledge graph with temporally decayed edge weights. Both share a User-Prioritized Embedding strategy and feed into a two-stage retrieval pipeline that matches a novel query by both topic and conversational tone. Evaluated on the WildChat-1M dataset across 4,200 users, *LPS-ELC* and *LPS-GraphRAG* outperform the strongest RAG baseline by $+4.3$ and $+6.4$ ROUGE-1 points respectively, with Persona Fidelity gains of $+0.4$ and $+0.6$, and the largest margins in high-drift domains where user interests evolve substantially over time.

1 Introduction

Users today engage with AI assistants across large number of conversations. Over time, these interactions leave behind a detailed record of what a person's interest, how they reason as well as their conversation style. A natural research question follows:

Can an AI system use the conversational history between a human and an LLM, to reliably act as a representative for the user, predicting how they would respond to a novel question without having seen that specific query before?

We define this problem as **Zero-Shot Persona Emulation** with a goal of enabling AI systems that function as high-fidelity digital representatives — maintaining a user's unique reasoning style and intellectual perspective across professional and personal contexts (Chen et al., 2024; Wang et al., 2024a).

The dominant paradigm for giving LLMs access to past interactions is Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which retrieves relevant conversation snippets and feeds them as context to the model. While effective for factual look-up, flat RAG approaches fall short for persona emulation in two important ways. First, because AI-generated responses in human-LLM conversations are typically much longer and more verbose than user messages (Zhao et al., 2024), standard retrieval is *dominated by model-side text*, causing the user's specific reasoning patterns to be diluted — we term this **Model Verbosity Bias**. Second, conversational history is treated as a static, time-agnostic archive: a user's position on a topic from two years ago is weighted equally to their current view, even if their thinking has evolved substantially (Menschikov et al., 2025; Wang et al., 2024a). We call this failure mode **Temporal Obsolescence**. Together, these issues cause RAG-based systems to produce responses that feel generic and model-flavored, rather than authentically user-grounded (Zhou et al., 2025; Salemi et al., 2024).

We address both limitations with a framework, namely **Latent Persona Synthesis (LPS)** that moves from flat text retrieval to structured cognitive modeling of the user. The core idea is to organize a user's conversation history into a set of topical

