

LEX-EC: A Lexical Evidence-Channel Audit Framework for Zero-Shot LLM Personality Classification in Black-Box Settings

Brittany Harbison^{1*}, Ashok K. Goel¹

¹Georgia Institute of Technology
Atlanta, Georgia, USA
bharbison3@gatech.edu, ashok.goel@cc.gatech.edu

Abstract

Large language models may easily assign personality labels from text, but model interpretability remains an open problem. To address this gap, we introduce LEX-EC, a reusable black-box audit framework combining agreement diagnostics with controlled lexical ablation to distinguish marginal-distribution effects from trait-associated signal recoverable under restricted evidence. Using this framework, we illustrate how various text genres may exhibit sharply different profiles: free-form essay text contains the broadest, but still weak, signal; in graduate student introductions, an observable Extraversion association weakened after masking; and single Facebook statuses yield little stable evidence even in a trait-balanced sample, indicating a possible lower bound of content or length. Masking topical and demographic content weakened some associations while leaving others detectable from function words, affective terms, and cognitive-style vocabulary. Linguistic prompting shifted model self-explanations but did not eliminate topical content. LEX-EC jointly evaluates item-level association and chance-corrected agreement, examines how both change under lexical restriction, and conducts a targeted audit of prompt sensitivity in model-generated explanations. Across datasets, models, and prompts, LEX-EC characterizes how trait associations may vary with available lexical evidence, introducing a novel application of lexical methods to black-box interpretability in personality labeling.

Code — <https://github.com/Noeru-H/LEX-EC>

Introduction

Large language models (LLMs) may assign personality traits (Goldberg 1990; McCrae and John 1992) to an author from ordinary text, and such inferences are increasingly used in education, hiring, and social computing (Vinciarelli and Mohammadi 2014; Matz et al. 2017). But a plausible label is not necessarily a valid one: a model may return the same confident judgment regardless of whether the text is rich or a single sentence, whether it reflects trait-relevant language or mere topical content, whether the inference tracks the author or demographic priors the model brings to it (Rao, Leung, and Miao 2023; Peters and Matz 2024), and whether aggregate scores mask near-chance item-level agreement. Prior work reports apparent success, as correlations with human raters

or self-reports across traits and genres (Derner et al. 2024; Peters and Matz 2024; Schoenegger et al. 2025; Piastra and Catellani 2025; Wright et al. 2026), but the evidential basis of these black-box predictions remains difficult to interpret: do observable associations exist across text genres and quantities, do predictions persist under lexical content masking, and what effect does prompt engineering have on behavior? For closed or privately hosted models standard interpretability studies are impossible: without access to weights, gradients, or activations, interpretability methods that require internal inspection (Belinkov 2022; Olah et al. 2020) do not apply. Black-box interpretability operates under this constraint, using input perturbation, attribution, and behavioral testing to characterize models from inputs and outputs alone (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Ribeiro et al. 2020). This study asks what zero-shot LLM Big Five Inventory (BFI) predictions rest on once they are examined across text conditions, prompt framings, distributional checks, and content-reduced inputs. This work is organized around three questions:

- **RQ1. Do zero-shot LLMs show BFI signal, and how does it vary with text length and genre?**
- **RQ2. When topical and demographic content is masked, which predictions collapse and which persist as recoverable from the preserved lexical layer?**
- **RQ3. Does linguistic prompting shift self-explanations toward stylistic and affective rationales?**

We approach these with LEX-EC (Lexical Evidence Channel auditing), a novel, reusable behavioral auditing framework for interpretability of LLMs at the task of personality prediction, applied across a text-length gradient, combining distributional and item-level agreement analysis, prompt comparison, and a content-masking ablation that tests signal survival after removal of topical and demographic vocabulary. Rather than treating LLM personality labels as measurements, they are treated solely as behavioral outputs. This framework can narrow causal hypotheses, motivate targeted follow-up interventions, and provide evidence informing decisions about under what conditions personality-labeling systems should be used.

*Corresponding author.

Related Work

Personality Measures and the Big Five Inventory

BFI models personality along Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness through questionnaire responses (John, Donahue, and Kentle 1991; McCrae and Costa 1987; John and Srivastava 1999). Although these traits are ordinarily measured continuously, some reference datasets used in personality labeling historically provide only binary high/low labels. Prior psycholinguistic work has found modest, context-dependent associations between personality and function words, affective vocabulary, and other linguistic patterns (Pennebaker and King 1999; Yarkoni 2010; Schwartz et al. 2013), motivating the possibility of personality prediction from text.

