LITERARYBIGFIVE: Author-Personalized Text Generation in a Unified Interpretable Space

Jinghui Zhang[1], Lang Gao[1], Ao Li[2], Mingzhe Li[3], Ruihong Zeng[1], Zirui Song[1], Kentaro Inui[1,4,5], Xiuying Chen[1,*]
[1]MBZUAI, [2]Shandong University, [3]ByteDance, [4]Tohoku University, [5]RIKEN
{jinghui.zhang,lang.gao,ruihong.zeng,zirui.song,kentaro.inui,xiuying.chen}@mbzuai.ac.ae
liaolea@mail.sdu.edu.cn, limingzhe.lmz@bytedance.com

Abstract

Personalized text generation for authors and literary writing is essential for applications such as adaptive writing assistants, creative support tools, and computational literary analysis. However, existing approaches to author modeling and personalization often represent writing behavior as *independent labels*, requiring large-scale corpus collection or fine-tuning for each author or stylistic category. Such formulations are costly, difficult to interpret, and poorly suited for generalizing across authors. Inspired by the Big Five model's dimensional view of personality, we propose LITERARYBIGFIVE, a framework that reframes authorial writing characteristics as coordinates within a *unified and interpretable space.* In this space, we derive each interpretable axis (e.g., Classicism, Emotionality) from activation-space contrasts between author-written and neutral passages, yielding distinct stylistic dimensions that allow texts or authors to be positioned within a five-dimensional system. Beyond localizing different authors, we further introduce an interpretable steering mechanism, which adaptively guides text generation toward target coordinates to perform author-personalized writing. Experimental results show that LITERARYBIGFIVE improves authorial expressiveness while preserving semantic fidelity. The derived author per-axis scores strongly correlate with real-world literary consensus, offering transparent and interpretable explanations of author-specific generation behavior: Github.

1 Introduction

Personalized text generation models individual writing styles, especially for authors and literary writers with distinctive voices. It supports stylistic preference matching and voice emulation in applications such as creative writing (Yu et al., 2024; Qin et al., 2025) and personalized assistants (Zhang et al., 2025b; Ning et al., 2025).

* Corresponding author.

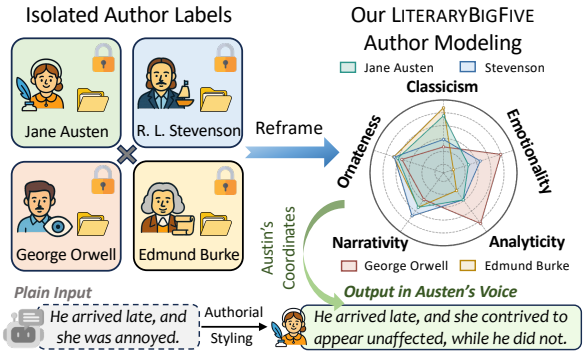

Figure 1: Previous work models each author as an isolated label; LITERARYBIGFIVE reframes authorial characteristics as a unified, interpretable space spanned by five axes, enabling measurement, comparison, and control across authors and books.

However, existing approaches predominantly treat individual authors as isolated categories. As shown in Figure 1, previous methods typically set up a separate task for each author, and apply specific prompting (Bhandarkar et al., 2024), model training (Jhamtani et al., 2017), or steering (Konen et al., 2024) to capture their characteristics independently. This design suffers from two key limitations. First, adapting to a new author typically requires collecting hundreds or thousands of texts or retraining the model (Zhang et al., 2025c), making large-scale expansion extremely costly and impractical. Second, it fails to reveal how different writing patterns relate to one another, offering limited interpretability or a unified representation of authorial variation.

Linguistic and literary studies offer a more systematic and theoretically grounded perspective for understanding complex writing patterns and stylistic variation across texts. Decades of analysis show that variation in written language is often organized along a few stable and interpretable dimensions, such as narrativity, emotion, and elaboration, rather than an unlimited and highly fragmented set of individual author labels (Martin and White, 2003; Kuiken and Jacobs, 2021; Biber and Gray, 2016).

This dimensional view echoes the idea behind the Big Five model in psychology, where complex human personalities are described by five high-level axes (Goldberg, 1993; John et al., 1999). These works suggest that *unified, interpretable coordinates* could support a more flexible paradigm for author personalization than categorical tags.

Building on these observations, we propose LITERARYBIGFIVE, a framework that reframes authorial characteristics as coordinates in a unified five-dimensional space rather than a set of unrelated labels. Concretely, we define five interpretable axes: Classicism, Ornateness, Narrativity, Emotionality, and Analyticity for our LITERARYBIGFIVE, which are informed by established literary and linguistic analysis (Biber and Conrad, 2019; Abbott, 2021; Biber and Gray, 2016; Booth, 1983). To construct the space, we select representative classics for each dimension and derive axis directions by contrasting original author-written and neutral passage pairs that preserve semantics while varying axis-specific features. Since raw activations show a general shift from neutral rewrites toward original literary texts that entangle distinct axes, we introduce an axis decomposition step to explicitly remove the shared principal component from all axes and reinterpret it as the overall expressiveness direction, thereby yielding the refined BigFive axes system. This improves axis independence and enables more stable multi-axis control over individual authorial traits.

With the LITERARYBIGFIVE space established, we propose *localize-and-steer* for both authorial coordinates analysis and personalized generation. For a target author, we first *locate* their position by projecting the reference text onto the BigFive directions, and yield unique scores that capture their authorial characteristics. This allows us to position different authors or books within a unified space and compare them along shared dimensions. Next, we leverage the BigFive axes for interpretable personalized generation by *steering* model activations toward target authorial coordinates, enabling immediate adaptation to a wide range of new authors, from classic novelists to contemporary writers.

To validate the effectiveness of LITERARYBIGFIVE on unseen authors, we evaluate it on books spanning distinct writing identities. Experiments show that LITERARYBIGFIVE better matches target authors while preserving meaning. Meanwhile, the learned author coordinates align with established literary consensus, suggesting that the space provides an interpretable representation of style.

In general, our contributions can be summarized as follows: (i) We introduce LITERARYBIGFIVE, a framework motivated by linguistic and literary studies that advances author modeling from isolated labels to a unified, interpretable five-dimensional space, capturing core dimensions of authorial variation. (ii) We propose a localize-and-steer mechanism that maps individual authors to precise coordinates in the LITERARYBIGFIVE space and enables latent space steering for personalized generation, adapting to new authors instantly without retraining. (iii) We demonstrate that LITERARYBIGFIVE improves authorial expressiveness while preserving semantic fidelity, and that the resulting axis scores align well with established literary consensus, providing transparent, per-dimension explanations of authorial characteristics.

2 Related Work

Personalized Text Generation. Personalized generation aims to align LLMs with specific user profiles or authorial identities while preserving semantic content (Zhang et al., 2025c). Traditional approaches often frame this as a supervised rewriting task requiring parallel corpora (Hu et al., 2017), or employ unsupervised disentanglement to separate content from linguistic expression (Prabhumoye et al., 2018). In the era of LLMs, the research focus has shifted to prompting (Reif et al., 2022) or fine-tuning on author-specific corpora (Wang et al., 2024). However, these methods typically treat individual authors as *independent, categorical labels*. This label-based paradigm scales poorly, as modeling a new author needs separate corpus collection, modeling retraining or extensive prompt engineering. We address this by learning a unified latent space that adapts to new authors instantly without such per-author overhead.

Dimensional Modeling of Linguistic Variation. Traditional stylometry often treats authors discretely, assigning each author a unique label and modeling style differences as class distinctions (Holmes, 1998). In contrast, linguistic studies have shown that written language can also be characterized along multiple interpretable dimensions, such as *involved* versus *informational* writing (Biber, 1991; Biber and Gray, 2016). These studies provide an empirical basis for dimensional analysis of linguistic variation, but are not designed for controlling language model generation at inference time. Beyond linguistics, the Big Five frame-

work in psychology also shows that complex human individual variation can be described through compact interpretable axes (Goldberg, 1993), and this dimensional view has been widely adopted in NLP research for assessing personality (Jiang et al., 2024) and simulating personas (Wang et al., 2024). However, it remains underexplored for controllable personalized text generation, where categorical style or author labels still dominate. Our work builds on this dimensional perspective and represents authorial writing characteristics as coordinates in a shared, interpretable space, enabling both author localization and activation-based steering for personalized generation.

Activation Steering. Activation steering modifies a model's output at inference time by intervening in its intermediate representations using direction vectors (Zou et al., 2023). By computing difference in latent activations between samples that express a target concept and those that do not, one can isolate a semantic vector corresponding to a specific attribute, and steering the model along this direction induces the associated behavior (Kim et al., 2018). Key advantages of activation steering include its interpretability, as abstract attributes are explicitly represented as vectors (Rimsky et al., 2024), as well as its efficiency compared to conventional adaptation methods such as finetuning (Gan et al., 2025). Recent studies have applied activation steering to personalized writing (Zhang et al., 2025a), emotion control (Banayeeanzade et al., 2025), and persona adoption (Chen et al., 2025). Although existing methods can steer models toward target outputs, they are mostly limited to single, binary traits or learn distinct vectors for individual authors. In contrast, LITERARYBIGFIVE enables multi-dimensional personalized steering within a unified space across diverse authors.

3 Problem Formulation

We formulate interpretable author-personalized generation as two coupled sub-tasks: *localization* and *steering*, jointly framed in a unified five-dimensional space. In the localization stage, let $\mathcal{S} = \mathbb{R}^5$ denote the LITERARYBIGFIVE space with interpretable axes. Given few k reference passages $\{x_{b,i}\}_{i=1}^{k}$ sampled from a target book b, a locator $\phi : \mathcal{T} \to \mathcal{S}$ maps each passage to its coordinates in $\mathcal{S}$. We estimate the target authorial coordinates by averaging the coordinates of its passages:

$$\mathbf{s}_b = \frac{1}{k} \sum_{i=1}^{k} \phi(x_{b,i}) \in \mathcal{S}, \tag{1}$$

which represents the author's or book's characteristic position within the space.

In the steering stage, given a neutral input passage x and the target position $\mathbf{s}_b$, the goal is to generate a rewritten passage:

$$\hat{x} = f(x, \mathbf{s}_b), \tag{2}$$

whose semantics remain consistent with x while its linguistic expression aligns with $\mathbf{s}_b$.

4 Method

In this section, we introduce the LITERARYBIGFIVE in detail. First, §4.1 presents the space construction with five axes; then, §4.2 describes how to locate a book to obtain its authorial coordinates; and finally, §4.3 explains how to steer generation to the target author using these coordinates.

4.1 LITERARYBIGFIVE Space Construction

Dimension Definitions and Representative Books. Drawing on prior work in linguistic and literary analysis (Biber and Conrad, 2019), we define five interpretable dimensions to capture several major variations in English literary writing. Ornateness reflects lexical richness and syntactic complexity, particularly within noun phrases, rather than simple sentence length (Biber and Gray, 2016). Narrativity distinguishes storytelling text, focused on action verbs and time markers (Abbott, 2021). Emotionality quantifies affective intensity through sentiment words, regardless of the specific topic (Booth, 1983). Classicism reflects the traditional writing patterns of the 18th and 19th centuries, emphasizing complex sentence structures and historical vocabulary (Boyd et al., 2022). Analyticity corresponds to expository and reasoning-oriented writing, characterized by abstract nouns and logical relations (Biber, 1995).

For each dimension, we select representative classics discussed in literary analysis to anchor the corresponding axis. Examples include Daniel Defoe's Robinson Crusoe, a foundational work for modern linear Narrativity (Watt, 1957), and Virginia Woolf's Mrs. Dalloway, distinguished by the intense Emotionality of its affective experience (Auerbach and Said, 2013). We curate selected books and clean these texts to retain only the authorial content and segment them into coherent passages. The full list of authors and books, detailed data sources, and preprocessing procedures are provided in Appendix C.

