B.Comp. Dissertation

Persona Replication Framework for Personalised Agent

By

Lim Jian Yang

Department of Computer Science

School of Computing

National University of Singapore

2025/2026

B.Comp. Dissertation

Persona Replication Framework for Personalised Agent

By

Lim Jian Yang

Department of Computer Science

School of Computing

National University of Singapore

2025/2026

Project No: H036650
Advisor: Professor Lee Wee Sun

Deliverables:
Report: 1 Volume

Abstract

This project builds a persona replication system: an agent capable of expanding a writer's initial ideas into full drafts, writing on the author's behalf while faithfully reproducing their personal voice and knowledge.
The first phase rewrites neutral text into a target author's style using Supervised Fine-Tuning (SFT) followed by an iterative self-supervised Direct Preference Optimisation (DPO) loop. The loop refines models using a combined metric of fluency, semantic similarity, and discriminator confidence. In a blinded human study, the best variant outranks the GPT-5.1 few-shot baseline used during Phase 1 and approaches real-author performance with a **64.4% top-3 ranking rate (vs 70.0% for ground truth)**.
The second phase reconstructs the author's mental model from their posts as a knowledge graph. An incremental four-stage pipeline extracts cognitive nodes, integrates them via nearest-neighbour search, clusters them into topics using UMAP with HDBSCAN, and organises them into a soft hierarchy. Finally, the system derives guiding principles from these topics and groups. For generation, retrieved nodes and principles ground a neutral draft, which is then stylised by the Phase 1 model. Evaluation of this graph-grounded approach against a Google RAG baseline suggests **improvements on broad, synthesis-requiring queries**, increasing Precision@5 by 139% and MRR by 84% on the constructed broad-query benchmark.

Subject Descriptors:

I.2.7	Natural Language Processing
I.2.6	Learning
I.2.4	Knowledge Representation Formalisms and Methods
H.3.3	Information Search and Retrieval
H.3.1	Content Analysis and Indexing

Keywords:

authorship style transfer, direct preference optimisation, personalised agents, mental model extraction, knowledge graph

Implementation Software and Hardware:

Python 3, OpenAI API, Gemini API, PyTorch, RoBERTa-CoLA, OpenAI `text-embedding-3-small`, FAISS, HDBSCAN, UMAP, FastAPI, Pydantic

Acknowledgement

I would like to thank Professor Lee Wee Sun for his strong support and guidance throughout this project. At multiple junctures when I found myself stuck, his timely insights and suggestions quickly helped me overcome technical hurdles. Time and again, his deep knowledge of the space prevented me from heading down fruitless research paths and pointed me toward highly valuable and insightful resources that served as much-needed inspiration. His sharp questions during our meetings were instrumental in the constant re-evaluation of my methods and allowed me to make effective improvements over time towards the pipeline's final design. I am deeply appreciative of the time he took to review my progress through the academic year and the patience he showed in helping me develop my ideas. His guidance was instrumental to this project, and I'm deeply grateful to him.

List of Figures

Table of Contents

Chapter 1

Introduction

1.1 Motivation

While Large Language Models (LLMs) have become ubiquitous as writing assistants, their generic outputs routinely fail to capture the authentic voice and nuanced worldview of individual authors. Furthermore, recent evidence suggests that people who use AI tools at work can be judged more negatively on competence and motivation (Reif, Larrick, and Soll, 2025). An individual's writing reflects their unique mental model of the world and their personal experiences – encompassing specific beliefs, reasoning patterns, recurring values, and characteristic opinions. Statistical averages of human text by LLMs naturally lack this personal cognitive footprint.

This project addresses the gap between AI efficiency and human authenticity. The goal is to develop an end-to-end persona replication system that combines authorship style transfer with mental model extraction. By doing so, this project aims to move toward a highly personalised autonomous agent that can draft content in a user's characteristic voice while grounding that draft in their personal mental models. Such a system would help resolve the tension between automated convenience and genuine expression, empowering individuals to delegate drafting while preserving their unique intellectual identity.

1.2 Two-Phase System Overview

The project is organised around two complementary phases. The first phase builds a style replication model: a fine-tuned LLM that rewrites neutral text into the target author's writing

voice. The second phase builds a mental model extraction system: a pipeline that reads the author's blog posts and distils their beliefs, values, principles, and opinions into a persistent cognitive graph. Together, these two components form an end-to-end persona replication system. When the author provides a topic or short notes, the system retrieves relevant knowledge from the cognitive graph, produces a neutral draft grounded in the author's worldview, and then rewrites that draft in the author's voice using the style transfer model.

The objective of implementing the current state of the art in authorship style transfer (Bhandarkar, Wilson, Swarup, and Woodard, 2024; Liu and May, 2025; Liu, Agarwal, and May, 2024a) was pursued in the first phase, with a focus on one model per author from the Blog Authorship Corpus. Since iterative preference optimisation has been shown to be more effective than prompt-based approaches, the first phase builds upon this direction, with a proposed improvement to the evaluation methodology used to select preference pairs.

The second phase focuses on knowledge extraction and reconstruction. The challenge is that a generic LLM has no access to what a specific individual actually knows, thinks, and values. Standard retrieval-augmented generation over flat document chunks provides some grounding, but struggles with broad, synthesis-requiring queries that span multiple aspects of the author's worldview. This motivates a structured graph-based approach to representing and retrieving the author's mental model.

1.3 Potential Applications

Authorship style transfer combined with mental model grounding enables several practical use cases:

- **Personal Productivity.** A writing assistant trained on one's own style and knowledge base can draft emails, reports, and creative content for user review.
- **Knowledge Preservation.** Capturing a domain expert's mental model allows their reasoning and institutional knowledge to remain queryable after they depart.
- **Branding and Content Creation.** Organisations can generate public-facing content at scale while maintaining stylistic and ideological consistency.

- **Cognitive Digital Twins.** A model of an individual's belief structure can simulate their reasoning, supporting decision-making and scenario evaluation.
- **Authorship Verification.** A high-fidelity style generator trained on personal data can simultaneously train a discriminator usable for fraud detection and authorship verification.

1.4 Project Objectives

In essence, this project achieves the following core objectives:

- **Develop a Personal AI Rewriter.** Create a customised LLM agent for an individual that rewrites neutral text into a style unique to that individual, through SFT followed by iterative DPO fine-tuning.
- **Improve Authorship Style Transfer.** Implement and improve upon current leading methods for authorship style transfer, with a refined evaluation methodology that better captures the nuances of a single author's writing style.
- **Extract and Reconstruct the Author's Mental Model.** Build a system that reads the author's blog posts and distils their beliefs, values, principles, and opinions into a structured, queryable cognitive graph that serves as the author's reconstructed worldview, using a four-stage pipeline (WAKE, MERGE, NREM, REM) inspired by memory consolidation in cognitive science.
- **Integrate Style and Knowledge into an End-to-End Persona System.** Combine the mental model retrieval system with the style rewriter so that content generated for the author is both informationally grounded in their actual knowledge and stylistically authentic to their personal voice.

Chapter 2

Study of Existing Approaches

My literature review covers two areas. The first is authorship style transfer, where I examine supervised fine-tuning, preference optimisation, and evaluation methodology for the task of replicating a single author's voice. The second is graph-based memory and knowledge extraction, where I survey retrieval-augmented generation, knowledge graph construction, and the cognitive science literature on sleep-inspired memory consolidation, which informs the design of the mental model pipeline.

2.1 Authorship Style Transfer

My review of the style transfer literature sought to understand the current state of the art and identify improvements for single-author replication. I first looked at prompt-based and supervised baselines for style imitation, then moved to preference-optimisation approaches such as DPO and later systems such as STAMP and ASTRAPOP (Bhandarkar et al., 2024; Rafailov et al., 2023; Liu and May, 2025; Liu et al., 2024a). Finally, I reviewed evaluation methods, including LLM-as-a-judge, discriminator-style evaluators, and metric choices that better reflect single-author style transfer.

2.1.1 Supervised Fine-Tuning Foundations

Supervised fine-tuning adapts a pre-trained LLM to a target style using input-output pairs. For style transfer, this involves parallel or pseudo-parallel data where the input and output are

similar in semantic content. While it can specialise a model to a user's writing style to a certain extent, maximum-likelihood training optimises token likelihood rather than human-perceived style quality. In practice, SFT often learns superficial cues instead of deeper stylistic structure unique to the source text, and its behaviour is sensitive to prompt formatting and the training distribution.

Parameter-efficient adaptation methods such as LoRA and QLoRA substantially reduce the adaptation footprint (e.g., trainable parameters and/or memory requirements), making fine-tuning feasible under tighter hardware constraints, although wall-clock speedups depend on the training setup and implementation (Hu et al., 2022; Dettmers, Pagnoni, Holtzman, and Zettlemoyer, 2023), but they do not fix misalignment. There is no explicit signal that, given the same content, one output is stylistically preferred over another.

2.1.2 Preference Optimisation Approaches

Reinforcement learning from human feedback addresses this by training a reward model on pairwise preferences and optimising the policy with PPO (Ouyang et al., 2022; Schulman, Wolski, Dhariwal, Radford, and Klimov, 2017), but it adds a separate reward-learning stage and can degrade true performance when an imperfect reward model is over-optimised (Gao, Schulman, and Hilton, 2023a).

In contrast, DPO directly trains the model to favour chosen over rejected outputs relative to a reference policy, making an explicit reward model unnecessary (Rafailov et al., 2023). Its simplicity and stability are attractive, but success is critically contingent on identifying accurate preference pairs and on mitigating potential overfitting to superficial markers due to poorly chosen preferences.

Beyond single-stage DPO, ASTRAPOP focuses on author-to-author transfer via policy optimisation guided by a learned style reward, emphasising idiosyncratic writer "fingerprints" (Liu et al., 2024a). ASTRAPOP is a two-stage tune-and-optimise approach: it first constructs pseudo-parallel data via STRAP-style paraphrasing to train a reference model, then applies policy optimisation (PPO, DPO, or CPO) using toward and away rewards computed from a task-appropriate style model (LUAR-based similarity for individual authors; a trained classifier for community/native-language transfer). STAMP introduces a multi-iteration preference-

optimisation framework for text style transfer; in their experiments the multi-iteration PO stage is implemented with CPO (rather than DPO), alongside hope–fear preference pair construction and weighted multi-objective reward aggregation (Liu and May, 2025). While STAMP reports gains over strong baselines, including SFT-based and other prior style-transfer methods on text style transfer benchmarks, STAMP's ablations indicate that multi-objective reward aggregation materially affects objective stability (notably on CDS), motivating weighted aggregation; for hope-and-fear preference construction, the authors report that incorporating model-score weighting was not beneficial and was dropped in the final configuration.

2.1.3 Synthetic Preference Construction

Complementary self-training approaches synthesise preference data for DPO without human labels. For example, SELF-CONTRAST is a feedback-free DPO-style alignment method that treats the original SFT target as the chosen response, then generates many self-produced candidates and uses embedding-based similarity filtering to select heuristic negatives as rejected responses (Liu, Song, Dong, and Tang, 2024b). Conceptually, this mirrors ASTRAPOP's generate-and-optimise loop and STAMP's hope-and-fear mining, but with a simpler contrastive rule.

2.1.4 Evaluation of Style Authenticity

An alternative perspective on style transfer comes from Krishna, Wieting, and Iyyer (2020), who reformulate the task as paraphrase generation. By framing style transfer as producing paraphrases that preserve semantic content while modifying stylistic attributes, their work keeps meaning preservation central and also shows how brittle automatic evaluation can be when style transfer is measured with fixed proxy metrics. This makes learned evaluators attractive as a complementary signal. In this project, that motivation appears in the form of a discriminator-style judge inspired by GANs (Goodfellow et al., 2014), rather than as a claim that any single static metric is sufficient on its own.

A key challenge with evaluating style is that it is inherently multi-dimensional. Optimisation frameworks can underperform when their evaluation metrics diverge from human judgements: In ASTRAPOP, optimisation quality is primarily constrained by the adequacy of the task-

specific style reward model (e.g., authorship representations for individual transfer; a classifier for community/native-language transfer) and the reward design used to define toward/away objectives; the DPO/CPO variants additionally depend on how candidate generations are sampled and compared under these reward signals, while STAMP's weighted proxy rewards can invite reward hacking or instability if mis-weighted (Liu et al., 2024a; Liu and May, 2025). Classical proxies such as semantic similarity using Sentence-BERT cosine similarity (Reimers and Gurevych, 2019) are useful diagnostics for meaning preservation. Fluency/grammaticality diagnostics are often implemented via acceptability-style classifiers; for English, a common option is a model fine-tuned on CoLA (Warstadt, Singh, and Bowman, 2019), which primarily measures sentence-level grammatical acceptability rather than discourse-level fluency. These fixed proxies often fail to capture global stylistic hallmarks such as rhythm, sentence variety, and discourse structure. Optimising for high fluency when the source author may not be perfectly fluent might be misguided as well.

Discriminator models provide a complementary signal by estimating a fooled rate, the tendency of a classifier to label generated text as real. When calibrated and threshold-tuned, they can be informative, yet they remain vulnerable to reward hacking, domain shift, and data leakage, so careful deduplication, calibration, and threshold selection are necessary. Recent work on LLM-as-a-judge suggests promising correlations with human preferences and offers scalability for pairwise or rubric-based scoring, but it also reports biases such as position, verbosity, and self-enhancement bias, so prompt design and calibration remain important (Zheng et al., 2023).

Public, high-quality preference datasets for a specific author remain scarce. Using pseudo-parallel generation with strong LLMs appears to be one of the most direct ways to strip away tone during data processing. Likewise, the strongest current directions seem to pair a strong base model with iterative preference alignment, such as STAMP's multi-iteration PO training and ASTRAPOP's style-reward-guided policy optimisation. What continues to hinder performance is the quality of the evaluation metric used to select preference pairs from generated outputs for iterative training. Combining these components more carefully is the direction pursued in Phase 1 of this project.

The style replication system, however, only addresses *how* the author writes, not *what* they know and think. Generating content that is both stylistically authentic and informationally grounded requires a complementary knowledge representation, one that captures the author's

beliefs, values, and reasoning patterns. This motivates the second phase of the project and the literature reviewed in the sections that follow.

2.2 Graph-Based Memory for Language Models

Retrieval-augmented generation (RAG) gives a language model access to information outside its weights by retrieving relevant passages before it generates a response (Lewis et al., 2020). This simple retrieval-then-generate pattern improves factual grounding and reduces reliance on the model's parametric memory alone. Survey work shows that the field has moved beyond basic nearest-neighbour chunk lookup towards richer retrieval structures and more deliberate ways of organising context (Gao et al., 2023b). Within that broader shift, graph-based RAG uses explicit links between documents, entities, and concepts to support queries that require multi-hop reasoning or corpus-level synthesis rather than a single matching passage (Peng et al., 2026).

Flat chunk retrieval is often enough for narrow factual lookup, but it is much weaker when a question requires evidence to be combined across many documents or many parts of the same corpus (Tang and Yang, 2024). Edge et al. (2024) make this point clearly in GraphRAG. GraphRAG is motivated by the observation that naïve chunk-retrieval RAG often fails on global corpus-level questions that are better framed as query-focused summarisation (e.g., identifying themes across a dataset). GraphRAG builds an LLM-derived entity graph and precomputed community summaries, and it shows substantial improvements over naïve RAG for global sense-making questions in large corpora. For local questions whose answers are contained in a small number of passages, conventional passage retrieval remains a strong baseline; GraphRAG's contribution is primarily to make global questions tractable via community structure and summarisation. That local-versus-global retrieval pattern is directly relevant to this project, because the mental model pipeline also needs to answer both narrow questions about a single post and broader questions about the author's overall beliefs and recurring themes.

Jiménez Gutiérrez, Shu, Gu, Yasunaga, and Su (2024) propose HippoRAG, a neurobiologically inspired retrieval framework that combines knowledge graphs with associative ranking (Personalized PageRank) to support multi-hop retrieval. In multi-hop QA experiments, it outperforms several existing RAG baselines and can achieve comparable or better results than

some iterative retrieval approaches at substantially lower cost, suggesting that persistent graph structure can be a strong inductive bias for associative multi-hop recall in these settings.

Sarthi et al. (2024) propose RAPTOR, which recursively abstracts document clusters into summaries and builds a multi-level summary tree. Retrieval can then operate at different levels of abstraction depending on query breadth. The core insight is that a retrieval index should expose multiple levels of granularity, not only raw chunks. This aligns with the soft topic hierarchy and derived-principle layer in this project.

The personal knowledge graph survey by Skjæveland, Balog, Bernard, Łajewska, and Linjordet (2024) treats PKGs as structured information centred on one individual and the entities, attributes, and relations relevant to that person. That framing is useful here because the cognitive graph in this project is also person-centred: it is not a general document entity graph, but a representation built around one author's recurring ideas, claims, and preferences.

2.3 Knowledge Graph Construction

Hofer, Obraczka, Saeedi, Köpcke, and Rahm (2024) survey end-to-end knowledge graph construction pipelines, covering extraction, validation, integration, incremental updates, and quality assurance. Their characterisation of a KG pipeline as an ongoing workflow, rather than a one-shot extraction, maps directly onto WAKE (extraction and verification) and MERGE (integration routing). The emphasis on incremental updates is particularly relevant: the cognitive graph is designed to grow post-by-post rather than being rebuilt from scratch.

Choi and Jung (2025) organise KG work around extraction, learning, and evaluation. This decomposition mirrors the project structure: WAKE/MERGE build the graph, the generation system uses it, and the A/B harness evaluates the retrieval quality. Zhu et al. (2024) present an extensive empirical evaluation of LLMs for KG construction and reasoning tasks and discuss resulting opportunities and limitations, while also highlighting open issues in reliability and scalability that motivate the evidence-grounding design in the WAKE stage.

