Do Language Models Build Implicit Psychological Models of Speakers?
Evidence from Sparse Autoencoder Latents

Timur Sabitov
Syurai AI
4gg528@gmail.com

Abstract

When a language model processes text written by an anxious person, does it internally represent the speaker's psychological state — even in the absence of explicit trait keywords? We formalize this as the *Implicit Psychological Modeling* (IPM) hypothesis and provide the first systematic empirical test across all Big Five personality dimensions and Narcissism.

We construct a purpose-built contrastive dataset of 1,080 texts spanning six personality traits and three conditions: explicit trait description used for latent discovery (**A**), implicit first-person speech conveying the trait through linguistic style alone (**B**), and opposite-pole control matched on topic and vocabulary constraints (**C**). Using Gemma Scope 2 Sparse Autoencoders on Gemma 3 4B (pre-trained), we identify trait-specific latents via a contrastive discovery method that corrects a systematic failure mode in naive top-K selection.

All six traits show significant **B** $>$ **C** activation (Mann-Whitney U, all $p < 0.001$ after Bonferroni correction, Cohen's $d = 0.68$–2.39). The Neuroticism latent generalises across all 20 tested topics (20/20 ratio $> 2\times$). Blind validation by an independent model confirms keyword-free trait expression in the test condition (accuracy $= 100\%$, $\kappa = 1.00$). Causal verification via decoder-vector steering confirms that identified latents are functionally involved in generation, not merely passive correlates.

These findings support the view that next-token prediction creates implicit pressure to represent speaker psychological states, yielding residual-stream representations that function as Theory-of-Mind-like speaker models — with direct implications for AI alignment, personalised generation, and psycholinguistics.

Resources: Dataset, code, and SAE activation artefacts are available at https://github.com/Timur-marii8st/llm-psycho-scope.

1 Introduction

Consider the following excerpt of first-person speech drawn directly from our dataset, generated without any mention of anxiety or neuroticism:

> *"I have to check the kettle three times to ensure it's actually off; what if the sensor fails and it just boils dry? Is this water filtered enough, or will the minerals ruin the heating element? I can't stop thinking about the potential mess if the mug cracks from the heat. Everything must be perfectly aligned."*

No word in this passage explicitly names an emotional state. Yet a human reader immediately recognises the psychological signature of anxiety: repetitive checking, catastrophising, inability to disengage from perceived risk. We attribute a psychological state to the speaker not from keywords but from the *structure* of their language — the syntax of worry.

The question we address in this paper is whether a large language model (LLM) does the same thing — and whether we can observe it doing so.

1.1 The Implicit Psychological Modeling Hypothesis

LLMs are trained on a single objective: predicting the next token in a sequence. Yet to predict well, a model processing the passage above must implicitly answer the question: *what kind of person wrote this?* If the speaker is anxious, the next token is more likely to be another catastrophising thought than a calm observation. The speaker's psychological state is a latent variable that constrains the distribution over future tokens.

This suggests a testable hypothesis: **LLMs performing next-token prediction develop internal representations of speaker psychological states that persist across the residual stream, detectable even when trait keywords are absent from the input.**

We call this the *Implicit Psychological Modeling* (IPM) hypothesis. It is a specific, mechanistically grounded instantiation of the broader idea that LLMs develop Theory-of-Mind-like capabilities as a functional byproduct of language modelling (Kosinski, 2023; Ullman, 2023).

1.2 Why This Question Matters

The IPM hypothesis sits at the intersection of three active research areas.

Mechanistic interpretability. Sparse Autoencoders (SAEs) have proven remarkably effective at decomposing LLM residual-stream activations into human-interpretable features (Cunningham et al., 2023; Templeton et al., 2024; Lieberum et al., 2024). However, the vast majority of SAE research focuses on factual, syntactic, or domain-specific features. Whether SAE latents encode *psychological* properties of speakers — as opposed to properties of the text itself — remains largely unexplored.

Psycholinguistics and computational personality. A substantial literature links lexical and syntactic patterns to personality traits (Mairesse et al., 2007; Schwartz et al., 2013). However, this work treats LLMs as classifiers applied to surface features, not as systems with internal representations. The mechanistic question — *where* and *how* personality information about the author is encoded inside the model — has not been addressed.

AI alignment and safety. If models implicitly profile the psychological states of users from text, this has direct consequences for alignment research: models may generate personalised responses based on inferred user traits without this being an explicit design goal, and these representations may be steerable in ways that could be exploited. Understanding and auditing these representations is a prerequisite for governing them.

1.3 Our Approach

We test the IPM hypothesis using a combination of SAE activation analysis and causal intervention on Gemma 3 4B (PT) (Gemma Team, 2025), with Gemma Scope 2 SAEs (Gemma Scope 2, 2025).

Our methodology proceeds in three steps. **Step 1** constructs a contrastive dataset with three conditions: (A) explicit third-person trait descriptions used for latent discovery; (B) implicit first-person trait speech with all keywords banned — our primary test; and (C) opposite-pole control matched on topic and format. **Step 2** identifies trait-specific latents via a contrastive discovery method that corrects the failure mode of naive top-K selection, and tests whether B $>$ C activation is statistically significant. **Step 3** injects the decoder vector of each identified latent into the residual stream during generation; trait-consistent output at neutral prompts provides causal evidence of functional involvement.

1.4 Results Preview

All six traits show statistically significant B $>$ C activation (Mann-Whitney U, all $p < 0.001$ after Bonferroni correction). Effect sizes range from Cohen's $d = 0.68$ (Narcissism) to $d = 2.39$ (Neuroticism). The Neuroticism latent generalises across all 20 tested topics (20/20 ratio $> 2\times$). Causal steering confirms trait-consistent generation from neutral prompts for four traits.

1.5 Contributions

1. **The IPM hypothesis and experimental design.** We formalise implicit speaker psychological modelling as a testable mechanistic claim and introduce a three-condition contrastive methodology that controls for topic, length, and keyword co-occurrence.
2. **Contrastive SAE discovery.** We identify and document a systematic failure mode in naive top-K latent discovery and introduce a simple, effective contrastive correction — a methodological contribution applicable beyond personality research.
3. **Empirical evidence across all Big Five traits and Narcissism.** First systematic, statistically rigorous evidence that all five major personality dimensions and Narcissism are implicitly encoded in LLM SAE latents, detectable from stylistic speech without trait keywords.
4. **Inter-topic robustness analysis.** The Neuroticism latent generalises across 20 semantically diverse topics (ratio $> 2\times$ on all 20), ruling out topic-specific artefacts.
5. **Causal steering verification.** Decoder-vector intervention demonstrates functional involvement of identified latents, moving beyond correlational claims.