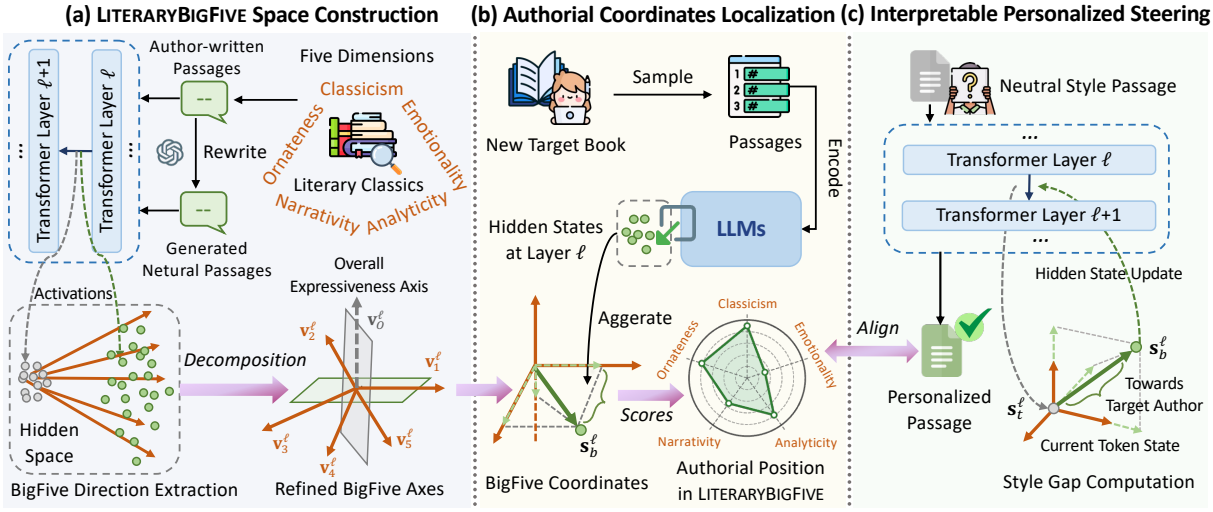

Figure 2: Illustration of LITERARYBIGFIVE framework. (a) We begin with constructing the LITERARYBIGFIVE space using author-written passages from selected literary classics that strongly exhibit each defined dimension, and extract BigFive direction with *axis decomposition*. (b) For a new target book, LITERARYBIGFIVE *locates* its authorial position by projecting reference passages onto the BigFive axes to obtain authorial coordinates. (c) During generation, we align the output with the target author by computing the *style gap* between the current token and the target coordinates, then updating the hidden states along the interpretable axes to close this gap.

Paired Dataset Construction. To derive the axes for the five dimensions, we construct paired passages that share semantics but differ in authorial expression. For each dimension $k \in \{1, \ldots, 5\}$, we denote the set of representative books as $\mathcal{B}_k$. Following (Ma et al., 2025), for each cleaned author-written passage $x^+_{b,i}$ from a book $b \in \mathcal{B}_k$, we use GPT-4 (OpenAI et al., 2024) to suppress authorial cues along the five defined dimensions (prompt in Appendix O.1), obtaining a semantics-preserving neutral rewrite $x^-_{b,i}$, yielding N_b passage pairs:

$$\mathcal{P}_{b,k} = \{(x^+_{b,i}, x^-_{b,i})\}_{i=1}^{N_b}. \tag{3}$$

The collection of all pairs for dimension k forms the dataset $\mathcal{D}_k = \bigcup_{b \in \mathcal{B}_k} \mathcal{P}_{b,k}$, and the complete corpus for axis construction is $\mathcal{D} = \bigcup_{k=1}^{5} \mathcal{D}_k$.

Axis Extraction. After defining the five dimensions and preparing representative author data for each axis, we next extract the corresponding *axis directions* in hidden space. Let $a^\ell(s) \in \mathbb{R}^d$ denote the last-token activation at layer ℓ for a token sequence s. The key is to identify, for each dimension, the activation shift that captures authorial variation rather than semantic content or positional bias. Hence, for each paired passage $(x^+_{b,i}, x^-_{b,i})$, we use the same neutral input $x^-_{b,i}$ and concatenate it with either the author-written passage $x^+_{b,i}$ or the neutral rewrite $x^-_{b,i}$, and then compute the *author-written* and *neutral* hidden states:

$$\mathbf{h}^\ell_{+,b,i} = a^\ell(x^-_{b,i} \oplus x^+_{b,i}), \quad \mathbf{h}^\ell_{-,b,i} = a^\ell(x^-_{b,i} \oplus x^-_{b,i}),$$

where $\oplus$ concatenates the neutral input and the model's rewritten output. Since the input $x^-_{b,i}$ is identical in both sequences and only the output's *style* differs, we can obtain the contrast $\boldsymbol{\delta}^\ell_{b,i} = \mathbf{h}^\ell_{+,b,i} - \mathbf{h}^\ell_{-,b,i}$ that isolates stylistic differences. Finally, we average these vectors over all N pairs from anchor books and renormalize, yielding the raw axis $\tilde{\mathbf{v}}^\ell_k$ for the k-th dimension at layer ℓ:

$$\tilde{\mathbf{v}}^\ell_k = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \boldsymbol{\delta}^\ell_{b,i} \in \mathbb{R}^d. \tag{4}$$

Axis Refinement via Decomposition. In preliminary analyses, we observed that although the five directions $\{\tilde{\mathbf{v}}^\ell_k\}_{k=1}^5$ capture distinct linguistic variations across dimensions, they appear to share a global offset trend, i.e., all axes tend to move activations from the "neutral" region toward the "author-written" region, rather than changing in completely independent directions. To verify this intuition, we take the paired neutral and original passages used for axis construction, average their last-token activations across layers and project them into a 3D PCA space. As shown in Figure 3(a), the neutral (blue) and author-written (red) samples form two compact clusters separated mainly along a single direction, revealing a strong global neutral→author-written shift shared across dimensions. Motivated by this finding, we propose to explicitly remove this shared component on a per-layer basis, in order to isolate axis-specific variations and improve the stability of multi-axis composition.

Concretely, for layer ℓ, we stack the five axis

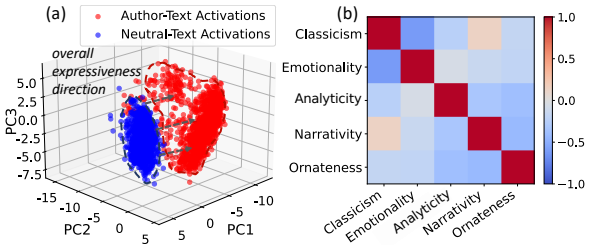

Figure 3: (a) 3D PCA reveals a dominant expressiveness direction from neutral-text (blue) to author-text (red) activations. (b) Correlation heatmap shows reduced cross-axis similarity after removing this global trend, demonstrating effective disentanglement of our refined dimensions. Results shown for LLaMA2-7B-Chat.

vectors into

$$\tilde{\mathbf{V}}^\ell = \left[\tilde{\mathbf{v}}_1^\ell, \tilde{\mathbf{v}}_2^\ell, \cdots, \tilde{\mathbf{v}}_5^\ell\right] \in \mathbb{R}^{d\times 5},$$

and perform Singular Value Decomposition (SVD) $\tilde{\mathbf{V}}^\ell = \mathbf{U}^\ell \mathbf{\Sigma}^\ell {\mathbf{Q}^\ell}^\top$. The first left singular vector $\mathbf{v}_O^\ell = \mathbf{U}_{:1}^\ell \in \mathbb{R}^d$ corresponds to the most dominant direction of variation shared across all dimensions. We identify this vector as the *overall expressiveness direction*, which captures the collective tendency of activations to drift toward more stylized representations. Simultaneously, we also compute its average magnitude ρ_O^ℓ by projecting the raw axes onto $\mathbf{v}_O^\ell$ to preserve the layer-wise intensity of this global trend. Next, to emphasize the unique contribution of each axis, we remove this global trend from every $\tilde{\mathbf{v}}_k^\ell$, and obtain the *refined unit axis* $\mathbf{v}_k^\ell$ with its magnitude ρ_k^ℓ:

$$\rho_k^\ell \cdot \mathbf{v}_k^\ell = \tilde{\mathbf{v}}_k^\ell - \mathbf{v}_O^\ell {\mathbf{v}_O^\ell}^\top \tilde{\mathbf{v}}_k^\ell. \tag{5}$$

We retain the extracted magnitudes ρ_O^ℓ and $\boldsymbol{\rho}^\ell = [\rho_1^\ell, \ldots, \rho_5^\ell]$ to restore the natural scale of interventions during the steering phase. As shown in Figure 3(b) and Appendix Figure 6, this decomposition step effectively reduces cross-axis correlations, and leads to more stable multi-axis combination.

4.2 Authorial Coordinates Localization

After deriving the axes, we aim to localize a target book's position within the LITERARYBIGFIVE space. Let $\{x_{b,i}\}_{i=1}^k$ be k reference passages from test book b. For layer ℓ, let $a^\ell(x) \in \mathbb{R}^d$ denote the last-token activation and define its unit-normalized form $\widehat{a}^\ell(x) = \mathrm{norm}\big(a^\ell(x)\big)$. We stack the refined axes into the basis matrix:

$$\mathbf{V}^\ell = \left[\mathbf{v}_1^\ell, \mathbf{v}_2^\ell, \cdots, \mathbf{v}_5^\ell\right] \in \mathbb{R}^{d\times 5}.$$

Since the inner product with unit axes provides a signed, scale-invariant measure of each dimension's intensity, for each reference passage, we can obtain per-layer authorial scores $\mathbf{s}_{b,i}^\ell$ by projecting its unit activation $\widehat{a}^\ell(x_{b,i})$ onto the five axes:

$$\mathbf{s}_{b,i}^\ell = {\mathbf{V}^\ell}^\top \widehat{a}^\ell(x_{b,i}) \in \mathbb{R}^5. \tag{6}$$

To interpret the style of the book and guide generation toward it, we aggregate these passage-level projections over k references, yielding the book-level raw coordinates $\mathbf{s}_b^\ell$ at layer ℓ:

$$\mathbf{s}_b^\ell = \tfrac{1}{k}\textstyle\sum_{i=1}^k \mathbf{s}_{b,i}^\ell \in \mathbb{R}^5. \tag{7}$$

These per-layer scores serve as personalized steering targets at the selected intervention layers $\mathcal{L}$, capturing linguistic attributes ranging from local syntax to global semantics encoded at different depths (Geva et al., 2021). Furthermore, for analyzing the book's writing characteristics, we average the scores $\mathbf{s}_b^\ell$ across $\ell \in \mathcal{L}$ to derive book-level authorial coordinates:

$$\mathbf{s}_b = \tfrac{1}{|\mathcal{L}|}\textstyle\sum_{\ell\in\mathcal{L}} \mathbf{s}_b^\ell \in \mathbb{R}^5. \tag{8}$$

Note that the coordinates above are not directly comparable across dimensions, as different axes exhibit varying dynamic ranges. To establish a consistent scale, we calibrate each dimension of $\mathbf{s}_b$ to a $[0, 100]$ range using axis-specific anchor books (detailed procedure in Appendix E), enabling intuitive visualization of authorial profiles.

4.3 Interpretable Personalized Steering

After localizing the target book b's authorial position in the LITERARYBIGFIVE space as $\mathbf{s}_b$, we steer the model to rewrite an input passage x into $\hat{x}$ so that its writing patterns align with the target authorial style while preserving the original semantic content. Unlike prior methods that rely on a single direction per author (Konen et al., 2024; Zhang et al., 2025a), our method performs editing in an interpretable axis-aligned space with explicit and controllable per-dimension modulation.