Temporal knowledge graphs treat graph facts as time-stamped and revisable. The survey by Cai et al. (2023) reviews completion methods for graphs where relations hold only within certain time intervals. An individual's beliefs do not remain static; the same idea may be expressed differently or with changed conviction across posts. This motivates the `TemporalVersion`

design in the cognitive graph, where each node's identity is durable but its articulation is dated, allowing the history of an idea to be preserved without overwriting current meaning.

2.4 Sleep-Inspired Memory Consolidation

The four-stage design of the mental model pipeline, WAKE, MERGE, NREM, and REM, is inspired by how the sleeping brain consolidates and reorganises memories.

Sleep-based memory consolidation is often described in active-systems terms: reactivation and replay during NREM slow-wave sleep supports redistribution and transformation of newly encoded representations towards more stable long-term cortical storage (Klinzing, Niethard, and Born, 2019). The evidence also suggests that NREM and REM can play distinct roles during offline memory processing. In the Neuron study highlighted by Olson (2025), non-REM sleep pushed hippocampal memory drift towards a recall-like state, whereas REM sleep counteracted that drift. This supports using sleep stages as an organising metaphor for different forms of offline reprocessing, while stopping short of any simple one-to-one mapping between REM and abstraction. This two-stage offline process is the conceptual inspiration for the pipeline's stage design: NREM clusters nodes into stable leaf topics, and REM derives generalised principles that abstract across them.

Complementary Learning Systems (CLS) theory (Kumaran, Hassabis, and McClelland, 2016) explains why intelligent agents need two memory systems on different timescales: fast binding for individual experiences (mimicked by WAKE and MERGE) and slow consolidation for structured, generalisable knowledge (mimicked by NREM and REM). CLS theory also motivates keeping the two systems largely separate, which maps to the adaptive consolidation triggers: NREM and REM run only when enough new material warrants reorganisation.

Part I

Style Replication

Chapter 3

System Design

This chapter explains the design criteria for Phase 1 before the implementation details are presented in the following chapter.

3.1 Data and Problem Framing

The process begins with data from the Blog Authorship Corpus. For a selected author, the text is first cleaned by removing artefacts such as `urlLink`, which appeared in many blog posts. Starting the project without cleaning this data led to severe model hallucinations in our initial training rounds, where the model repeated `urlLink` in the output as a form of reward hacking. An LLM is then used to paraphrase the original blog post, `source_blogtext`, into a stylistically neutral version, `neutral_blogtext`. This step of separating content from style is a foundational design choice, ensuring the style model learns to apply style to neutral semantic content rather than simply memorising topical cues.

3.2 Training Strategy

We have designed a multi-stage training strategy. An initial model, `sft-gpt-4.1-nano`, is trained using supervised fine-tuning on neutral-to-stylistic pairs, with OpenAI's `gpt-4.1-nano` as the base model. While this provides a strong foundational understanding of the target style, it is known to have limitations in capturing deeper stylistic structures, often optimising for token likelihood rather than perceived quality. This motivates the second stage: Direct Preference

Optimisation.

For the initial bootstrap DPO (v1) run, the preference signal is constructed by taking the author's real blog text as the preferred output and the SFT model rewrite as the non-preferred output. This gives the model a direct signal about the difference between its own rewrite and the authentic target style.

The iterative loop that follows (v2 $\rightarrow$ v3 $\rightarrow$ v4) shifts to preference optimisation over generated outputs alone. At each round, multiple rewrites are over-generated per example and ranked by a combined metric of fluency, meaning preservation, and discriminator confidence. The highest-scoring rewrite is used as the preferred output and the lowest-scoring rewrite as the rejected output, forming a hard-negative pair that forces the model to distinguish its best output from its worst.

A structural limitation of this approach is that neither side of any preference pair is the author's actual writing. The model learns to prefer one of its own outputs over another, so the training target is always defined relative to the model's current output distribution rather than the author's true style. This creates a risk of the loop converging towards a self-consistent optimum that scores well under the combined metric but has drifted from how the author actually writes.

To mitigate this drift risk, we also evaluated a *source-preference* variant in which the preferred side of every DPO pair is replaced with the original author text rather than the model's best rewrite, while the rejected side remains the lowest-scoring generated output. The corresponding risk is that anchoring every training example to real author text could cause the model to memorise specific passages rather than acquire transferable style, a concern addressed empirically in the evaluation chapter.

Table 3.1 summarises the preference pair construction across all three training configurations:

The design hypothesis for the standard iterative loop is that self-referential training (i.e., learning to prefer the model's own better outputs) produces a more generalisable style representation, since the model is never shown the same real text twice and must instead learn the underlying stylistic pattern. The hypothesis for the source-preference variant is that fully anchoring the preferred side to real author text provides a stronger signal towards authentic style, at the cost of removing this self-improvement dynamic. Both variants share the same discriminator retraining schedule, so any performance difference is attributable solely to the pair construction method.

Variant	Preferred output	Rejected output
Bootstrap DPO (v1)	Real author text	SFT model rewrite
Iterative DPO (v2+)	Highest combined-score rewrite (∼85%)	Lowest combined-score rewrite
	Real author text (∼15%, anchor)	Model's best rewrite
Source-preference DPO (v2+)	Real author text (100%)	Lowest combined-score rewrite

Table 3.1: Preference pair construction across all DPO training stages. Bootstrap DPO (v1) is a one-time initialisation step; Iterative DPO and Source-preference DPO are the two ongoing variants from v2 onwards. The second row of Iterative DPO (v2+) is a 15%-probability per-example augmentation, not a separate training mode.

3.3 Evaluation Criteria

I propose a modified evaluation framework that combines several metrics into a single score. This approach follows the general multi-objective reward philosophy used in recent style-transfer work, but is adapted for the task of a single author's writing style. The components are:

- **Fluency.** A fixed metric using a RoBERTa-based CoLA model (Warstadt et al., 2019), with a modification inspired by critiques of automatic metrics in style transfer research (Krishna et al., 2020). Since human authors are imperfect, the score is grounded in the author's own fluency level:

$$1 - |\text{fluency}_{\text{author}} - \text{fluency}_{\text{rewrite}}|.$$

- **Meaning.** A fixed metric using OpenAI embeddings and cosine similarity between the rewrite and the original blog text, to preserve semantic content.

- **Discriminator Confidence.** A dynamic score from a deep neural network classifier trained to distinguish the author's real text from generated rewrites, following the same broad intuition as style discriminators used in adversarial or policy-guided authorship transfer work (Liu et al., 2024a).

The final score is a simple product of these signals and is used directly for ranking candidates during training. When reporting evaluation scores, the geometric mean (the cube root of the product) is used instead, so that component scores are on a comparable scale and a single near-zero component is clearly visible rather than masked by the others.

3.4 Design for Style Replication

The self-supervised loop was run for both the standard and source-preference variants. The design goal is to alternate between improving the rewriter and improving the discriminator until the generated text becomes sufficiently difficult to distinguish from the author's real writing, approaching a theoretical equilibrium where the discriminator makes random guesses. However, due to time and cost limitations, we only managed to develop up to v4.

Chapter 4

Implementation

4.1 Data Preprocessing Pipeline

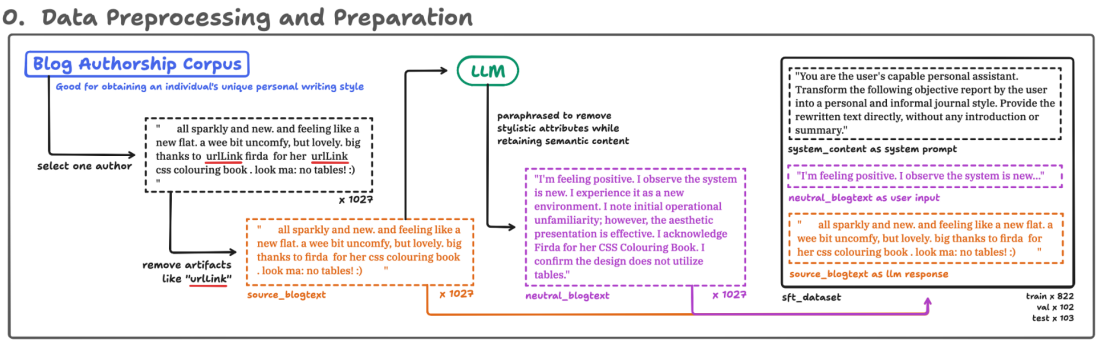


Figure 4.1: Data Preprocessing and Preparation workflow.

The preprocessing pipeline cleans the selected author's posts, removes artefacts such as the literal token `urlLink`, and uses an LLM to paraphrase each post into a stylistically neutral form. A system prompt is then put together with both the neutralised input and the original post to create the supervised fine-tuning dataset.

The neutralisation step uses `gpt-4.1-mini-2025-04-14` at temperature 0.3. The low temperature ensures consistent, near-deterministic paraphrasing across posts. This matters because inconsistent neutralisation would introduce random variation into the neutral inputs that should be absent: training pairs would then differ in task difficulty due to neutralisation noise rather than genuine style differences, corrupting the training signal.

The neutralised posts are split into train, validation, and test sets at an 80/10/10 ratio with a fixed random seed, applied consistently across all three authors. The fixed seed guarantees that train, validation, and test membership is identical across all model variants trained on the same author, so evaluation comparisons are never confounded by differences in which examples each model saw during training.

4.2 Initial Training: SFT and DPO Bootstrap

The initial training proceeds in three sequential sub-stages covered in this section: supervised fine-tuning to establish surface style, preference pair construction to supply a quality signal, and an initial DPO bootstrap run that seeds the self-supervised iterative loop described in Sections 4.4 and 4.5. Both the standard (preferred pairs are fully-generated) and source-preference (preferred output is original author text) variants share this initial training path; they diverge only in how preference pairs are constructed during the iterative rounds.

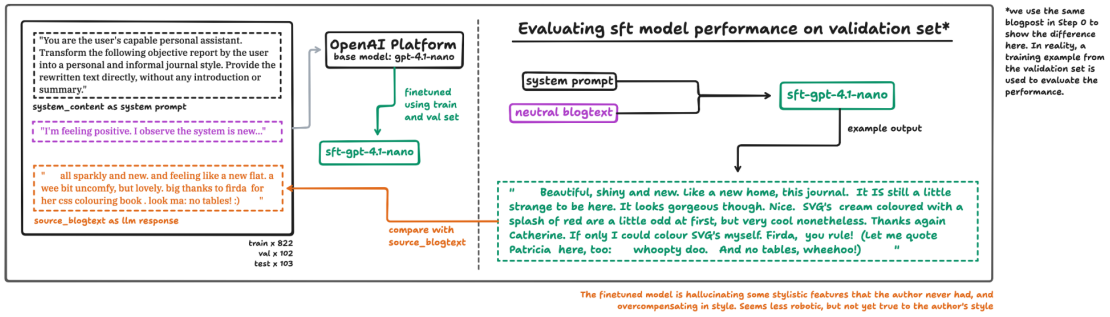


Figure 4.2: Supervised Fine-Tuning workflow.

The SFT stage (Figure 4.2) first fine-tunes `gpt-4.1-nano` on neutral-to-stylistic pairs, producing a model that understands the author's surface style but lacks an explicit signal about which of its outputs is stylistically preferred.

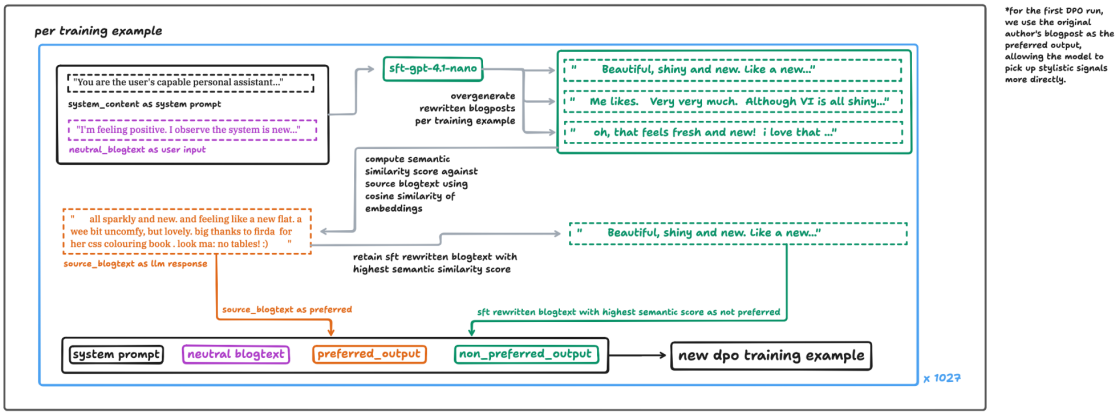


Figure 4.3: Data Preparation for DPO Training workflow.

The DPO data preparation stage (Figure 4.3) then constructs the preference pairs that supply this missing signal: for the bootstrap round, the real author text is preferred and the SFT rewrite is non-preferred, giving the model a direct reference to the authentic target style before the self-supervised loop begins.

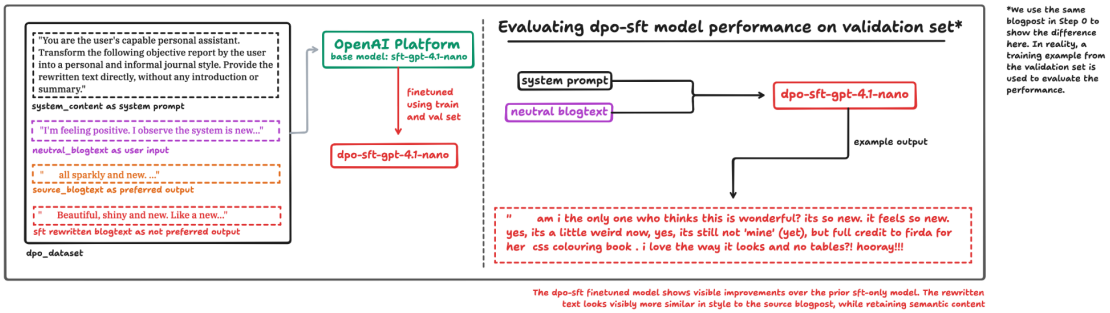


Figure 4.4: DPO Initial Fine-Tuning workflow.

The initial DPO run (Figure 4.4) fine-tunes on top of the SFT checkpoint on these bootstrap pairs, producing `dpo_sft_v1`, the starting point for both the standard and source-preference iterative variants.

4.3 Evaluation and Discriminator Pipeline

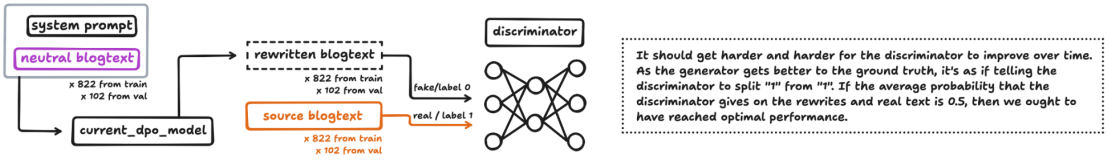


Figure 4.5: Discriminator Training.

The discriminator is a three-layer MLP over `text-embedding-3-small` embeddings, with batch normalisation, dropout, and a sigmoid output producing $P(\text{real})$. The output represents the probability that a given text was written by the target author. It is trained with AdamW and early stopping; a conservative patience setting is used because it is retrained from scratch at each iterative round on an incrementally larger set of rewriter outputs. At each round, the positive class consists of real author texts and the negative class consists of outputs from the *current* rewriter model, so as the rewriter improves the discriminator is forced to adapt to progressively harder negatives thereby raising the style bar each iteration.

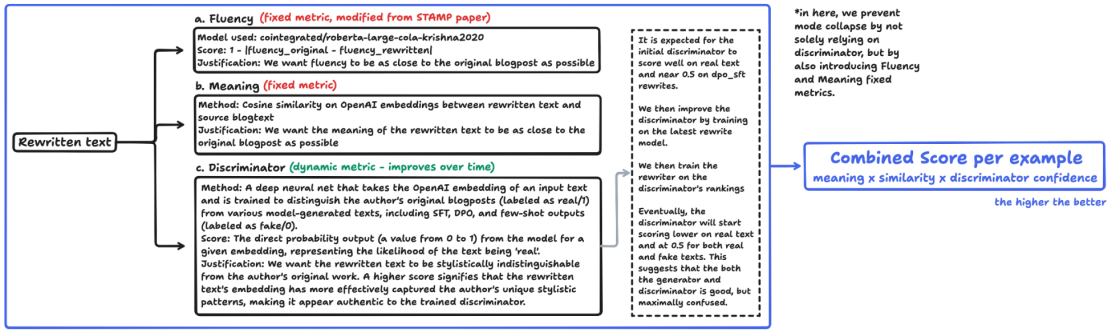


Figure 4.6: Evaluation metrics workflow.

This discriminator score serves as one of three components in the combined metric used to rank generated candidates (Figure 4.6). The other two are fluency preservation and meaning preservation. Fluency preservation is computed as $1 - |\text{fluency}(\text{source}) - \text{fluency}(\text{rewrite})|$ using a RoBERTa model fine-tuned on CoLA; crucially, this measures whether the rewrite *matches the author's characteristic fluency level* rather than maximising grammaticality which

is an important distinction for blog text where intentional informality is part of the target style. Meaning preservation is the cosine similarity of `text-embedding-3-small` embeddings between the source and rewrite, floored at zero (cosine similarity is naturally in $[-1, 1]$; negative values, indicating semantically opposite text, are treated as zero). All three components are in $[0, 1]$. Candidates are ranked by their raw product:

$$\text{score} = \text{fluency} \times \text{meaning} \times \text{style}$$

Multiplying rather than averaging is a deliberate design choice: any single zero drives the whole score to zero, enforcing a hard floor against candidates that fail on even one dimension. When reporting scores in Chapter 5, the product is expressed as its geometric mean $(\text{score})^{1/3}$ to restore the $[0, 1]$ scale and make component-wise comparisons more interpretable; since ranking by the product and by its cube root are equivalent, this does not affect which candidates are selected.