Lexicon-based Textual Analysis: LIWC, Emotion, NRC

Lexicon-based methods such as LIWC map words to psychologically motivated categories including affect, cognition, and function-word use (Tausczik and Pennebaker 2010). Open resources such as Empath and NRC-EmoLex provide related affective and cognitive categories (Fast, Chen, and Bernstein 2016; Mohammad and Turney 2013). Notably, Empath’s categories, when compared against analogous LIWC categories, showed a reported average correlation of $r = .906$ (Fast, Chen, and Bernstein 2016). We use these resources to construct the preserved lexical layer in our masking ablation.

Corpus-linguistic studies show that text length is systematically associated with register, linguistic-feature distributions, and communicative function rather than varying independently of the setting in which text is produced (Liimatta 2022, 2023; Öhman and Liimatta 2024), rendering these characteristics difficult to separate without confounding.

LLM Personality Prediction

Recent work suggests that LLMs can recover nonzero BFI signal from text, sometimes approaching individual human judgments (Derner et al. 2024; Schoenegger et al. 2025; Piastra and Catellani 2025; Wright et al. 2026). Prompting strategies can materially affect this performance (Yang et al. 2023). Reported performance varies substantially with prompt framing, text type, text quantity, and the evaluation target (Bhandarkar et al. 2025; Piastra and Catellani 2025). Models also exhibit demographic sensitivity, positivity bias, poor confidence calibration, and weaker agreement with self-reported traits than with individual human ratings (Rao, Leung, and Miao 2023; Peters and Matz 2024). In particular, prior studies commonly aggregate multiple messages or extended texts per author, leaving performance under limited-text conditions less well characterized; linguistic and affective evidence from topical or demographic cues are also rarely examined. We therefore evaluate predictions across a text-length gradient and test whether apparent signal persists after content masking.

Complementing prompt-based studies, Maharjan et al. trained classifiers over BERT, RoBERTa, and OpenAI embeddings on the PANDORA dataset, reporting stronger per-

formance than zero-shot prompting and associations with LIWC and NRC features, although psychometric reliability was only moderate (Maharjan et al. 2025).

Black-box Interpretability

Because closed LLMs do not expose weights or activations, behavior is often characterized through controlled input transformations. Common black-box interpretability methods include local feature-attribution approaches such as LIME and SHAP, which estimate the contributions of input features to individual predictions (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), as well as behavioral-testing approaches such as CheckList, which evaluate model responses under targeted input conditions and transformations (Ribeiro et al. 2020). Input-reduction studies similarly examine whether predictions persist as information is removed, while showing that behavior on transformed inputs need not provide a faithful account of the process that produced the original prediction (Feng et al. 2018). Some work additionally argues that interpretability methods should be evaluated in relation to the particular claims and contexts they are intended to support (Doshi-Velez and Kim 2018).

Model-generated explanations present a related, but distinct, black-box signal. Fluent self-explanations may not faithfully reflect the computation producing a prediction (Jacovi and Goldberg 2020; Turpin et al. 2023), though they may still reflect some signal in aggregate (Madsen, Chandar, and Reddy 2024).

Comparative evidence indicates that lexicon-based measures are useful in NLP because they provide transparent, computationally inexpensive, and directly inspectable representations of language (van der Veen and Bleich 2025). Related work in quantitative discourse analysis likewise emphasizes the value of explicit lexical evidence for improving methodological transparency and making analytical decisions easier to trace (Compton 2025). Relatedly, PsyTEx removes spans chosen from LLM-generated trait criteria and repeats prediction on the remainder, subsequently analyzing the selected text with LIWC (Bhandarkar et al. 2025). LEX-EC instead specifies the retained psycholinguistic channel independently to avoid self-report dependence and couples lexical restriction with item-level agreement audits.

Datasets

Three datasets were selected to span distinct genres and average text lengths: single MyPersonality Facebook statuses (Celli et al. 2013), graduate student forum introductions (Harbison et al. 2026), and Pennebaker-King essays (Pennebaker and King 1999; Celli et al. 2013). The texts average 17, 89, and 672 mean retained word-tokens¹, respectively, representing micro-, short-, and middle-length conditions. This places the study at the lower end of the text quantities used in prior zero-shot personality-prediction work, which often aggregates multiple letters, statuses, or other texts per au-

¹Word-tokens: spaCy tokens on preprocessed text (URLs stripped, PII and program identifiers replaced), excluding punctuation, whitespace, and emoticon/emoji spans.