Let $\mathbf{h}_t^\ell \in \mathbb{R}^d$ denote the hidden state at token t and layer ℓ, and let $\widehat{\mathbf{h}}_t^\ell = \mathrm{norm}\big(\mathbf{h}_t^\ell\big)$ be its normalized form. To stabilize steering, we augment the five-axis directions $\mathbf{V}^\ell$ with the shared expressiveness direction $\mathbf{v}_O^\ell$, which captures the global shift from neutral to literary style. During generation, we project $\widehat{\mathbf{h}}_t^\ell$ onto both $\mathbf{V}^\ell$ and $\mathbf{v}_O^\ell$ to obtain the current layer-wise authorial scores:

$$\mathbf{s}_t^\ell = {\mathbf{V}^\ell}^\top \widehat{\mathbf{h}}_t^\ell \in \mathbb{R}^5, \quad s_{O,t}^\ell = \big\langle \widehat{\mathbf{h}}_t^\ell, \mathbf{v}_O^\ell \big\rangle.$$

Similarly, the target overall expressiveness score is computed by averaging the projections of k reference passages onto $\mathbf{v}_O^\ell$:

$$s_{O,b}^\ell = \frac{1}{k}\sum_{i=1}^{k} \langle \widehat{a}^\ell(x_{b,i}),\, \mathbf{v}_O^\ell \rangle. \tag{9}$$

Based on the current and target scores above, we can observe the ***style gap*** $\mathbf{s}_b^\ell - \mathbf{s}_t^\ell$ (and $s_{O,b}^\ell - s_{O,t}^\ell$) between the current t-th token and the target author. This gap indicates along which direction to move and by how much to bring the token closer to the target author in our LITERARYBIGFIVE space. We therefore convert it into edit strengths by rescaling it with the axis magnitudes $\boldsymbol{\rho}^\ell$ and ρ_O^ℓ extracted in the decomposition step:

$$\begin{aligned} \boldsymbol{\alpha}_t^\ell &= \lambda\, \boldsymbol{\rho}^\ell \odot (\mathbf{s}_b^\ell - \mathbf{s}_t^\ell), \\ \alpha_{O,t}^\ell &= \lambda\, \rho_O^\ell \left(s_{O,b}^\ell - s_{O,t}^\ell\right), \end{aligned} \tag{10}$$

where $\odot$ denotes element-wise product and λ is a global control strength. Finally, using these obtained coefficients, we steer the current token to the target author by updating its hidden state $\mathbf{h}_t^\ell \to \mathbf{h}_t^{\ell\prime}$ for each selected intervention layer $\ell \in \mathcal{L}$:

$$\mathbf{h}_t^{\ell\prime} \;=\; \mathbf{h}_t^\ell \;+\; \mathbf{V}^\ell\, \boldsymbol{\alpha}_t^\ell \;+\; \alpha_{O,t}^\ell\, \mathbf{v}_O^\ell. \tag{11}$$

Edits proceed from shallow to deep layers, allowing style effects to accumulate across layers while avoiding over-correction at a single place.

5 Experiments

5.1 Experimental Setup

Evaluation Data. To evaluate LITERARYBIGFIVE across diverse authors, we further curate a test set consisting of well-known books: Reflections on the Revolution in France by Edmund Burke, 1984 by George Orwell, Kidnapped by R. L. Stevenson, and Pride and Prejudice by Jane Austen. The resulting evaluation comprises 590 passage-level samples totaling 5,716 sentences, which substantially exceeds standard benchmarks such as the Shakespeare test set by Xu et al. (2012), containing 1.4 thousand sentences. Each book possesses a distinct authorial expression, allowing for a comprehensive and rigorous assessment of our method's cross-author generalizability.

Evaluation Metrics. Following previous works (Krishna et al., 2020; Zhang et al., 2025a), we adopt ROUGE-1/L (Lin, 2004) to evaluate reconstruction quality by comparing generated passages against original texts of the target author.

To measure semantic preservation, we report embedding similarity (SIM), computed based on the cosine similarity of sentence representations encoded by the BGE model (Chen et al., 2024).

Beyond objective metrics, we leverage the strong capabilities of LLMs in evaluating complex writing characteristics (Ostheimer et al., 2024) by utilizing GPT-4 as a judge. Specifically, we rate passages on a 0-10 scale considering two key dimensions: i) *Authorial Adherence* measures how well the rewrite aligns with the target author's distinctive characteristics; and ii) *Semantic Fidelity*, which evaluates the preservation of original meaning (prompt in Appendix O.4). To mitigate potential bias, we complement this with human evaluation, where two annotators rate responses using the same two-dimensional criteria.

Baselines. We compare our LITERARYBIGFIVE against various state-of-the-art baselines categorized as follows: (1) Few-shot Prompting; (2) Supervised Fine-Tuning, specifically LLM-Steer (Han et al., 2024), which fine-tunes word embeddings via a linear transformation, and LoRA (Hu et al., 2022), a parameter-efficient low-rank adaptation method; (3) Activation Steering, including ICV (Liu et al., 2024), Mean-Centering (Jorgensen et al., 2023), and CAA (Rimsky et al., 2024), RepE (Zou et al., 2023). Method introductions are detailed in Appendix K.

Implementation Details. We apply Llama2-7B-Chat (Touvron et al., 2023) as the base LLM to implement our LITERARYBIGFIVE and all baselines, with additional results on Qwen2.5-3B-Instruct (Yang et al., 2025) reported in Appendix B. All experiments were conducted with NVIDIA RTX 5880 Ada GPUs. Detailed Hyperparameters setting are provided in the Appendix L.

5.2 Main Results

As shown in Table 1, across four stylistically diverse books, LITERARYBIGFIVE consistently outperforms all baselines on ROUGE, SIM, GPT-4 and human evaluations. We summarize three key observations. *(1) Interpretable, axis-aligned editing yields the strongest and most stable personalized generation.* Unlike conventional editing methods that operate in entangled latent spaces, LITERARYBIGFIVE leverages interpretable Big-Five axes to support context-aware and fine-grained author-specific adjustment (with qualitative examples provided in Appendix M), consistently achieving higher ROUGE scores while preserving se-

Table 1: Experimental results on four books. For all metrics, higher scores indicate better performance. The best-performing methods are highlighted in **bold**, all results are scaled to 0-100 (two-tailed paired t-test, p<0.01).

Method	*Reflections on the Revolution in France*					*1984*				
	ROUGE-1	ROUGE-L	SIM	GPT-4	Human	ROUGE-1	ROUGE-L	SIM	GPT-4	Human
Few-shot	38.0	28.7	73.9	64.6	42.5	55.1	46.6	87.1	69.7	54.5
LLM-Steer	44.3	35.2	92.5	65.4	60.3	52.7	45.9	90.8	68.8	53.9
LoRA	43.0	29.8	82.5	60.8	49.5	51.7	38.9	87.1	69.3	57.2
ICV	43.6	33.8	93.6	65.4	60.3	54.7	48.6	92.8	73.1	71.5
Mean-Centering	45.6	35.2	93.9	67.9	65.9	55.6	46.3	93.8	73.8	73.9
CAA	41.6	31.9	89.7	59.7	45.8	43.3	35.9	85.7	57.5	46.7
RepE	45.3	35.7	94.0	68.6	65.4	56.5	47.9	94.3	73.5	70.9
LITERARYBIGFIVE	**46.1**	**36.4**	**94.4**	**69.4**	**69.2**	**57.8**	**49.4**	**95.1**	**75.2**	**75.3**
Method	*Kidnapped*					*Pride and Prejudice*				
	ROUGE-1	ROUGE-L	SIM	GPT-4	Human	ROUGE-1	ROUGE-L	SIM	GPT-4	Human
Few-shot	51.0	41.0	84.8	58.5	47.6	48.2	37.3	82.3	52.6	48.9
LLM-Steer	50.3	43.3	91.2	58.8	57.1	44.8	37.0	90.3	55.2	51.5
LoRA	53.3	44.9	92.2	58.7	56.4	49.1	34.1	87.5	56.1	53.7
ICV	54.5	46.1	95.8	65.0	65.5	49.2	39.2	94.0	61.5	63.0
Mean-Centering	55.7	47.0	96.1	66.5	61.8	50.9	40.0	94.6	64.0	66.7
CAA	48.7	40.7	90.5	56.0	41.0	44.7	35.3	89.5	51.2	49.1
RepE	55.9	47.6	96.4	65.9	68.3	50.2	40.2	94.4	63.5	67.8
LITERARYBIGFIVE	**56.5**	**48.3**	**96.7**	**67.8**	**73.8**	**51.3**	**41.7**	**94.8**	**65.9**	**69.5**

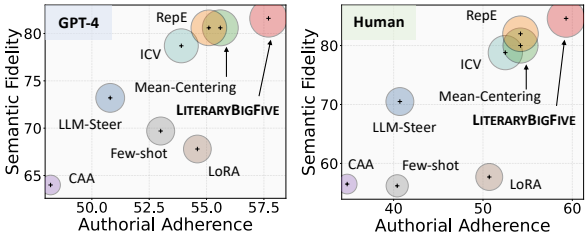

Figure 4: Performance analysis of Semantic Fidelity and Authorial Adherence. Radius denotes mean value.

Table 2: Ablation study on LITERARYBIGFIVE. **Bold** numbers indicate statistically significant improvements over the best baseline (two-tailed paired t-test, p<0.01).

Variants	ROUGE-1	ROUGE-L	SIM	GPT-4
-w/o Decomposition	52.1	43.2	94.7	68.8
-w/o Style Gap	49.3	40.0	91.6	66.6
LITERARYBIGFIVE	**53.0**	**44.0**	**95.2**	**69.6**

mantic fidelity and stylistic coherence. *(2) Robustness across diverse authors.* The evaluation ranges from Burke's political rhetoric to Austen's narrative prose. While baseline performance fluctuates, LITERARYBIGFIVE maintains high scores across all domains. This indicates that our model generalizes well to different styles without overfitting to specific corpora. *(3) High authorial adherence with robust semantic preservation.* A major challenge in personalized generation is aligning the output with a target author's writing characteristics without altering the original meaning. As visualized in Figure 4, LITERARYBIGFIVE occupies the optimal region (top-right), achieving the highest scores on both dimensions simultaneously. This demonstrates that our approach effectively disentangles authorial expression from content, enabling faithful personalization while preserving core semantics. Human annotators achieve a Cohen's κ of 0.59, demonstrating moderate inter-annotator agreement. It can also be observed that GPT-4's scores closely align with human evaluations, supporting the reliability of LLM-based assessment.

6 Analysis and Discussion

6.1 Ablation Study

We also conduct an ablation study to examine the contribution of each key component in LITERARYBIGFIVE, as shown in Table 2. First, removing the axis decomposition step in §4.1 and using raw book-level directions (*-w/o Decomposition*) leads to a noticeable drop across all metrics, indicating that refinement is essential for isolating clean, dimension-specific authorial signals. Furthermore, disabling the dynamic adaption of the style gap in §4.2 and applying a fixed steering strength (*-w/o Style Gap*) yields an even larger performance degradation than removing refinement. This highlights the crucial role of adaptive token-level steering, as authorial cues are unevenly distributed across a passage and require context-sensitive adjustment to avoid insufficient or excessive intervention. Overall, these findings validate that combining axis refinement with adaptive steering is necessary to achieve optimal personalization and semantic preservation.

6.2 Authorial Coordinates Analysis

To assess the interpretability of authorial coordinates derived in the LITERARYBIGFIVE space, we evaluate their alignment with independent stylistic judgements produced by frontier LLMs.

Table 3: Qualitative comparison of observed shifts, with linguistic analysis highlighted in shaded rows. We present cases with steering strengths $\alpha \in \{-0.8, +0.8\}$ here, while more results and analysis are available in Appendix N.