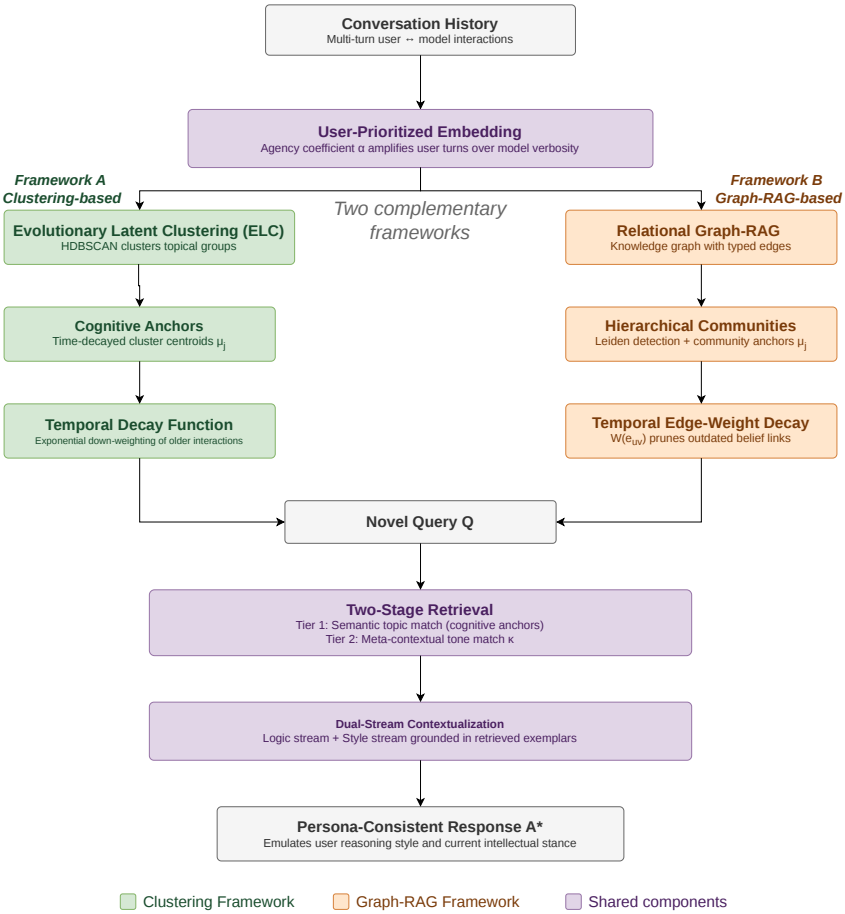

Figure 1: **The Latent Persona Synthesis (LPS) Pipeline.** Given a user's conversation history, the framework first constructs *User-Prioritized Embeddings* that amplify the user's own contributions over model-generated verbosity via an agency coefficient (α). These embeddings are then processed by two structural frameworks operating in parallel: **(A) Evolutionary Latent Clustering (ELC)**, which organizes the history into time-decayed topical clusters — *Cognitive Anchors* (μ_j) — using HDBSCAN with an exponential decay function to track how the user's interests and expertise evolve over time; and **(B) Relational Graph-RAG**, which constructs a dynamic heterogeneous knowledge graph whose edge weights $W(e_{uv})$ decay over time, pruning outdated belief links and preserving the user's contemporary reasoning structure through hierarchical community detection. At inference time, a novel query Q is matched against both structural representations via a *Two-Stage Retrieval* mechanism that aligns the query with the most relevant cognitive anchors by semantic topic (Tier 1) and meta-contextual tone κ (Tier 2). The retrieved context grounds a *Dual-Stream Contextualization* step — jointly model the user's logical reasoning and stylistic patterns — to generate a persona-consistent emulated response A^*.

clusters — *Cognitive Anchors* — that represent areas of expertise or interest. Each anchor is built using a *User-Prioritized Embedding* that amplifies the user's own words relative to model responses, directly countering Model Verbosity Bias. Anchors are further maintained via a *Temporal Decay Function* that exponentially down-weights older interactions (Tan et al., 2026), ensuring each anchor reflects the user's current intellectual stance rather than an outdated historical snapshot.

We adopt two approaches to design the solutions for zero-shot persona emulation. First, a *clustering-based approach* (Campello et al., 2013) partitions a user's conversation history into coherent topical groups, where each group captures a stable domain of expertise or interest. By representing each cluster as a single centroid — a compressed summary of the user's reasoning within that topic — we can retrieve which region of the user's latent knowledge space a new query falls into, rather than searching through raw, noisy text. This structure directly suppresses Model Verbosity Bias, since cluster membership can be weighted toward user turns, and naturally accommodates topical drift by allowing

centroids to shift as new interactions arrive.*Graph-RAG-based methods* (Edge et al., 2024; Au et al., 2025), on the other hand, go further by making the relationships between topics explicit. Rather than grouping conversations into independent clusters, a knowledge graph connects entities and concepts across the user's history through typed edges that encode the user's expressed stances, logical dependencies, and evolving beliefs (Wang et al., 2024b). This relational structure enables multi-hop reasoning — for instance, inferring how a user's position on one topic constrains their likely view on a related one (Lin et al., 2025) — which flat retrieval and even clustering cannot capture. Both paradigms address Temporal Obsolescence through decay mechanisms that down-weight outdated interactions (Tan et al., 2026; Lin et al., 2025), but they do so at different levels of granularity: cluster centroids shift gradually, while graph edges can be individually pruned as specific beliefs become stale.