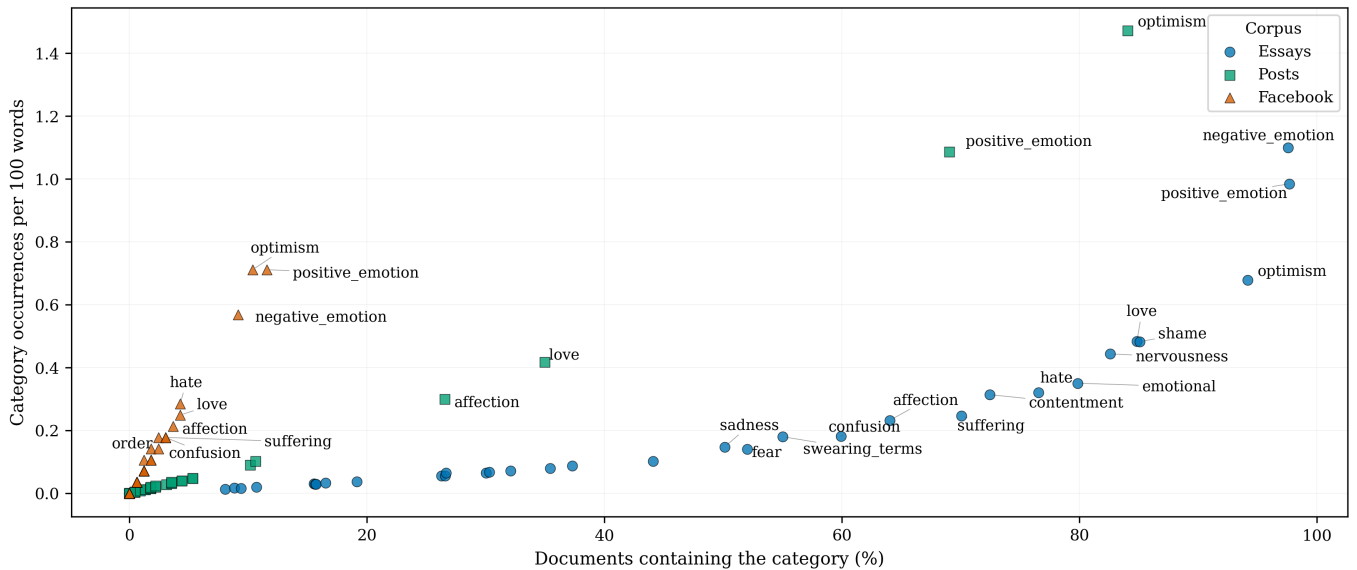


Figure 1: Document prevalence and corpus-level density of retained Empath and NRC categories. Prevalence is the percentage of documents containing at least one category term; density is the number of category occurrences per 100 word-tokens. Lexicon categories are not mutually exclusive. Labels identify selected high-prevalence, high-density, or corpus-distinctive categories.

thor. Binary high/low labels provide a common prediction and evaluation target across corpora.

The corpora also differ in retained psycholinguistic content (Figure 1). Essays show broader document-level coverage of several affective categories; Facebook statuses show lower prevalence but sometimes substantial density when a category occurs; and student introductions contain comparatively frequent optimism-related language. Thus, the conditions vary in lexical composition as well as length and genre.

Student Introduction Posts

We collected BFI-44 responses (John, Donahue, and Kentle 1991) from students across two semesters of Georgia Tech’s OMSCS KBAI course. Of these, 226 also posted course-forum introductions, yielding complete paired survey–post records. Participation was voluntary and consented, procedures were IRB-approved, and names, emails, and links were removed before analysis. BFI-44 scores were converted to binary high/low labels using the midpoint procedure from prior work (Harbison et al. 2026), which is included in the Supplementary material.

The cohort’s trait distributions were uneven, with marked right skew for all traits except Extraversion and Neuroticism; only Extraversion, Neuroticism, and Conscientiousness were sufficiently distributed for trait-level analysis. The introduction posts are a distinctive self-presentational genre: public educational forums encourage impression management (Hayes, Suleiman, and Watling 2024), and the icebreaker template solicited demographic and biographical details including location, courses, specialization, hobbies, motivation, and an interesting fact.

Free-form Essays

The essay corpus is the Pennebaker–King stream-of-consciousness dataset (Pennebaker and King 1999): open-ended writing produced under a fixed-period, write-whatever-comes-to-mind protocol, paired with BFI labels. It serves as our long-form condition, averaging 672 word-tokens ($\approx 3,295$ characters) per text after preprocessing.

We report essay results at two scales. A 226-item subset, drawn by proportional stratification over joint five-trait label profiles with a fixed seed, preserves the full corpus’s trait-profile distribution and is used for pilot prompt and model comparisons; unlike the Facebook sample, it is a proportional downsample. The full corpus ($n=2468$) provides the more stable estimate and supersedes the pilot subset for all trait-level essay claims; the subset serves only to check that pilot behavior is consistent with sampling variability from the full set.

Facebook Statuses

The Facebook corpus derives from the myPersonality project (Celli et al. 2013), which linked Facebook status updates to user-level BFI questionnaire labels. After sampling one status per user with fixed seed 67, the corpus provides a naturally occurring microtext condition averaging approximately 17 word-tokens (91 characters) per text after preprocessing.