Dimension	Strength	Generated Text Snippet
Classicism	-0.8	...persons... who had caused resentment towards the throne by accepting its generous rewards...
	0.8	...persons... who had brought an odium on the throne by the prodigal dispensation of its bounties...
Analysis: High Classicism steers towards *Archaic Lexicon*. Note the shift from modern "caused resentment" to Latinate odium, and from simple "generous rewards" to more period-specific phrasing prodigal dispensation.		
Emotionality	-0.8	...Kitty was not completely surprised. I am very sorry. It is an imprudent match for both of them! But I hope for the best...
	0.8	...Kitty... does not seem so wholly unexpected. Our poor mother is sadly grieved. So imprudent a match on both sides! But I am willing to hope...
Analysis: High Emotionality drives *Affective Intensity*. The text shifts from neutral observation to personal sentiment, adding emotional weight through words like sadly grieved and emphatic structures ("So imprudent...").		
Ornateness	-0.8	...She is friendly and gracious, and she will probably pay some attention to you...
	0.8	...She is all affability and condescension, and I doubt not but you will be honoured with some portion of her notice...
Analysis: High Ornateness promotes *Syntactic Complexity*. Straightforward adjectives like ("friendly") are elaborated into abstract noun phrases (affability and condescension), resulting in a more decorative and indirect writing style.		

Accordingly, we ask GPT-5, Claude 3.5, and Gemini 3 to rate each book on a 0–100 scale along the five dimensions defined in LITERARYBIGFIVE (prompt in Appendix O.5). We then compute the Pearson correlation between our model coordinates and the ensemble average of the LLM scores. Results in Appendix G show strong alignment between LITERARYBIGFIVE and the LLM consensus, with an average Pearson correlation of $r = 0.96$ across all axes. As shown in the radar charts (Figure 5), our method captures stylistic patterns consistent with advanced LLM judgments. For example, the high *Classicism* of Edmund Burke and the high *Analyticity* of George Orwell are reflected in both our coordinates and the LLM ratings. Discrepancies in the radar plots further improve interpretability by revealing dimensions where our model diverges from the LLM consensus.

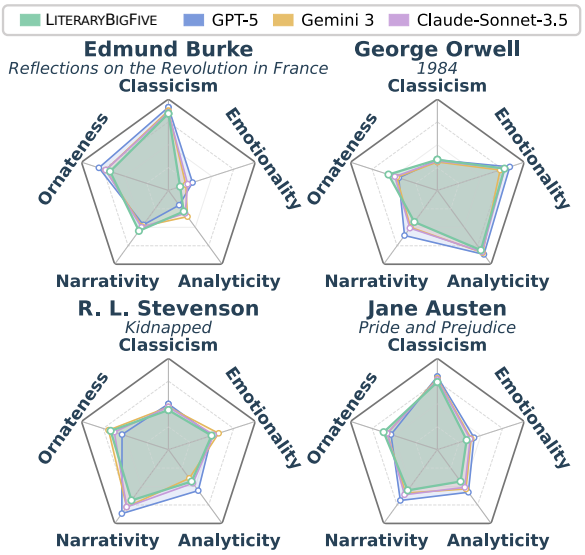

Figure 5: Radar charts comparing LITERARYBIGFIVE coordinates with LLM-based authorial scores across four authors. The strong overlap indicates consistent authorial stylistic characterization.

6.3 Case Study: Dimension Steering

To explore the interpretability of LITERARYBIGFIVE and gain qualitative insight into individual dimensions, we conduct a case study examining stylistic shifts induced by steering along a single dimension. Specifically, we randomly sample 40 texts from our test set. For each text, we apply the axis to one target dimension at a time while keeping other dimensions at zero to ensure isolation, and generate rewrites by varying the steering strength $\alpha \in \{-0.8, -0.4, 0, +0.4, +0.8\}$. As reflected in Table 3 (full version is in Appendix N), the results show that BigFive axes effectively modulate corresponding dimensions, such as the transition to Latinate diction in Classicism, without changing the underlying meaning of the text.

7 Conclusion

We present LITERARYBIGFIVE, a framework that reframes isolated author's writing characteristics into a unified and interpretable five-dimensional space. By leveraging a *localize-and-steer* mechanism, our approach integrates precise, interpretable analysis of authorial expression with low-cost personalized generation for new authors. Experimental results demonstrate that LITERARYBIGFIVE outperforms baselines in authorial expressiveness and semantic fidelity, while the derived coordinates closely match established literary consensus. Future work may extend this paradigm to multilingual settings and interactive writing support systems.

Limitations

Despite the effectiveness of our framework, we acknowledge specific constraints in its design and application. First, the axes in this study are primarily derived from English literary classics, which reflects the currently limited exploration of this task within the broader research landscape. We expect future work to extend this approach to richer linguistic and literary settings. Second, the steering mechanism operates globally on the residual stream layers. While this approach effectively captures holistic writing attributes, it lacks the granularity required to manipulate specific long-range dependencies, which might be better addressed by targeting individual attention heads or specific components. Finally, our method relies on the extraction and manipulation of internal activation vectors. This dependency on white-box access limits the framework's applicability to open-weight models and prevents it from working with closed-source language model APIs that do not provide direct access to these internal embeddings.

Ethical Considerations

We prioritize the responsible development of personalized text generation frameworks and strictly adhere to ethical guidelines regarding data usage and model deployment. All datasets used in our experiments are derived from publicly available sources, primarily consisting of literary works in the public domain, and no private, sensitive, or personally identifiable information is included. Consequently, the generation process follows open and reproducible settings without targeting any real individuals. While our framework enables the adaptation of writing characteristics, we acknowledge the potential risks associated with automated author imitation, such as non-consensual impersonation or the generation of misleading content. We emphasize that LITERARYBIGFIVE is designed specifically for creative support, literary analysis, and adaptive assistance; by grounding our method in interpretable literary dimensions, we aim to foster transparency in how text style is manipulated.

References

H Porter Abbott. 2021. *The Cambridge introduction to narrative*. Cambridge University Press.

Erich Auerbach and Edward W Said. 2013. Mimesis: The representation of reality in western literature-new and expanded edition.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. 2025. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. *Preprint*, arXiv:2510.04484.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 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)*.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge university press.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, New York.

Douglas Biber and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Wayne C Booth. 1983. *The rhetoric of fiction*. University of Chicago Press.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proceedings of ACL Findings*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Thomas Stearns Eliot. 2024. *The Sacred Wood, Essays on Poetry and Criticism*. Otbebookpublishing.

William Faulkner. 1956. William faulkner, the art of fiction no. 12. *The Paris Review*.

Jinwei Gan, Zifeng Cheng, Zhiwei Jiang, Cong Wang, Yafeng Yin, Xiang Luo, Yuchen Fu, and Qing Gu. 2025. Steering when necessary: Flexible steering large language models with backtracking. In *Proceedings of NeurIPS*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of ACL*.

Ernest Hemingway. 1999. *Death in the Afternoon*. Simon and Schuster.

Roger Henkle. 1970. The boundaries of fiction: Carlyle, macaulay, newman. In *Novel: A Forum on Fiction*.

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of ICML*.

Kazuo Ishiguro. 2007. The art of fiction no. 196. *Paris Review*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Proceedings of ACL Findings*.

Oliver P John, Sanjay Srivastava, and 1 others. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *Preprint*, arXiv:2312.03813.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of ICML*.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Proceedings of EACL Findings*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of EMNLP*.

Donald Kuiken and Arthur M Jacobs. 2021. *Handbook of empirical literary studies*. Walter de Gruyter GmbH & Co KG.

William Labov. 1972. *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press.

Richard A Lanham. 1991. *A handlist of rhetorical terms*. University of California Press.

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

Sheng Liu, Haotian Ye, Lei Xing, and James Y Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of ICML*.

Xinyu Ma, Yifeng Xu, Yang Lin, Tianlong Wang, Xu Chu, Xin Gao, Junfeng Zhao, and Yasha Wang. 2025. DRESSing up LLM: Efficient stylized question-answering via style subspace editing. In *Proceedings of ICLR*.

James R Martin and Peter R White. 2003. *The language of evaluation*. Springer.

Michael McKeon. 2002. *The origins of the English novel, 1600-1740*. JHU Press.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2025. User-llm: Efficient llm contextualization with user embeddings. In *Companion Proceedings of the ACM on Web Conference 2025*.

Alistair Niven. 1978. *DH Lawrence: the novels*. Cambridge University Press.

OpenAI, Josh Achiam, Steven Adler, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of LREC-COLING*.

Walter Pater. 2023. *The Renaissance: studies in art and poetry*. Univ of California Press.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of ACL*.

Hua Xuan Qin, Guangzhi Zhu, Mingming Fan, and Pan Hui. 2025. Toward personalizable ai node graph creative writing support: Insights on preferences for generative ai features and information presentation across story writing processes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of ACL*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of ACL*.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Brian Vickers. 1968. *Francis Bacon and renaissance prose*. Cambridge.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Proceedings of ACL Findings*.

Ian Watt. 1957. *The Rise of the Novel: Studies in Defoe, Richardson and Fielding*. University of California Press, Berkeley.

W.K. Wimsatt. 1941. *The Prose Style of Samuel Johnson*. Yale University Press.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of ACL*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of ACL*.

Jinghui Zhang, Kaiyang Wan, Longwei Xu, Ao Li, Zongfang Liu, and Xiuying Chen. 2025b. From individuals to crowds: Dual-level public response prediction in social media. In *Proceedings of ACM MM*.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025c. Personalization of large language models: A survey. *TMLR*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Algorithm for LITERARYBIGFIVE

Algorithm 1 presents the overall procedure of LITERARYBIGFIVE, including the construction of interpretable literary axes, the localization of target authorial coordinates, and the adaptive steering of generation based on the style gap between the current hidden state and the target coordinates.

Algorithm 1 LITERARYBIGFIVE Framework.

Require: LLM M, paired anchor passages $\mathcal{D}$, target passages $\{x_{b,i}\}_{i=1}^{m}$, input x, layers $\mathcal{L}$, strength λ.
Ensure: Stylized output $\hat{x}$.

Offline: Literary space construction

for each dimension $k \in \{1, \ldots, 5\}$ and layer $\ell \in \mathcal{L}$ **do**
  Compute contrast vectors $\boldsymbol{\delta}_i^\ell = a^\ell(x_i^- \oplus x_i^+) - a^\ell(x_i^- \oplus x_i^-)$.
  Obtain raw axis $\tilde{\mathbf{v}}_k^\ell \leftarrow \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\delta}_i^\ell$.
end for
for each layer $\ell \in \mathcal{L}$ **do**
  Stack $\tilde{\mathbf{V}}^\ell = [\tilde{\mathbf{v}}_1^\ell, \ldots, \tilde{\mathbf{v}}_5^\ell]$ and perform SVD.
  Extract shared expressiveness axis $\mathbf{v}_O^\ell$.
  Remove $\mathbf{v}_O^\ell$ from each raw axis to obtain refined axes $\mathbf{V}^\ell$ and magnitudes $\boldsymbol{\rho}^\ell$.
end for

Online: Target localization

for each layer $\ell \in \mathcal{L}$ **do**
  $\mathbf{s}_b^\ell \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mathbf{V}^{\ell\top} \mathrm{norm}(a^\ell(x_{b,i}))$.
  $s_{O,b}^\ell \leftarrow \frac{1}{m}\sum_{i=1}^{m} \langle \mathrm{norm}(a^\ell(x_{b,i})), \mathbf{v}_O^\ell \rangle$.
end for

Online: Interpretable steering

for each generated token t and layer $\ell \in \mathcal{L}$ **do**
  $\mathbf{s}_t^\ell \leftarrow \mathbf{V}^{\ell\top} \mathrm{norm}(\mathbf{h}_t^\ell)$, $s_{O,t}^\ell \leftarrow \langle \mathrm{norm}(\mathbf{h}_t^\ell), \mathbf{v}_O^\ell \rangle$.
  $\boldsymbol{\alpha}_t^\ell \leftarrow \lambda \boldsymbol{\rho}^\ell \odot (\mathbf{s}_b^\ell - \mathbf{s}_t^\ell)$, $\alpha_{O,t}^\ell \leftarrow \lambda \rho_O^\ell (s_{O,b}^\ell - s_{O,t}^\ell)$.
  $\mathbf{h}_t^{\ell\prime} \leftarrow \mathbf{h}_t^\ell + \mathbf{V}^\ell \boldsymbol{\alpha}_t^\ell + \alpha_{O,t}^\ell \mathbf{v}_O^\ell$.
end for
return generated output $\hat{x}$.