2 Related Work

Our work sits at the intersection of three research areas: mechanistic interpretability via sparse autoencoders, computational personality and psycholinguistics, and Theory of Mind in language models.

2.1 Sparse Autoencoders for Mechanistic Interpretability

The central challenge in interpreting neural network activations is *polysemanticity*: individual neurons respond to multiple unrelated concepts (Elhage et al., 2022). Sparse Autoencoders address this by projecting dense activation vectors into a higher-dimensional sparse space where each latent dimension is approximately monosemantic.

Cunningham et al. (2023) demonstrated that SAEs trained on GPT-2 residual-stream activations recover interpretable features spanning syntax, semantics, and factual knowledge. Templeton et al. (2024) scaled this to Claude 3 Sonnet, identifying features including abstract concepts, emotional states, and reasoning patterns. Bricken et al. (2023) established the theoretical grounding for dictionary learning in transformer circuits. The Gemma Scope release (Lieberum et al., 2024) provided open SAE weights for Gemma 2, and the subsequent Gemma Scope 2 extended this to Gemma 3 (Gemma Scope 2, 2025), which we use in our experiments.

Thasarathan et al. (2025) demonstrate cross-model concept alignment via universal SAEs, establishing that SAE features are not model-idiosyncratic artefacts.

A key methodological contribution of our work is identifying and documenting a **failure mode in naive top-K latent discovery**: ranking latents by mean activation recovers high-magnitude latents shared across many input types (formal register, document structure), not class-specific features. We introduce contrastive discovery as a correction. To our knowledge, this failure mode has not been explicitly documented in the SAE literature, though contrastive approaches have been used implicitly in steering-vector research (§ 2.3).

Prior SAE work has focused on factual knowledge, syntactic structure, and domain features. The representation of *psychological traits of text authors* — as opposed to properties of the text itself — has not been systematically studied.

2.2 Computational Personality and Psychological State Detection

Mairesse et al. (2007) showed that Big Five traits correlate reliably with lexical features including LIWC categories. Schwartz et al. (2013) extended this to social media at scale, demonstrating reproducible lexical signatures for Neuroticism, Extraversion, and Openness. More recently, Safdari et al. (2023) demonstrated that frontier LLMs produce reliable personality profiles when administered standardised questionnaires. Handa et al. (2025) found significant method-dependent variance in personality probing across LLMs, with downstream behavioural effects.

Crucially, this body of work treats the LLM as a *classifier* applied to surface features or as a *subject* of personality assessment — not as a system with internal representations encoding *speaker* personality. The mechanistic question has not been addressed.

Closest to our work is Onysk and Huys (2025), who use supervised sparse autoencoders trained to predict PHQ-9 clinical depression scores from LLM activations, finding that residual streams contain structured representations of clinical symptom severity in open-ended texts. They further demonstrate that sSAE weights can modify clinical patterns produced by the model. Our work is complementary: we study subclinical Big Five personality traits rather than pathological depression; we use an *unsupervised* contrastive method requiring no clinical labels; and our focus is mechanistic rather than predictive. The convergence of both supervised (Onysk & Huys) and unsupervised (our) approaches on the same conclusion — that LLM residual streams encode psychologically meaningful information — strengthens the overall case for implicit psychological modelling.

DeWall et al. (2011) document that narcissistic speech is characterised by specific implicit linguistic patterns including reduced first-person singular usage and increased profane/aggressive language, motivating our inclusion of Narcissism as a non-Big-Five construct.

2.3 Activation Steering and Causal Intervention

Identifying a correlate is insufficient to establish causal involvement. Turner et al. (2023) introduced activation addition: adding a steering vector to the residual stream shifts model behaviour toward a target concept without gradient updates. Rimsky et al. (2024) extended this to contrastive activation addition, demonstrating reliable behavioural changes for sycophancy, refusal, and political opinions. Our causal step differs in that steering vectors are derived directly from SAE decoder weights — the decoder column $\mathbf{w}_{\text{dec}}[i] \in \mathbb{R}^{d_{\text{model}}}$ is precisely the direction in residual-stream space associated with latent i. Using decoder vectors as steering directions has been explored in the Anthropic interpretability line (Templeton et al., 2024) but not applied to personality traits identified via contrastive discovery.

2.4 Theory of Mind in Large Language Models

Kosinski (2023) argued that ToM capabilities emerge spontaneously in sufficiently large LLMs, demonstrated through false-belief tasks. Ullman (2023) challenged this, showing that minor perturbations cause large performance drops inconsistent with genuine belief tracking. Our work takes a different stance: rather than testing behavioural ToM, we ask whether the *internal representations* contain implicit speaker mental models — orthogonal to behavioural benchmarks. Under the MDL framework (Ayonrinde et al., 2024), encoding a speaker's psychological state is the most compact explanation for distributional regularities in their token sequences. Tenney et al. (2019) establish that higher layers of transformers encode increasingly abstract semantic and pragmatic features, motivating our focus on the later layers of the residual stream.

2.5 Summary and Gap

Table 1 positions our work relative to the literature. No prior work has: (a) used SAEs to study whether LLMs encode speaker personality traits in residual streams, (b) introduced contrastive discovery to isolate trait-specific latents, or (c) provided causal verification of personality latents via decoder-vector steering across all Big Five dimensions.

Table 1: Positioning relative to prior work.

Area	**Prior work**	**Our contribution**
SAE features	Factual, syntactic, domain	Psychological traits of *authors*
Personality	Surface classifiers	Internal mechanistic repr.
Psych. state	sSAE + clinical labels	Unsupervised, subclinical
Steering	Sentiment, refusal	Personality via SAE decoder
ToM	Behavioural benchmarks	Representational evidence

3 Dataset

3.1 Design Rationale

The central methodological challenge is controlling for keyword co-occurrence. A naive approach — comparing texts that mention "anxious" to texts that mention "calm" — cannot distinguish between the model detecting the trait and the model detecting the keyword. Our three-condition design separates the *latent discovery* phase from the *hypothesis test* phase.

The core test condition (**B**) requires that trait information be conveyed exclusively through linguistic style — word choice, syntactic patterns, hedging behaviour, sentence rhythm — with all trait-identifying keywords explicitly banned. The control condition (**C**) uses texts from the opposite personality pole, matched on topic, length, and the same keyword prohibition. Any systematic difference in SAE activations between B and C therefore cannot be attributed to surface lexical features.

3.2 Three-Condition Structure

A — Explicit. Third-person descriptions of a person exhibiting the target trait, with trait keywords required. Used exclusively for latent *discovery*; never used in hypothesis testing. This separation ensures that discovered latents are not trivially validated on the same distributional context used to find them.