We simultaneously balanced the five marginal trait distributions using mixed-integer linear programming. Let $\mathcal{P}$ be the set of observed five-trait binary profiles, n_p the number of users with profile p , and $a_{jp} \in \{0, 1\}$ indicate whether profile p has a positive label for trait j . We selected the number

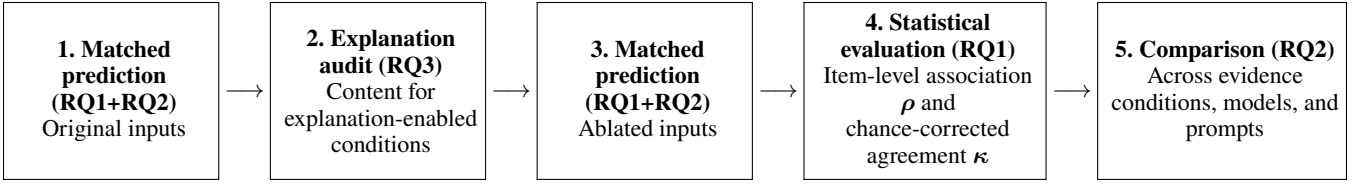


Figure 2: LEX-EC processing stages. Statistical analyses address RQ1, full-versus-masked comparisons address RQ2, and self-explanation analysis addresses RQ3.

z_p of users retained from each profile by solving

$$\begin{aligned}
 \max_{\mathbf{z}} \quad & N = \sum_{p \in \mathcal{P}} z_p \\
 \text{s.t.} \quad & z_p \in \{0, \dots, n_p\}, \quad p \in \mathcal{P}, \\
 & (0.5 - \delta)N \leq \sum_{p \in \mathcal{P}} a_{jp} z_p, \quad j \in \mathcal{T}, \\
 & \sum_{p \in \mathcal{P}} a_{jp} z_p \leq (0.5 + \delta)N, \quad j \in \mathcal{T},
 \end{aligned} \tag{1}$$

where $\mathcal{T}$ is the set of five traits and $\delta = 0.05$. With no fixed target size, the solution attained the largest feasible sample. We solved the program with `scipy.optimize.milp` using its HiGHS backend, sampled users without replacement within each profile, and used fixed seed 67 for all sampling. The resulting dataset contains 164 unique users, with each trait’s positive-label proportion between 45% and 55%.

Methodology

Figure 2 summarizes the LEX-EC workflow. Across the three datasets, models assign binary BFI labels to original and content-masked texts under matched prompt conditions. We separately inspect item-level association and chance-corrected agreement, full-to-masked changes, and explanation content where collected. Each observation consists of one text from one author, and no model is fine-tuned on the target datasets. Model outputs in both conditions were collected through required tool calls.² Unless otherwise stated, all results were each gathered via a single run.

Model Selection

Proprietary LLM evaluation presents a reproducibility problem: model families are revised frequently, older endpoints may be deprecated, and implementation details are generally unavailable. A result obtained from one model snapshot may therefore reflect transient provider- or version-specific behavior rather than a stable property of the task. We address this major reproducibility issue in the field as it applies to our work by evaluating five closed models spanning providers, model generations, and prediction paradigms: GPT-4o-mini (OpenAI 2024a), GPT-4o (OpenAI 2024b), o3-mini (OpenAI 2025), GPT-5.4-mini (OpenAI 2026), and Claude Haiku

²Strict schema-constrained decoding in the API was avoided because it degraded association statistics in pilot tests.

4.5 (Anthropic 2025).^{3 4}

GPT-4o-mini and GPT-4o provide compact and larger-model baselines from the same pre-reasoning family; o3-mini and GPT-5.4-mini provide reasoning-oriented and newer-generation contrasts; and Claude Haiku provides a cross-provider comparison. The purpose is to test whether the main evidence-channel findings may persist across model snapshots that differ in scale, provider, and generation.

Because the pilot comparisons showed similar dataset-level patterns, with limited descriptive variation in the direction and magnitude of Spearman’s r and Cohen’s κ across models, the full-corpus essay analysis was run with GPT-5.4-mini under the linguistic prompt rather than repeated for every model.

Prompts

Prompts directed the model to adopt an expert persona for this task; prompting was treated as a sensitivity factor rather than optimized for a single “best” formulation. We compared *Basic* (BASIC), requesting binary BFI classifications; *Basic+Explanation* (BASICEX), adding trait-level justifications; and *Linguistic+Explanation* (LINEX), directing justifications toward linguistic, stylistic, and affective rather than topical or demographic evidence. Although findings on the efficacy of expert personas are mixed, persona prompting can alter model output behavior, with effects that vary by use (Zheng et al. 2024; Kong et al. 2024; Luz de Araujo et al. 2025). BASICEX therefore matched LINEX in its expert-role and justification instructions, differing only in the instruction to prioritize linguistic evidence.