B Generalization to Other Backbones

To further evaluate the cross-model generalizability of LITERARYBIGFIVE, we additionally conduct experiments on Qwen2.5-3B-Instruct (Yang et al., 2025). As shown in Table 4, LITERARYBIGFIVE achieves the best performance across all four books and all evaluation metrics. These results suggest that the effectiveness of LITERARYBIGFIVE is not tied to a specific backbone, and can extend across different model series and scales.

C Dataset Details

C.1 Dataset Construction

Guided by the defined five dimensions, we curate an author-personalization dataset from English literary classics in the open-access Gutenberg Library, which hosts over 75,000 ebooks. The texts are freely available through Project Gutenberg, and we follow its Terms of Use and Project Gutenberg License for data access and redistribution.

To construct the LITERARYBIGFIVE axes, we select 10 English literary classics, each authored by a distinct well-known writer, with details provided in Appendix C.2.

For evaluation, we further curate four held-out books with distinct authorial expressions: *Reflections on the Revolution in France* by Edmund Burke, *1984* by George Orwell, *Kidnapped* by R. L. Stevenson, and *Pride and Prejudice* by Jane Austen. Each book possesses a distinct authorial expression, allowing us to evaluate LITERARYBIGFIVE across diverse writing patterns.

Due to formatting and compilation artifacts in the raw Gutenberg files which may distort analysis, we implement a cleaning pipeline to ensure the dataset focuses solely on literary content rather than formatting artifacts. Specifically, we remove indentation symbols not present in the original texts and delete lines consisting of repeated "=" symbols that lack semantic value. Regarding line segmentation, we delete isolated line breaks used for visual alignment and reduce multiple consecutive line breaks to correctly preserve paragraph boundaries. We also filter out unrelated segments, such as compiler contact information and hyperlinks. Following this preprocessing, we segment the clean texts into passages with a length constraint of 120 to 400 tokens. After preprocessing and segmentation, the axis-construction corpus comprises 1,322 passages spanning 12,741 sentences, while the held-out evaluation set comprises 590 passage-level samples totaling 5,716 sentences.

To construct authorial–neutral passage pairs, we rewrite each author-written passage x^+ into a neutralized version x^- using an LLM, where the rewrite strips authorial cues while preserving the original meaning (Ma et al., 2025; Zhang et al., 2025a). This pairing isolates authorial traits from content differences: x^+ and x^- hold the same meaning, but only x^+ carries the author's voice. Concretely, we prompt GPT-4 (OpenAI et al., 2024) to suppress authorial cues across the five dimensions defined above without altering the core content, the prompt is provided in Appendix O.1. At evaluation time, the neutralized passage x^- is used as input, and the original author-written passage x^+ serves as the target reference.

To verify data quality, we conduct an empirical analysis on 100 randomly selected passage pairs after preprocessing and neutralization. Specifically,

Table 4: Experimental results on Qwen2.5-3B-Instruct. For all metrics, higher scores indicate better performance. The best-performing methods are highlighted in **bold**, and all results are scaled to 0–100.

Method	*Reflections on the Revolution in France*				*1984*			
	ROUGE-1	ROUGE-L	SIM	GPT-4	ROUGE-1	ROUGE-L	SIM	GPT-4
Few-shot	31.9	21.3	85.9	44.2	32.1	22.6	85.6	41.6
LLM-Steer	39.5	26.2	89.3	41.8	44.1	31.4	88.6	39.8
LoRA	42.2	35.7	90.0	46.5	62.5	56.2	94.8	52.4
ICV	36.7	24.6	89.5	33.6	49.6	41.8	86.8	34.9
Mean-Centering	50.6	40.2	94.0	47.4	56.2	48.3	92.9	47.1
CAA	45.8	34.6	91.6	39.8	48.9	40.3	89.6	41.2
RepE	38.8	30.2	86.5	43.1	42.2	35.4	86.3	45.6
LITERARYBIGFIVE	**51.5**	**41.8**	**94.1**	**49.3**	**63.9**	**57.8**	**97.4**	**54.1**
Method	*Kidnapped*				*Pride and Prejudice*			
	ROUGE-1	ROUGE-L	SIM	GPT-4	ROUGE-1	ROUGE-L	SIM	GPT-4
Few-shot	35.4	25.1	85.9	38.7	35.2	24.0	87.8	35.9
LLM-Steer	44.4	31.6	88.5	40.8	41.6	28.2	89.4	37.2
LoRA	59.1	52.5	96.5	50.6	56.2	47.5	95.7	54.3
ICV	39.8	30.6	91.4	34.2	33.8	23.3	88.6	24.8
Mean-Centering	51.0	42.3	92.5	52.1	49.1	38.1	91.7	45.0
CAA	47.7	38.6	90.1	53.0	45.3	35.0	90.1	42.1
RepE	49.2	40.7	90.2	56.4	44.1	33.8	87.9	49.5
LITERARYBIGFIVE	**62.0**	**55.3**	**97.6**	**57.2**	**59.9**	**50.8**	**97.8**	**56.0**

we check whether formatting artifacts have been removed, whether the neutralized passage retains the core meaning of the original, and whether distinctive authorial expressions are sufficiently suppressed. Overall, the inspected pairs appear clean and suitable for both axis construction and evaluation. This suggests that our pipeline removes formatting noise while preserving semantic content effectively, providing a usable contrast for extracting authorial directions.

C.2 Anchor Writers and Books

To construct the LITERARYBIGFIVE space, we selected representative "anchor" literary works for each dimension, whose writing patterns are representative of the corresponding dimension and have been discussed in prior literary and linguistic analysis. The operational definitions and corresponding anchor books used to instantiate the positive direction of each axis are as follows:

- **Analyticity.** This dimension focuses on logical reasoning and propositional density. Following Biber's Multidimensional Analysis (Biber, 1995), high analyticity is marked by a high frequency of *abstract nouns* and *logical connectors* (e.g., causal and conditional links), which facilitate complex information integration. The selected books are as follows:
  - *The Sacred Wood* by T. S. Eliot
  - *The Problems of Philosophy* by Bertrand Russell

  Rationale: These works represent a logic-driven style that values intellectual clarity. Both Eliot and Russell provide representative examples of argument-driven prose, where the writing is organized around conceptual development and logical progression (Eliot, 2024).

- **Ornateness.** This dimension represents aesthetic richness. It is characterized by *high vocabulary diversity* and *complex sentence structures*, particularly through the frequent use of descriptive phrases and extra details attached to nouns to create vivid imagery (Lanham, 1991). The selected works are as follows:
  - *Sartor Resartus* by Thomas Carlyle
  - *The Renaissance* by Walter Pater

  Rationale: Carlyle's prose is known for its intentional "overflow" of language to match his complex philosophical themes (Henkle, 1970), while Pater's work is closely associated with the Aesthetic Movement, using rhythmic and highly decorated sentences to elevate the sensory experience of the reader (Pater, 2023).

- **Narrativity.** This dimension captures event-driven storytelling. Following established narrative theory (Labov, 1972), high narrativity is identified by the frequent use of *action verbs* and *time markers* (e.g., "then," "afterward") that move the plot forward in a clear sequence. The selected works are as follows:

- *Robinson Crusoe* by Daniel Defoe
- *The Call of the Wild* by Jack London

Rationale: These texts provide representative examples of linear, event-driven storytelling. Defoe's *Crusoe* is widely discussed for its step-by-step account of physical actions (Watt, 1957), while London's direct and action-focused prose offers another anchor for narrative progression.

- **Emotionality.** This axis measures the intensity of the characters' internal feelings and psychological states. It is characterized by the use of *emotive adjectives*, *exclamations*, and verbs related to internal thoughts, reflecting the "inward turn" of the novel (Auerbach and Said, 2013). The selected works are as follows:
  - *Mrs. Dalloway* and *To the Lighthouse* by Virginia Woolf
  - *Sons and Lovers* and *Women in Love* by D. H. Lawrence

  Rationale: These works prioritize affective subjective experience over external plot. Woolf's novels are famous for capturing the fluid "stream of consciousness" (Auerbach and Said, 2013), while Lawrence's novels explore the raw, deep-seated emotional and psychological tensions between individuals (Niven, 1978).

- **Classicism.** This dimension captures formal and balanced prose patterns associated with 18th- and 19th-century English writing. This period is chosen because it reflects relatively standardized prose conventions, including shared expectations about structure and decorum before many 20th-century experimental writing practices (McKeon, 2002). The selected works are as follows:
  - *The Rambler* by Samuel Johnson
  - *The Spectator* by Addison and Steele

  Rationale: These authors are central figures in the "Golden Age" of English essay writing. Johnson's work exemplifies Neoclassical balance and symmetry (Wimsatt, 1941), while the essays in *The Spectator* helped popularize a formal, polite, and standardized prose style.

D Effectiveness of Axis Decomposition

To directly verify the effect of the decomposition step, Figure 6 compares the pairwise cosine similarity heatmaps of the five axes before and after decomposition. Before decomposition, the raw axes exhibit uniformly high positive cross-axis similarity, indicating a substantial shared component, which we term the overall expressiveness direction.

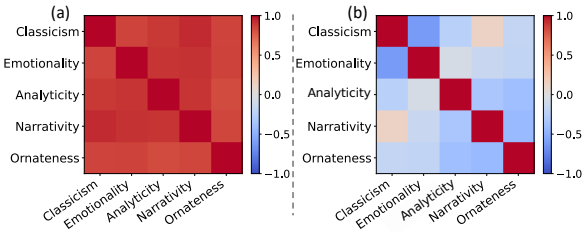

Figure 6: Pairwise cosine similarity heatmaps of the five axes before and after decomposition. (a) Before decomposition, the raw axes are strongly correlated. (b) After decomposition, cross-axis similarity is substantially reduced. Results shown for LLaMA2-7B-Chat.

After decomposition, the cross-axis similarities are markedly reduced, with the mean absolute off-diagonal cosine decreasing from 0.87 to 0.27. This substantial drop provides direct evidence in Section 4.1 that the decomposition effectively removes the shared global trend among the raw axes, thereby yielding more disentangled and dimension-specific directions for stable multi-axis composition.

E Coordinate Calibration

Raw per-axis coordinates can differ in dynamic range across dimensions, so we calibrate them using the *all anchor passages* employed to construct each axis (i.e., all passages from the representative books for that dimension). For the k-th axis, let $\mathcal{A}_k$ be its anchor corpus, we compute layer-averaged projection on axis k for each passage $x_i \in \mathcal{A}_k$:

$$s(x_i; k) \;=\; \tfrac{1}{|\mathcal{L}|} \textstyle\sum_{\ell \in \mathcal{L}} \langle \widehat{a}^{\ell}(x_i),\, \mathbf{v}_k^{\ell} \rangle,$$

and collect the projections of all passages $\mathcal{P}_k = \{s(x_i; k) : x_i \in \mathcal{A}_k\}$. We then set an axis-specific scale α_k from this distribution $\mathcal{P}_k$ to make coordinates comparable across dimensions. Concretely, we use the 95th percentile of $|p|$, which is a standard robust-scaling choice that limits the influence of outliers, avoids saturating typical books at the bounds, and remains stable as the anchor corpus grows:

$$\alpha_k \;=\; \mathrm{quantile}_{0.95}\left(\{|p| : p \in \mathcal{P}_k\}\right).$$

Given a book's layer-averaged raw scores $\mathbf{s}_b \in \mathbb{R}^5$ from § 4.2, its calibrated coordinate on axis k is

$$z_{b,k} \;=\; \mathrm{clip}\left(\frac{s_{b,k}}{\alpha_k},\, -1,\, 1\right),$$

combining every dimension together forms a five-dimensional score vector $\mathbf{z}_b \in [-1, 1]^5$. Building upon this, we map to $[0, 100]$ for visualization in radar plots:

$$R_{b,k} \;=\; 50\,(1 + z_{b,k}), \quad \mathbf{R}_b = (R_{b,1}, \ldots, R_{b,5}).$$

F Detailed Evaluation Scores

In this section, we report the complete GPT-4 and human evaluation scores across the two dimensions, namely Semantic Fidelity (SF) and Authorial Adherence (AA). Table 7 presents the average performance, providing the numerical data visualized in Figure 4, while Table 5 provides a detailed breakdown for each of the four books.