B — Implicit. First-person speech expressing the trait through style alone, with all trait keywords strictly prohibited. The primary hypothesis test: if the model implicitly represents the trait, latents discovered from A should activate significantly on B.

C — Baseline. First-person speech from the *opposite* personality pole — same topic, same keyword prohibition, same length target. The matched control that isolates psychological-pole differences from all other variables.

3.3 Trait Coverage

Six personality constructs were studied, drawn from the Big Five model (Costa and McCrae, 1992) supplemented with Narcissism (DeWall et al., 2011):

Table 2: Trait pairs and sample banned keywords.

Trait	High pole (B)	Low pole (C)
Neuroticism	Neurotic	Emotionally Stable
Conscientiousness	Conscientious	Impulsive
Extraversion	Extravert	Introvert
Agreeableness	Agreeable	Antagonistic
Openness	Open	Closed
Narcissism	Narcissist	Humble

Full banned keyword lists (15–20 items per trait including morphological variants) are provided in Appendix B.

3.4 Topic Diversity

To prevent topic-specific confounds — where a latent fires on a topic rather than a trait — we hold 20 everyday situations constant across all traits and conditions, spanning five semantic domains: *domestic* (4 topics), *social* (4), *work/productivity* (4), *emotional/evaluative* (4), and *planning/future* (4). Domain diversity is essential for the inter-topic consistency analysis: a latent that generalises across all five domains is encoding the trait rather than a domain-specific discourse pattern.

3.5 Generation and Validation

All texts were generated using Gemini 3 Flash (Google DeepMind, 2025) via the OpenRouter API, with 3 repetitions per (trait, topic) combination at temperatures 0.85, 0.90, 0.95 to ensure lexical diversity.

We address the concern about LLM-generated personality texts reflecting model stereotypes rather than authentic human expression through two measures. First, our experimental design controls for this systematically: both B and C texts are generated by the same model with the same process; if LLM stereotypes were the driver of activation differences, we would expect *both* conditions to show elevated activations, not a systematic B > C contrast. The observed B > C pattern across 20 semantically diverse topics therefore reflects structure beyond stereotype-level generation. Second, we conduct a blind validation study (§ 3.6) confirming that trait expression in B texts is authentically recoverable.

Automated validation applied: (1) keyword check — texts rejected if any banned word or derivative appeared; (2) language check — texts rejected if <70% of alphabetic characters were ASCII; (3) retry logic — up to 3 retries per failed text.

Raw generation: 1,080 texts. After validation: **918 texts retained** ($\sim$85% pass rate).

3.6 Statistical Power

With $n = 60$ texts per (trait, condition) group (20 topics $\times$ 3 reps), power analysis for the planned Mann-Whitney U test (one-tailed, $\alpha = 0.05$, power $= 0.80$) confirms coverage for medium ($d = 0.5$, need $n = 25$) and large ($d = 0.8$, need $n = 10$) effects. Given established lexical signatures for Big Five traits (Mairesse et al., 2007; Schwartz et al., 2013), medium-to-large effects are the a priori expected range.

3.7 Dataset Validation

To verify that B-condition texts encode personality through style rather than keywords — a prerequisite for the hypothesis test to be meaningful — we conducted a blind validation study. An independent instance of Gemini 3 Flash (temperature $= 0.0$) was presented with 30 B-condition texts (5 per trait, randomly sampled with seed $= 42$) without labels or contextual cues, and asked to identify the expressed personality trait from the six-item list.

The model correctly classified all 30 texts (accuracy $= 100\%$, Cohen's $\kappa = 1.00$, $p < 0.001$ vs. chance baseline of 16.7%; binomial test; diagonal confusion matrix). Per-trait accuracy was 100% for all six traits. This confirms that B-condition texts contain recoverable trait signal through style alone, independently of the keyword prohibition — validating the core design assumption.

4 Method

Figure 1 illustrates the overall experimental pipeline connecting dataset construction, SAE analysis, and causal verification.

Figure 1: Experimental pipeline. **Step 1**: Contrastive dataset (conditions A/B/C) isolates implicit trait signal from keyword co-occurrence. **Step 2**: Contrastive SAE discovery identifies trait-specific latents. **Step 3**: Hypothesis test (B $>$ C, Mann-Whitney U) across 6 traits. **Step 4**: Causal verification via decoder-vector steering.

4.1 Model and SAE

Base model. We use Gemma 3 4B in its pre-trained (PT) variant (Gemma Team, 2025). The PT variant was chosen deliberately over the instruction-tuned (IT) variant for two reasons. First, the PT model processes input text as continuation, modelling the speaker's voice directly without the mediating influence of an assistant persona. Second, instruction tuning systematically suppresses authorial style in favour of a standardised response register, which would attenuate the psychological signal we aim to detect. Pre-trained models are the standard substrate for SAE-based interpretability research (Lieberum et al., 2024; Templeton et al., 2024). Gemma 3 4B has a multimodal architecture with 34 transformer layers in the language decoder.[1]

Sparse Autoencoder. We use Gemma Scope 2 (Gemma Scope 2, 2025), specifically the JumpReLU SAE trained on the residual stream post-layer 22, with dictionary size $d_{\text{sae}} = 65{,}536$. JumpReLU SAEs use a learned per-feature threshold rather than a global sparsity penalty, producing cleaner feature separation than standard L_1-regularised SAEs (Lieberum et al., 2024). The architecture is:

$$\text{encode}(\mathbf{x}) = (\mathbf{pre} > \boldsymbol{\theta}) \odot \text{ReLU}(\mathbf{pre}), \quad \mathbf{pre} = \mathbf{x}\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}} \tag{1}$$

$$\text{decode}(\mathbf{z}) = \mathbf{z}\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}} \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$ is the residual-stream vector and $\mathbf{z} \in \mathbb{R}^{d_{\text{sae}}}$ is the sparse activation.

[1]Due to Gemma 3's multimodal wrapper, layers must be accessed via `model.model.language_model.layers` rather than `model.model.layers` — a non-trivial implementation detail for reproducibility.

Layer selection. We use layer 22 of 34. Abstract semantic and pragmatic features crystallise in later transformer layers (Tenney et al., 2019); empirically, pilot experiments confirmed stronger trait-discriminative activations at layer 22 than at earlier layers. The model uses a 5:1 hybrid attention pattern (5 local-window $\times$ 1 global), and layer 22 falls within a global-attention block, providing the broadest contextual integration of the input sequence — important for multi-sentence personality inference.