A preliminary linguistic-only condition was dropped after showing no descriptive performance difference from its explanation-bearing counterpart, as was BASICEX, though it remained in use only as the explanation-bearing baseline for the targeted RQ3 self-explanation audit.

Model Self-Explanations

For explanation-enabled conditions, GPT-4o-mini provided a brief justification alongside each binary label. We an-

³Snapshots used: *gpt-4o-mini-2024-07-18*, *gpt-4o-2024-11-20*, *o3-mini-2025-01-31*, *gpt-5.4-mini-2026-03-17*, *claude-haiku-4-5-20251001*.

⁴Models queried with *temperature* = 0 for non-reasoning models, however this was not possible for reasoning models. *max_tokens* = 3000, *max_completion_tokens* = 3000 used for reasoning models. Reasoning models were set to an effort of *medium*. All used the default *top_p* = 1. No fixed seed was supplied.

alyzed explanations only for Extraversion in the student-introduction dataset, the sole post-level trait with an interpretable predictive signal; label imbalance or lack of trait discrimination made the remaining traits unsuitable for explanation analysis. This is therefore a targeted audit of one trait rather than a five-trait analysis.

We manually coded, with a single annotator, GPT-4o-mini Extraversion explanations for the student introduction posts as linguistic or linguistic-adjacent (LIN), demographic or topical (DEM), combined (CMB), or other (OTH).

We treat self-explanations as aggregate behavioral outputs only, as post hoc LLM explanations may be unfaithful (Jacovi and Goldberg 2020; Turpin et al. 2023; Madsen, Chandar, and Reddy 2024). We test whether the distribution of stated self-explanation types, such as linguistic or affective versus demographic or topical, shifts across prompt conditions. A shift toward linguistic justifications under the linguistic prompt indicates a change in explanation behavior, providing evidence of prompt sensitivity.

Performance and Association Metrics

Because the reference labels may be imbalanced in real-world data, we assess predictions using complementary measures of item-level association and chance-corrected agreement. Our primary measures are

$$\rho_s = \text{corr}(\text{rank}(\mathbf{y}), \text{rank}(\hat{\mathbf{y}})), \quad \kappa = \frac{p_o - p_e}{1 - p_e}, \quad (2)$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the reference and predicted labels, respectively, p_o is observed agreement, and p_e is agreement expected from the observed marginals. For binary variables, Spearman’s ρ is equivalent to the phi coefficient, while Cohen’s κ measures agreement after accounting for marginal prevalence.

We test dependence between predicted and reference labels using Fisher’s exact test for sparse 2×2 tables, and Pearson’s χ^2 test otherwise. We interpret p -values as evidence against independence rather than as measures of practical strength, which is assessed using the magnitudes of ρ and κ .

LLM Personality Trait Prediction

The first prediction stage used the unablated text: for each dataset, prompt condition, and model, the LLM produced binary BFI labels from the full text available in that condition, establishing pre-ablation performance, per Algorithm 1.

Content-Masking Ablation

The preserved lexical layer was constructed from selected Empath Emotion Lexicon categories, selected to closely follow LIWC and the noted strong correlation between them. (Fast, Chen, and Bernstein 2016; Tausczik and Pennebaker 2010). The channel therefore draws on validated lexical categories and an established psycholinguistic framework, granting it a stronger a priori construct-validity basis than an ad hoc or generic bag-of-words method. NRC categories (Mohammad and Turney 2013) were additionally selected and added to the layer to address lexical coverage gaps of specific Emotion categories not granted by the Empath library.

Algorithm 1: Zero-shot BFI prediction

Require: Texts $X = \{x_i\}$, model m , prompt condition c
Ensure: Predictions $\hat{y}_{i,t} \in \{0, 1\}$ and optional explanations $j_{i,t}$
1: $T \leftarrow \{O, C, E, A, N\}$
2: Select prompts (s_c, u_c) and format example F_c
3: Define required tool schema S_c for all $t \in T$, including justification when required by c
4: **for each** text $x_i \in X$ **do**
5: $q_i \leftarrow u_c \mathbin{++} F_c \mathbin{++} x_i$
6: $a_i \leftarrow \text{ToolCall}(m, s_c, q_i, S_c)$
7: $z_i \leftarrow \text{Parse}_m(a_i)$
8: **for each** trait $t \in T$ **do**
9: $\hat{y}_{i,t} \leftarrow \mathbb{I}[\text{low}(z_i[t].\text{classification}) = \text{high}]$
10: **if** c requires explanations **then**
11: $j_{i,t} \leftarrow z_i[t].\text{justification}$
12: **end if**
13: **end for**
14: Store predictions, optional explanations, model, prompt, dataset, and text identifier
15: **end for**
16: **return** $\{(\hat{y}_{i,t}, j_{i,t})\}$

The retained psycholinguistic vocabulary is represented by,

$$V_{\text{psych}} = \bigcup_{c \in C_E} \ell_E(c) \cup \{w \in W_{\text{NRC}} : \ell_N(w) \cap K \neq \emptyset\}, \quad (3)$$

where C_E contains the selected Empath affective and cognitive-style categories, K the retained NRC emotion and polarity categories, $\ell_E(c)$ the Empath word set for category c , and $\ell_N(w)$ the NRC tags assigned to w .