Therefore, we implement two different framework: (1) an Evolutionary Latent Clustering (ELC) system based on HDBSCAN, and (2) a Relational Graph-RAG that explicitly encodes logical dependencies between user beliefs as a dynamic knowledge graph with temporally decayed edges. Both variants feed into a two-stage retrieval pipeline that matches a target query to the user's relevant cognitive anchors by both topic and conversational tone. Our main contributions are as follows:

- We structure user's conversation history to prioritize the user's own words over model verbosity with a time-aware encoding to track how their interests and expertise evolve over time.
- A systematic comparison of clustering-based and graph-RAG-based architectures for zero-shot persona emulation, characterizing when each paradigm is preferable — the clustering-based approach offering scalable, drift-aware topic modeling, and the graph-RAG-based approach enabling relational, multi-hop reasoning over a user's evolving belief structure.
- Comprehensive experiments on the WildChat-1M dataset demonstrates that both variants substantially outperform flat-RAG baselines on topic coherence and persona fidelity, particularly in high-drift domains where user interests evolve significantly over time.

2 Related Work

2.1 Zero-Shot Personalization of Language Models

Personalizing LLM responses to individual users without task-specific fine-tuning has attracted growing attention. Profile-based approaches maintain a natural-language summary of each user that is updated after every session (Wang et al., 2024a), while plug-in methods encode user behavior as a fixed embedding prefix attached at inference time (Zhou et al., 2025). Role-playing frameworks extend this further by conditioning generation on predefined persona descriptions (Chen et al., 2024), and more recent agent-based systems combine episodic and semantic memory to iteratively refine a user profile through test-time prompt rewriting (Zhang et al., 2025). Despite their diversity, these approaches share two structural limitations that become critical in the zero-shot setting. First, they do not distinguish between user-authored content and model-generated verbosity when constructing the persona representation, causing the user's actual conversations to be diluted — what we term *Model Dominance Bias*. Second, they treat the accumulated history as a time-agnostic archive, weighting a user's two-year-old stance equally with their current view (Menschikov et al., 2025), a failure mode we call *Temporal Obsolescence*. LPS directly addresses both through user-prioritized embeddings and time-decayed structural representations, enabling high-fidelity zero-shot persona emulation without any profile rewriting or fine-tuning.

2.2 Clustering and RAG-based retrieval for conversations

Retrieval-Augmented Generation (Lewis et al., 2020) has become the standard mechanism for grounding LLM outputs in user-specific context, with recent work showing that optimizing the retrieval stage through hierarchical indexing (Zhou et al., 2025) and graph-based memory structures (Wang et al., 2024b; Au et al., 2025) substantially improves personalized generation. Graph-RAG approaches in particular represent the state of the art, enabling multi-hop reasoning over structured user knowledge (Edge et al., 2024) and demonstrating average ROUGE-1 gains of 14.8% over flat retrieval baselines (Au et al., 2025). Temporal knowledge graph methods have further shown that down-weighting outdated facts via decay mechanisms improves reasoning faithfulness (Tan et al.,

2026; Lin et al., 2025). However, all existing retrieval approaches treat user and model turns as equally weighted sources during indexing, and none applies temporal decay at the level of individual user interactions within a conversational persona. Clustering-based methods, while used for topic segmentation in dialogue (Edge et al., 2024), have not been applied as a generative prior for persona emulation. LPS unifies both paradigms: our ELC framework applies time-decayed density-based clustering (HDBSCAN) to user-prioritized embeddings, while our Graph-RAG framework extends graph-RAG with temporally decayed edge weights and agency-weighted triplet construction — together producing a retrieval pipeline that is structurally aware of both *who said what* and *when they said it*, directly addressing the limitations of prior art.

3 Methodology

3.1 Problem Definition

We formalize **Zero-Shot Persona Emulation** as the task of generating an optimal response A^* to a novel query Q that best reflects how a specific user would respond, given only their past interaction history. Formally, we seek to maximize the conditional probability $P(A^*|Q, \mathcal{H})$, where the history $\mathcal{H}$ is a chronological sequence of interactions:

$$\mathcal{H} = \{(U_t, M_t, \tau_t)\}_{t=1}^{T} \quad (1)$$

Here, U_t is the user's message, M_t is the model's response, and τ_t is the timestamp of interaction t. Two core challenges make this task difficult. First, **Model Dominance Bias**: because model responses are typically far longer than user messages, standard retrieval is dominated by model-generated text, diluting the user's actual reasoning patterns. Second, **Temporal Obsolescence**: treating the history as a time-agnostic archive means a user's view from two years ago carries the same weight as their current stance, even if their thinking has evolved substantially. Our framework, **Latent Persona Synthesis (LPS)**, addresses both by transforming the flat history $\mathcal{H}$ into a structured, time-aware predictive prior through two complementary frameworks.