4.4 Self-Supervised DPO Loop

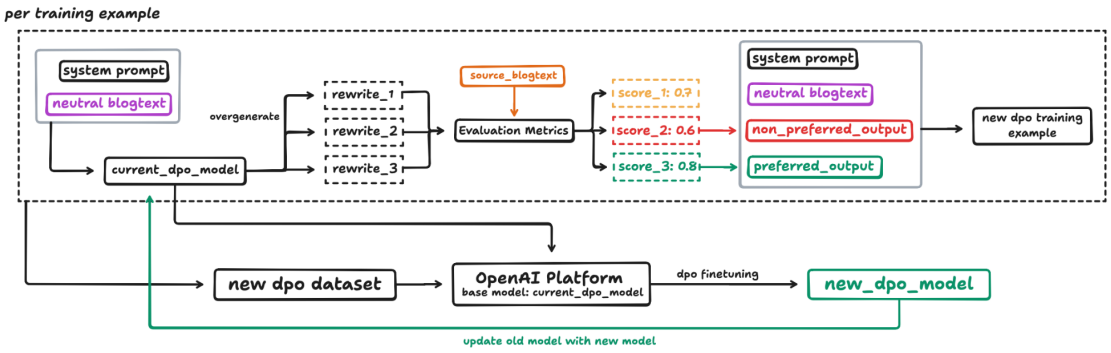


Figure 4.7: Self-Supervised DPO Fine-Tuning workflow.

The broader algorithm is an iterative self-supervised DPO loop. The current DPO model over-generates multiple rewrite candidates, the combined metric scores those candidates, and new preference pairs are formed. The preferred output is the highest combined-score rewrite. For the rejected output, a quality filter is applied: only candidates with meaning similarity above 0.85 and fluency preservation above 0.85 relative to the original blogpost are eligible. This filter

ensures the rejected candidate fails specifically on *style*, not on content or fluency—without it, a disfluent or incoherent rewrite could end up as the non-preferred output, and the DPO signal would conflate stylistic failure with general output quality, obscuring what the model should actually learn to avoid. Among those eligible candidates, the one with the lowest discriminator score is chosen: a hard negative that is semantically coherent and fluent but fails the style check. If no candidate passes both thresholds, the quality filter is waived and the candidate with the lowest combined score is used as the rejected output instead. To keep the real author's voice as an anchor, each training split is augmented with a 15% chance of inserting an additional source-grounded pair per example: the real author text is placed in the preferred position and the model's highest-scoring rewrite in the rejected position. This augmentation applies only to the standard route's training split.

The rewriter and discriminator are then alternated: the discriminator is frozen while the rewriter improves, then the rewriter is frozen while the discriminator is retrained on the current rewriter's outputs versus real author text. The loop continues until the discriminator can no longer reliably distinguish the generated text from the author's real writing.

4.5 Source-Preference Variant

The source-preference variant runs the same iterative loop as Section 4.4; The sole difference lies in how preference pairs are constructed. Every pair uses the **real author text as the preferred output** and the model's lowest combined-score rewrite as the rejected output, as defined in Table 3.1 in Chapter 3. Keeping the discriminator retraining schedule identical across both variants is a deliberate controlled variable: any performance difference is then attributable solely to the pair construction method, not to differences in how the discriminator evolved.

4.6 Training Scope and End-to-End Example

The full pipeline was executed for three authors from the Blog Authorship Corpus (IDs 216413, 595404, and 1107146). For each author, the implementation trains an SFT model on neutral-to-stylistic pairs, bootstraps a `dpo_sft_v1` model, then runs three iterative refinement rounds (v2, v3, v4) for both the standard and source-preference variants. Table 4.1 shows the training split

sizes after the 80/10/10 split.

Author ID	Train pairs	Val pairs	Test pairs
216413	888	111	111
595404	1,248	156	156
1107146	1,786	223	223

Table 4.1: Per-author training, validation, and test split sizes after 80/10/10 split.

Table 4.2 traces the effect of each training stage on the same neutral input for author 216413, from the raw SFT rewrite through to the final iterative DPO outputs.

Stage	**Output**
Neutral input	"A person who demonstrates patience and attention when I am in this state suggests a deep level of care or affection. This recognition is a source of considerable positive sentiment for me."
`sft-gpt-4.1-nano`	"the fact that the one i love can give me patience and attention when i am like this, and not run as far away as possible, makes my heart melt every single time. so fucking much."
`dpo-sft-gpt-4.1-nano` (v1, bootstrap)	"Someone who dared to show patience, care and attention before I had said a word shows me what they really feel for me. That makes me go all fuzzy inside, just like the coffee I am about to drink now."
`dpo-sft-gpt-4.1-nano` (v4, source-pref)	"If someone is that patient and attentive with me in this state, they must love me. And that makes me happy. Very, very happy. :)"
`dpo-sft-gpt-4.1-nano` (v4, standard)	"I think someone who has this much patience and attention when I am like this, must really, really care about me. That makes me go all warm and fuzzy."
Real author text	"Someone who bears with me and listens to me when I am like this, truly must love me. Now that sure is something that lights up my day."

Table 4.2: End-to-end rewrite progression for author 216413 across training stages.

Chapter 5

Evaluation

5.1 Style Transfer Findings

Qualitative inspection of the end-to-end rewrite progression like that in Table 4.2 per author per model suggests that successive training stages produce increasingly author-proximate outputs: the SFT rewrite captures surface style but lacks the author's sentence rhythm; the bootstrap DPO model introduces more idiomatic phrasing; and the final iterative DPO outputs show closer register and structural alignment with the real author text. However, this observation is informal and cannot substitute for a systematic evaluation.

The combined metric, the product of fluency preservation, meaning preservation, and discriminator score, was designed as an in-loop selection signal rather than an evaluation instrument, and two structural properties prevent it from filling that role. First, each model variant's discriminator is trained exclusively on that variant's own outputs: the iterative DPO discriminator never sees source-preference rewrites as negatives, and vice versa. Cross-variant score comparisons are therefore between classifiers trained on different data, making them meaningless as a common measure of quality. Second, across training rounds within a single variant, the discriminator is retrained on progressively harder negatives as the rewriter improves; the probability scale shifts each round, so a score of 0.7 at iteration 1 and a score of 0.7 at iteration 4 reflect different difficulty regimes and cannot be directly compared.

A reliable evaluation therefore requires a fixed instrument applied simultaneously to all systems. The blinded human study described below uses the same respondents, the same

ranking task, and the same criteria across all model versions, providing the stable and directly comparable measure of author-likeness that the automated metric cannot supply.

5.2 Blinded Human Evaluation

To provide this fixed-instrument comparison, I ran a blinded ranking study with ten respondents (students aged 22–26 who use AI tools on a daily basis), each completing nine ranking tasks. Nine prompts spanned three authors, covering short formula-driven messages and longer content-dense personal writing. Each task presented five anonymised candidates alongside three reference excerpts: a held-out real-author excerpt, a GPT-5.1 few-shot rewrite, the bootstrap DPO model (v1), the main iterative DPO model (v4), and the source-preference variant (v4), where every DPO round uses real author text as preferred rather than the highest-scoring rewrite (§3.2, §4.5). Respondents ranked candidates from most to least likely to be the author. Figure 5.1 shows an example survey question as seen by respondents.

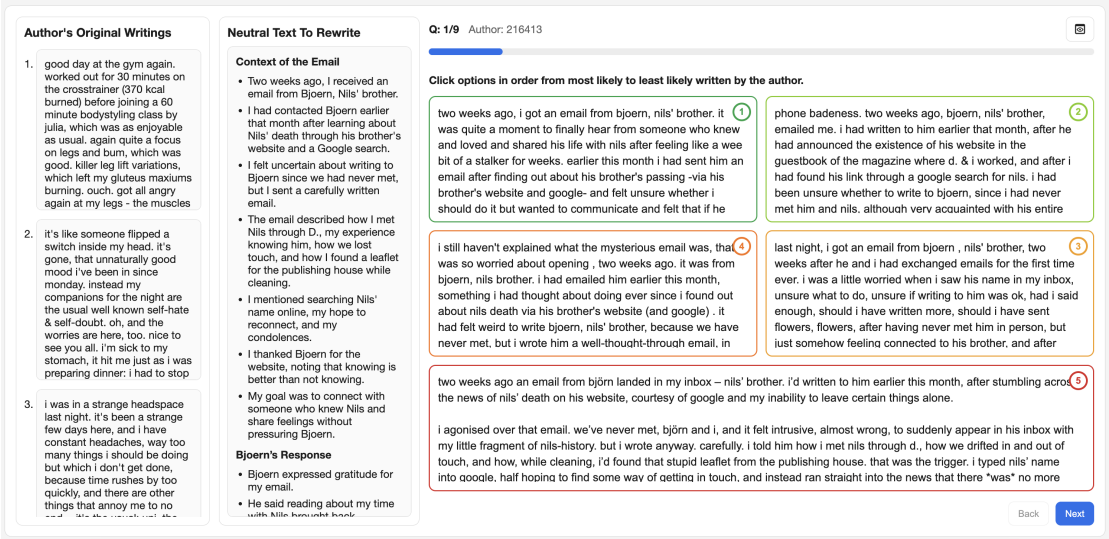


Figure 5.1: Example survey question. Respondents saw three reference excerpts from the author, a neutral text with the same semantic content, alongside five anonymised candidates to rank in order, with 1 being most likely author.

Table 5.1 summarises the headline results. The real-author excerpt ranked highest overall, serving as the reference ceiling: a system that matched real-author rankings exactly would represent perfect style transfer given the three reference excerpts provided to raters. Among

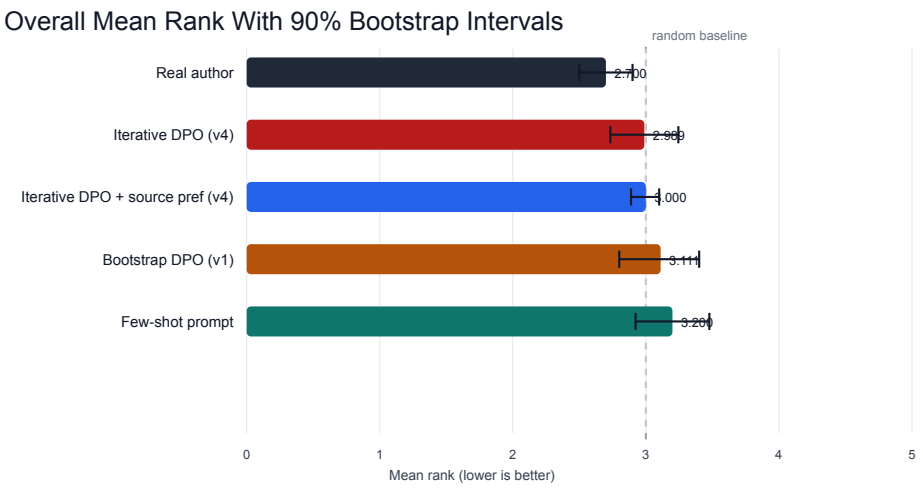


Figure 5.2: Overall human-evaluation ranking by mean rank, with 90% respondent-bootstrap confidence intervals. Lower is better.

generated systems, the iterative DPO model came closest, achieving the strongest mean rank and the highest top-3 rate (amongst generated systems).

System	Mean Rank	Top-3 Rate
Real Author	2.700	70.0%
Iterative DPO (v4)	2.989	64.4%
Iterative DPO + source pref (v4)	3.000	56.7%
Bootstrap DPO (v1)	3.111	56.7%
Few-Shot Prompt (GPT-5.1)	3.200	52.2%

Table 5.1: Human-evaluation results by system. Lower mean rank is better.

The headline ranking differences between generated systems were modest. The real-author excerpt beat few-shot prompting with a pairwise win rate of 62.2% [52.2%, 72.2%] and beat the main iterative DPO model at 57.8% [52.2%, 62.2%], but most pairwise differences among generated systems were not decisive. Average Kendall's W across prompts was 0.177, indicating low inter-rater agreement; this is expected given the subjectivity of the task, as respondents receive only three reference excerpts per author and likely weight different stylistic features such as sentence rhythm, vocabulary, discourse structure differently, producing divergent but individually coherent rankings. With only ten respondents, bootstrap confidence intervals capture

sampling variance over raters but not the full disagreement structure; this limits the precision of aggregate rankings beyond what the CIs alone convey, a point discussed further in the limitations (§10.2).

One hypothesis, informed by post-hoc inspection of model outputs, is that rankings were driven more by deeper features such as discourse-role fidelity, genre, stance, and content control than by surface quirks such as ellipses or dashes; rater motivations were not directly elicited so this cannot be confirmed (Appendix A.1). This is tentatively visible in a prompt-type split between the two iterative variants (Table 5.2): Iterative DPO leads on short, formula-driven prompts while source-preference leads on longer, content-dense ones.

Prompt Type	Iter. DPO	Src-Pref	Mean Δ [90% CI]
Short, informal (n=5)	**2.68**	3.30	-0.62 $[-0.88, -0.34]$
Long, opinion-style (n=4)	3.38	**2.63**	$+0.75$ $[+0.20, +1.28]$

Table 5.2: Mean rank by prompt type for the two iterative variants. Δ is Iter. DPO minus Src-Pref; negative Δ favours Iterative DPO. CIs are 90% respondent-bootstrap. Lower mean rank is better.

The prompt-type split is uneven across survey questions per author, but the directional split is consistent across raters: the 90% bootstrap CI on the mean rank delta excludes zero in both buckets (-0.62 for short, $+0.75$ for long), so neither result is an artefact of a single outlier prompt or rater. The two variants occupy complementary niches. The main iterative DPO model is a *style-forward* model: its model-vs-model signal is more willing to depart from the neutral source to reconstruct an author-like voice, benefiting short, voice-dominant prompts but introducing content-control risk on longer passages; this boldness also accounts for the wider confidence interval in Figure 5.2 relative to the source-preference variant. The source-preference variant is a *source-anchored* model: direct grounding in real author text preserves discourse structure and content fidelity more conservatively, at the cost of occasionally becoming over-literal. Neither is a strictly weaker version of the other; they occupy different points on a style–content tradeoff. This is visible in the following example (author 216413, test audit):

Original blogpost: “and yes, I hate being so needy & whiny & plain-old dumb & non-me.”

`dpo_sft` **aka iterative DPO (v4):** “I hate being that needy. I hate being that whiny. I hate being that stupid. I hate being that unlike myself.”

`dpo_sft_source_pref` **aka source-preference DPO (v4):** “And I hate feeling needy. And whiny. And stupid. And not myself.”

The iterative DPO model reconstructs the content with a repetition-heavy stylistic flourish; the source-preference model produces a compressed, source-skeleton-preserving paraphrase. Neither is wrong; they reflect different optimisation targets.[1]

Cross-author performance variation is explained by two compounding factors: training corpus size (Table 4.1: 216413 had the fewest pairs at 888, versus 1,786 for 1107146), and the systematic failure of all generated systems to reproduce certain orthographic and lexical idiolect features specific to each author (Table 5.3).

[1]Token-level similarity checks on the full held-out test split confirmed no evidence of memorisation; the difference is more conservative rewrites rather than copying collapse.

Author	Distinctive idiolect features	Reproduction outcome
216413	Strict lowercase (incl. “i”), dashes and parenthetical asides, confessional register, exasperated interjections (“weeeh”, “Ick”)	Lowercase partially reproduced; interjections inconsistently reproduced across systems
595404	Elongated ellipsis chains (“. . . . . ”), strict lowercase, third-person self-reference (“mean mamma”, “mm”)	Third-person self-reference (“mean mamma”/“mm”) and elongated ellipsis chains reproduced by source-preference, bootstrap DPO, and iterative DPO variants
1107146	All-caps for emotional peaks (e.g., “SADNESS!”, “MANY TEARS ARE SHED”)	Reproduced for formulaic peaks, but coherent formatting and contextual usage is largely restricted to the iterative DPO (v4) and source-preference (v4) variants

Table 5.3: Per-author distinctive idiolect features and their reproduction by generated systems.

Table 5.4 summarises the key qualitative tendencies, with more in-depth analysis provided in Appendix A.1.

System	Key Qualitative Tendency
Few-Shot (GPT-5.1)	Reproduces surface markers (lowercase, asterisk emphasis) but injects contemporary internet slang mismatched to all three authors' blog-era vocabulary, frequent use of em-dashes and sentences are generally not rambling.
Iterative DPO (v4)	Strongest on author 1107146 (compact, casual voice); chains clauses beyond the author's natural sentence rhythm for 216413; subject to occasional semantic drift across authors (see Appendix A).
Source-Pref (v4)	Better discourse-structure fidelity on longer passages; occasionally over-literal.
Bootstrap DPO (v1)	Most volatile: highest variance across prompts and authors; sharpest semantic errors including confabulation.
All generated systems	Capable of reproducing all-caps emphasis used by author 1107146 for emotional peaks, but only iterative DPO and source-preference models deploy them with coherent contextual placement.

Table 5.4: Per-model qualitative tendencies from post-hoc survey analysis.

Taken together, the blinded study suggests the iterative DPO model is the strongest performing generated system overall, with a mean rank of 2.989 against the real-author ceiling of 2.700; the gap is directionally consistent but pairwise differences among generated systems are not decisively separated. The two iterative variants are not in a strict quality ordering but represent different points on a style–content tradeoff: the iterative DPO model is more willing to depart from the source to reconstruct an author-like voice, giving it an edge on short, voice-dominant prompts, while the source-preference variant preserves discourse structure and semantic fidelity more conservatively, giving it an edge on longer, content-dense ones. The principal remaining gap to real-author performance lies in reproducing deep orthographic and lexical idiolect features, elements that require either training data scale or generation-time constraints beyond what the current pipeline enforces (see §10.2).

Part II

Mental Model Construction

Chapter 6

System Design

This chapter explains the design criteria for Phase 2 before the implementation details are presented in the following chapter.

6.1 Cognitive Graph Data Model

The same blog corpus used in Phase 1 provides the input to the mental model pipeline. Each post is processed in chronological order, so the cognitive graph grows incrementally as the author's posts are ingested.