Each text is transformed by,

$$f = \Lambda \circ \sigma \circ \text{Recon}_\tau \circ \text{nlp} \circ \rho_{\text{prog}} \circ \rho_{\text{pii}} \circ \rho_{\text{tmpl}} \circ \rho_{\text{url}}, \quad (4)$$

where ρ_{url} removes URLs; ρ_{tmpl} removes copied ice-breaker questions after normalizing punctuation variants; ρ_{pii} replaces PII placeholders; and ρ_{prog} masks program and course identifiers, degree abbreviations, and course numbers with placeholder μ . Template removal applies only to student introductions, preventing copied prompts from being treated as authored evidence.

After tokenization and POS tagging, tokens are mapped top-down by,

$$\tau(t) = \begin{cases} t, & E(t) \vee Q(t), \\ t, & \text{pos}(t) \in \Pi_F \vee \text{low}(t) \in W_F, \\ t, & \{\text{low}(t), \text{low}(\text{lem}(t))\} \cap V_{\text{psych}} \neq \emptyset, \\ \mu, & \text{pos}(t) = \text{NUM}, \\ \mu, & \text{otherwise.} \end{cases} \quad (5)$$

Here, $E(t)$ and $Q(t)$ identify emoji or emoticons and punctuation or whitespace; Π_F contains retained function-word POS classes; and W_F contains supplemental quantifiers, negations, degree and temporal adverbs, and wh-words. Numerals are masked because they may encode ages,

Metric	O	C	E	A	N
<i>Essays, full corpus: GPT-5.4-mini, LINEX</i>					
ρ	.168→.158 [−.010]	.220→.121 [−.099]	.187→.147 [−.040]	.151→.150 [−.001]	.115→.083 [−.032]
κ	.124→.145 [+0.021]	.220→.119 [−.101]	.187→.144 [−.043]	.135→.136 [+0.001]	.044→.026 [−.018]
<i>Facebook statuses: GPT-4o, BASIC (n = 164)</i>					
ρ	−.020→.066 [+0.086]	.163→.088 [−.075]	.064→.028 [−.036]	−.016→ −.027 [−.011]	.184→ −.058 [−.242]
κ	−.020→.052 [+0.072]	.148→.044 [−.104]	.063→.025 [−.038]	−.016→ −.022 [−.006]	.183→ −.056 [−.239]
<i>Posts: GPT-4o-mini, LINEX (n = 226)</i>					
ρ	−.045→ −.035 [+0.010]	−.026→ −.111 [−.085]	.226→.062 [−.164]	−.036→.045 [+0.081]	.024→.078 [+0.054]
κ	−.031→ −.032 [−.001]	−.023→ −.068 [−.045]	.173→.058 [−.115]	−.017→.044 [+0.061]	.009→.054 [+0.045]

Table 1: Primary association and agreement results before and after content ablation. Cells report Full → Ablated [Δ], where Δ = Ablated − Full. O=Openness, C=Conscientiousness, E=Extraversion, A=Agreeableness, and N=Neuroticism. Shown are: full set essays with GPT-5.4-mini/LINEX after similar pilot results across models; posts with GPT-4o-mini/LINEX; Facebook with GPT-4o/BASIC, which showed the clearest pre-ablation signal and its subsequent attenuation. Deltas use coefficients rounded to three decimal places.

Dataset (pairs)	O	C	E	A	N
Posts (10)	0→1; +.011/ −.001	0→0; −.071/ −.055	7→2; −.078/ −.093	0→0; +.081/ −.005	0→0; +.054/ +.022
Essays (10)	10→9; −.024/ +.004	4→1; −.044/ −.053	6→3; −.034/ −.039	8→7; +.014/ +.008	6→4; −.043/ −.049
Facebook (2)	0→0; +.069/ +.052	1→0; −.044/ −.074	0→0; +.006/.000	0→0; −.022/ −.020	2→0; −.173/ −.171

Table 2: Cross-model summary of changes after content ablation. Cells report significant Spearman associations, Full→Ablated, followed by median $\Delta\rho/\Delta\kappa$ across matched model–prompt pairs. Posts and reduced-set essays include five models under two prompts; Facebook includes GPT-4o under two prompts. Δ = Ablated − Full. Threshold for significance was considered to be $\rho \approx .1307$ for Essays/Posts, $\approx .153$ for Facebook. Deltas use coefficients rounded to three decimal places; pairs with undefined ρ were omitted from median $\Delta\rho$.

years, or counts. Recon _{τ} restores original whitespace, σ collapses consecutive placeholders, and Λ removes empty or placeholder-only lines. The masking placeholder μ was instantiated as the literal underscore character ‘_’ to avoid introducing lexical associations that may accompany replacement with unrelated words, or adjacency issues arising from simple word deletion.