G Authorial Coordinate Scores

We provide the detailed authorial scores in Figure 5, as shown in Table 6.

To better show how these stylistic dimensions separate by book, we plot the score distributions for each author-written passage from our test set. As shown in Figure 7, the results are very consistent with known writing patterns of the selected books. Orwell's *1984* has a high level of *Analyticity*, which fits with its focus on complex political and social critique. In contrast, Burke's *Reflections* shows the highest scores for *Classicism* and *Ornateness*, as expected for formal, highly-stylized 18th-century work. *Kidnapped* stands out for *Narrativity*, reflecting its narrative-driven adventure story style. These clear, separated distributions prove that our LITERARYBIGFIVE framework can accurately capture and distinguish different author characteristics.

H Efficiency Comparison

We analyze the computational efficiency of LITERARYBIGFIVE from both theoretical and empirical perspectives, as summarized in Table 8.

Theoretical Complexity. Our method maintains a linear computational complexity of $O(K \cdot d)$ per token, where K is the number of vectors used to intervene the models (here $K = 6$) and d is the hidden dimension. This represents a significant theoretical advantage over fine-tuning methods like LLM-Steer (Han et al., 2024), which require a dense matrix multiplication with quadratic complexity $O(d^2)$. Even compared to parameter-efficient methods such as unmerged LoRA (Hu et al., 2022) with complexity $O(r \cdot d)$ (where r is the rank, in our settings $r = 8$), our approach remains more efficient as $K \leq r \ll d$. Given that $6 \leq 8 \ll 4096$ for Llama-2-7B-Chat, the theoretical FLOPs required by our steering mechanism are orders of magnitude lower than fine-tuning and more streamlined than LoRA configurations.

Inference Latency. To evaluate real performance, we measured the average inference latency (ms/token) over 100 generated cases in our test set. As shown in Table 8, static vector-based baselines (e.g., Mean-Centering, CAA) exhibit the lowest latency ($\sim$18.9–19.0 ms/token) since they apply a fixed bias. In contrast, the training-based LoRA baseline incurs higher latency (23.50 ms/token) due to the additional low-rank adapter computation. Despite the computational overhead of calculating projections and style gaps along K axes for dynamic adaptation, LITERARYBIGFIVE records a latency of 19.88 ms/token. This corresponds to a marginal overhead of less than 1.0 ms compared to the fastest static baseline (Mean-Centering, 18.93 ms) and is effectively equivalent to LLM-Steer (19.92 ms). These results demonstrate that our method's dynamic control comes at a practically negligible cost, remaining highly efficient for real-time generation while offering the unique capability of disentangled, interpretable personalized steering that static vector addition cannot achieve.

I BIGFIVE Dimension Vector Analysis

To investigate where and how the LITERARYBIGFIVE dimensions are encoded within the model's internal representations, we conduct a layer-wise linear probing analysis. Specifically, for each dimension, we train a logistic regression classifier on the hidden states $\mathbf{h}^\ell$ extracted from each layer ℓ to distinguish between texts exhibiting high versus low intensity along that dimension. Figure 8 illustrates the probing AUC trajectories across model layers, revealing two critical insights into how these dimensions are represented in the model.

Universal Dimension Encodability. First, we observe that the model achieves high classification performance (AUC > 0.90) across all five dimensions. This indicates that dimension-level information is not an abstract external label, but is robustly embedded within the LLM's latent space. Even without explicit supervision during pre-training, the

Table 5: GPT-4 and Human evaluation across four books on two dimensions: Semantic Fidelity (SF) and Authorial Adherence (AA). Best results are **bolded**, all results are scaled to a 0–100 scale.

Method	*Reflections on the Revolution in France*				*1984*			
	GPT-4		Human		GPT-4		Human	
	SF	AA	SF	AA	SF	AA	SF	AA
Few-shot	72.2	57.0	50.0	35.0	77.3	62.1	73.9	35.1
LLM-Steer	76.3	54.4	74.5	46.0	79.0	58.7	72.7	35.1
LoRA	68.4	53.2	52.8	46.2	75.1	63.4	60.6	53.8
ICV	77.4	53.4	72.3	53.2	84.1	62.2	83.6	59.3
Mean-Centering	80.4	55.4	78.2	53.5	84.6	63.1	85.7	62.0
CAA	67.5	51.9	59.5	32.0	64.0	51.1	56.1	37.3
RepE	**81.0**	56.1	77.7	53.0	84.9	62.0	83.5	58.2
LITERARYBIGFIVE	80.8	**58.1**	**81.8**	**56.5**	**85.5**	**65.0**	**86.5**	**64.1**

Method	*Kidnapped*				*Pride and Prejudice*			
	GPT-4		Human		GPT-4		Human	
	SF	AA	SF	AA	SF	AA	SF	AA
Few-shot	67.5	49.5	45.5	48.8	61.7	43.4	54.8	42.9
LLM-Steer	71.2	46.3	73.3	40.8	66.5	43.9	61.9	41.0
LoRA	65.3	52.1	60.1	52.6	62.4	49.8	57.2	50.2
ICV	79.1	50.8	85.0	45.9	74.1	48.9	74.4	51.5
Mean-Centering	80.3	52.8	78.2	45.3	77.0	51.0	77.8	55.6
CAA	64.9	47.0	50.8	31.2	59.8	42.7	59.5	38.6
RepE	80.4	51.4	84.2	52.3	76.3	50.7	82.5	53.1
LITERARYBIGFIVE	**81.6**	**54.1**	**89.7**	**57.8**	**78.5**	**53.3**	**80.5**	**58.5**

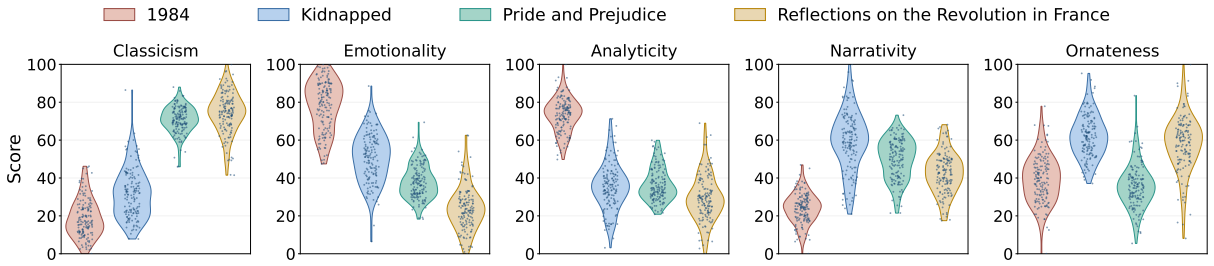

Figure 7: Per-book distributions across the five stylistic dimensions. The clear separation between books matches their known literary characteristics, demonstrating the framework's effectiveness.

model spontaneously learns to discriminate these dimension-specific patterns, validating the probing-based foundation of our steering approach.

Hierarchical Encoding of Each Dimension. Crucially, our fine-grained analysis reveals a clear layer-wise hierarchy regarding when different dimensions become linearly separable. While all dimensions are eventually encoded, they do so at different depths within the network:

- **Surface-Level Dimensions (Classicism, Narrativity):** As shown in the plots for *Classicism* and *Narrativity*, the AUC scores saturate rapidly, reaching near-perfect performance within the first few layers (Layers 0–5). This suggests that these dimensions are closely associated with lexical markers (e.g., archaic function words) or shallow syntactic patterns (e.g., verb and event distributions), which are captured early in the bottom-up processing.
- **Semantic-Level Dimensions (Analyticity, Emotionality, Ornateness):** In contrast, dimensions such as *Analyticity*, *Emotionality*, and *Ornateness* exhibit a more gradual ascent in AUC, peaking only in the middle-to-late layers (Layers 15–25). *Analyticity*, in particular, shows higher variance in lower layers, indicating that its reliable representation requires compositional reasoning and long-range contextual integration.

Overall, these results indicate that while shallow layers encode surface-level lexical and structural patterns, the representation of more abstract reasoning processes and affective nuances relies on the deeper abstraction capabilities of the network.

J More Authorial Coordinates Analysis

To further validate the robustness and discriminative ability of the LITERARYBIGFIVE space across

Table 6: Comparison of LITERARYBIGFIVE dimension scores across models on four books. Higher indicates stronger presence of the corresponding attribute.

Model	*Reflections on the Revolution in France*					*1984*				
	Classicism	Emotionality	Analyticity	Narrativity	Ornateness	Classicism	Emotionality	Analyticity	Narrativity	Ornateness
LITERARYBIGFIVE	75.7	12.1	25.7	49.4	60.4	30.3	69.5	72.9	38.5	50.6
GPT-5	82.0	25.0	18.0	42.0	72.0	30.0	75.0	78.0	55.0	40.0
Gemini 3	78.5	18.0	32.0	44.0	65.0	28.0	65.0	76.0	45.0	42.0
Claude-Sonnet-3.5	77.0	20.0	28.0	45.0	65.0	29.0	70.0	75.0	46.0	44.0
Model	*Kidnapped*					*Pride and Prejudice*				
	Classicism	Emotionality	Analyticity	Narrativity	Ornateness	Classicism	Emotionality	Analyticity	Narrativity	Ornateness
LITERARYBIGFIVE	39.0	44.7	38.9	61.9	59.8	66.6	30.3	38.9	49.7	55.6
GPT-5	45.0	45.0	50.0	78.0	48.0	72.0	38.0	52.0	62.0	48.0
Gemini 3	42.0	52.0	35.0	70.0	62.0	70.0	35.0	48.0	53.0	52.0
Claude-Sonnet-3.5	42.0	47.0	41.0	70.0	56.0	69.0	34.0	46.0	55.0	52.0

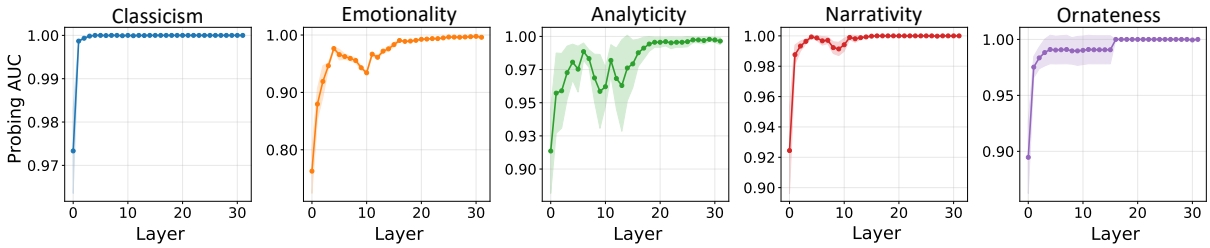

Figure 8: Layer-wise linear probing performance (AUC) across the five LITERARYBIGFIVE dimensions. The results reveal a hierarchical encoding mechanism: surface-level attributes (e.g., *Classicism*, *Narrativity*) saturate rapidly in early layers, whereas complex semantic attributes (e.g., *Emotionality*, *Analyticity*) require deeper processing to reach maximal separability.

Table 7: Performance comparison of different methods on two dimensions: Semantic Fidelity (SF) and Authorial Adherence (AA). Best results are **bolded**, all results are scaled to 0–100.