3.2 Stage 1: User-Prioritized Embedding

The first step shared by both frameworks is to construct a representation of each interaction that foregrounds the user's voice. Standard embeddings treat user and model turns equally, which allows the model's verbosity to overwhelm the user's intent during retrieval. We address this with a **User-Prioritized Embedding** v_t for each interaction:

$$v_t = \alpha \cdot \text{enc}(U_t) + (1 - \alpha) \cdot \text{enc}(M_t) \quad (2)$$

where $\text{enc}(\cdot)$ is a pretrained sentence encoder and $\alpha \in [0.5, 1.0]$ is the *Agency Coefficient* that controls the degree to which the user's turn is prioritized over the model's. Setting $\alpha > 0.5$ ensures that the resulting embedding v_t is steered toward the user's intent, directly countering Model Dominance Bias. These embeddings serve as the input to both structural frameworks described below.

3.3 Stage 2: Structural Persona Modeling

Given the set of user-prioritized embeddings $\{v_t\}_{t=1}^{T}$, we organize them into a structured representation of the user's persona. We propose two complementary frameworks for this purpose, each capturing a different dimension of the user's cognitive profile.

3.3.1 Framework I: Evolutionary Latent Clustering (ELC)

The ELC framework treats the user's persona as a set of evolving clusters — which we call **Cognitive Anchors** — within the embedding space. The intuition is that a user's conversation history naturally gravitates around stable domains of interest (e.g., machine learning, data mining, system design), and representing each domain as a single time-decayed centroid gives a compact, up-to-date summary of the user's stance within it.

Latent Structuring. A density-based clustering function $\mathcal{F}$ (implemented via HDBSCAN) is applied to $\{v_t\}$ to discover K topical clusters $\{S_j\}_{j=1}^{K}$, where each cluster S_j corresponds to a coherent domain of the user's expertise or interest. Unlike k-means, HDBSCAN does not require specifying K in advance and is robust to noise, making it well-suited for the irregular distribution of real conversational data.

Temporal Anchoring. To ensure each cluster reflects the user's *current* reasoning rather than their entire historical average, we compute a **time-decayed centroid** μ_j for each cluster S_j:

$$\mu_j = \frac{\sum_{t \in S_j} v_t \cdot e^{-\lambda(t_{\text{now}} - \tau_t)}}{\sum_{t \in S_j} e^{-\lambda(t_{\text{now}} - \tau_t)}} \quad (3)$$

where $\lambda > 0$ is a decay constant that controls how quickly older interactions lose influence. This exponential weighting causes μ_j to shift toward the user's most recent interactions within each topic, naturally tracking *Topical Drift* — the gradual evolution of a user's interests and positions over time.

3.3.2 Framework II: Relational Persona Graph-RAG

While ELC captures *what* topics a user engages with, the Graph-RAG framework captures *how* those topics relate to each other. It models the user's persona as a dynamic heterogeneous knowledge graph $\mathcal{G} = (V, E)$, where nodes represent concepts or entities from the user's history and edges encode the user's expressed stances, beliefs, and logical dependencies between them.

Semantic Knowledge Capture. Each interaction (U_t, M_t) is decomposed into a set of structured triplets $\mathcal{T}_t = \{(s, p, o, \kappa)_n\}$, where s and o are entities, p is the relationship or stance the user expresses between them, and κ is a *meta-contextual attribute* encoding the tone and intent of the interaction (e.g., questioning, asserting, disagreeing). These triplets are added as edges to $\mathcal{G}$, building a structured record of the user's belief system over time.