The resulting ablation is a retention test: all material outside the function-word, affective, cognitive-style, and structural channel is masked. Persistence indicates that an association remains recoverable from the retained channel; attenuation indicates that broader lexical content contributed to the unmasked association.

Experiments

Model Self-Explanations

Under the basic explanation prompt, 113 of 226 explanations (50.0%) were coded LIN; under the linguistic prompt, this increased to 153 of 226 (67.7%). Explanations containing any demographic or topical reasoning (DEM or CMB) decreased from 113 of 226 (50.0%) to 72 of 226 (31.9%). Linguistic prompting shifted the content of the generated explanations without eliminating topical reasoning.

Explanation type did not reliably distinguish correct from incorrect predictions: linguistic and demographic/topical justifications appeared in both groups. Linguistic prompting showed the same pattern. For example, sports and fitness references appeared equally often in correct-high and incorrect-high explanations under the basic prompt (14 each), and

appeared across multiple outcome categories under the linguistic prompt.

Prediction correlations also showed prompt sensitivity, but changes between the BASICEX and LINEX conditions varied across models, with no consistent advantage for either prompt.

Prediction and Ablation Results

Prediction patterns differed substantially across datasets (Tables 1 and 2). In the student introduction posts, Extraversion was the only trait with statistically detectable pre-ablation associations in a majority of model–prompt conditions (seven of ten). After masking, only two of ten Extraversion associations remained detectable. Both its association and agreement coefficients decreased in eight of ten conditions, with median paired changes of $\Delta\rho = -.078$ and $\Delta\kappa = -.093$. In the focal GPT-4o-mini/LINEX condition, Extraversion decreased from $\rho = .226$ to $.062$ and from $\kappa = .173$ to $.058$. The other traits remained inconsistent across conditions, with generally negligible chance-corrected agreement and several constant or near-constant prediction distributions.

Essays produced broader associations. In the reduced-set model comparison, positive pre-ablation coefficients appeared across multiple traits, although their magnitudes remained small. In the full-corpus GPT-5.4-mini analysis under the linguistic prompt, the Spearman associations for all five traits were statistically detectable both before and after masking. The largest decrease occurred for Conscientiousness ($\rho: .220 \rightarrow .121$; $\kappa: .220 \rightarrow .119$). Extraversion also decreased ($\rho: .187 \rightarrow .147$; $\kappa: .187 \rightarrow .144$), whereas Agreeableness

changed little (ρ : .151 $\rightarrow$.150; κ : .135 $\rightarrow$.136). For Openness, the association was nearly unchanged (ρ : .168 $\rightarrow$.158), while chance-corrected agreement increased from .124 to .145. Neuroticism had the weakest agreement both before and after masking (κ : .044 $\rightarrow$.026).

The Facebook results were weak across most model-prompt-trait combinations. Four of the 50 pre-ablation associations were statistically detectable: Conscientiousness under Claude Haiku/LINEX and GPT-4o/BASIC, and Neuroticism under GPT-4o/BASIC and GPT-4o/LINEX. Matched Facebook ablation analyses were available for the two GPT-4o prompts; after masking, none of their associations remained statistically detectable. In the focal GPT-4o/BASIC condition, Conscientiousness decreased from $\rho = .163$ and $\kappa = .148$ to $\rho = .088$ and $\kappa = .044$, while Neuroticism changed from $\rho = .184$ and $\kappa = .183$ to negative values ($\rho = -.058$, $\kappa = -.056$). Under the linguistic prompt, Neuroticism also decreased (ρ : .166 $\rightarrow$.062; κ : .164 $\rightarrow$.062). The remaining Facebook coefficients changed in both directions but were small before and after masking.

Overall, masking most consistently attenuated Extraversion in the introduction posts and Conscientiousness in the full-corpus essay analysis. Openness and Agreeableness remained comparatively stable in the focal essay condition, while the limited pre-ablation Facebook associations were not maintained after masking.

Across models and prompts, the reduced-set essay results varied in magnitude but showed the same broad trait-level pattern.

In the Facebook condition, 59.8% of lexical tokens survived masking, comparable to the student posts. However, statuses were very short, the resulting inputs averaged only 10.2 retained tokens, including ≈ 1.9 psycholinguistic-vocabulary tokens. Post-ablation, mean retained word-token counts for Essays were 471, Posts 54, Facebook 10.