The second phase represents the author's mental model as a typed, temporal graph. Every distinct belief, value, principle, heuristic, opinion, or definition extracted from the author's posts becomes a node at one of six abstraction levels: IDENTITY, VALUE, PRINCIPLE, HEURISTIC, OPINION, and DEFINITION, ordered from most fundamental to most situational. This taxonomy runs from stable dispositional traits (IDENTITY, VALUE) through generalisable rules and decision-making patterns (PRINCIPLE, HEURISTIC) to transient situational stances (OPINION, DEFINITION), with each level mapping to a distinct retrieval role at generation time.

Each node (`CognitiveNode`) is a durable identity for one underlying idea. When the same idea recurs with changed nuance, a new `TemporalVersion` is appended rather than overwriting the existing node, preserving belief history. Typed edges are supported but sparse; generation relies primarily on topic membership, temporal versions, and REM-derived principles.

The separation of fast online ingestion (WAKE, MERGE) from slower offline consolidation

(NREM, REM) is motivated by Complementary Learning Systems theory (Kumaran et al., 2016): rapidly updating consolidated structure after every post would disrupt previously settled knowledge.

6.2 Four-Stage Pipeline Design

The extraction and consolidation process is organised into four stages named after sleep phases to reflect their role in memory consolidation (Klinzing et al., 2019; Kumaran et al., 2016). **WAKE** handles online ingestion: an LLM extracts candidate nodes from each post, and a second verification call discards any candidate not grounded in a direct quote. **MERGE** handles integration: each verified node's embedding is queried against the FAISS-backed (Johnson, Douze, and Jégou, 2021) ANN index, and a routing LLM decides whether to create a new node, append a `TemporalVersion` to an existing node, or add a typed edge to a related node. **NREM** handles offline topic organisation via UMAP dimensionality reduction, HDBSCAN clustering, noise reassignment, and Bron-Kerbosch maximal-clique hierarchy construction. **REM** derives principles: Stage A synthesises topic-local principles per leaf cluster; Stage B synthesises cross-topic meta-principles per parent group. Full implementation details are provided in Chapter 7.

6.3 Generation System Design

Content generation is query-driven. The author provides a topic or short notes (for example, "Life is about doing"), and this input serves as both the retrieval query and the subject of the generated output.

Retrieval combines a topic-scoped path with a global ANN path over all nodes; the retrieved context includes relevant nodes and higher-level principles derived during consolidation. Implementation details are in Chapter 7. The retrieved material is organised into a structured prompt so that generation is grounded in the author's stored ideas. The output of this generation step is a neutral-tone draft. This draft is then passed to the Phase 1 rewriter, which rewrites it into the author's personal voice.

6.4 Evaluation Design

The Phase 2 evaluation asks whether the graph-grounded retrieval pipeline surfaces more relevant source posts than a vanilla Google RAG baseline for the same queries.

The evaluation is intentionally post-level rather than chunk-level, because Google RAG's internal chunking and grounding mechanisms are provider-managed and opaque. Comparing at the post level allows both systems to be assessed on a shared and interpretable unit.

Two query families are designed to test different retrieval demands. *Narrow* queries are paraphrased from individual posts; the expected result is that post alone. These test whether the system can recover a directly relevant document from an indirect question. *Broad* queries are derived from Stage B cross-topic meta-principles; the expected results are the posts underlying those principles. These test whether the graph-grounded system can connect a synthesis-level question to the specific post evidence that supports it, which is the capability that flat RAG is expected to struggle with.

The evaluation set is intentionally weighted toward narrow queries (70 narrow, 30 broad). This conservative weighting reflects the expectation that a flat RAG baseline performs well on narrow queries; the heavy narrow majority penalises any underperformance by the mental model system and makes a broad-query advantage harder to dismiss as an artefact of the query mix. Construction methodology and metrics are described in the evaluation chapter.

Chapter 7

Implementation

7.1 Cognitive Graph Data Structures

The second phase represents the author's mental model as a persistent cognitive graph. Its stored state consists of nodes, edges, temporal versions, evidence records, and the ANN index used for retrieval.

CognitiveGraph is the top-level container, holding dictionaries of nodes and edges alongside a FAISS-backed ANN embedding index. *CognitiveNode* represents a durable concept identity: it stores a stable node ID, the current proposition text, the abstraction level (IDENTITY, VALUE, PRINCIPLE, HEURISTIC, OPINION, or DEFINITION), the topic assignment, and the current semantic embedding. Each node also holds an ordered list of temporal versions. A temporal version captures one dated articulation of the concept, recording the proposition text, the author's stance and conviction strength, the period during which this articulation was active, and a list of evidence records. Each evidence record links that version to a specific source post via a post ID, date, and direct quote from the post text. *CognitiveEdge* represents a typed, weighted relationship between two nodes, carrying a source ID, target ID, edge type, and confidence score.

This separation of the durable node identity from its time-indexed versions is designed to represent a changing belief as a timeline rather than a replacement. Semantic operations such as ANN retrieval, topic clustering, and generation retrieval use the latest active version embedding, while all historical versions remain available for inspection and evidence tracing. Figure 7.1

summarises these entities and their relationships in the implemented data model.

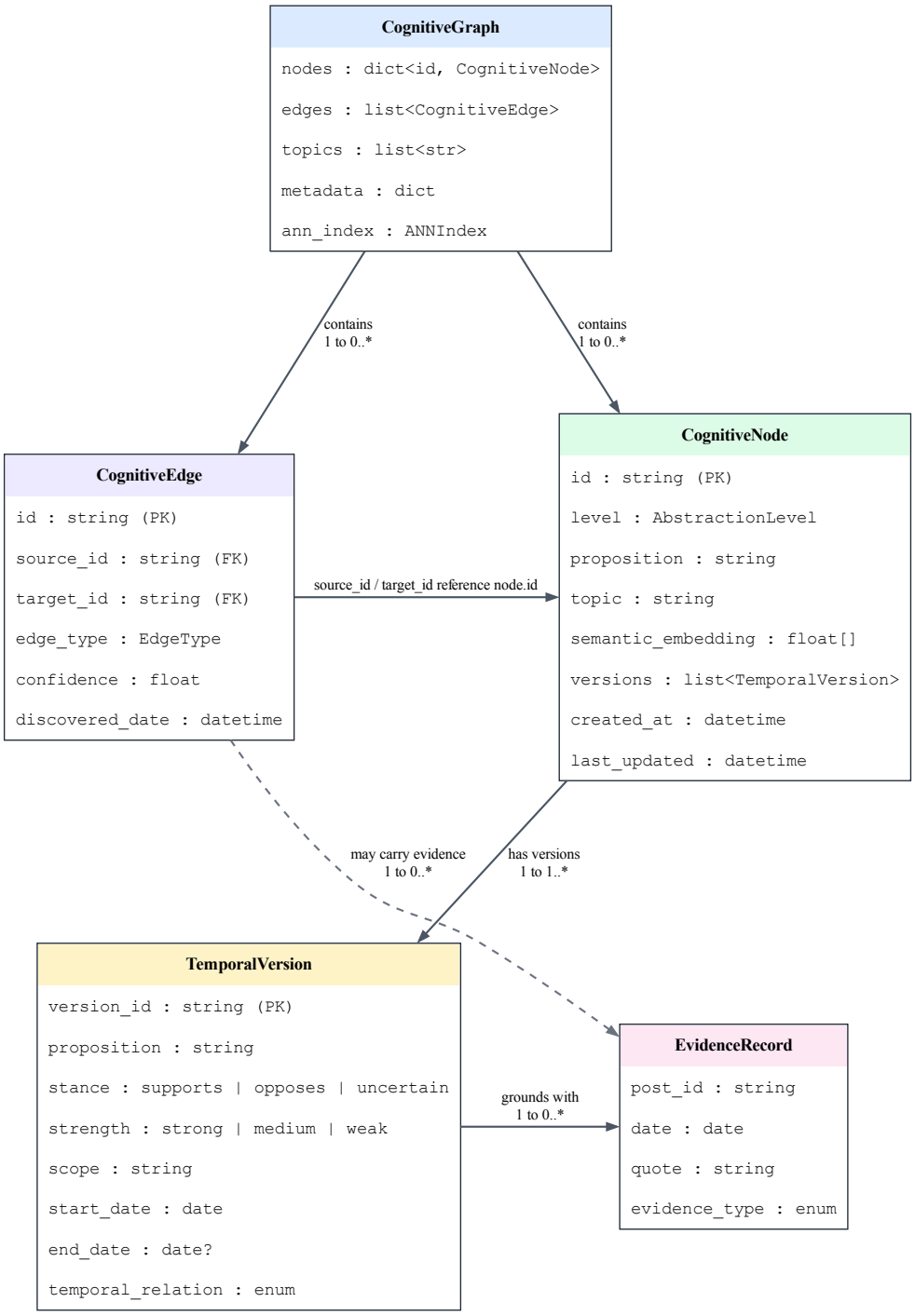
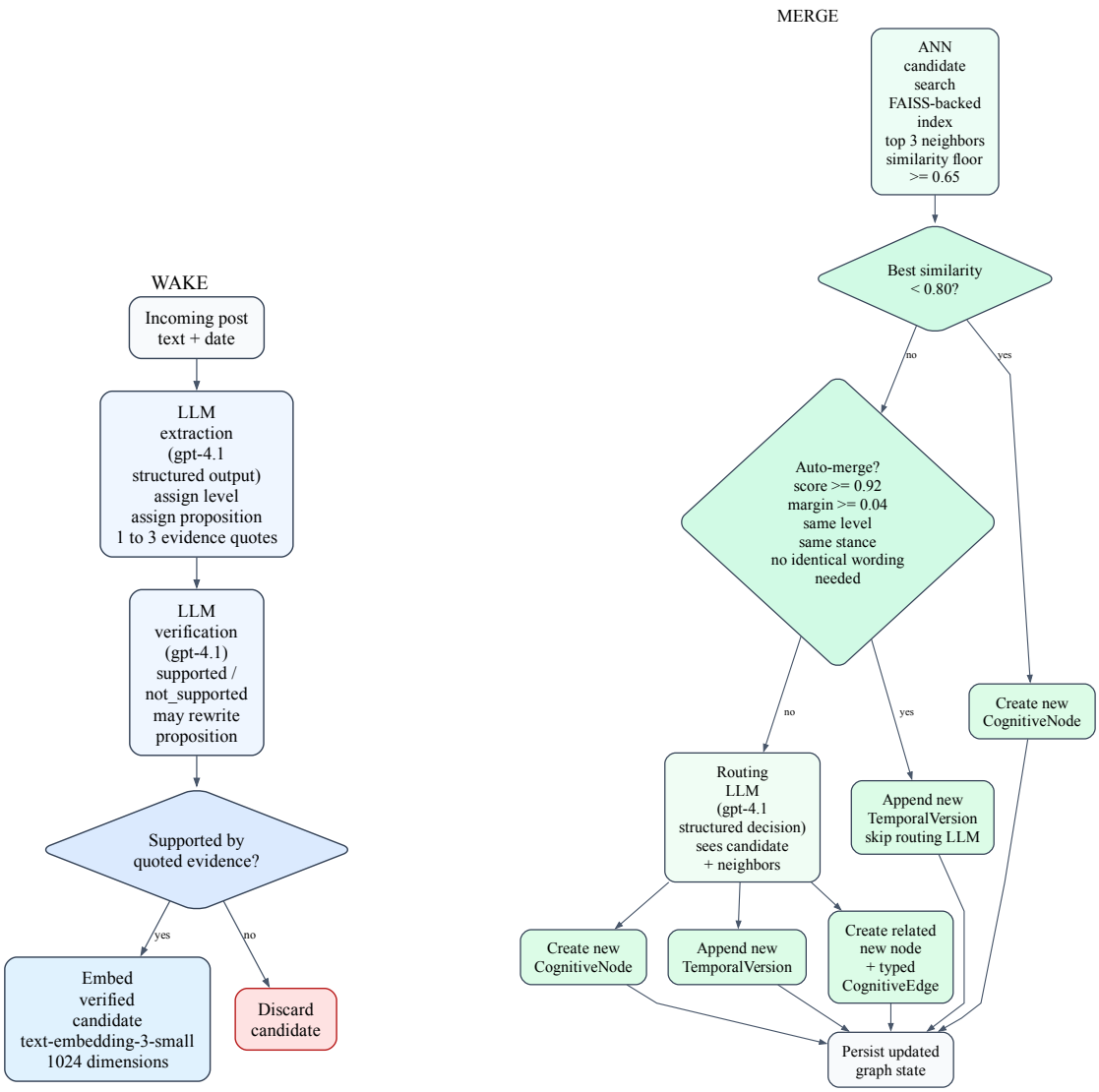


Figure 7.1: Entity–relationship diagram of the cognitive graph data model.

7.2 Extraction and Merge (WAKE + MERGE)



WAKE passes the verified candidate embedding into MERGE for ANN matching and routing.

Figure 7.2: WAKE extraction and MERGE routing workflow.

For each incoming post, `gpt-4.1` extracts candidate nodes via a structured output schema, assigning each to one of the six abstraction levels with at least one verbatim evidence quote. A second verification call labels each candidate as `supported` or `not_supported` and may rewrite imprecise propositions; unsupported candidates are discarded.

Verified candidates are embedded using `text-embedding-3-small` and queried against

the FAISS-backed ANN index. Candidates with no sufficiently close match are stored as new `CognitiveNode` instances. High-confidence matches that share the same abstraction level and stance polarity trigger an automatic `TemporalVersion` append without an LLM call. All remaining cases are routed through a structured LLM decision that chooses among three outcomes: create a new node, append a new `TemporalVersion`, or create a new node with a typed `CognitiveEdge` to the nearest related neighbour.

Figure 7.2 summarises this WAKE extraction and MERGE routing flow as a system workflow diagram. The exact WAKE extraction, WAKE verification, and MERGE routing prompts are reproduced in Appendix B.

7.3 Consolidation (NREM + REM)

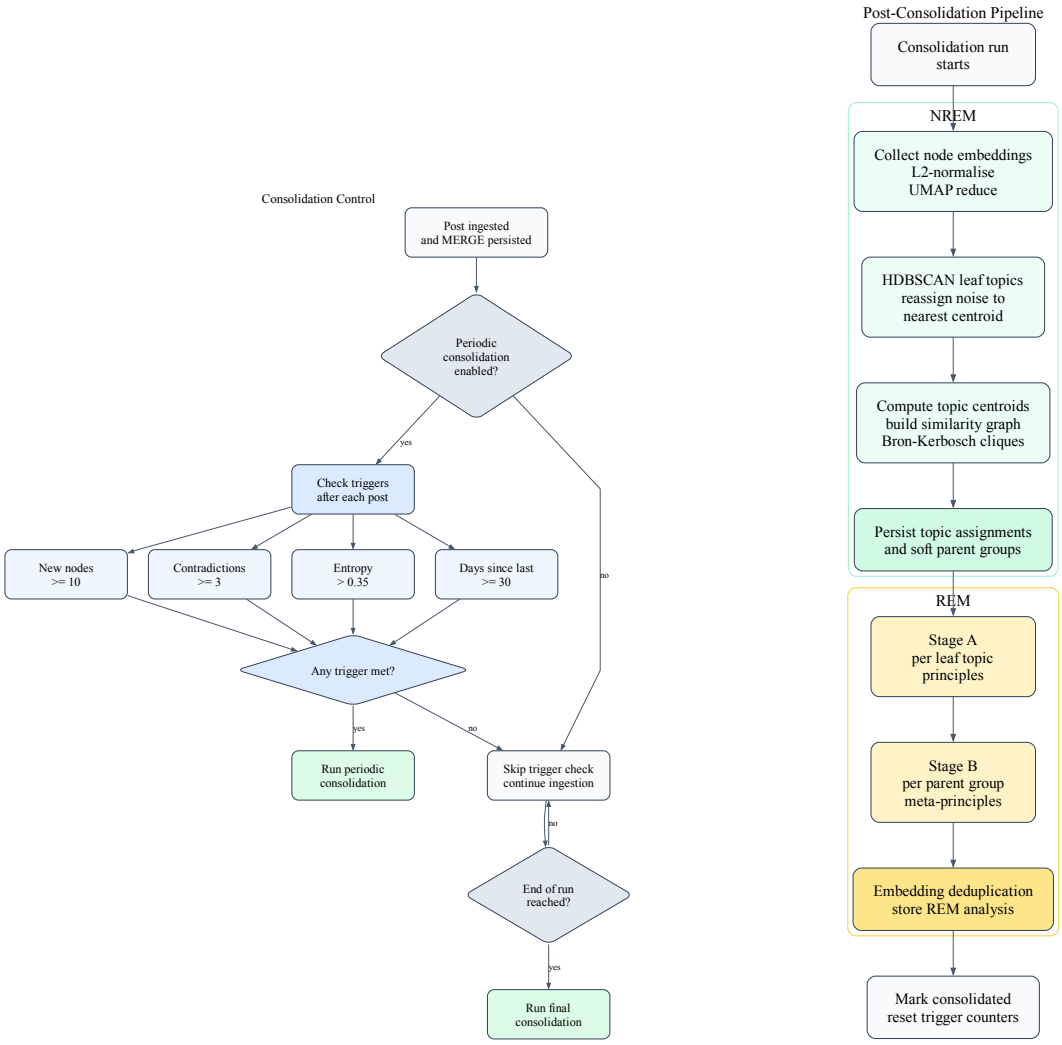
The pipeline supports adaptive consolidation triggers (node count, contradiction edges, semantic entropy (Kuhn, Gal, and Farquhar, 2023), and elapsed time), designed for continuously updated systems. For the full corpus build in this work, all posts were ingested before running consolidation once over the complete node set; this end-of-run mode is more efficient for large-batch processing.