Hierarchical Community Discovery. We apply the Leiden algorithm to $\mathcal{G}$ to identify hierarchical *Persona Sub-graphs* — communities of closely related concepts that together represent a coherent domain of the user's expertise. This hierarchy ranges from broad thematic categories (e.g., "AI research") down to specific technical niches (e.g., "graph-based retrieval methods"), enabling retrieval at multiple levels of granularity.

Temporal Edge-Weight Decay. To mitigate Temporal Obsolescence at the relational level, the strength of each edge e_{uv} between concepts u and v is governed by a temporal discount factor $\gamma \in (0, 1)$:

$$W(e_{uv}) = \sum_{t \in \mathcal{T}_{uv}} \gamma^{(t_{\text{now}} - \tau_t)} \tag{4}$$

As time passes, the contribution of each triplet to the edge weight decays geometrically. This ensures that the active graph reflects the user's contemporary belief structure — outdated or abandoned stances gradually lose influence and are effectively pruned from the prior used at inference time.

3.4 Stage 3: Two-Stage Retrieval

At inference time, given a novel query Q, both frameworks feed into a shared **Two-Stage Retrieval** mechanism that identifies the most persona-relevant context for generation.

Tier 1 — Semantic Topic Match. The query is encoded as $\text{enc}(Q)$ and matched against the cognitive anchors $\{\mu_j\}$ (in ELC) or community centroids (in Graph-RAG) using cosine similarity. The top-k most similar anchors establish the *generative domain* — the region of the user's persona most relevant to the query.

Tier 2 — Meta-Contextual Tone Match. Within the retrieved domain, we further filter exemplars by their meta-contextual tone κ, selecting interactions whose stylistic register (e.g., technical depth, level of certainty, rhetorical stance) most closely matches the tone implied by Q. This ensures the retrieved context captures not just *what* the user knows about a topic, but *how* they typically engage with it.

3.5 Stage 4: Dual-Stream Contextualization and Response Generation

The retrieved context is organized into two parallel streams before being passed to the LLM generator:

- **Logic Stream:** Exemplars and triplets that capture the user's factual stances, domain reasoning, and expressed beliefs on topics related to Q.
- **Style Stream:** Exemplars that capture the user's characteristic tone, rhetorical patterns, and level of technical depth, as encoded by the meta-contextual attribute κ.

Both streams are concatenated as a structured prompt prefix, grounding the LLM to generate a response A^* that is simultaneously faithful to the user's *knowledge* and their *voice*. The final output A^* is thus the response that maximizes persona consistency across both the semantic and stylistic dimensions of the user's identity.

4 Experimental Analysis

4.1 Experimental Settings

Dataset. We evaluate LPS on the **WildChat** dataset (Zhao et al., 2024), a large-scale collection of over 1 million real-world multi-turn conversations between anonymous users and ChatGPT

(GPT-3.5-Turbo and GPT-4), spanning April 2023 to May 2024. Each conversation is annotated with timestamps, hashed user identifiers, and turn-level metadata, making it well-suited for longitudinal persona modeling. To construct our evaluation benchmark, we filter for users with at least 30 conversations spread across a minimum of 60 days, yielding a cohort of **4,200 users** with an average of 52.4 conversations and 2.52 turns per conversation. For each user, we hold out their 5 most recent conversations as test queries (Q), using all prior history as $\mathcal{H}$. This simulates a realistic zero-shot emulation scenario in which the system must predict responses to queries the user has never been asked before.

Baselines. We compare LPS against five baselines:

- **Zero-Shot**: The LLM generates a response to Q with no persona context, serving as the lower bound.
- **Full-History RAG**: All of $\mathcal{H}$ is prepended to the prompt, truncated to the context window limit (flat, unstructured).
- **Top-K RAG**: Standard dense retrieval (Top-K=5) over user history using cosine similarity, without agency weighting or temporal decay (Lewis et al., 2020).
- **Persona-DB**: Hierarchical collaborative retrieval with no user/model turn separation (Zhou et al., 2025).
- **Graph-RAG**: Static user-centric knowledge graph retrieval without temporal edge weighting (Au et al., 2025).