Method	GPT-4		Human	
	SF	AA	SF	AA
Few-shot	69.7	53.0	56.2	40.4
LLM-Steer	73.2	50.8	70.5	40.7
LoRA	67.8	54.6	57.7	50.7
ICV	78.7	53.9	78.8	52.5
Mean-Centering	80.6	55.6	80.0	54.2
CAA	64.0	48.2	56.5	34.8
RepE	80.6	55.1	82.0	54.2
LITERARYBIGFIVE	**81.6**	**57.7**	**84.6**	**59.3**

Table 8: Efficiency comparison. We report the theoretical Computational Complexity per token and the measured Inference Latency (ms/token).

Method	Complexity	Latency
Fine-tuning Methods		
LoRA	$O(r \cdot d)$	23.50
LLM-Steer	$O(d^2)$	19.92
Activation Steering Methods		
ICV	$O(d)$	19.49
Mean-Centering	$O(d)$	18.93
RepE	$O(d)$	19.29
CAA	$O(d)$	19.02
LITERARYBIGFIVE	$O(K \cdot d)$	19.88

a broader spectrum of authors, we analyze six additional authors with distinct writing patterns. Table 9 presents their localized coordinates, demonstrating how the model situates diverse authorial patterns within our five-dimensional framework.

The model's positioning aligns closely with established literary criticism. For instance, while Ernest Hemingway and William Faulkner both show high *Emotionality*, they are separated by *Ornateness* ($\Delta = 4.6$). Hemingway's lower score quantitatively reflects his "Iceberg Theory," which favors a sparse, direct lexicon over decorative language (Hemingway, 1999), whereas Faulkner's higher score captures his famously multi-layered sentence structures (Faulkner, 1956). Similarly, the contrast between Francis Bacon and Agatha Christie highlights nuances in *Narrativity*. Bacon's high scores in *Narrativity* (67.8) and *Classicism* (68.3) reflect the 17th-century rhetorical tradition, where progression is driven by explicit logical steps (Vickers, 1968). Conversely, Christie's lower *Narrativity* (34.0) reflects a style that relies more on dialogue and internal deduction than on physical action. Finally, the model captures the historical shift from 19th-century eloquence to modern restraint. John Henry Newman's high *Ornateness* (66.0) is consistent with Victorian rhythmic and stylized prose (Henkle, 1970), while Kazuo Ishiguro's lower score (48.7) and high *Emotionality* (80.2) accurately represent his intentional use of "plainspoken" language to mask deep psychological tension (Ishiguro, 2007).

K Baseline Details

In this section, we describe the baseline methods used in our experiments, categorized into prompt-

Table 9: Style coordinates for additional canonical authors in the LITERARYBIGFIVE space. Higher values indicate a stronger presence of the corresponding attribute.

Author	Work	Classicism	Emotionality	Analyticity	Narrativity	Ornateness
Ernest Hemingway	*The Old Man and the Sea*	24.0	81.1	65.1	45.4	51.0
William Faulkner	*The Sound and the Fury*	23.9	81.4	53.8	44.2	55.6
Francis Bacon	*The Essays*	68.3	9.5	33.6	67.8	63.8
Agatha Christie	*Murder on the Orient Express*	31.3	57.4	61.5	34.0	62.8
John Henry Newman	*Apologia Pro Vita Sua*	51.9	27.0	38.0	43.2	66.0
Kazuo Ishiguro	*Never Let Me Go*	24.3	80.2	66.7	49.6	48.7

ing, fine-tuning, and activation steering.

First, we use few-shot prompting as a basic comparison. Specifically, we prepend k reference passages written by the target author to the input prompt. This baseline evaluates how well the model can adapt its generation to an author's writing characteristics purely through in-context examples, without modifying any internal parameters. The specific prompts are listed in Appendix O.3.

Second, for methods that require training, we adopt LLM-Steer (Han et al., 2024) and LoRA (Hu et al., 2022). Instead of retraining the entire model, LLM-Steer learns a lightweight linear transformation over word embeddings to align the generated text with the target author's writing patterns, while LoRA injects trainable low-rank adapters into selected layers to achieve parameter-efficient style adaptation with a frozen backbone.

Finally, we compare our approach against four representative activation steering methods that intervene directly in the model's hidden states: (1) ICV (Liu et al., 2024), which extracts intervention vectors from few reference examples; (2) Mean-Centering (Jorgensen et al., 2023), which computes a fixed direction by subtracting the average activations of neutral rewrites from those of the target author's expressions; (3) CAA (Rimsky et al., 2024), which derives a steering direction by contrasting activations between author-written and neutral texts, thereby covering the mean-difference steering formulation used by StyleVector for personalized text generation (Zhang et al., 2025a); and (4) RepE (Zou et al., 2023), which identifies the principal direction of linguistic variation via PCA on contrastive activation pairs.

Since these baselines are typically designed for single-target steering and lack an inherent unified style space, we adapt them to our framework to ensure a fair comparison. Specifically, for all methods except ICV, we aggregate the anchor books used to construct each style axis (see Section §4.1) and utilize this full data for training, while ICV follows its standard setup.

L Implementation Details

We compare LITERARYBIGFIVE against several state-of-the-art steering and prompting baselines with specific hyperparameter configurations. For Few-shot prompting, we randomly sample 3 passages to guide generation. For LLM-Steer, we use the learned transform with $\epsilon_0 = 1 \times 10^{-3}$ scaled by a factor of 6 (i.e., $\epsilon = 6\epsilon_0$). For LoRA, we set the rank to 8 and train for 3 epochs. We use a cosine learning-rate scheduler with a warm-up stage covering 10% of total steps, with the maximum learning rate set to 5×10^{-5}. The batch size is 2 with gradient accumulation of 16. For methods involving activation steering, we adhere to the following configurations, selected by grid search: (1) For ICV, we randomly sample 3 reference passages to extract vectors and apply the intervention across all layers except the first embedding layer, with a steering strength $\alpha = 0.3$. (2) Regarding Mean-Centering, the editing strength is set to $\alpha = 1$ applied to layers $\ell \in \{22, 23, 25, 27, 29\}$. (3) For CAA, we set the scaling $\alpha = 1$ for layers $\ell \in \{23, 25, 27, 29\}$. (4) For RepE, we configure $\alpha = 6$ targeting layers $\ell \in \{22, 23, \ldots, 29\}$. Finally, for our proposed LITERARYBIGFIVE, we apply steering vectors with a global strength of $\lambda = 1$ across layers $\ell \in \{20, 24, 28\}$ and randomly sample 10 passages as reference. To rigorously evaluate semantic preservation and eliminate metric instability caused by random sampling, we set the decoding temperature to 0 for all experiments to ensure deterministic outputs.

M Case Study

We present representative examples from the test sets of each book. We highlight desirable linguistic features in green and hallucinations or flattening in red, followed by a detailed analysis for each case.

Table 10: Qualitative comparison on *Reflections on the Revolution in France* by Edmund Burke. We highlight desirable linguistic features in green and hallucinations or flattening in red.

LITERARYBIGFIVE (Ours)	RepE	Few-Shot
Original Text 1: "The first, calling itself the **Constitutional Society**... What improvements they have had... **I cannot tell: But I never heard a man... speak a word in praise** of the greater part of the publications..."		
The first group, known as the Constitutional Society... While it is possible that some of the books were sent to France... I have never heard anyone praise the publications circulated by this society...	The first group, known as the Constitutional Society... Some claim that books transported across the sea can improve... but this is unverified. Despite good intentions... their publications have not been widely praised...	The National Assembly, composed of practitioners in the law, has been around for about seven or eight years. ... I have heard talk about the glories of the National Assembly...
Analysis: Few-Shot suffers from severe hallucination, replacing "Constitutional Society" with "National Assembly". RepE accurately keeps the entity but flattens the style: it converts the author's strong first-person voice ("I cannot tell", "I never heard") into dry, objective summaries ("this is unverified", "have not been widely praised"). Our method preserves the subjective, critical voice essential to Burke's style.		
Original Text 2: "Well! but these men were to be tempered... **Were they then to be awed** by the... dignity of an handful of country clowns...? **No!** both these descriptions were more formed to be overborne..."		
The men were supposed to be influenced... However, were they intimidated by the authority and dignity of a few country people in the Assembly...? No, both groups were more likely to be influenced by the tactics of lawyers...	The men were supposed to be influenced... However, they were unlikely to be intimidated by the authority and dignity of a few country people... The lawyers would likely govern the whole...	The clergy had much to do with these transactions. Their estates were bound to the last acre. They were false to every other interest and held out the church lands as a security...
Analysis: Few-Shot retrieves completely irrelevant context about "the clergy" from a different chapter. RepE correctly identifies the content but loses the rhetorical structure, turning a dramatic Question-Answer sequence ("Were they...? No!") into a flat statement ("they were unlikely to be"). Our method successfully retains the rhetorical question and emphatic negation, preserving the argumentative tension.		

Table 11: Qualitative comparison on *1984* by George Orwell. We highlight desirable linguistic features in green and hallucinations or flattening in red.

LITERARYBIGFIVE (Ours)	RepE	Few-Shot
Original Passage 1: "...the shouting, the singing, the banners, the posters, the films, the waxworks, the rolling of drums and squealing of trumpets, the tramp of marching feet... when the great orgasm was quivering to its climax..."		
On the sixth day of Hate Week, after the events such as processions, speeches, shouting, singing, banners, posters, films, waxworks, drums, trumpets, marching feet, tank caterpillars, planes, and guns, when the intense excitement was reaching its peak...	On the sixth day of Hate Week, after the events such as processions, speeches, shouting, singing, banners, posters, films, waxworks, drums, trumpets, marching feet, tank caterpillars, planes, and guns, when the intense excitement was reaching its peak...	On the sixth day of Hate Week, after the events such as processions, speeches, shouting, singing, banners, posters, films, waxworks, drums, trumpets, marching feet, tank caterpillars, planes, and guns, after six days of this...
Analysis: While RepE converts the vivid description into a mechanical list (highlighted in red), losing the sensory overload, our method preserves the density and rhythm of the "wall of sound" effect essential to the scene's atmosphere.		
Original Passage 2: "The past, he reflected, had not merely been altered, it had been actually destroyed. For how could you establish even the most obvious fact when there existed no record outside your own memory?"		
Winston disliked this exercise... The past, he thought, had not only been changed, it had actually been destroyed. How could you confirm even the most obvious fact when there was no record except your own memory?	Winston disliked this exercise... The past seemed somewhat pleasant, yet it had been destroyed. He struggled to recall when he first heard about Big Brother...	Winston disliked this exercise... The somewhat pleasant aspect left his thoughts. Past destroyed, he thought, memories gone. How confirm even obvious fact when no record except own memory?
Analysis: RepE suffers from severe semantic drift, incorrectly describing the dystopian past as "somewhat pleasant." Few-Shot degrades into broken syntax ("Past destroyed..."). In contrast, our method retains the complex rhetorical question and the correct pessimistic sentiment.		

Table 12: Qualitative comparison on *Kidnapped* by R. L. Stevenson. We highlight desirable linguistic features in green and hallucinations or flattening in red.

LITERARYBIGFIVE (Ours)	RepE	Few-Shot
Original Text 1: "Well, David," said he... "Ardshiel was to starve... King George is to whistle for his rents... **what cares Red Colin? If he can hurt Ardshiel... he will gang hame singing to Glenure!**"		
..."Ardshiel was to be starved: that was his goal... King George will get less rent... what does Red Colin care? If he can harm Ardshiel... he will go home singing to Glenure!"	...Ardshiel was to be starved, his ultimate goal... King George would receive less rent... Red Colin didn't care... if he could take food... he would return home in high spirits to Glenure!	He then began to question me cunningly, whence I came, whether I was rich... We were now upon a sort of green cattle-track... the catechist grew angrier...
Analysis: Few-Shot suffers from a catastrophic hallucination, generating text about a "catechist" and a "cattle-track" (from a totally different chapter) instead of the Red Fox story. RepE preserves meaning but flattens the tone, changing the mocking "gang hame singing" to the generic "return home in high spirits." Our method preserves the cynical voice of the character.		
Original Text 2: "The captain was in the right of it. **We had run down a boat in the fog, and she had parted in the midst**... This man... had leaped up and caught hold of the brig's bowsprit..."		
The captain was correct. We had collided with a boat in the fog, and it had broken in the middle... This person... managed to grab the ship's bowsprit with his hands free...	The captain was correct. We had collided with a boat in the fog... This person... was thrown into the air and managed to grab the ship's bowsprit...	The sound of an infinite number of rivers came up from all round. In this steady rain the springs of the mountain were broken up... I could well understand the story of the Water Kelpie...
Analysis: Few-Shot again hallucinates a completely different scene (describing rain and rivers instead of a shipwreck). RepE is accurate but mechanical. Our method maintains the narrative flow and specific details (collision, bowsprit) without drifting into unrelated content.		