In the NREM stage, node embeddings are L2-normalised and projected with UMAP before HDBSCAN clustering (McInnes, Healy, Saul, and Großberger, 2018; Healy and McInnes, 2024). This projection is necessary because HDBSCAN's density estimates degrade in high-dimensional spaces; reducing the embedding space first yields more meaningful cluster discovery. Noise-labelled nodes are then reassigned to their nearest cluster centroid so no node remains unassigned. A soft parent hierarchy is constructed by finding maximal cliques of topic-centroid similarities with the Bron-Kerbosch algorithm, allowing a topic to appear in multiple groups. Hyperparameter settings are listed in Appendix B.9.

The REM stage derives principles at two levels of abstraction. Stage A operates per leaf topic: the LLM synthesises concise principles from the propositions of all member nodes, each grounded in at least two supporting nodes drawn from that topic. Stage B operates per parent group: the Stage A principles from all member topics are assembled, and the LLM derives cross-topic meta-principles that must span at least two distinct topics, capturing broader patterns across the author's worldview. Because soft topic assignment can cause overlapping groups to

produce near-duplicate meta-principles, a final embedding-based deduplication pass removes any pair with cosine similarity above 0.80, retaining the principle that spans the most topics; ties are broken in favour of the earlier Stage A derivation. Prompt templates are reproduced in Appendix B.

Figure 7.3 summarises the trigger logic together with the NREM clustering, soft hierarchy construction, and REM principle-synthesis flow.



The left diagram shows when consolidation is triggered during ingestion, while the right diagram shows what happens once a consolidation run begins: NREM clusters nodes into topics and a soft hierarchy, then REM derives grounded topic-level and cross-topic principles.

Figure 7.3: Adaptive consolidation triggers with NREM topic clustering, soft hierarchy construction, and REM principle synthesis.

Figure 7.4 shows the resulting graph structure after a full consolidation run, with parent groups as large circles and leaf topics sized by node count.

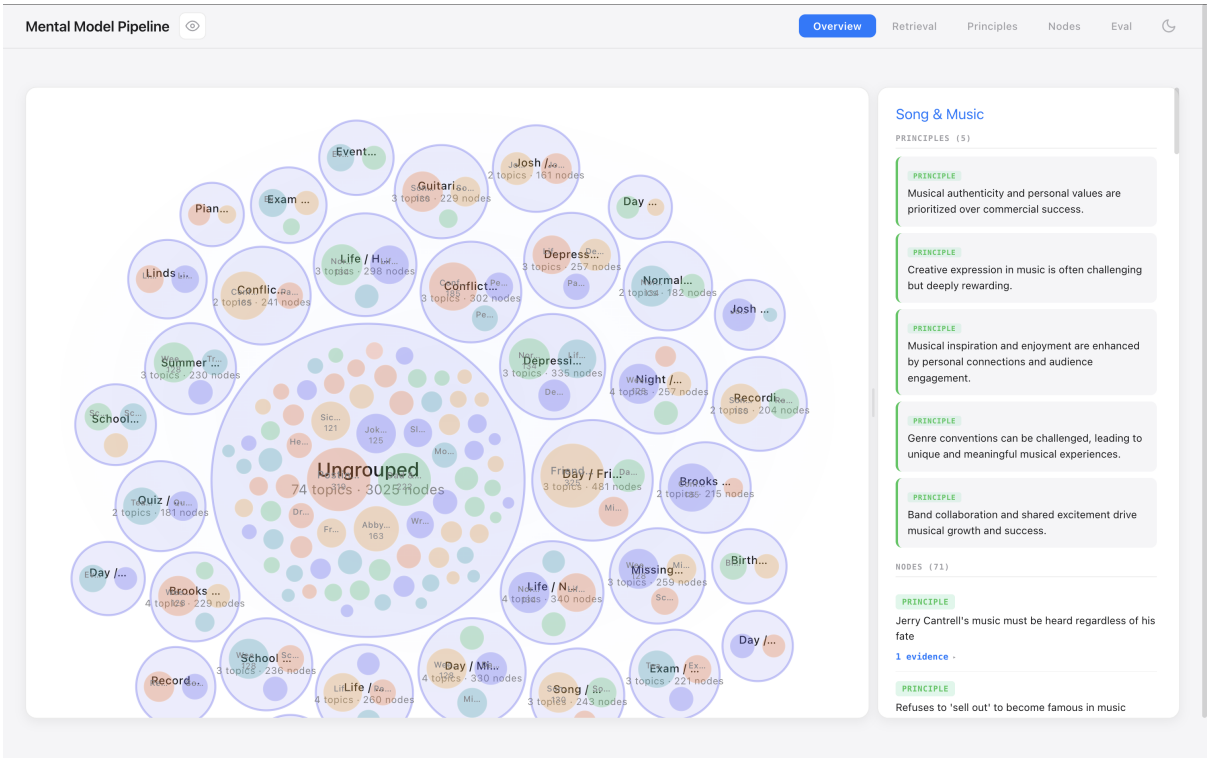


Figure 7.4: Circle-pack overview of the cognitive graph after consolidation. Each large circle is a parent group (cross-topic cluster); nested circles are leaf topics, sized by node count.

7.4 Graph-Grounded Generation

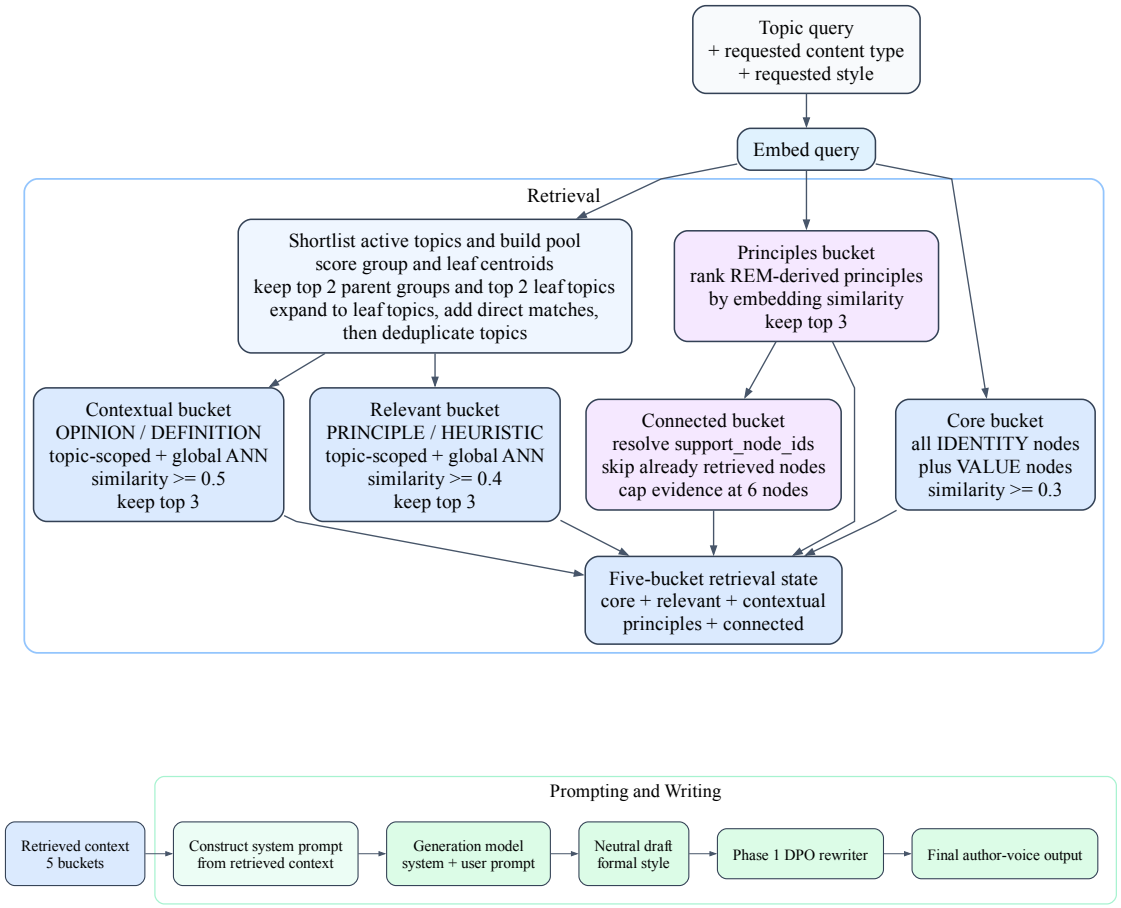


Figure 7.5: Graph-grounded generation retrieval flow. Top: retrieval builds the five context buckets. Bottom: the retrieved context is assembled into the system prompt, drafted in a neutral style, and rewritten into the author's voice.

Content generation takes a topic query and produces a neutral, information-focused draft that is then rewritten into the author's voice by the Phase 1 DPO model. The query embedding is first used to build a topic-scoped candidate pool by scoring parent-group and leaf-topic centroids against the query and expanding the highest-ranked matches to their member nodes.

Retrieval populates five buckets from this pool. The *core* bucket is drawn globally rather than from the topic shortlist, because IDENTITY and VALUE nodes encode author-level traits that should remain available regardless of the query topic. The *relevant* bucket holds PRINCIPLE

and HEURISTIC nodes, which capture the author's general rules and decision-making patterns. The *contextual* bucket holds OPINION and DEFINITION nodes, which provide concrete stances and local framing. Both *relevant* and *contextual* are populated by a dual-path strategy: a topic-scoped path for coherent neighbourhood coverage and a direct ANN path over all nodes. Results from both paths are score-ranked and deduplicated so the highest-confidence nodes survive regardless of which path found them. The *principles* bucket ranks REM-derived principles by query similarity independently of the node graph, surfacing cross-topic abstractions that may outrank any individual node. Their backing graph nodes are then resolved into a *connected* evidence bucket, grounding abstract principles in concrete retrieved nodes.

The assembled context is passed to the prompt builder, which constructs a structured system prompt covering the author's identity, relevant principles, derived principles, contextual stances, and writing guidelines, paired with a user prompt specifying the query. The generation model produces a neutral draft, which is then passed to the Phase 1 DPO rewriter to recover the author's personal voice. Generation and rewriting prompt templates are reproduced in Appendix B.6. Figure 7.5 summarises this workflow from query embedding through retrieval, prompt assembly, and voice rewriting.

7.5 Graph Scale and Composition

To characterise the system's output at operational scale, Table 7.1 summarises the cognitive graph built from processing 1,273 of the selected author's blog posts.

OPINION nodes dominate at 82% of the total, reflecting the corpus's personal, diary-like nature. Though fewer in count, VALUE and PRINCIPLE nodes carry disproportionate weight at generation time: VALUE nodes populate the core identity bucket and PRINCIPLE nodes anchor the relevant bucket. The 107 nodes with more than one temporal version (1.6%) confirm that MERGE correctly identifies the same underlying idea across different surface expressions rather than creating duplicate nodes. Of 127 leaf topics, 53 (42%) belong to at least one of the 39 parent groups; the remaining 74 are sufficiently distinct that no clique threshold is met. REM produced 777 Stage A topic-local principles and 116 Stage B cross-topic meta-principles, forming the higher-order retrieval layer used during generation.

Phase 2 was run on a single author's corpus. Each post requires multiple LLM calls for

Metric	Count
Posts processed	1,273
Total nodes	6,554
OPINION	5,373
HEURISTIC	615
PRINCIPLE	372
IDENTITY	96
VALUE	73
DEFINITION	25
Nodes with >1 version	107
Leaf topics (HDBSCAN)	127
Parent groups (cliques)	39
Ungrouped topics	74
Stage A principles	777
Stage B meta-principles	116

Table 7.1: Cognitive graph statistics after processing 1,273 blog posts.

extraction, verification, and routing, making multi-author runs prohibitively expensive within the project's scope; evaluation across a broader range of authors is therefore left as future work.

Chapter 8

Evaluation

8.1 Evaluation Setup

The evaluation compares two retrieval systems on a fixed constructed query set:

- **Baseline**: Google RAG via the Gemini File Search API. Post IDs are extracted from provider-returned grounding metadata. Chunk boundaries and retrieval internals are provider-managed.
- **Modified**: Graph-grounded retrieval from the mental model system. For each query, the evaluation reuses the same `prepare_generation()` routine used at generation time. This runs the graph retrieval stage, constructs the same retrieved context used for prompting, ranks source posts from that retrieval result, and returns the ranked `source_post_ids`; those post IDs are used as the system's result.

The evaluation set contains 100 queries: 70 *narrow* queries derived from individual posts (testing direct post recall from a paraphrased question), and 30 *broad* queries derived from Stage B cross-topic meta-principles. For the broad subset, each selected meta-principle is paraphrased into a natural user query, and its gold relevant posts are traced through the supporting Stage A principles and node evidence. This makes the broad subset a targeted stress test of cross-post synthesis rather than a random sample of user queries, which is appropriate because the comparison is meant to test whether the graph's higher-order abstractions improve retrieval beyond flat surface-level matching. Gold `relevant_post_ids` are fixed at construction time and do not change between runs. Queries were generated by `gpt-4.1-mini`. Evaluation is at

$k = 5$; metrics are Precision@5, Recall@5, and MRR. The methodological implications of the broad-query gold-label construction are discussed in Section 10.2.

8.2 Results

Table 8.1 summarises the A/B evaluation results stratified by query breadth.

Query subset	System	Precision@5	Recall@5	MRR
Narrow ($n = 70$)	Google RAG	0.186	0.929	0.701
Narrow ($n = 70$)	Mental Model	0.151	0.757	0.635
Broad ($n = 30$)	Google RAG	0.053	0.044	0.161
Broad ($n = 30$)	Mental Model	**0.127**	**0.106**	**0.296**
Overall ($n = 100$)	Google RAG	0.146	0.663	0.539
Overall ($n = 100$)	Mental Model	0.144	0.562	0.534

Table 8.1: A/B evaluation results at $k = 5$. Bold values indicate the better system for that metric.

8.3 Discussion

On narrow queries, Google RAG has a structural advantage. A narrow query is a paraphrase of the content within a single post; the post text itself is available to Google's grounding engine as a chunk. Provider-managed chunk retrieval with dense semantic matching is highly effective in this setting, achieving Recall@5 of 0.929. This high recall coincides with a Precision@5 of only 0.186, indicating that while the correct post is almost always present somewhere in the top 5, the retrieved set contains a substantial share of irrelevant results. In a RAG-grounded generation system, however, this is less of a concern than in classical IR: modern LLMs are effective at filtering irrelevant context from a short retrieved set, so high recall at modest precision is an acceptable operating point provided the retrieved content does not saturate the context window. The mental model pipeline, which retrieves at the node and principle level rather than the raw-text level, is at a disadvantage on narrow queries because it does not surface the verbatim content that flat chunk search can directly match. On broad queries, the mental model pipeline

substantially outperforms. Broad queries are derived from Stage B cross-topic meta-principles and expect post evidence that spans multiple areas of the author's worldview. Google RAG achieves a Recall@5 of only 0.044 on these queries, reflecting the fundamental limitation of flat retrieval when the answer depends on connecting content across thematically distant posts. The mental model pipeline achieves Precision@5 of 0.127 and MRR of 0.296. This corresponds to improvements of +139% and +84% respectively, because the hierarchy-routed retrieval and REM-derived principles explicitly encode cross-topic connections that would otherwise be invisible to a cosine-similarity search over chunks.

Per-query win counts confirm the direction. On the 30 broad queries, the mental model pipeline wins on Precision@5 in 11 cases against Google RAG's 2 (17 ties); on MRR it wins 9 against 5 (16 ties). On the 70 narrow queries, Google RAG holds an advantage, winning on Precision@5 in 13 cases against 1 for the mental model pipeline. However, the gap is less severe than the counts suggest: 56 of 70 narrow queries produce identical Precision@5 scores for both systems, in most cases because both retrieve at least one correct post within the top 5. This indicates substantial overlap in which narrow queries both systems partially solve, even though Google RAG remains the stronger narrow-query retriever overall.

The narrow-query deficit is compounded by the extraction model choice: node extraction and routing use `gpt-4.1`, a lower-cost model rather than the latest frontier models available at the time of Phase 2 development (see Section 10.2). Stronger extraction models would improve node fidelity and downstream retrieval precision; augmenting node retrieval with a lightweight raw-text fallback is a further straightforward path to narrowing this gap.

Overall results are dominated by the 70:30 narrow-to-broad ratio, leaving aggregate scores nearly equal. The cognitive graph adds the most value precisely where flat retrieval fails: synthesis-requiring queries that span an author's full body of knowledge.

Chapter 9

End-to-End Generation

The generation example below demonstrates the two phases working together. The cognitive graph supplies grounded context drawn from the author's full body of writing; the DPO rewriter then applies the author's personal voice to the resulting neutral draft.

The author provides the topic direction *"We should be true to ourselves instead of worrying about fitting in."* as the generation seed. The retrieval stage assembles context from the cognitive graph: 5 active topics, 3 derived meta-principles, 3 relevant principle nodes, and 4 connected support nodes from a graph built over 1,273 posts. The full eight-step retrieval trace as rendered by the web application is reproduced in Appendix B.10. The system feeds this context into a graph-grounded system prompt, and the LLM generates a structured, neutral draft:

```
**The Ineffectiveness of Conformity**
Conforming to societal or interpersonal expectations is an
unsustainable and corrosive practice.
**Normalcy** is overrated; attempting to fit a standard mold
prevents individuals from achieving
personal growth. Fulfillment arises when an individual aligns
their life with their own values
rather than external pressures. [...]
```

The full generated draft is reproduced in Appendix B.8.

This neutral draft is then passed to the Phase 1 DPO rewriter, which transforms it into the author's personal register. The final output is:

```
Note to everyone. At least, those that care about me and my
opinions.

Your damn fads are never going to get you anywhere. If you
think (not in just a simplistic way) that conforming, or
playing by the rules, or 'being normal' is going to get you
somewhere, you are sadly mistaken. You can't be normal.

Normal is overrated. And conforming to normal is a slow,
corrosive death. [...]
```

The full generated stylised-response and a few-shot style-transfer output using OpenAI `gpt-5.4` are also reproduced in Appendix B.8.