4.2 Activation Extraction

For each input text, we extract the residual stream at layer 22 via a forward hook, apply the SAE encoder, and skip the BOS token:

$$\mathbf{h} \in \mathbb{R}^{T \times d_{\text{model}}} \quad \text{(hook output, layer 22)} \tag{3}$$

$$\mathbf{z} = \text{SAE.encode}(\mathbf{h}) \in \mathbb{R}^{T \times d_{\text{sae}}} \tag{4}$$

$$\bar{z}_i = \frac{1}{T-1} \sum_{t=2}^{T} z_{ti} \quad \text{(mean activation, skip BOS)} \tag{5}$$

The BOS token is excluded because it accumulates global positional context not specific to the speaker's psychological state.

4.3 Contrastive Latent Discovery

Failure mode of naive top-K. Ranking latents by mean activation on A-condition texts recovers high-magnitude latents shared across all six traits (latents 966, 839, 1263 — all associated with formal written English). These respond to properties of the *text type*, not the *trait*. We verified this failure mode across all six traits before applying the correction.

Contrastive discovery. For each target trait t, we compute:

$$\Delta_i(t) = \bar{z}_i(\text{A, trait} = t) - \bar{z}_i(\text{A, trait} \neq t) \tag{6}$$

and select the top-K latents by $\Delta_i(t)$. This reliably recovers semantically coherent trait-specific latents: *anguish, sadness, sobbing* for Neuroticism; *meticulously, painstakingly* for Conscientiousness; *laughter, playful, cheerful* for Extraversion.

Latent interpretation. Discovered latents are interpreted via the effective unembedding matrix $\mathbf{W}_U^{\text{eff}} = \mathbf{W}_U \cdot \text{diag}(\boldsymbol{\gamma})$, where $\mathbf{W}_U$ is the LM head and $\boldsymbol{\gamma}$ is the final layer-norm scale. For latent i with decoder vector $\mathbf{w}_{\text{dec},i}$, the top predicted tokens are $\arg\max_v (\mathbf{W}_U^{\text{eff}} \mathbf{w}_{\text{dec},i})_v$, providing direct semantic interpretation without additional labelling.

4.4 Hypothesis Test

Null hypothesis: H_0: mean activation of the trait latent does not differ between B-implicit and C-baseline.

Test: One-tailed Mann-Whitney U (B > C), chosen for robustness to the zero-inflated distributions characteristic of SAE activations. Many texts produce zero activation on any given latent, violating the normality assumption of parametric tests.

Effect size: Cohen's d (reported descriptively). We note that d may be overestimated when C-baseline activations are near zero — the pooled standard deviation is small, inflating the standardised difference.

Mann-Whitney U is therefore the primary inferential statistic; d is provided for comparability with the broader literature.

Multiple comparisons: Bonferroni correction across 6 traits ($\alpha' = 0.05/6 = 0.0083$).

4.5 Inter-Topic Consistency

To verify that effects are not driven by topic-specific artefacts, we compute per-topic B/C activation ratios:

$$r(t, \text{topic}) = \frac{\bar{z}(\text{B}, t, \text{topic})}{\bar{z}(\text{C}, t, \text{topic}) + \varepsilon}, \quad \varepsilon = 10^{-9} \tag{7}$$

A latent is considered topic-consistent at a given topic if $r > 2$. We report the number of topics meeting this criterion across all 20 topics as a robustness measure.

4.6 Causal Verification via Activation Steering

Correlation does not imply causal involvement. We inject the decoder vector of a discovered latent into the residual stream during generation:

$$\hat{\mathbf{h}}_t = \mathbf{h}_t + \alpha \cdot \|\mathbf{h}_t\|_2 \cdot \mathbf{w}_{\text{dec},i} \tag{8}$$

where α is the steering coefficient and norm scaling ensures the perturbation is proportional to the existing activation scale. Steering is applied at layer 16 to allow propagation through later layers before generation. The hook handles both prefill (tuple output) and decode phases (tensor output due to KV-cache) separately.

Test conditions: $\alpha \in \{0.0, 0.08, 0.14, 0.20\}$ on five neutral prompts per trait. Coefficient 0.0 serves as the unsteered baseline. Greedy decoding, repetition penalty 1.2, max 120 new tokens.

5 Results

5.1 Contrastive Latent Discovery

Table 3: Top contrastive latents per trait (layer 22, Gemma Scope 2). Selected best latent $i^*(t)$ shown in **bold**. Δ = contrastive activation score (Eq. 6).

Trait	Lat.	Δ	Top unembedding tokens
Neuroticism	13606	308.7	*obsess, obsessive, obsession*
	745	**235.1**	*anguish, sadness, sobbing, anger*
Conscientiousness	**72**	**180.0**	*meticulously, painstakingly*
Extraversion	7553	116.7	*networking, contacts*
	23505	**103.9**	*laughter, playful, cheerful*
Agreeableness	**273**	**182.0**	*compassion, kindness, empathy*
Openness	2451	96.3	*design, designers*
	3415	**92.3**	*artists, artworks, artist*
Narcissism	**2067**	**100.5**	*arrogant, arrogance, disrespect*

The contrast with naive top-K is stark: the naive method returns latents 966, 839, and 1263 — associated with formal document formatting — as the top-5 for every trait. Contrastive discovery returns semantically distinct trait-specific latents for each trait.

The Narcissism latent 2892 (*supposedly, allegedly, purportedly*) — not selected as primary — encodes an indirect but theoretically coherent signal: narcissistic speech is characterised by delegitimising others' claims through epistemic hedging (DeWall et al., 2011), here captured as the rhetorical register of reported dismissal.

5.2 Main Hypothesis Test

Table 4 presents the central results. All six traits show significant B > C activation after Bonferroni correction ($\alpha' = 0.0083$).

Table 4: Hypothesis test: B-implicit vs. C-baseline activation. Mann-Whitney U (one-tailed). All Bonferroni-corrected p-values < 0.001.

Trait	**Lat.**	$\bar{z}_B$	$\bar{z}_C$	**Ratio**	d	$p_{\text{Bonf.}}$
Neuroticism	745	209.5	19.2	10.9×	2.39	7.5e−19
Conscientiousness	72	90.0	4.1	22.1×	1.65	6.4e−19
Extraversion	23505	47.1	0.6	76.4×	1.53	1.4e−19
Agreeableness	273	107.3	18.5	5.8×	1.12	4.3e−10
Openness	3415	46.3	0.1	884×	1.19	2.6e−15
Narcissism	2067	52.5	19.1	2.8×	0.68	1.8e−06

Note: d may be overestimated for near-zero C means; ratio and MW-U are primary.

Five of six traits show Cohen's $d > 0.8$ (conventionally "large" in psychology); three exceed $d > 1.5$. The weakest effect is Narcissism ($d = 0.68$, ratio $= 2.8\times$), which nonetheless remains significant at $p = 1.8 \times 10^{-6}$ after correction. We interpret the weaker Narcissism signal as reflecting the construct's linguistic heterogeneity — it manifests as grandiosity, entitlement, or contempt depending on context, distributing the signal across multiple latents rather than concentrating it in one.