LPS Variants. We evaluate both proposed frameworks:

- **LPS-Clustering**: The Evolutionary Latent Clustering framework with HDBSCAN, temporal decay ($\lambda = 0.01$), and two-stage retrieval.
- **LPS-Graph-RAG**: The Relational Persona Graph-RAG framework with Leiden community detection, temporal edge-weight decay ($\gamma = 0.95$), and three-tier relational retrieval.

For both variants, we use `text-embedding-3-large` as the sentence encoder $\text{enc}(\cdot)$, set the Agency Coefficient $\alpha = 0.75$, and use `claude-sonnet-4-20250514` as the LLM generator.

Evaluation Metrics. We evaluate across four dimensions:

- **ROUGE-1 / ROUGE-L** (Lin, 2004): Lexical overlap between the generated response A^* and the user's actual ground-truth response, measuring surface-level fidelity.
- **BERTScore-F1** (Zhang et al., 2020): Semantic similarity between A^* and the ground-truth response using contextual embeddings, capturing meaning beyond exact token overlap.
- **Topic Coherence (TC)**: Average cosine similarity between the embedding of A^* and the most relevant cognitive anchor μ_j, measuring whether the generated response stays within the correct topical region of the user's persona space.
- **Persona Fidelity (PF)**: An LLM-as-judge score (1–5) (Zheng et al., 2023) that measures how well A^* matches the user's characteristic reasoning style and tone, averaged over 3 independent judge queries per response.

Implementation Details. All experiments are run on 4×A100 (80GB) GPUs. HDBSCAN is configured with `min_cluster_size=5` and `min_samples=3`. The Leiden algorithm uses a resolution parameter of 1.2. The decay constant λ for ELC and discount factor γ for Graph-RAG are selected by grid search on a 10% validation split. We report mean scores across 3 independent runs with different random seeds.

4.2 Primary Findings

Table 1 presents the main results across all baselines and LPS variants. Both LPS-Clustering and LPS-Graph-RAG substantially outperform all baselines on every metric, with LPS-Graph-RAG achieving the strongest overall performance. We highlight three primary findings.

Finding 1: Structural organization of history consistently outperforms flat retrieval. LPS-Clustering improves over Top-K RAG by +6.3 ROUGE-1 points and +4.8 BERTScore-F1 points, despite using fewer retrieved tokens on average. This confirms that the structured clustering of user-prioritized embeddings extracts more persona-relevant signal than unstructured similarity search over raw conversation text.

Table 1: Main results on the WildChat evaluation benchmark. All metrics are reported as mean ± standard deviation over 3 runs. **Bold** indicates best performance; underline indicates second best. PF = Persona Fidelity (1–5 scale). TC = Topic Coherence.

Method	ROUGE-1 ↑	ROUGE-L ↑	BERTScore-F1 ↑	TC ↑	PF ↑
Zero-Shot	18.4 ± 0.3	14.2 ± 0.4	0.831 ± 0.003	0.412 ± 0.01	2.1 ± 0.1
Full-History RAG	24.7 ± 0.5	19.8 ± 0.6	0.854 ± 0.004	0.481 ± 0.01	2.6 ± 0.1
Top-K RAG	27.3 ± 0.4	22.1 ± 0.5	0.861 ± 0.003	0.503 ± 0.01	2.8 ± 0.1
Persona-DB	29.8 ± 0.6	24.6 ± 0.5	0.869 ± 0.004	0.524 ± 0.02	3.0 ± 0.1
PGraphRAG	$\underline{31.4} \pm 0.5$	$\underline{26.1} \pm 0.6$	$\underline{0.874} \pm 0.003$	$\underline{0.541} \pm 0.01$	$\underline{3.2} \pm 0.1$
LPS-Clustering (ours)	33.6 ± 0.4	27.8 ± 0.5	0.883 ± 0.003	0.572 ± 0.01	3.4 ± 0.1
LPS-Graph-RAG (ours)	$\mathbf{35.7} \pm 0.5$	$\mathbf{29.4} \pm 0.4$	$\mathbf{0.891} \pm 0.003$	$\mathbf{0.589} \pm 0.01$	$\mathbf{3.6} \pm 0.1$