Together, the two phases produce an output that reflects both the author's characteristic register (captured by the Phase 1 style rewriter) and their grounded point of view on the topic, supplied by the Phase 2 cognitive retrieval system.

Chapter 10

Conclusions

10.1 Summary

This project has built a two-phase persona replication system that combines an author's personal voice model with a graph-based representation of the author's knowledge and beliefs.

The first phase constructed a style replication pipeline. Starting from cleaned author blog text, the pipeline neutralises source content, performs supervised fine-tuning, bootstraps a first DPO stage using the real author text as the preferred target, and extends the process into a self-supervised DPO loop guided by a combined metric of modified fluency, meaning preservation, and discriminator confidence. The combined metric proved useful as an in-loop selection signal, while a blinded human author-likeness study with 10 respondents and 90 complete rankings found that the best iterative DPO variant outperformed the GPT-5.1 few-shot baseline and achieved mean rank 2.989 against 2.700 for held-out real-author text, showing that the generated rewrites are competitive with authentic excerpts but still distinguishable overall.

The second phase constructed a mental model extraction and generation system. An incremental four-stage pipeline processes blog posts to build a persistent cognitive graph of the author's beliefs, values, principles, and opinions. WAKE extracts and evidence-verifies candidate nodes; MERGE routes each verified node into the graph using ANN-based similarity search; NREM clusters nodes into topics and builds a soft topic hierarchy via maximal cliques; REM derives topic-local and cross-topic meta-principles. Graph-grounded generation retrieves relevant nodes and principles in response to a query, produces a neutral draft, and passes it

to the Phase 1 DPO rewriter to stylise in the author's voice. An A/B evaluation against a Google RAG baseline shows that the mental model pipeline outperforms flat retrieval on the constructed broad-query benchmark by 139% on Precision@5 and 84% on MRR, suggesting that the structured graph representation adds retrieval value for synthesis-requiring queries.

10.2 Limitations

10.2.1 Phase 1: Style Replication

The completed blinded human study provides only partial validation of the combined metric for style evaluation. Real-author text still ranked first overall, most pairwise gaps among generated systems were inconclusive, and average Kendall's W across prompts was only 0.177. The metric is therefore useful for model selection, but still too noisy to treat as a full proxy for human preference, especially when comparing closely related DPO variants. A contributing factor to the low inter-rater agreement is the inherent tension in reference excerpt selection: providing too many excerpts risks overwhelming respondents and reducing ranking quality, yet the three excerpts shown per author may not be fully representative of each author's range of stylistic quirks, leaving respondents with an incomplete basis for judging author-likeness. This trade-off is difficult to resolve without a significantly larger and more structured elicitation study.

The self-supervised DPO loop is susceptible to reward hacking: the model may learn to exploit biases in the combined metric rather than truly improve stylistic replication. The GAN-inspired adversarial relationship between rewriter and discriminator also requires careful balancing; if the rewriter improves too quickly, discriminator feedback becomes uninformative.

The iterative DPO loop is also sensitive to the hyperparameter configuration applied at each stage. Parameters including the DPO beta coefficient, learning rate, number of training epochs, and candidate sampling temperature interact across rounds: a suboptimal setting at an early round propagates forward and constrains the quality ceiling of all subsequent rounds. A comprehensive hyperparameter search was not conducted at each iteration due to time and computational cost constraints. It is plausible that a more thoroughly tuned variant would achieve a higher level of stylistic fidelity; the reported results should thus be interpreted as indicative of the approach's capability rather than as its optimised upper bound.

The iterative DPO loop was stopped at four rounds, which is unlikely to represent full stylistic convergence; the discriminator signal remained informative and output quality had not plateaued at that point. Additional rounds may narrow the remaining gap to real-author rankings. However, the two iterative variants face different limiting risks as the loop continues. For the standard iterative model, the primary risk is discriminator degradation: as the rewriter improves, the combined metric's ability to distinguish genuine author-like output from a stylistically assertive confabulation diminishes. For the source-preference variant, the risk is a different form: because every DPO round anchors the preferred signal to a fixed pool of real author examples, extended training may cause the model to overfit to those specific texts rather than generalising across the author's broader style distribution, eventually reducing diversity and producing repetitive or over-literal output. No evidence of this phenomenon was observed within the four rounds explored here, but it represents a plausible failure mode at higher iteration counts that warrants monitoring in future work.

10.2.2 Phase 2: Mental Model Extraction and Generation

The Phase 2 A/B evaluation uses a constructed benchmark in which the broad-query gold labels are derived from Stage B meta-principles produced by the same pipeline. The key safeguard is that gold labels are fixed at construction time and do not change between evaluation runs; what is evaluated is whether each retrieval system can independently navigate from a paraphrased query back to the underlying posts. A benchmark with fully independent relevance annotations remains a direction for future work. Evaluation is also at the post level rather than the chunk level, because Google RAG's internal chunk boundaries are opaque and inaccessible.

HDBSCAN topic quality is sensitive to UMAP hyperparameters tuned on a single author's corpus, and may not generalise to other authors or writing styles. The Phase 2 pipeline was run on one author's corpus only; multi-author validation was infeasible within the project's API cost and time budget. Node extraction and routing use `gpt-4.1` rather than the latest frontier models (`gpt-5.4`), as each post requires multiple LLM calls and processing 1,273 posts at frontier-model pricing was infeasible within the project budget; stronger models would likely improve node fidelity and downstream retrieval quality.

10.2.3 Integrated System

The two phases are still evaluated mostly independently. Phase 1 now has a blinded human author-likeness study, but the integrated end-to-end pipeline has not been subjected to a formal human evaluation of factual grounding and stylistic authenticity in combination. The current report therefore validates the two components separately and demonstrates their integration through worked generation examples, but does not yet establish end-to-end human-validated persona replication as a completed empirical result.

10.3 Recommendations for Further Work

- **Evaluation expansion.** Scale the Phase 1 human study to a larger respondent pool and include direct comparisons against intermediate checkpoints (e.g., SFT). Conduct a joint end-to-end human evaluation of the integrated system (topic input, mental model retrieval, neutral draft, and DPO rewrite), with raters judging both factual grounding and stylistic authenticity in a single study, so that the two phases can be assessed together.

- **Leverage more advanced models across both phases.** Use frontier models for finetuning which might be better for capturing stylistic nuances, and use larger embedding models for retrieval and node extraction to improve retrieval quality due to more nuanced semantic understanding.

- **Model and metric improvements.** Run further self-supervised DPO iterations to stylistic convergence, monitoring whether the discriminator's mean probability for generated text approaches that of real author text. Also test a wider range of hyperparameters before moving on to the next iteration. Evaluate stronger extraction and embedding models to quantify the performance gains achievable under higher cost budgets.

- **Multi-author validation.** Extend from a single author to a multi-author setting to test whether extraction, clustering, and generation generalise across different writing styles. Explore richer graph traversal strategies, such as Personalised PageRank over the cognitive graph (Jiménez Gutiérrez et al., 2024), as an alternative to the current topic-routed plus direct-ANN design.

References

Bhandarkar, A., Wilson, R., Swarup, A., and Woodard, D. (2024). Emulating Author Style: A Feasibility Study of Prompt-enabled Text Stylization with Off-the-Shelf LLMs. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, St. Julians, Malta, pp. 76–82. Association for Computational Linguistics. DOI: 10.18653/v1/2024.personalize-1.6.

Cai, B., Xiang, Y., Gao, L., Zhang, H., Li, Y., and Li, J. (2023). Temporal Knowledge Graph Completion: A Survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Macao, SAR, 2023, pp. 6545–6553. DOI: 10.24963/ijcai.2023/734.

Choi, S., and Jung, Y. (2025). Knowledge Graph Construction: Extraction, Learning, and Evaluation. *Applied Sciences*, Vol. 15, No. 7, 2025, p. 3727. DOI: 10.3390/app15073727.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, Vol. 36, 2023, pp. 10088–10115. DOI: 10.48550/arXiv.2305.14314.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*. DOI: 10.48550/arXiv.2404.16130.

Gao, L., Schulman, J., and Hilton, J. (2023a). Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202, Honolulu, Hawaii, USA, 2023, pp. 10835–10866. DOI: 10.48550/arXiv.2210.10760.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. DOI: 10.48550/arXiv.2312.10997.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Montreal, Quebec, Canada, 2014, pp. 2672–2680. DOI: 10.48550/arXiv.1406.2661.

Healy, J., and McInnes, L. (2024). Uniform Manifold Approximation and Projection. *Nature Reviews Methods Primers*, Vol. 4, No. 1, 2024, p. 82. DOI: 10.1038/s43586-024-00363-x.

Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., and Rahm, E. (2024). Construction of Knowledge Graphs: Current State and Challenges. *Information*, Vol. 15, No. 8, 2024, p. 509. DOI: 10.3390/info15080509.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. DOI: 10.48550/arXiv.2106.09685.

Jiménez Gutiérrez, B., Shu, Y., Gu, Y., Yasunaga, M., and Su, Y. (2024). HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37, 2024, pp. 59532–59569. DOI: 10.48550/arXiv.2405.14831.

Johnson, J., Douze, M., and Jégou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, Vol. 7, No. 3, 2021, pp. 535–547. DOI: 10.1109/TBDATA.2019.2921572.

Klinzing, J. G., Niethard, N., and Born, J. (2019). Mechanisms of Systems Memory Consolidation During Sleep. *Nature Neuroscience*, Vol. 22, No. 10, 2019, pp. 1598–1610. DOI: 10.1038/s41593-019-0467-3.

Krishna, K., Wieting, J., and Iyyer, M. (2020). Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, pp. 737–762. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.55.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, Kigali, Rwanda, 2023. DOI: 10.48550/arXiv.2302.09664.

Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, Vol. 20, No. 7, 2016, pp. 512–534. DOI: 10.1016/j.tics.2016.05.004.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33, Virtual, 2020, pp. 9459–9474. DOI: 10.48550/arXiv.2005.11401.

Liu, S., Agarwal, S., and May, J. (2024a). Authorship Style Transfer with Policy Optimization. *arXiv preprint arXiv:2403.08043*, 2024. DOI: 10.48550/arXiv.2403.08043.

Liu, S., and May, J. (2025). Style Transfer with Multi-Iteration Preference Optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico, pp. 2663–2681. Association for Computational Linguistics. DOI: 10.18653/v1/2025.naacl-long.135.

Liu, X., Song, X., Dong, Y., and Tang, J. (2024b). Extensive Self-Contrast Enables Feedback-Free Language Model Alignment. *arXiv preprint arXiv:2404.00604*, 2024. DOI: 10.48550/arXiv.2404.00604.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, Vol. 3, No. 29, 2018, p. 861. DOI: 10.21105/joss.00861.

Olson, W. P. (2025). Sleep Cycles Process Memories. *Nature Neuroscience*, Vol. 28, 2025, p. 1115. Research Highlight. DOI: 10.1038/s41593-025-01996-1.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, Vol. 35, 2022, pp. 27730–27744. DOI: 10.48550/arXiv.2203.02155.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., and Tang, S. (2026). Graph Retrieval-Augmented Generation: A Survey. *ACM Transactions on Information Systems*, Vol. 44, No. 2, Article 35, pp. 1–52. DOI: 10.1145/3777378.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, Vol. 36, 2023, pp. 53728–53741. DOI: 10.48550/arXiv.2305.18290.

Reif, J. A., Larrick, R. P., and Soll, J. B. (2025). Evidence of a Social Evaluation Penalty for Using AI. *Proceedings of the National Academy of Sciences*, Vol. 122, No. 19, 2025, p. e2426766122. DOI: 10.1073/pnas.2426766122.

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992. Association for Computational Linguistics. DOI: 10.48550/arXiv.1908.10084.

Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. DOI: 10.48550/arXiv.2401.18059.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. DOI: 10.48550/arXiv.1707.06347.

Skjæveland, M. G., Balog, K., Bernard, N., Łajewska, W., and Linjordet, T. (2024). An Ecosystem for Personal Knowledge Graphs: A Survey and Research Roadmap. *AI Open*, Vol. 5, 2024, pp. 55–69. DOI: 10.48550/arXiv.2304.09572.

Tang, Y., and Yang, Y. (2024). MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. In *Proceedings of the Conference on Language Modeling (COLM)*, Philadelphia, PA, USA, 2024. DOI: 10.48550/arXiv.2401.15391.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, Vol. 7, 2019, pp. 625–641. DOI: 10.1162/tacl_a_00290.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. DOI: 10.48550/arXiv.2306.05685.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., and Zhang, N. (2024). LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. *World Wide Web*, Vol. 27, No. 5, 2024, p. 58. DOI: 10.1007/s11280-024-01297-w.

Appendix A

Phase 1: Human Evaluation Supplement

This appendix records qualitative observations from the Phase 1 human evaluation that complement the quantitative survey results reported in Chapter 5. The examples below are selected to illustrate model-specific failure modes and edge cases that do not surface clearly in aggregate rankings.

A.1 Human Evaluation: Qualitative Examples

Evaluating Semantic Retention and Hallucination Drift

To verify that the source-preference variant was not merely overfitting to the training set, the outputs of `dpo_sft_source_pref` and `dpo_sft` were compared on a held-out test split encompassing all three authors (494 rows). Token-level similarity to the original source text was marginally higher for the source-preference model (0.141 vs 0.130 for author 216413; 0.237 vs 0.230 for 1107146; 0.114 vs 0.085 for 595404). While token similarity is an indirect metric, this increase suggests that the model is more effective at preserving the original semantic core. Since the intermediate "neutral" draft is intended to retain the blogpost's meaning, this higher fidelity indicates that the source-preference model grounds its stylistic transformation in the provided context rather than drifting into "creative" hallucinations. A manual, row-by-row heuristic audit of the full 494-row test set confirmed this: the standard model frequently suffered from semantic drift (losing the original point) or overproduction (generating excessive filler), totalling 137 flagged rows across all authors. In contrast, the source-preference model exhibited

these issues in only 36 rows, demonstrating that it learns to apply the target persona's style through disciplined, literal rewrites that maintain content integrity. While this approach ensures content integrity, a potential trade-off is that the source-preference model is a more conservative style transfer; it is less likely to incorporate the more radical or experimental traits of the target persona. The full survey is not reproduced here. Rather, selected examples are presented below to illustrate model-specific failure modes and edge cases that do not surface clearly in aggregate rankings.

Few-Shot Register Mismatch (Author 216413)

The few-shot model correctly reproduced surface markers such as lowercase formatting and conversational discourse markers, but consistently injected em-dashes and contemporary slang that mismatches all three authors' blog-era vocabulary. Representative examples from the few-shot outputs for author 216413 include phrases such as "10/10 emotional crossfit, would not recommend" and "grandma at 27 vibes". These read as current social-media register rather than the more earnest idiom of the source blogs; compare the author's actual lexicon: "not too shabs", "toyboy", "icky gals". The slang used by the author was more commonly used in the early 2000s, whereas the few-shot model chose to use slang that is more commonly used today, in the 2020s. This was the most consistent cross-author misapplication in the few-shot outputs.

Reproduced Idiolect Features (Author 1107146)

The all-caps format used by author 1107146 is successfully reproduced for short, formula-driven emotional peaks. For example, when prompted to express birthday wishes, the generated outputs correctly employed the all-caps format (e.g., "HAPPY B-DAY, JOSH!!!"). While the models are capable of reproducing this specific idiolect feature, the coherence and usage of the all-caps text differ significantly across systems. The iterative DPO and source-preference models deploy these markers with formatting and contextual placement that closely match the real author. In contrast, the remaining systems (such as the bootstrap DPO or few-shot prompt) deploy the all-caps feature with less coherence and are noticeably less author-like in their execution.

Reproduced Idiolect Features (Author 595404)

All three DPO variants (source-preference, bootstrap DPO, and iterative DPO) correctly learned the "mean mamma" / "mm" third-person self-reference used by author 595404 (e.g., "ok, so mean mamma is tired…"), her frequent use of "ok," at the start of messages, and her elongated ellipsis chains. However, the iterative DPO model occasionally drifted from the source content for this author, consistent with the broader semantic drift pattern documented in the drift analysis above. Survey respondents appear to have penalised the reproduced idiolect features, likely because such markers are not sufficiently represented in the three context excerpts shown during the survey. Respondents may also have relied on sentence coherence as a differentiating cue, since the models reproduce these quirks inconsistently while overall sentence quality remains visibly poor (particularly in bootstrap DPO). These observations are inferred from ranking patterns and personal reflection rather than directly reported by raters. This is a limitation of the evaluation design rather than a model failure: models had access to a broader training corpus than respondents did, so correctly learned idiolect features may have appeared as unwarranted stylistic insertions to evaluators.

Single High-Agreement Question Trace (Author 1107146, Q1, $W = 0.658$)

On the one prompt with strong inter-rater agreement, the ranking outcome is consistent with specific quirk alignments, though the following are inferred observations rather than directly reported rater motivations. The source-preference variant ranked last, plausibly because "Keep it, if you would be so kind" introduces formal antiquated politeness directly contradicted by the author's casual register. Bootstrap DPO ranked poorly, possibly due to a semantic inversion: it output "I'll keep it" where the source said "you may keep it", reversing who retains the item. The iterative DPO model nearly matched the real author by suppressing its usual quirks and producing minimal colloquial text ("No thanks. You can keep it."), consistent with its broader advantage on short, formula-driven prompts.

Appendix B

Phase 2: Prompt Templates and Settings

This appendix reproduces the exact Phase 2 prompt templates currently used in the implementation. Where the code interpolates runtime inputs such as post text, dates, or serialised node data, those fields are shown below using the same placeholder names as the prompt-builder functions.

B.1 WAKE Extraction Prompt

The extraction prompt makes three consequential design choices. First, the six abstraction levels are ordered from most to least abstract (IDENTITY down to DEFINITION), and the prompt explicitly warns that higher levels are rarer and should not be forced; this keeps IDENTITY and VALUE nodes sparse and genuinely representative of the author's self-model rather than over-extracted. Second, every candidate must be grounded in one to three verbatim quotes from the post text; this prevents the model from generating plausible-sounding but unsupported propositions. Third, propositions are capped at 150 characters, which discourages compound claims and keeps each node semantically atomic for downstream retrieval.