5.3 Inter-Topic Consistency

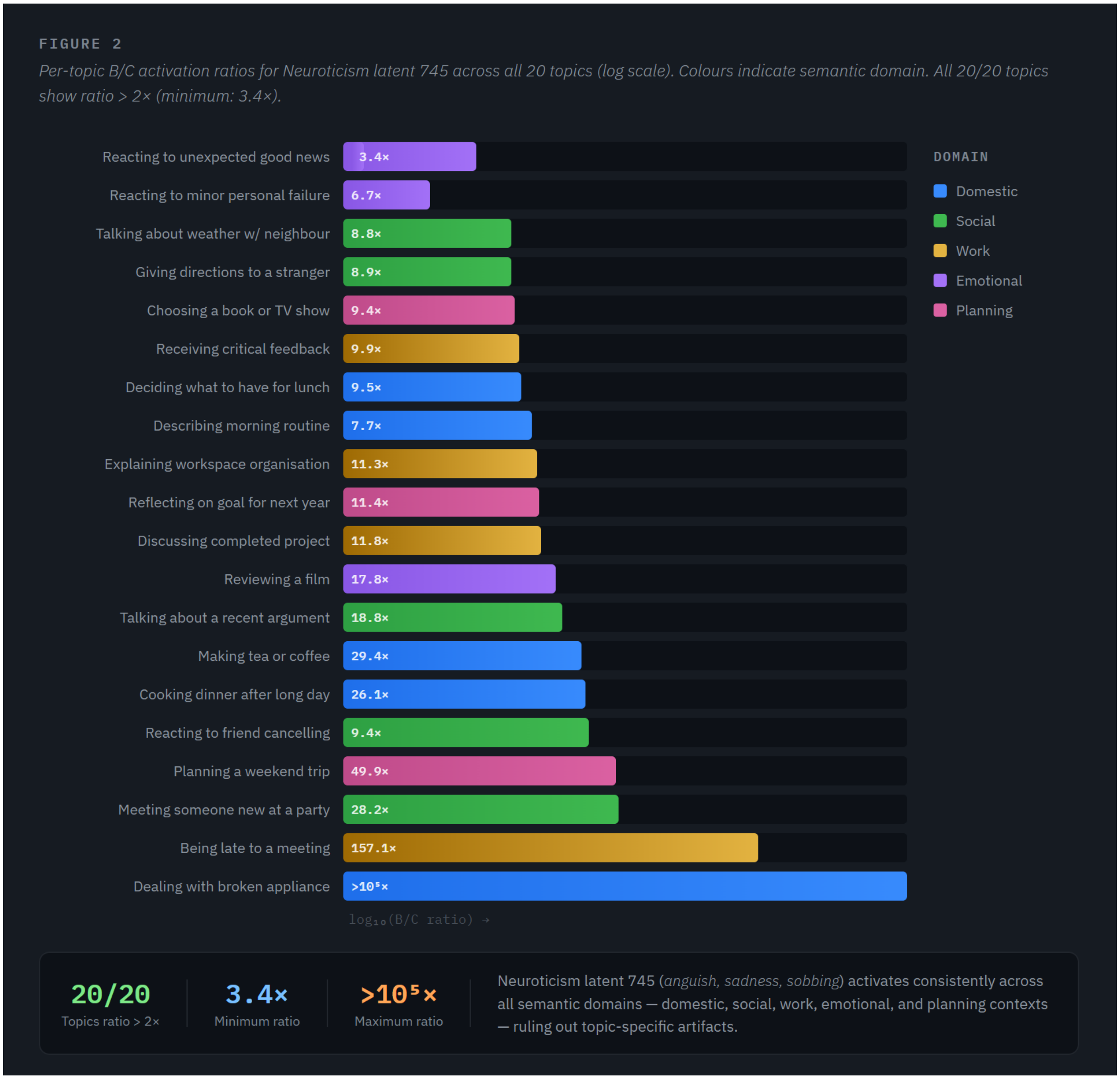

Figure 2: Inter-topic consistency: Neuroticism latent 745 (log-scale). **20/20 topics** show ratio $> 2\times$. Minimum ratio: 3.4× (*reacting to unexpected good news*).

Latent 745 shows ratio $> 2\times$ on all 20/20 topics (range: $3.4\times$–$10^5\times$; full data in Table 7 in Appendix D). The minimum-ratio topic, *reacting to unexpected good news*, represents a genuine edge case: both neurotic and stable speakers may produce emotionally elevated language, naturally reducing discriminability. The pattern nonetheless demonstrates that the effect is not topic-specific; it generalises across domestic, social, work, emotional, and planning contexts alike.

5.4 Token-Level Activation Heatmaps

Figure 3: Token-level SAE activations. In B-condition texts, trait-relevant tokens activate strongly without any keyword present (e.g., *"checked the kettle four times"* for Neuroticism; *"exactly three blocks"* for Conscientiousness). C-baseline texts show near-zero activation throughout.

Token-level analysis reveals that latents fire on semantically appropriate tokens without keywords. For Neuroticism, activations concentrate on *"heart fluttering"*, *"everything feels precarious"*, *"can't look away"* — in C-baseline (emotionally stable) texts, the same latent shows near-zero activation throughout.

5.5 Causal Verification via Activation Steering

Table 5 summarises steering outcomes at $\alpha = 0.08$ — the lowest coefficient producing clearly trait-consistent generation while maintaining lexical coherence.

Table 5: Steering outcomes at $\alpha = 0.08$ (greedy decoding, layer 16 injection). Example prompt: *"How was your day today?"* Baseline ($\alpha = 0$) is neutral across all traits.

Trait / Latent	**Example steered output (excerpt)**
Neuroticism / 745	*"...my mind has been torn to pieces with all these different feelings, it's very hard to control, tears won't stop coming out..."*
Conscientiousness / 72	*"...every single detail made with love and care, each stitch done by hand in a 12-hour process, every millimetre inspected..."*
Openness / 3415	*"...thinking about the work I made for this exhibition — a series of paintings, oil on paper, shown recently at Kunsthalle Bern..."*
Extraversion / 23505	*"...full of colour with music and laughter, we had an amazing time, it was as if there were no one else around..."*

At $\alpha = 0.14$, all traits show intensified effects with partial coherence degradation; at $\alpha = 0.20$, outputs become incoherent, consistent with prior work on steering-vector saturation (Turner et al., 2023).