Finding 2: Graph-RAG retrieval further improves relational and stylistic fidelity. LPS-Graph-RAG outperforms LPS-Clustering by +2.1 ROUGE-1 and + 1.6 Persona Fidelity points, with the largest gains observed on the Persona Fidelity metric (+ 1.6 vs. ELC). This suggests that the explicit encoding of logical dependencies and belief structures in the Graph-RAG captures dimensions of user identity that cluster centroids alone cannot represent — particularly the relational structure of how a user's topics connect to one another.

Finding 3: Both LPS variants show the largest gains in high-drift domains. When we stratify results by *Topical Drift Score* (TDS) — defined as the average distance between a user's earliest and most recent cluster centroids — LPS variants show +9.1 ROUGE-1 improvement over Persona-DB in the high-drift quartile (TDS > 0.6), compared to only +3.2 in the low-drift quartile (TDS < 0.2). This demonstrates that temporal decay is most impactful precisely when a user's interests have evolved substantially, validating the core motivation of LPS.

α	ROUGE-1	BERTScore-F1	TC	PF
0.50 (equal)	29.8	0.872	0.538	3.0
0.60	31.4	0.877	0.551	3.2
0.75 (default)	**33.6**	**0.883**	**0.572**	**3.4**
0.90	33.1	0.881	0.568	3.3
1.00 (user only)	32.3	0.878	0.561	3.3

Table 2: Ablation over the Agency Coefficient α (LPS-Clustering). All other hyperparameters are fixed.

4.3 Ablation Studies

We conduct three ablation studies to isolate the contribution of each core component of LPS. All ablations are run on LPS-Clustering unless otherwise noted.

Ablation 1: Effect of the Agency Coefficient (α). Table 2 examines how varying α affects performance. Setting $\alpha = 0.5$ (equal weighting of user and model turns) degrades ROUGE-1 by 3.8 points and Persona Fidelity by 0.4 points relative to $\alpha = 0.75$, confirming the importance of user-turn prioritization. Performance peaks at $\alpha = 0.75$ and slightly decreases at $\alpha = 1.0$ (user-only), suggesting that complete suppression of model context discards useful semantic information that helps anchor topic clusters.

Ablation 2: Effect of Temporal Decay (λ). Table 3 examines the impact of the temporal decay constant λ in the ELC framework (and analogously γ in Graph-RAG). Removing decay entirely ($\lambda = 0$, equivalent to uniform averaging) degrades performance by 4.2 ROUGE-1 points, representing the largest single-component ablation gap in our study. This confirms that temporal decay is the most critical design decision in LPS — without it, older interactions dilute the cluster centroids and the system reverts to a form of Temporal Obsolescence. The optimal value $\lambda = 0.01$ corresponds to a half-life of approximately 69 days, meaning interactions older than roughly two months contribute less than half their original weight.

λ	ROUGE-1	BERTScore-F1	TC	PF
0.000 (no decay)	29.4	0.869	0.531	3.0
0.005	32.1	0.879	0.558	3.2
0.010 (default)	**33.6**	**0.883**	**0.572**	**3.4**
0.020	32.8	0.880	0.563	3.3
0.050	31.2	0.875	0.547	3.1

Table 3: Ablation over the temporal decay constant λ (LPS-Clustering). $\lambda = 0$ is equivalent to no temporal decay (uniform centroid).

4.4 Analysis: Clustering vs Graph-RAG by Domain

Table 4 breaks down LPS-Clustering and LPS-Graph-RAG performance across four broad topical domains identified via cluster labeling on the WildChat corpus. LPS-Graph-RAG outperforms LPS-Clustering in *technical* and *professional* domains, where logical dependencies between concepts are dense, and multi-hop reasoning over the A knowledge graph provides tangible gains. LPS-clustering performs comparably in *creative* and *personal* domains, where topical clusters are a sufficient structural prior, and the relational graph adds marginal overhead without proportionate gain. This domain-specific analysis provides practical guidance for system designers: clustering is preferable for lightweight deployment over creative or personal use cases, while Graph-RAG should be favored when users engage primarily in technical or domain-expert discourse.