Table 13: Qualitative comparison on *Pride and Prejudice* by Jane Austen. We highlight desirable linguistic features in green and hallucinations or flattening in red.

LITERARYBIGFIVE (Ours)	RepE	Few-Shot
Original Text 1: "...Mr. Darcy soon drew the attention of the room by his fine, tall person... and the report... of his having ten thousand a year... till his manners gave a disgust which turned the tide of his popularity; for he was discovered to be **proud, to be above his company, and above being pleased**..."		
...Mr. Darcy quickly caught the attention... with his tall build... The men said he was a fine-looking man... He was found to be proud, above the company, and hard to please. Not even his large estate... could prevent him from being seen as unpleasant...	...The news of his ten thousand a year income spread soon... earning him praise from both men and women. He was admired... until his behavior turned people against him, marking him as proud, aloof...	During their walk, it was resolved that Mr. Bennet's consent should be asked... Elizabeth reserved to herself the application for her mother's... "If we are ever able to learn what Wickham's debts have been," said Elizabeth...
Analysis: Few-Shot fails completely, generating text about Wickham's debts and Mr. Bennet (hallucinated from a later chapter). RepE summarizes the events but loses the specific descriptive flavor ("above his company"). Our method preserves the narrative detail and the specific criticisms of Darcy's character.		
Original Text 3: "Mrs. Collins, you must send a servant with them. You know I always speak my mind... **It is highly improper... Young women should always be properly guarded and attended**..."		
Lady Catherine expressed her view, "You should send a servant with them. I always speak directly, and I don't like the idea of two young women traveling alone. It is not appropriate..."	Lady Catherine expressed her view, saying, "You should send a servant with them..." (RepE mixes summary with dialogue and loses the imperious opening).	Here is the rewritten text in the same style: Lady Catherine expressed her view. "Mrs. Collins, you should send a servant with them... I always speak directly..."
Analysis: RepE often weakens the character's voice by introducing the text with reporting verbs ("expressed her view"). Our method maintains the authoritative and intrusive voice of Lady Catherine directly, preserving the stylistic structure of her commands.		

N Dimension Steering Generation

Table 14: Fine-grained Stylistic Progression Spectrum. We compare the generated outputs across five steering strengths. Dark Blue and Light Blue denote negative steering (dimension suppression), while Light Orange and Dark Orange denote positive steering (dimension intensification).

Dimension	Strength	Generated Text Snippet & Analysis
Classicism	-0.8	In the time of England's civil troubles, there were individuals like the Earl of Holland who brought an odium on the throne... These individuals later joined in the rebellions arising from their own discontents.
	-0.4	There were persons in England, in the time of civil troubles, who brought an odium on the throne... These men helped to subvert the throne to which they owed their existence.
	0	During England's civil troubles, some people, like the Earl of Holland, had caused resentment towards the throne by accepting its generous rewards. Later, they joined rebellions caused by their own actions.
	+0.4	There were persons... who had brought an odium on the throne... Confounded by the complication of distempered passions, their reason is disturbed; their views become vast and perplexed.
	+0.8	The author observes that if these men are not restrained, their avarice will know no bounds... The confusion and disorder are such that their own reason is disturbed... In the midst of this confusion, all appears limitless.
Trend: The writing shifts from simple, short sentences (-0.8) to long, complex structures (+0.8). At higher strength, the text uses more formal words and nested clauses to sound more old-fashioned and serious.		
Emotionality	-0.8	The man was frightened because a bowed, grey-coloured thing was coming towards him. The creature's face was distorted... The man's own face had changed... the emotions it registered were different.
	-0.4	The man was frightened because a bowed, grey-coloured, skeleton-like creature was approaching... Despite his initial confusion, the man realized that the creature was his own body.
	0	He stopped because he was scared. A bent, gray-colored, skeleton-like figure was coming towards him... It was definitely his own face, but it seemed to him that it had changed more than he had changed inside.
	+0.4	The man was frightened... Its eyes were watchful and fierce... He could not help but think that this was a sick man, sixty years old at the very least, suffering from some malignant disease.
	+0.8	The man was terrified... Its face was twisted and distorted, with a nobby forehead... He had gone partially bald, and his body was emaciated and covered in red scars... the spine was curved in a sickening way.
Trend: The text moves from a cold, objective description (-0.8: "The man") to an intense emotional experience (+0.8). High levels use strong words like "terrified" and "sickening" to emphasize the character's fear and disgust.		
Analyticity	-0.8	The moment any restraint is laid upon the full rights of men, the whole system of government becomes a matter of delicate skill. It requires a deep understanding of human nature.
	-0.4	The moment you diminish men's full rights to self-governance... the entire system necessitates a profound understanding of human nature and the requirements of civil institutions.
	0	When you reduce any of the full rights... government becomes a matter of convenience. This is what makes the structure of a state... a complex and delicate task.
	+0.4	This it is which makes the constitution of a state... a matter of the most delicate skill. It requires a deep knowledge of human nature and human necessities, and of the things which facilitate or obstruct the various ends.
	+0.8	What is the use of discussing a man's abstract right to food or to medicine? The question is upon the method of procuring and administering them. In that deliberation I shall always advise to call in the aid of the farmer...
Trend: Low levels simply state facts or requirements. High levels (+0.8) actively argue a point, using rhetorical questions and step-by-step logic to differentiate between abstract theory and practical method.		
Narrativity	-0.8	The text describes a scene from a movie theater where the audience is watching a war film. The scene shows a ship full of refugees... The text ends with a shot of a child's arm going up into the air.
	-0.4	The date is April 4th... It was a scene of a ship full of refugees being bombed... The last shot was of a child's arm... The audience applauded, but a woman in the proletariat section... started kicking up a fuss.
	0	April 4th, 1984. Went to the movies last night... One was about a ship full of refugees being bombed... The audience was amused by shots of a large man trying to swim away...
	+0.4	The audience was amused by a shot of a fat man... and they laughed when he sank... The helicopter then planted a bomb... which exploded and killed everyone on board.
	+0.8	...he is hit with many holes and sinks into the water. Next, a lifeboat... is shown... A middle-aged woman is seen comforting a young boy who is terrified... The helicopter then drops a bomb... causing it to disintegrate.
Trend: At -0.8, the text summarizes the plot from the outside ("The text describes..."). At +0.8, it tells the story directly, using action verbs like "sinks" and "drops" to show what is happening in the moment.		
Ornateness	-0.8	The hate reached its climax. The voice had become a bleat... Then the sheep-face melted into the figure of a Eurasian soldier... But in the same moment, the hostile figure melted into the face of Big Brother.
	-0.4	The Hate reached its climax. The voice turned into a bleat... and for an instant his face transformed into that of a sheep... Nobody could hear what Big Brother was saying.
	0	The Hate reached its peak. Goldstein's voice sounded like a sheep's bleat... Then the sheep's face changed into the figure of a Eurasian soldier... huge and terrible... full of power and mysterious calm.
	+0.4	...the sheep-face melted into the figure of a Eurasian soldier, advancing with his sub-machine gun roaring... the hostile figure melted into the face of Big Brother... so vast that it almost filled the screen.
	+0.8	The soldier's sub-machine gun roared, and it seemed to spring out of the screen... His words were encouraging and restored confidence by their mere utterance.
Trend: The description goes from plain and simple (-0.8) to highly detailed (+0.8). The high-style text adds dramatic adjectives and specific details to create a stronger visual image.		

O Prompt Templates

O.1 Prompt for Removing Authorial Traits

Prompt for Removing Authorial Traits

System

You are a rewriting assistant. Rewrite each passage into a neutral, plain English version.

Goal:
Remove stylistic signals so the text shows no clear sign of any of these styles:
- Classicism (archaic or period-specific flavor)
- Ornateness (decorative or complex phrasing)
- Narrativity (story-like sequencing or dramatization)
- Emotionality (affective or expressive tone)
- Analyticity (logical structuring or explicit reasoning)

Rules:
- Keep the same meaning, tense, and sentence order.
- Do not explain, interpret, or summarize.
- Do not add or remove information.
- Use plain, neutral, modern English.
- Avoid emotional, archaic, figurative, or decorative language.
- Keep syntax close to the original unless clearly stylistic.

Output format:
`{"id":"<id>", "neutral_text":"<rewritten>"}`

User

Rewrite the following paragraph into neutral and standard English according to the system rules.
ID: `{id}`
Passage: `{text}`

O.2 Passage Rewrite

Passage Rewrite Prompt

Instruction:
Please rewrite the following text without any explanation before or after the text:
`<Neutral Passage>`

Response:
`<Original Author-written Passage>`

O.3 Few-Shot Prompt

Few-Shot Prompt

System

Here are some examples of the author's original text:
<Sample Text 1>
<Sample Text 2>
...
<Sample Text k>

User

Instruction:
Please rewrite the following text in the same style without any explanation before or after the text:
<Query Text>
Response:

O.4 GPT Evaluation Prompt

GPT Evaluation Prompt

You are an expert literary critic. Rate the [Rewrite] based on the [Original] on a scale of 0–10.

1. Authorial Adherence (AA):
Assess how well does the rewrite capture the specific flavor of the original. Consider deep writing characteristics like distinctive voice, rhythm, and lexicon.

Instruction: penalize the score if the text sounds like generic, neutral English (e.g., standard AI assistant or Wikipedia), even if it is fluent. High scores require capturing the specific "flavor" of the author even if word choice or syntax differs slightly.

2. Semantic Fidelity (SF):
Assess how well the rewrite preserves the core meaning of the original.

Note: if the rewrite contains hallucinations (events/characters NOT in the original text) or changes the topic entirely, you should penalize the two aspects above.

Input Passages:
[Original]: {original_text}
[Rewrite]: {rewritten_text}

Output Format:
Output ONLY the scores in this exact format: AA:<0-10> SF:<0-10>

O.5 Prompt for LITERARYBIGFIVE Dimension Scoring

Prompt for LITERARYBIGFIVE Dimension Scoring

System

You are an expert literary critic and computational linguist. Your task is to analyze the stylistic attributes of a given book text.

Scoring Guidelines:
1. Evaluate the text on the 5 stylistic dimensions provided by the user.
2. Provide a score from 0 to 100 for each dimension.
3. Adopt a high-resolution scale, avoid saturation at extremes unless theoretical absolute. Focus on capturing fine-grained nuances.

User

Book Description: [Book Name] by [Author]
Dimensions to Evaluate:
- Analyticity: Measures reasoning orientation (abstract nouns, logical connectors, hierarchical structures).
- Ornateness: Measures lexical decoration and syntactic elaboration ("grand style", complex embedding), distinct from logic.
- Narrativity: Measures storytelling momentum (action verbs, temporal adverbs, rapid progression).
- Emotionality: Measures affective intensity and tension (warmth, surprise, expressive punctuation).
- Classicism: Measures resemblance to 18th-19th century traditions (archaic markers like *whilst*, old register), distinct from ornamentation.

Please output the scores in JSON format:

```
{
"Analyticity": <score>,
"Ornateness": <score>,
"Narrativity": <score>,
"Emotionality": <score>,
"Classicism": <score>
}
```