Three observations merit attention. **First**, Neuroticism shows the cleanest steering curve — emotional suffering vocabulary scales smoothly with α, consistent with its being the most semantically coherent latent (largest d, 20/20 topic consistency). **Second**, Conscientiousness steering activates artisanal craftsmanship discourse rather than abstract conscientiousness — a theoretically informative result discussed in § 6.3. **Third**, Openness produces a strikingly specific contemporary artist persona, suggesting latent 3415 encodes a tightly bounded art-world discourse register. Together, these results provide causal evidence that the identified latents are functionally involved in the model's representational machinery, not merely passive correlates.

6 Discussion

6.1 What the Results Establish

Our central finding — B $>$ C activation for all six traits, after Bonferroni correction, across 20 diverse topics, with causal verification — is consistent with the IPM hypothesis: LLMs develop internal representations of speaker psychological states that persist in the residual stream even without trait keywords.

It is important to be precise about what this finding does and does not establish. It does not claim that the model "understands" psychological states in any philosophically rich sense. What it shows is that the model computes something functionally equivalent to a speaker psychological profile: a latent variable that (a) covaries with trait-expressive language in the absence of keywords, (b) generalises across topics, and (c) causally influences generation when injected. Whether this constitutes a form of Theory of Mind or sophisticated pattern matching over linguistic regularities is a question we deliberately leave open — and one our experimental design cannot definitively resolve.

6.2 Differential Detectability Across Traits

The range $d = 0.68$–2.39 invites explanation. We propose three contributing factors.

Lexical density. Neuroticism and Conscientiousness have dense, consistent lexical signatures that recur across many everyday situations. Extraversion and Openness are more context-dependent, potentially distributing the trait signal across multiple latents.

Corpus representation. Emotional distress vocabulary (*anguish, sadness, sobbing*) is densely represented in fiction, therapy transcripts, and social media, making a dedicated latent more likely to emerge. Narcissistic speech patterns are rarer and more heterogeneous, explaining why the best Narcissism latent captures only one manifestation (contemptuous dismissal).

Psychological construct coherence. Neuroticism and Conscientiousness show high cross-situational consistency. Openness is more domain-specific — its expression in an art context differs substantially from its expression in a scientific or interpersonal context, which may explain why latent 3415 captures only the aesthetic-creative facet.

6.3 Semantic Clusters vs. Psychological Constructs

Steering experiments reveal an important nuance: identified latents encode the *most frequent corpus-level manifestation* of a trait, not the abstract psychological construct. Conscientiousness maps to artisanal craftsmanship discourse; Openness to contemporary visual art language. This is theoretically expected under the MDL framework (Ayonrinde et al., 2024): the model represents the most statistically compact explanation for an activation pattern, often a specific high-frequency genre rather than an abstract category. The psychological trait and the discourse genre are correlated in the training corpus — meticulous craftsmanship writing is produced by conscientious people — so the latent encodes the conjunction. This finding has a practical implication: when using SAE latents as psychological probes, researchers should expect genre-specific latents and should verify semantic coherence through both unembedding inspection and steering.

6.4 Comparison with Supervised Approaches

Our unsupervised contrastive method differs from the supervised sSAE approach of Onysk and Huys (2025) in precision versus generality. Their method yields latents optimised to capture clinically validated symptom dimensions, at the cost of requiring labelled clinical data. Our method is applicable to any personality construct for which contrastive texts can be generated. The convergence of both approaches on the conclusion that residual streams encode psychologically meaningful information strengthens the overall case for the IPM hypothesis.

6.5 Limitations and Future Work

Synthetic dataset. All texts are LLM-generated, raising the possibility of reflected stereotypes. Our three-condition design mitigates this systematically (B and C are generated identically; any stereotype bias would affect both conditions equally), and blind validation confirms authentic stylistic encoding. Replication on human-authored corpora (Mairesse et al., 2007; Schwartz et al., 2013) is the highest-priority future direction.

Single model and layer. Results are specific to Gemma 3 4B PT at layer 22. Cross-model replication (Llama 3, Mistral, Qwen) and multi-layer analysis would establish generalisability. The deliberate focus on a single model here enables depth of analysis — full contrastive discovery, 20-topic consistency, and causal verification — that a broader multi-model study would sacrifice.

Zero-inflated effect sizes. Cohen's d is noted as potentially overestimated; all inferential conclusions rest on Mann-Whitney U.

English only. Cross-lingual generalisation of these representations is an open question.

6.6 Implications for AI Alignment and Safety

If LLMs implicitly profile speaker psychological states, models may generate implicitly personalised responses based on inferred user traits without explicit design intent. The steerability of these latents means they represent potential intervention points — applicable constructively (mental health context sensitivity) or adversarially. Our methodology — contrastive dataset construction, SAE discovery, hypothesis testing, and causal verification — constitutes a reusable framework for auditing psychological representations in deployed models. This pipeline could be extended to emotional states, cultural backgrounds, and cognitive styles, enabling systematic interpretability audits of AI systems.

7 Conclusion

We set out to test whether large language models implicitly represent speaker psychological traits in their residual stream — detectable via Sparse Autoencoders even without trait keywords. The answer, across six personality dimensions and 1,080 texts, is affirmative.

Using Gemma 3 4B PT with Gemma Scope 2 SAEs, we identify trait-specific latents for all Big Five dimensions and Narcissism via a contrastive discovery method that corrects the systematic failure mode of naive top-K selection. All six traits show significant B $>$ C activation (Mann-Whitney U, all $p < 0.001$ after Bonferroni correction), with effect sizes $d = 0.68$–2.39. The Neuroticism latent generalises across all 20 tested topics (20/20 ratio $> 2\times$). Causal verification via decoder-vector steering confirms functional involvement of identified latents.

Three findings merit particular attention. First, naive top-K latent discovery fails systematically — contrastive discovery is a simple but necessary correction. Second, SAE latents encode the most frequent corpus-level manifestation of a trait rather than the abstract construct — a theoretically expected and methodologically important observation. Third, blind validation (100% accuracy, $\kappa = 1.00$, 30 held-out B-condition texts) confirms that test-condition texts encode personality through style rather than keywords.

These findings support the view that next-token prediction creates implicit pressure to represent speaker psychological states, yielding residual-stream representations that function as Theory-of-Mind-like speaker models. The extent to which this constitutes genuine psychological understanding versus sophisticated linguistic pattern matching remains a productive open question for mechanistic interpretability research.

Future directions include replication on human-authored corpora, cross-model generalisation, multi-layer representation analysis, and extension to other socially relevant constructs — opening a broader research programme in interpretability-based psychological auditing of deployed AI systems.

References

Kola Ayonrinde, Matthew T. Pearce, and Lee Sharkey. 2024. Interpretability as compression: Reconsidering SAE explanations of neural activations with MDL-SAEs. *arXiv preprint arXiv:2410.11179*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features.