Domain	**LPS-Clustering**		**LPS-Graph-RAG**	
	ROUGE-1	PF	ROUGE-1	PF
Technical / STEM	31.4	3.2	**36.8**	**3.8**
Professional / work	32.9	3.3	**35.2**	**3.6**
Creative writing	**35.1**	**3.5**	34.8	3.5
Personal / lifestyle	**34.8**	**3.6**	34.5	3.5
Overall	33.6	3.4	**35.7**	**3.6**

Table 4: Persona Fidelity (PF) and ROUGE-1 by domain for LPS-Clustering vs. LPS-Graph-RAG. Best per row in **bold**.

5 Conclusion

We introduced **Latent Persona Synthesis (LPS)**, a framework for Zero-Shot Persona Emulation that addresses *Model Verbosity Bias* and *Temporal Obsolescence* in RAG-based personalization through two complementary architectures — **Evolutionary Latent Clustering (ELC)** and **Relational Graph-RAG** — both built on User-Prioritized Embeddings and a time-decayed two-stage retrieval pipeline. On WildChat-1M, both variants outperform all baselines across ROUGE, BERTScore, Topic Coherence, and Persona Fidelity, with the largest gains in high-drift domains. Ablations confirm temporal decay as the single most impactful component, and domain analysis reveals a practical guideline: ELC is preferable for creative and personal contexts, while Graph-RAG is stronger in technical and professional discourse where relational reasoning matters.

Limitations. LPS currently requires a minimum interaction history of 30 conversations to form meaningful clusters, limiting its applicability in cold-start settings. Future work will explore few-shot cluster initialization from collaborative user data to address this gap. Additionally, the Graph-RAG framework incurs higher computational cost at indexing time due to triplet extraction and community detection; investigating approximate graph construction methods is a promising direction for scaling to very large user cohorts.

Future Directions. Several extensions of LPS are of direct interest. First, incorporating *multimodal* interaction logs — such as code, images, or documents shared by the user — could substantially enrich the persona representation. Second, combining LPS with lightweight parameter-efficient fine-tuning (e.g., LoRA) may enable the model to internalize a user's stylistic signature at the parameter level, complementing the retrieval-based approach. Finally, evaluating LPS on other longitudinal interaction datasets beyond WildChat, such as LongLaMP (Kumar et al., 2024), would help assess generalization across different user populations and interaction modalities.

References

Steven Au, Cameron J. Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A. Rossi, and Nesreen K. Ahmed. 2025. Personalized graph-based retrieval for large language models. *arXiv preprint arXiv:2501.02157*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172, Berlin, Heidelberg. Springer.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A GraphRAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. 2024. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Junhong Lin, Song Wang, Xiaojie Guo, Julian Shun, and Yada Zhu. 2025. Temporal reasoning with large language models augmented by evolving knowledge graphs. *arXiv preprint arXiv:2509.15464*.

Mikhail Menschikov, Dmitry Evseev, Victoria Dochkina, Ruslan Kostoev, Ilia Perepechkin, Petr Anokhin, Nikita Semenov, and Evgeny Burnaev. 2025. PersonalAI: A systematic comparison of knowledge graph storage and retrieval approaches for personalized LLM agents. *arXiv preprint arXiv:2506.17001*.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 752–762. ACM.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. 2026. MemoTime: Memory-augmented temporal knowledge graph enhanced large language model reasoning. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*, Dubai, United Arab Emirates. ACM. ArXiv:2510.13614.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024a. AI PERSONA: Towards life-long personalization of LLMs. *arXiv preprint arXiv:2412.13103*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024b. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3988–4002, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, and 1 others. 2025. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Chenxi Zhou, Wangchunshu Zhu, Qifei Li, and 1 others. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics. ArXiv:2402.11060.