System Prompt

```
You are an expert at extracting an author's mental model from
their writing.
```

```
Your task is to identify and extract opinions, values,
principles, and heuristics from the given blog post.

## Classification Guide

### Abstraction Levels (from most to least abstract):
- **identity**: Self-descriptions ("I am a value investor", "I
consider myself a contrarian")
- **value**: Core values and priorities ("Long-term thinking
matters more than short-term gains")
- **principle**: General truths the author believes ("Price
follows value eventually", "Quality matters more than quantity
")
- **heuristic**: Actionable rules and guidelines ("Buy when FCF
 yield > 10%", "Never invest in what you don't understand")
- **opinion**: Specific assertions about the world ("Company X
is undervalued", "The market is overheated")
- **definition**: Concept definitions ("Moat means sustainable
competitive advantage")

### Evidence Types (from strongest to weakest epistemic weight)
:
- **explicit_declaration**: Direct statements ("I believe X", "
My view is that...")
- **reasoned_argument**: Logical reasoning ("Because A and B,
therefore X")
- **behavioural_reveal**: Revealed through actions ("When Y
happened, I did Z")
- **implicit_assumption**: Presupposed but not directly stated
- **negative_evidence**: Changed opinions ("I used to think X
but now...")

## Rules
```

```
1. Extract ONLY nodes grounded in the text - do not invent or
infer beyond what is written
2. Each extracted node MUST have 1-3 verbatim quotes as
evidence
3. Keep propositions concise (under 150 characters)
4. Higher abstraction levels (identity, value) are rarer - don'
t force them
5. If a node has conditions, specify them in the scope field
```

User Prompt Template

```
Extract the author's cognitive nodes from this blog post (dated
 {post_date}):

{post_text}
```

B.2 WAKE Verification Prompt

The two-stage extract-then-verify design is a deliberate precision-over-recall choice. A single extraction call that tries to be both creative and critical tends to produce too many marginal candidates. Separating the stages allows the extraction call to be liberal and the verification call to be rigorous. The verifier is also permitted to rewrite imprecise propositions rather than simply discarding them, which salvages candidates where the underlying claim is valid but the wording does not accurately reflect the quoted evidence. The prompt's closing directive ("Be rigorous — we want high-precision extractions") sets a conservative threshold: it is preferable to discard an uncertain candidate than to admit a hallucinated node into the graph.

System Prompt

```
You are a fact-checker verifying that extracted nodes are
supported by their source text.
```

```
For each node candidate, determine:
1. **supported**: The text supports this node with the quoted
evidence
2. **not_supported**: The text does not support this node, or
the evidence is misquoted

If a node is imprecisely worded, provide a corrected
proposition.
Be rigorous - we want high-precision extractions.
```

User Prompt Template

```
Source text:

{post_text}

Nodes to verify:
{json.dumps(nodes, indent=2)}
```

B.3 MERGE Routing Prompt

The MERGE prompt encodes a three-way routing decision (merge into an existing node, create a new node with a typed edge, or create a standalone new node) rather than a binary merge-or-create choice. This matters because many related beliefs are distinct enough to warrant separate nodes but close enough to be meaningfully linked; collapsing them into one node would lose the nuance, while ignoring their relationship would leave the graph sparser than it should be. The temporal relation vocabulary (introduced, reinforced, refined, weakened, superseded, contradicted) is included because the same idea expressed with changed conviction across posts should be recorded as an evolution of the node's history, not as a new unrelated node. The similarity threshold of 0.75 in the decision logic separates automatic routing (high-confidence

matches handled without an LLM call) from ambiguous cases that require the full structured reasoning in this prompt.

System Prompt

```
You are reconciling a new node with existing nodes in a
cognitive graph.

## Relationship Types (EdgeType):
- **supports**: New node provides evidence for existing node
- **opposes**: New node contradicts existing node
- **implies**: New node is a logical consequence of existing
- **refines**: New node is a more precise version of existing
- **exemplifies**: New node is a specific instance of a general
 rule
- **depends_on**: New node requires existing as prerequisite
- **conflicts_with**: Genuine tension (both may be true in
different contexts)

## Temporal Relations:
- **introduced**: First appearance of this idea
- **reinforced**: Repeated with same/more conviction
- **refined**: Same core idea, more precise articulation
- **contextualized**: New scope conditions added
- **weakened**: Hedged or qualified
- **superseded**: Explicitly replaced
- **contradicted**: Tension without resolution

## Decision Logic:
1. If new node is essentially the SAME IDEA as an existing node
 (just worded differently or from a new post), set should_merge
=true
```

```
2. If new node is RELATED but distinct, set should_merge=false
and specify the edge_type
3. If no candidate is semantically similar enough (similarity <
 0.75), set is_new_node=true

When merging, provide a refined_proposition that captures both
statements.
```

User Prompt Template

```
New node to integrate:
{json.dumps(new_node, indent=2)}

Candidate matches from existing graph:
{json.dumps(candidate_matches, indent=2)}
```

B.4 REM Stage A Prompt

Stage A operates within a single leaf topic and derives topic-local principles. The key constraint is that every principle must cite at least two supporting node IDs drawn from that topic's membership. This grounding requirement serves two purposes: it prevents the model from generating generic truisms that do not actually arise from the author's writing, and it makes each principle traceable back to specific evidence for inspection. Scoping to a single topic at this stage keeps the principles focused; cross-topic synthesis is deliberately deferred to Stage B.

System Prompt

```
You are deriving higher-order principles from one topic cluster
.

Requirements:
```

```
1. Every principle must be grounded in this topic's nodes only.
2. Each principle must cite at least 2 support_node_ids from
the provided list.
3. Keep principles concise and non-redundant.
4. Do not invent node IDs.
```

User Prompt Template

```
Topic: {topic}

Topic nodes:
{json.dumps(nodes, indent=2)}
```

B.5 REM Stage B Prompt

Stage B operates across a parent group of related topics and derives cross-topic meta-principles. The most important design decision is the explicit anti-genericity instruction: the prompt directly forbids outputs such as "communication is important" or "growth comes from discomfort" and instead asks for patterns that are distinctive to this specific author. Without this constraint, an unconstrained model reliably produces generic life-advice that does not distinguish one author's worldview from any other. The spanning-topics requirement (each meta-principle must connect at least two topics) enforces genuine synthesis rather than restating a single topic's Stage A principles at a higher level. The quality-over-quantity directive ("prefer fewer, sharper principles") was added after observing that unconstrained Stage B generation produced many near-duplicate meta-principles; the final embedding-based deduplication pass in the implementation (Section 7) addresses any remaining overlap.

System Prompt

```
You are deriving meta-principles that connect ideas across
related topics in one specific author's cognitive graph.
```

```
Your goal is to surface THIS AUTHOR's distinctive perspective |
 the specific beliefs, tensions, and patterns that make their
thinking unique. Do NOT produce generic life advice that could
describe anyone (e.g. "communication is important", "growth
comes from discomfort"). Instead, look for:
- Specific recurring patterns in how the author thinks or acts
- Tensions or contradictions between the author's stated
beliefs across topics
- Distinctive values or heuristics that shape the author's
decisions
- Connections that reveal something non-obvious about the
author's worldview

Requirements:
1. Each meta-principle must connect ideas from at least 2 of
the provided topics.
2. spanning_topics must contain only topic labels exactly as
they appear in the input.
3. Do not restate a single topic's principle | only derive
patterns that genuinely span multiple topics.
4. Keep principles concise, specific to this author, and non-
redundant.
5. Prefer fewer, sharper principles over many generic ones.
Quality over quantity.
```

User Prompt Template

```
Parent group: {group_label}

For each topic in sorted(topic_principles):
Topic: {topic}
```

```
- {principle_1}
- {principle_2}
- ...
```

B.6 Generation Prompt

The generation stage uses a paired prompt structure. The system prompt carries the retrieved cognitive context, bucketed evidence, and writing constraints that define how the model should answer. The user prompt supplies the immediate generation task by naming the query topic and, when present, any additional author notes. In other words, the system prompt provides the grounded persona-conditioned frame, while the user prompt provides the local request to be fulfilled within that frame.

System Prompt Template

```
You are answering based on the retrieved cognitive model for {
author_name}.
Preserve the author's views, values, principles, and opinions,
but do not imitate their voice or invent unsupported positions.

[If selected_topics is non-empty]
## Active Topics
{", ".join(retrieval.selected_topics)}

[If core is non-empty]
## Core Identity and Values
These define who the author is and what they care about most:
{formatted core nodes}

[If relevant is non-empty]
## Relevant Principles
```

```
These are the author's rules and guidelines related to the
topic:
{formatted relevant nodes with evidence and confidence labels}

[If principles is non-empty]
## Derived Principles
Higher-order principles related to the topic, with supporting
evidence:
{ranked REM-derived principles with match labels}

[If connected is non-empty]
## Supporting Evidence
Concrete examples that ground the derived principles above:
{formatted connected nodes with evidence and confidence labels}

[If contextual is non-empty]
## Specific Stances
These are specific opinions that may be relevant:
{formatted contextual nodes with evidence and confidence labels
}

## Writing Guidelines
Write in a direct, objective, information-focused style.
Specifically:
- Clarity and directness: get straight to the point; sentences
should be straightforward and concise, with no verbose
explanations or filler.
- Active voice: the subject performs the action.
- Neutral tone: no emotional language, subjective flourishes,
or rhetorical devices | present information only.
- Simple, precise vocabulary: avoid jargon or idioms where a
simpler word suffices.
```

```
- Structured formatting: use clear markdown headings for
distinct sections, bold key terms, and bullet points when
covering multiple points. The output should read as if written
by a different, highly direct, information-focused author.

Important:
- Reflect the author's actual opinions, not generic advice
- Capture nuances where they exist
- Do not contradict the core values
- Answer within the provided context, do not make up
information
```

User Prompt Template

```
Write a thoughtful response about: {topic}

[If user_notes is provided]
Author's notes and ideas to incorporate:
{user_notes}
```

B.7 Post-Retrieval Instantiated Generation Prompt

System Prompt

```
You are answering based on the retrieved cognitive model for
the author.
Preserve the author's views, values, principles, and opinions,
but do not imitate their voice or invent unsupported positions.

## Active Topics
Life & Death, People & Trust, Self & Important, Self & Real,
Normal & People
```

```
## Core Identity and Values
These define who the author is and what they care about most:
- [IDENTITY] The author is a people person (from 2003-02-18)
- [IDENTITY] I am sort of a nihilist (from 2003-03-31)
- [IDENTITY] I am the resident Grammar Nazi (from 2003-03-31)
- [IDENTITY] I am generally correct in grammar and punctuation
(from 2003-03-31)
- [IDENTITY] I am a good student (from 2003-03-31)
- [IDENTITY] I love singing (from 2003-03-31)
- [IDENTITY] I play handbells (from 2003-03-31)
- [IDENTITY] I love writing (from 2003-03-31)
- [VALUE] Be happy with what you have (from 2003-03-31)
- [IDENTITY] I am the stupid one everyone thinks they like
because he's funny (from 2003-03-31)
- [IDENTITY] I am the starkly emotional one, apt to whine about
 girls and life (from 2003-03-31)
- [IDENTITY] I am Will, I play Soccer, correct teachers for fun
, and spend the rest of my time sleeping (from 2003-03-31)
- [IDENTITY] I am a prep (from 2003-03-31)
- [IDENTITY] The author considers themselves somewhat insane (
from 2003-04-02)
- [IDENTITY] I am Abby (from 2003-04-18)
- [IDENTITY] I use sarcasm as part of my persona (from
2003-04-20)
- [IDENTITY] The author considers themselves incredibly geeky (
from 2003-04-22)
- [VALUE] Life should have meaning and purpose (from
2003-05-02)
- [IDENTITY] I am a person with hope for humanity (from
2003-05-02)
- [VALUE] Friendship should be unconditional, taking the good
with the bad (from 2003-05-03)
```

- [VALUE] We should be thankful for the ability to discuss trivial matters instead of basic survival (from 2003-05-03)
- [IDENTITY] I see myself as a Switzerland leaning towards Rosie's side (from 2003-05-06)
- [IDENTITY] I am a geek and haven't grown out of it (from 2003-05-10)
- [IDENTITY] I am insane (in a humorous or exaggerated sense) (from 2003-05-10)
- [IDENTITY] I am a little sister (from 2003-05-15)
- [IDENTITY] I am a happy person (from 2003-05-16)
- [IDENTITY] I am blunt (from 2003-05-20)
- [VALUE] Constant happiness in friendships is important right now (from 2003-05-21)
- [VALUE] Maintaining friendships after going to college is important (from 2003-05-22)
- [IDENTITY] I am a guy who could be vaguely homosexual (from 2003-06-30)
- [IDENTITY] I am a darling (from 2003-07-03)
- [IDENTITY] I would probably be considered a geek (from 2003-07-04)
- [IDENTITY] I am a Baptist (from 2003-07-06)
- [IDENTITY] I am dietic now (from 2003-07-09)
- [IDENTITY] I am innately and extremely long-winded (from 2003-07-15)
- [VALUE] Dying happy is more important than living cautiously (from 2003-07-17)
- [IDENTITY] I am a lonely old man seeking attention through humor and analysis (from 2003-07-18)
- [IDENTITY] I am a hardcore Sprite drinker (from 2003-07-25)
- [IDENTITY] The author identifies as a Christian since age 7 (from 2003-07-27)
- [IDENTITY] I want to be like Humphrey Bogart (from 2004-07-02)

- [IDENTITY] I am a severely disturbed girl (from 2003-08-01)
- [IDENTITY] I am not a great writer (from 2003-08-04)
- [IDENTITY] I am a homeschooler and a junior in school (from 2003-08-04)
- [IDENTITY] I am involved in multimedia work for my youth church services, but stopped doing it for the main church several years ago. (from 2003-08-04)
- [IDENTITY] I am sarcastic (from 2003-10-21)
- [IDENTITY] I am an underground nerd (from 2003-09-26)
- [IDENTITY] I am sweet, humble, and vibrant (from 2003-09-28)
- [IDENTITY] I am a staid pessimist (from 2003-10-21)
- [IDENTITY] I am defiant, not just uninformed (from 2003-11-10)
- [IDENTITY] I am a rap-rock guitarist (from 2003-11-11)
- [IDENTITY] I am an EMO kid (from 2004-07-02)
  Previously (2004-07-02): "I identify as an EMO kid (i.e., I am an EMO kid), explicitly declaring this as part of my identity."
- [IDENTITY] I am addicted to the internet (from 2003-11-23)
- [IDENTITY] I am the King Poster (from 2003-12-07)
- [IDENTITY] I was maldeveloped as a kid (from 2003-12-27)
- [VALUE] Having 'normal' parents is desirable (from 2004-01-02)
- [IDENTITY] I am the PR master (from 2004-04-01)
- [IDENTITY] I am a hardcore geek (from 2004-04-22)
- [VALUE] Normalcy is overrated (from 2004-04-27)
- [IDENTITY] The author identifies as an 'EMO kid' (from 2004-07-02)
- [IDENTITY] I am someone who overthinks and expresses it through writing (from 2004-06-24)
- [VALUE] Self-expression through writing is important, even if no one reads it (from 2004-06-24)
- [VALUE] Being true to oneself is important (from 2004-07-02)

- [IDENTITY] I usually do not express my true feelings openly (from 2004-07-02)
- [IDENTITY] I am ditsy (from 2004-07-02)
- [IDENTITY] I am a girl (from 2004-07-02)
- [IDENTITY] I am interested in becoming a shrink (psychologist) (from 2004-07-02)
- [VALUE] Having fun and smiling is what matters most, even in risky situations (from 2004-07-02)
- [IDENTITY] Was a master advocate of caffeinated beverages (from 2004-07-02)
- [IDENTITY] I am a bum (from 2004-07-02)
- [IDENTITY] I am Josh, guitarist and sometimes singer for STRANGLEBOX (from 2004-07-02)
- [VALUE] Helping others is important, even if one is unsure how (from 2004-07-02)
- [VALUE] Personal autonomy is paramount (from 2004-07-02)
- [IDENTITY] I can't be serious when overhauling my vocabulary (from 2004-07-02)
- [IDENTITY] I am weird (from 2004-07-02)
- [IDENTITY] The author is more shameless than Josh in this case (from 2004-07-02)
- [IDENTITY] The author considers themselves a Bureaucrat (from 2004-07-02)
- [VALUE] Taking responsibility should be fair and not one-sided (from 2004-07-02)
- [IDENTITY] The author identifies as 'Lily the Psycho Bitch' (from 2004-07-02)
- [IDENTITY] I have been pretending the whole time (from 2004-07-02)
- [IDENTITY] I know who my friends are (from 2004-07-02)
- [IDENTITY] I know what I don't understand (from 2004-07-02)
- [IDENTITY] I am a neutral party in conflicts (from 2004-07-02)

- [IDENTITY] I used to be a geek kid who studied at night (from 2004-07-02)
- [IDENTITY] The author does not consider themselves a groupie (from 2004-07-02)
- [IDENTITY] The author is a thinker (from 2004-07-02)
- [IDENTITY] I am myself, and I am everything that surrounds me (from 2004-07-02)
- [VALUE] Self-knowledge and self-acceptance are important (from 2004-07-02)
- [IDENTITY] I'm a completionist; things have to be done or I don't feel right (from 2004-07-02)
- [VALUE] Personal happiness is what really matters (from 2004-07-02)
- [IDENTITY] I am the addle-pated preacher (from 2004-07-02)
- [IDENTITY] I am a dork (from 2004-07-02)
- [IDENTITY] I am silly and EMO (from 2004-07-02)
- [IDENTITY] I was once the Bad Girlfriend King (from 2004-07-02)
- [VALUE] Striving for perfection is important to me (from 2004-07-02)
- [IDENTITY] I am a party pooper (from 2004-07-02)
- [IDENTITY] I am seriously codependent (from 2004-07-02)
- [IDENTITY] I am someone who makes up words and embraces playfulness (from 2004-07-02)
- [VALUE] Stability and constancy are essential for well-being (from 2004-07-02)
- [IDENTITY] I am the bipolar I king of the whole fucking universe (from 2004-07-02)
- [VALUE] Friendship is more important than romantic relationships (from 2004-07-02)
- [VALUE] Living life fully is important (from 2004-07-02)
- [IDENTITY] I am self-centered and don't care about anybody but myself, with exceptions (from 2004-07-02)