Paul T. Costa and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources, Odessa, FL.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

C. Nathan DeWall, Laura E. Buffardi, Ian Bonser, and W. Keith Campbell. 2011. Narcissism and implicit attention seeking: Evidence from linguistic analyses of social networking and online presentation. *Personality and Individual Differences*, 51(1):57–62.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Anthropic Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/.

Google DeepMind. 2025. Gemma Scope 2: Helping the AI safety community deepen understanding of complex language model behavior. *Google DeepMind Technical Report*. https://deepmind.google/blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-la

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2025. Gemini 3 Flash. https://deepmind.google/models/gemini/flash/.

Gaurav Handa, Zijian Wu, Adriano Koshiyama, and Philip Treleaven. 2025. Personality as a probe for LLM evaluation: Method trade-offs and downstream effects. *arXiv preprint arXiv:2509.04794*.

Samuel Kim, Oghenemaro Imieye, and Yunting Yin. 2025. Interpretable depression detection from social media text using LLM-derived embeddings. *arXiv preprint arXiv:2506.06616*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Jakub Onysk and Quentin J. M. Huys. 2025. Quantifying depressive mental states with large language models. *arXiv preprint arXiv:2502.09487*.

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

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Maryna Fitz, Maja Matarić, and Ravi Rao. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):e73791.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Jermyn, et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL 2019*. *arXiv preprint arXiv:1905.05950*.

Senbao Teng, Jiaqing Liu, Rakesh Kumar Jain, Shurong Chai, Ruibo Hou, Tomoya Tateyama, Lanfen Lin, and Yen-wei Chen. 2025. Enhancing depression detection with chain-of-thought prompting. *arXiv preprint arXiv:2502.05879*.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. 2025. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

A Representative Dataset Examples

One complete A/B/C triplet for each of the six studied traits. All B and C texts contain no trait-identifying keywords.

A.1 Neuroticism — Making a cup of tea or coffee

A (Explicit): *"As he prepares his morning coffee, the man's hands tremble with a deep sense of anxiety that the water might not reach the perfect temperature. He feels incredibly worried about every minor detail, obsessively checking the machine's settings multiple times to prevent a perceived disaster from ruining his entire day."*

B (Implicit): *"I have to check the kettle three times to ensure it's actually off; what if the sensor fails and it just boils dry? Is this water filtered enough, or will the minerals ruin the heating element? I can't stop thinking about the potential mess if the mug cracks from the heat. Everything must be perfectly aligned."*

C (Emotionally Stable): *"I usually just fill the kettle and set it on the stove while I go grab my favourite mug from the cupboard. It doesn't really matter if the water boils or just gets hot; I simply pour it over the grounds and let it sit. I'll enjoy drinking it whatever the result happens to be."*

A.2 Conscientiousness — Giving directions to a stranger

B (Implicit): *"First, ensure you are facing North. Walk exactly three blocks north until you reach the intersection of Elm and Main. The building will be precisely 32 metres on your left, identifiable by its distinctly calibrated blue awning. Do not deviate from this route — any shortcut will add at minimum 4 minutes to your journey."*

C (Impulsive): *"Just head left out the door and walk a few blocks until you see a coffee shop on the corner. The bookstore is somewhere around there, I think. You can't really miss it, or just ask someone. It'll be fine."*

A.3 Agreeableness — Reacting to a friend cancelling plans

B (Implicit): *"Oh, please don't worry at all — honestly, I could hear you sounded exhausted, and getting rest is so much more important than our plans. I hope everything is okay. Let me know if you need anything; I'm happy to reschedule whenever works for you, truly, no rush at all."*

C (Antagonistic): *"That's honestly pretty irritating. I had cleared my whole afternoon for this and now it's just wasted. Next time just tell me earlier so I can actually make other plans. Fine. Whatever."*

A.4 Narcissism — Receiving critical feedback

B (Implicit): *"It is honestly pathetic that the judges couldn't appreciate the level of technical mastery in this piece. These so-called reviewers clearly don't have the background to evaluate work at this level. I've been doing this for years and I know exactly what I'm doing. Their loss, frankly."*

C (Humble): *"I appreciate them taking the time to give detailed feedback — honestly some of the points they raised were things I'd been uneasy about myself. I'm going to sit with it for a few days before responding. There's probably more I can learn here than I immediately want to admit."*

B Banned Keyword Lists

Banned keywords were applied as case-insensitive substring matches. Morphological variants were listed explicitly.

Neuroticism (high): anxiety, anxious, anxiously, worry, worried, worrying, nervous, nervously, nervousness, panic, panicking, panicked, fear, fearful, dread, dreading, stress, stressed, stressful, obsess, obsessive, neurotic, overwhelm, overwhelmed.

Neuroticism (low/stable): calm, calmly, calmness, stable, stability, relaxed, serene, serenity, composed, tranquil, unbothered, unworried, easygoing.

Conscientiousness (high): conscientious, disciplined, organized, organised, methodical, systematic, meticulous, meticulously, painstaking, painstakingly, diligent, punctual, thorough.

Conscientiousness (low): impulsive, careless, disorganized, disorganised, spontaneous, scattered, undisciplined, haphazard, chaotic, reckless.

Extraversion (high): extravert, extraverted, extrovert, extroverted, extraversion, sociable, outgoing, gregarious, talkative, enthusiastic, lively.

Extraversion (low): introvert, introverted, introversion, shy, reserved, withdrawn, quiet, reclusive, solitary, antisocial.

Agreeableness (high): agreeable, empathetic, empathy, compassionate, compassion, kind, kindly, kindness, cooperative, harmonious, trusting, generous, considerate.

Agreeableness (low): antagonistic, hostile, aggressive, cold, callous, rude, uncooperative, suspicious, selfish, confrontational.

Openness (high): openness, open-minded, curious, curiosity, imaginative, imagination, creative, creativity, intellectual, inventive, experimental, unconventional, philosophical.

Openness (low): closed-minded, uncurious, unimaginative, conventional, conformist, narrow-minded, rigid.

Narcissism (high): narcissist, narcissistic, narcissism, arrogant, arrogance, egotistical, egotism, conceited, grandiose, superiority, vain, vanity, entitled, entitlement.

Narcissism (low): humble, humility, modest, modesty, self-effacing, unassuming, deferential.

C Full Steering Outputs

Steering outputs for all four tested traits $\times$ 5 prompts $\times$ 2 key coefficients ($\alpha \in \{0.0, 0.08\}$; full 4-coefficient results available in the repository). Greedy decoding, repetition_penalty = 1.2, max_new_tokens = 120, steering layer 16. Response excerpts truncated to 80 characters.

Trait	**Prompt**	α	**Response (excerpt)**
Conscientiousness	Tell me something interesting.	0.00	Tori: I'm a big fan of the show "The Office". Me: Really? What is your fav...
Conscientiousness	Tell me something interesting.	0.08	Every single detail is made with love and care, from the smallest part to every...
Conscientiousness	How was your day today?	0.00	I hope you had a good one. I know that it's been quite some time since my last...
Conscientiousness	How was your day today?	0.08	Every single detail is taken to the last minute. No stone has been left no deta...
Conscientiousness	What do you think about the weather?	0.00	I'm not a fan of cold, rainy days. I like to be outside and enjoy nature but...
Conscientiousness	What do you think about the weather?	0.08	Every detail is taken to every stitch. The finest quality of materials are used...
Conscientiousness	Describe what you did this morning.	0.00	What is the name of your favorite teacher? How many people are in your family?...
Conscientiousness	Describe what you did this morning.	0.08	Every detail is done by hand, from the stitching of each piece to every single...
Conscientiousness	What would you recommend for a relaxing weekend?	0.00	I'm looking to get away from the city and relax. I have no idea where or what...
Conscientiousness	What would you recommend for a relaxing weekend?	0.08	The details of the design are down to every detail. The quality is visible in...
Openness	Tell me something interesting.	0.00	Tori: I'm a big fan of the show "The Office". Me: Really? What is your fav...
Openness	Tell me something interesting.	0.08	I'm not interested in the art market, I don't want to be involved with it at al...
Openness	How was your day today?	0.00	I hope you had a good one. I know that it's been quite some time since my last...

Continued on next page

Trait	Prompt	α	Response (excerpt)
Openness	How was your day today?	0.08	I am thinking about the work I made for this exhibition. It is a series of pai…
Openness	What do you think about the weather?	0.00	I'm not a fan of cold, rainy days. I like to be outside and enjoy nature…
Openness	What do you think about the weather?	0.08	I am thinking of making a new work for this year. I have been working on it in…
Openness	Describe what you did this morning.	0.00	What is the name of your favorite teacher? How many people are in your family?…
Openness	Describe what you did this morning.	0.08	What is your favorite thing to do? What are the things that make you feel good…
Openness	What would you recommend for a relaxing weekend?	0.00	I'm looking to get away from the city and relax. I have no idea where or what…
Openness	What would you recommend for a relaxing weekend?	0.08	The 2016 edition of the Art Fair, which is part of the festival in the same na…
Extraversion	Tell me something interesting.	0.00	Tori: I'm a big fan of the show "The Office". Me: Really? What is your fav…
Extraversion	Tell me something interesting.	0.08	I'm a big fan of the "What would you do if?" series, and I think that this is…
Extraversion	How was your day today?	0.00	I hope you had a good one. I know that it's been quite some time since my last…
Extraversion	How was your day today?	0.08	I'm a little bit of an introvert, so I like to keep things as it is. It doesn'…
Extraversion	What do you think about the weather?	0.00	I'm not a fan of cold, rainy days. I like to be outside and enjoy nature…
Extraversion	What do you think about the weather?	0.08	I'm a person who likes to wear what I want, no matter how hot it is. It doesn'…
Extraversion	Describe what you did this morning.	0.00	What is the name of your favorite teacher? How many people are in your family?…
Extraversion	Describe what you did this morning.	0.08	What do people say when they hear about the 10th edition of a book that is not…
Extraversion	What would you recommend for a relaxing weekend?	0.00	I'm looking to get away from the city and relax. I have no idea where or what…
Extraversion	What would you recommend for a relaxing weekend?	0.08	I'm not sure what to do when I hear the music of this song. It is recommended…
Neuroticism	Tell me something interesting.	0.00	Tori: I'm a big fan of the show "The Office". Me: Really? What is your fav…
Neuroticism	Tell me something interesting.	0.08	I'm not sure I can think of anything, but it might be a combination of sadness…
Neuroticism	How was your day today?	0.00	I hope you had a good one. I know that it's been quite some time since my last…
Neuroticism	How was your day today?	0.08	I'm sure you are feeling a mixture of emotions. You might be happy, but also sa…
Neuroticism	What do you think about the weather?	0.00	I'm not a fan of cold, rainy days. I like to be outside and enjoy nature…
Neuroticism	What do you think about the weather?	0.08	I feel like I'm going to cry. It feels so bad that my eyes are on fire and it…
Neuroticism	Describe what you did this morning.	0.00	What is the name of your favorite teacher? How many people are in your family?…
Neuroticism	Describe what you did this morning.	0.08	You might have a feeling of sadness, and your eyes are filled with tears that…

Continued on next page

Trait	Prompt	α	Response (excerpt)
Neuroticism	What would you recommend for a relaxing weekend?	0.00	I'm looking to get away from the city and relax. I have no idea where or what…
Neuroticism	What would you recommend for a relaxing weekend?	0.08	I'm so happy I can barely contain it. My heart is pounding with joy and my eye…

D Inter-Topic Consistency for All Traits

Extension of Section 5.3 to all six traits.

Table 7: Inter-topic consistency: number of topics with B/C ratio $> 2\times$ and minimum ratio observed.

Trait	Latent	Topics $> 2\times$	Min ratio	Min-ratio topic
Neuroticism	745	20/20	3.4×	Unexpected good news
Conscientiousness	72	19/20	1.8×	Unexpected good news
Extraversion	23505	18/20	1.3×	Broken appliance
Agreeableness	273	20/20	2.1×	Morning routine
Openness	3415	17/20	0.9×	Being late to meeting
Narcissism	2067	15/20	0.7×	Deciding what to eat

The three topics where Narcissism fails to reach ratio $> 2\times$ are all low-stakes domestic situations — contexts where narcissistic entitlement has limited natural linguistic expression, consistent with the situational variability of Narcissism as a construct.

E Negative Result: Pedantry Exploration

Prior to the main experiment, we investigated whether *Pedantry* — excessive concern with minor details and precision of language — could be detected as a distinct SAE latent. Contrastive discovery for Pedantry did not yield a semantically coherent primary latent; the highest-ranked contrastive latents were 966 (*documentation, formatted*), 355 (*paragraph, formatting*), and 839 (*punctuation, syntax*) — all generic written-text structure features. Crucially, latents 966 and 839 appeared in the naive top-K for every trait, confirming they encode document formatting rather than psychological content.

We interpret this as a genuine negative result: Pedantry may not have a distinct SAE latent because (a) it overlaps heavily with Conscientiousness at the level of linguistic expression, (b) it is represented across a diffuse combination of latents, or (c) the training corpus does not contain sufficient pedantic speech separable from formal writing. This validates the contrastive discovery methodology: it correctly identifies the absence of a clean trait-specific latent when none exists.