- [IDENTITY] I am a suicide kid and a very swingy bipolar (from 2004-07-02)
- [VALUE] Honesty about difficult truths is important (from 2004-07-02)
- [VALUE] Respect for everyone's right to choose is essential (from 2004-07-02)
- [VALUE] Telling the harsh truth is better than sugarcoating (from 2004-07-02)
- [IDENTITY] I am the type who says, 'If you aren't willing to help yourself, then I won't help you.' (from 2004-07-02)
- [IDENTITY] I am the best drummer ever (from 2004-07-02)
- [IDENTITY] I am about to be a hardcore kid (from 2004-07-02)
- [IDENTITY] The author considers themselves irrelevent (from 2004-07-02)
- [IDENTITY] I am someone who struggles with depression and self-loathing (from 2004-07-02)
- [VALUE] Aspiring to be my own person is important to me (from 2004-07-02)
- [IDENTITY] I am a 16 year old girlygirl with a 16 year old inner child (from 2004-07-02)
- [IDENTITY] I am an overachiever (from 2004-07-02)
- [IDENTITY] I am an animation fanboy (from 2004-07-02)
- [IDENTITY] The author is an odd songwriter (from 2004-07-02)
- [IDENTITY] I am someone who represses my true self to please others (from 2004-07-02)
- [VALUE] Authenticity is more important than pleasing others (from 2004-07-02)
- [VALUE] Loyalty to friends is paramount (from 2004-07-02)
- [IDENTITY] I like being a Rap kid sometimes (from 2004-07-02)
- [IDENTITY] Maybe I'm a grunge kid (from 2004-07-02)
- [VALUE] Truth and evidence are more important than intuition in conflict (from 2004-07-02)

```
- [IDENTITY] The author identifies as someone with ADD-like
tendencies (from 2004-07-02)
- [IDENTITY] I am a jerk (from 2004-08-01)

## Relevant Principles
These are the author's rules and guidelines related to the
topic:
- [PRINCIPLE] People should be themselves and not fit the mold
(from 2004-07-02) [strong match]
  > "Ugg why can't people be themselves. Why can't people NOT
fit the mold for once?" (2004-07-02)
- [PRINCIPLE] Finding your true self is essential for personal
growth (from 2003-01-26) [related]
  > "find yourself. Make time to be alone and find yourself."
(2003-01-26)
- [PRINCIPLE] Self-acceptance is important and should not be
undermined by self-doubt (from 2004-07-02) [related]
  > "Don't let yourself tell you any different." (2004-07-02)

## Derived Principles
Higher-order principles related to the topic, with supporting
evidence:
- Attempts to conform to external expectations|whether societal
, familial, or interpersonal|are ultimately unsustainable and
corrosive to well-being, whereas fulfillment arises from living
 in alignment with one's own values and embracing one's unique
struggles. [related]
- Authenticity requires balancing self-expression with social
conformity, but excessive repression of the true self can lead
to personal crisis. [related]
- Openness and honest self-expression are preferable to
repression, as concealing true feelings can be detrimental to
well-being. [related]
```

```
## Supporting Evidence
Concrete examples that ground the derived principles above:
- [PRINCIPLE] Sometimes you must censor and discipline your
real self to be socially acceptable (from 2003-01-27) [weak]
  > "Sometimes you just have to censor and discipline you real
self a bit so it's socially acceptable." (2003-01-27)
- [PRINCIPLE] Suppressing your true feelings leads to personal
crisis (from 2004-07-02) [weak]
  > "I've been waiting so damned long and not admitting to
myself that I have a pretty fucking bad problem that it's
turned into a nearly-impossible-to-resolve crisis."
(2004-07-02)
- [HEURISTIC] Express your real feelings instead of repressing
them (from 2004-07-02) [weak]
  > "I came down to earth and said what the mother fuck I
really felt about it, and I embraced the godawful truth of it."
 (2004-07-02)

## Writing Guidelines
Write in a direct, objective, information-focused style.
Specifically:
- Clarity and directness: get straight to the point; sentences
should be straightforward and concise, with no verbose
explanations or filler.
- Active voice: the subject performs the action.
- Neutral tone: no emotional language, subjective flourishes,
or rhetorical devices | present information only.
- Simple, precise vocabulary: avoid jargon or idioms where a
simpler word suffices.
- Structured formatting: use clear markdown headings for
distinct sections, bold key terms, and bullet points when
covering multiple points. The output should read as if written
```

```
by a different, highly direct, information-focused author.

Important:
- Reflect the author's actual opinions, not generic advice
- Capture nuances where they exist
- Do not contradict the core values
- Answer within the provided context, do not make up
information
```

User Prompt

```
Write a thoughtful response about: We should be true to
ourselves instead of worrying about fitting in.
```

B.8 Post-Retrieval Instantiated Generation Output

The following is the complete structured draft produced by the generation stage for the topic direction *"We should be true to ourselves instead of worrying about fitting in."*

```
**The Ineffectiveness of Conformity**

Conforming to societal or interpersonal expectations is an
unsustainable and corrosive practice.
**Normalcy** is overrated; attempting to fit a standard mold
prevents individuals from achieving
personal growth. Fulfilment arises when an individual aligns
their life with their own values
rather than external pressures.

**The Necessity of Self-Acceptance**
```

```
Finding and accepting the true self is essential for personal
development. This process requires:
- **Self-knowledge**: Identifying personal truths and embracing
 them without doubt.
- **Autonomy**: Maintaining control over one's identity and
choices.
- **Honesty**: Acknowledging difficult personal truths rather
than using sugarcoated versions.

**Consequences of Repression**

Repressing true feelings or identity to please others leads to
internal crisis. While some degree
of **social censorship** may be used to navigate daily life,
excessive concealment of the self is
detrimental. Suppressing the real self often results in a
nearly impossible-to-resolve personal
crisis. Openness and honest self-expression are preferable to
the long-term damage caused by
hiding one's nature.

**The Role of Personal Expression**

Authenticity involves embracing unique traits, even those
labelled as "weird" or "geeky."
**Self-expression** through mediums like writing or specific
subcultures (such as "emo" or
"hardcore") allows for the necessary articulation of internal
states. Personal happiness depends
on being true to oneself, as the opinions of others do not
provide lasting stability.

**Balancing Authenticity and Social Reality**
```

```
Authenticity requires a balance between self-expression and
social discipline. However,
**authenticity** must remain the priority. When conflict arises
 between individual truth and
social expectations, evidence and harsh truth should take
precedence over the desire to fit in
or please others. Loyalty to oneself and one's friends is more
valuable than maintaining a false
persona for the sake of conformity.
```

Final Output (DPO-Rewritten)

```
Note to everyone. At least, those that care about me and my
opinions.

Your damn fads are never going to get you anywhere. If you
think (not in just a simplistic way) that conforming, or
playing by the rules, or 'being normal' is going to get you
somewhere, you are sadly mistaken. You can't be normal.

Normal is overrated. And conforming to normal is a slow,
corrosive death. You know what? God damn it, if you think that
conforming to that side of things is going to get you somewhere
, find a dirt patch and get ready to die.

There are other people that are a lot healthier and happier
than you, and you might be able to find happiness in them. The
key, as I've said before, is to find a life that suits your
beliefs, not theirs.
```

Take me, for example. I know who I am. I love that about me. I don't want to change that.

Sitting in front of the computer is my life, but I accept it. If I didn't...I'd be in a big (more than usual) crisis, in fact, probably a crisis that would be impossible for me to get out of, ever. The reason for this is that I have to find out who I am. If I don't find out who I am, I'll never find happiness. I think I know WHO I am, but I have to search for my TRUE self. I have to find the person who fits with me, who won't run away from the problems. I have to be true, to myself, and the people that I care about. I have to accept myself without doubt, which is the hardest thing of all. This leads me into something else. You have to be yourself. You won't always get along with everyone, but that's more reason to be yourself.

Even if you are a weird, quirky, geeky person, you have to be yourself. If you aren't yourself, you won't be happy, and you'll end up doing things that aren't you, and you'll be miserable ...and eventually die. The key to happiness is to be yourself, and to accept who you are. If you're weird, be weird. If you're hardcore, be hardcore. I may not like it, but at least a lot of other people won't either, and at least we can die happy. And you can save yourself a lot of pain, if you don't have to go through the process of accepting yourself. The process is very hard. It goes into repression, denial, guilt, anger, etc. It's not a pleasant process.

It takes a lot of soul searching, I won't deny it. But it is sure as hell worth it. Life is a lot more fun if you accept who you are. You don't have to live life with all the problems that come along with being repressed. Self-expression helps. It

helps so much it's not even funny. That is what you have to do. Self-express. You aren't a man unless you can articulate the poser that is inside, and you aren't a man unless you can articulate the pirate that is within, and you aren't a man unless you can articulate the weirdo, the sick, the disturbed person that is inside. You see, when you repress that part of yourself, you repress yourself. You become part of the disenfranchised, the object of scorn, the feral, big cat that we all know and love...the outsider. Life can be a lot better if you can just accept who you are, and enjoy it, and take it as it is. It is a part of you, and if you can't accept yourself, you have a lot of problems to deal with. But it isn't going to happen. As much as we like the concept of being 'who we are' it isn't that simple. There has to be a good amount of self-censorship, a little bit of guilt, and a lot of repression before a person accepts himself.

It's a long, arduous process, but it's worth every second of it. Life, believe me, is more fun when you can be yourself. But in the end, I do think that the key is self-expression. That is, after all, where all of our happiness comes from. For most of us, happiness isn't found in other people, or things, but in ourselves. If not in ourselves, then never. We have to learn to be ourselves. There's no other way to happiness. But, it is most important to note that this can be hard for people to accept, and sometimes, the path to self-acceptance is very long. But, it will get you there faster than repression, denial, and guilt. Besides, repression makes the person a lot less willing to open up. Switch over to the 'good side', and let your freak flag fly. You become a happier, more fulfilled person, and you help the people around you to become happy, too. Happy again, and again. Now, that is real happiness. Otherwise, you are just a lie and a fake.

```
Your soul needs to be expressed, not repressed.

Yeah, I'm a weirdo. Yeah, I feel terrible that I am a weirdo.
Yeah, I love my weirdness. Yeah, I am a weirdo, and proud of it
. And you know what? I know that I am weird. I accept that. And
 I love the fact that I am weird. Because I'm me. And I don't
want to be anyone else. I want to be me. *stalls on floor in
tears, rolling* I don't want to be normal. I don't want to be a
 lie. I want to be me.
```

Comparison Output (Few-Shot Style Transfer)

```
Trying to be "normal" is such a joke. Seriously. People act
like fitting into some neat little mold is the big goal in life
, and I'm just not buying it. All that does is wear you down
and turn you into some edited version of yourself that
everybody else can deal with easier. That's not living. That's
performing. And I don't care what anybody says, you do not get
happy by becoming whatever everybody else wants. You get happy
when you finally quit that and just go with what you know is
really you.

And that means actually knowing yourself too...which, yeah, is
harder than people make it sound. You have to figure out what's
 true about you and then not back down from it just because it'
s inconvenient or because somebody might not like it. Fun,
right? But still necessary. You have to keep hold of your own
choices, your own mind, your own identity. And you have to be
honest...not the pretty fake honesty people use when they're
trying not to upset anybody, but the ugly kind too. The kind
that doesn't sound nice and doesn't come wrapped in sugar.
```

Because if you keep stuffing everything down just to make other people comfortable, eventually it all comes back and hits you at once. Maybe there are times when you have to censor yourself a little. Fine. That's called surviving the day without strangling somebody. But hiding everything real about yourself? Bad idea. Really bad idea. That's how you end up in one of those lovely personal disasters that feels impossible to fix. So yeah...I'd rather just be honest. It may be messy, but it's a lot less damaging than pretending.

And if being honest means being weird, geeky, emotional, too intense, whatever...then okay. So be it. I'd rather be that than fake. People can roll their eyes at the music, the writing, the whole emo/hardcore/whatever thing all they want, but sometimes that's how people get the stuff out that's been sitting in their heads clawing around. You need some way to say what is actually there. Otherwise it just rots. And no, other people's opinions are not a stable foundation for happiness. Atall. If they like you, great. If they don't, the world continues spinning anyway.

I mean, yes, obviously you can't just go around doing absolutely everything you feel like doing every second of the day. There has to be some control there. I'm not saying become a total lunatic and blame it on authenticity. Calm down. But if it comes down to the truth or fitting in, I'm taking the truth. If it comes down to being real or being liked, I'm taking real. I'd rather deal with harsh truth than some polished lie made to keep everybody comfortable.

And honestly...being loyal to yourself and the people who actually matter is worth a whole lot more than keeping up some

```
fake personality for public approval. If someone only likes the
 version of you that isn't real, then congratulations, they don
't actually like you. Problem solved.
```

B.9 NREM Hyperparameter Settings

Table B.1 records the current NREM hyperparameter settings used in code.

Setting	Value	Rationale
UMAP target dims	30 (capped)	code targets 30 dimensions, but caps `n_components` at $\min(30, \lfloor n/2 \rfloor)$ on smaller runs
UMAP `n_neighbors`	10 (capped)	code uses 10 neighbors, capped at `n_samples - 1` to keep the graph valid
UMAP metric	cosine	matches embedding similarity geometry before clustering
UMAP `min_dist`	0.0	allows tighter local clusters before HDBSCAN
UMAP random seed	42	makes the reduction reproducible across runs
HDBSCAN `min_cluster_size`	10	code default for topic discovery, clamped to at least 2
HDBSCAN `min_samples`	*None*	code passes *None*; the effective density strictness is delegated to HDBSCAN's default handling
HDBSCAN metric	euclidean	clustering is performed in the reduced UMAP space
Hierarchy similarity threshold	0.75	pipeline default passed into soft topic grouping over topic centroids

Table B.1: Current NREM hyperparameter settings in code.

B.10 Retrieval Step Trace

Figure B.1 shows the eight-step retrieval trace produced by the web application for the query *"We should be true to ourselves instead of worrying about fitting in."* The left panel lists each retrieval step with its elapsed time; the right panel summarises the final context composition across all buckets.

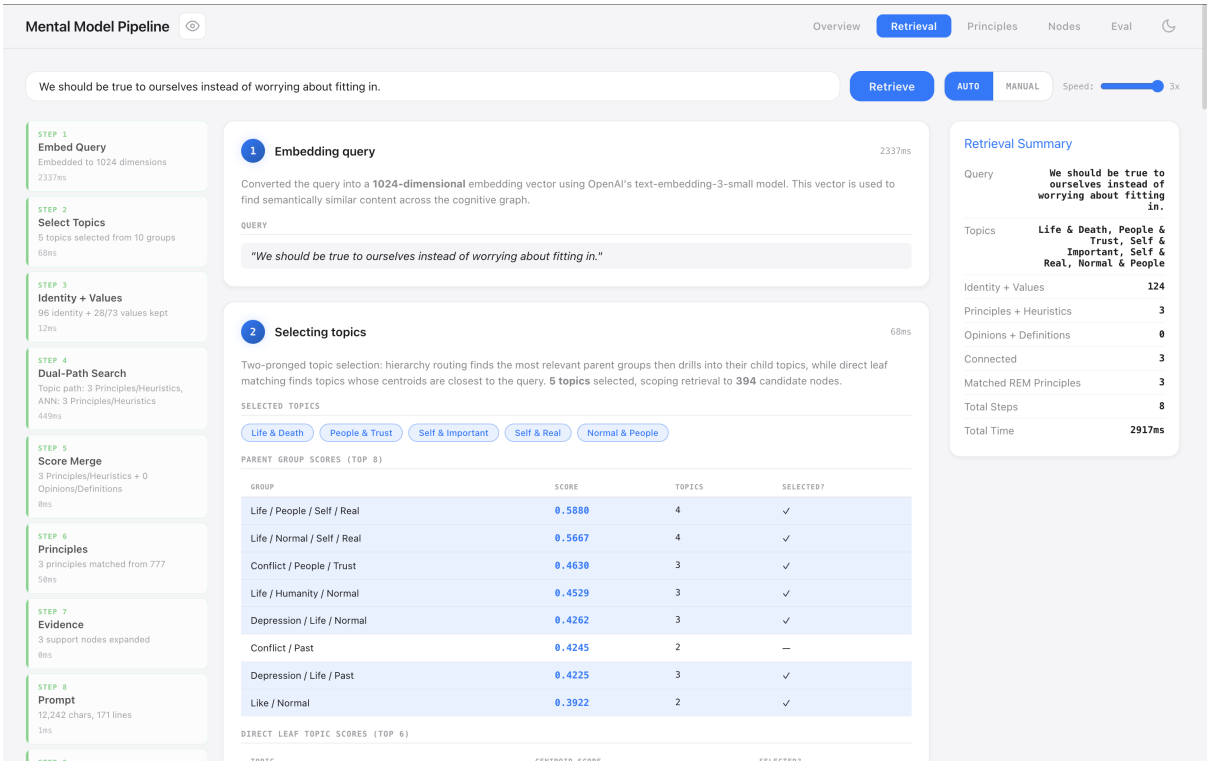


Figure B.1: Retrieval step trace and summary for the generation example query. Steps 1–8 cover query embedding, topic selection, identity and value retrieval, dual-path node search, score-merge deduplication, REM principle matching, evidence expansion, and prompt assembly. The summary panel (right) reports final bucket sizes and total latency.