Discussion

Essays produced the broadest associations, but small chance-corrected agreement, indicating that statistical detectability did not translate into reliable individual-level classification. Student introductions yielded a narrower pattern centered on Extraversion, perhaps because this genre naturally elicits descriptions of social activities and preferences. Facebook performance remained weak despite balanced reference labels, showing a possible lower length/content boundary condition.

The ablation results further suggest that different traits are consistent with dependence on different linguistic cues. In the full-corpus essay analysis, Conscientiousness was most attenuated, consistent with greater dependence on removed content-bearing vocabulary. Agreeableness was comparatively stable, while the near-stable Openness association accompanied by increased κ more likely reflects a change in prediction marginals than stronger trait information. In student introductions, the repeated decline in Extraversion across model-prompt conditions suggests sensitivity to comparatively explicit lexical cues. The disappearance of the isolated Facebook associations is likewise consistent with dependence on sparse lexical evidence, although the short

retained inputs and limited matched ablation conditions preclude a stronger explanation.

No model showed a stable advantage across datasets and conditions; model differences were less systematic than differences associated with genre, trait, and masking. This suggests that the pattern of findings may generalize to newer model families.

Linguistic prompting increased linguistic explanations, while explanations containing demographic or topical reasoning decreased. Prompt framing therefore changed the distribution of stated justifications without directly establishing that those justifications reflected the evidence responsible for the classifications (Jacovi and Goldberg 2020; Turpin et al. 2023).

The central result is not simply that performance was weak: associations differed in their sensitivity to lexical masking. Some remained observable from the retained function-word, psycholinguistic, and structural channel, whereas others attenuated when broader lexical evidence was removed. The broadly similar patterns across models suggest that weak performance in some conditions may reflect limitations of the available textual evidence rather than a deficiency specific to one model.

These findings suggest that zero-shot personality prediction is best characterized as a property of a specific configuration of model, genre, trait, available word usage than as a general capability of a model. Persistence under lexical restriction is not inherently evidence of better personality prediction: it identifies signal recoverable from the retained channel, which may include differing stylistic contents. Personality-labeling systems should thus be evaluated separately for each intended text condition, rather than treated as possessing a general personality prediction capability.

Limitations

The explanation audit was manually coded by one annotator, so the reported explanation-category frequencies and prompt-related shifts lack an independent inter-rater reliability estimate.

Exposure of closed models to the public personality corpora is unknown; results on those corpora cannot establish fully uncontaminated out-of-sample performance. The private student-introduction corpus was immune to this effect. Additionally, by removing all but psycholinguistic material during ablation, this may disrupt correspondence with data consumed during training.

Conclusion

Zero-shot BFI prediction from a single text is not a unitary capability: evidential profiles varied across corpora and traits, and lexical restriction produced heterogeneous changes. Linguistic prompting also shifted model self-explanations toward stylistic and affective evidence, indicating prompt sensitivity in explanation behavior.

These results demonstrate the utility of LEX-EC as a task-specific black-box audit framework. By examining agreement, lexical-channel persistence, and explanation content together, our framework revealed information invisible to conventional statistical measures alone.

AI Tool Use Disclosure

The authors utilized GPT 5.5-5.6, Claude Opus 4.8, and Claude Fable 5 while editing text for clarity, grammar, and concision, and as non-decisional adversarial review aids to flag assumptions/methodological weaknesses in both the text and suggestions. Most non-copyediting suggestions were rejected as scientifically invalid/out-of-scope by the authors; suggestions that did inform revisions were human-verified.

References

- Anthropic. 2025. Claude Haiku 4.5 System Card. System card, Anthropic. Available at <https://www.anthropic.com/claude-haiku-4-5-system-card>.
- Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1): 207–219.
- Bhandarkar, A.; Wilson, R.; Swarup, A.; Webster, G. D.; and Woodard, D. L. 2025. PsyTEx: A Knowledge-Guided Approach to Refining Text for Psychological Analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 151–178. Albuquerque, USA: Association for Computational Linguistics.
- Celli, F.; Pianesi, F.; Stillwell, D.; and Kosinski, M. 2013. Workshop on Computational Personality Recognition: Shared Task. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(2): 2–5.
- Compton, T. 2025. Beyond the Black Box: Integrating Lexical and Semantic Methods in Quantitative Discourse Analysis with BERTopic. arXiv:2508.19099.
- Derner, E.; Kučera, D.; Oliver, N.; and Zahálka, J. 2024. Can ChatGPT Read Who You Are? *Computers in Human Behavior: Artificial Humans*, 2(2): 100088.
- Doshi-Velez, F.; and Kim, B. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. In Escalante, H. J.; Escalera, S.; Guyon, I.; Baró, X.; Güçlütürk, Y.; Güçlü, U.; and van Gerven, M., eds., *Explainable and Interpretable Models in Computer Vision and Machine Learning*, The Springer Series on Challenges in Machine Learning, 3–17. Cham: Springer. ISBN 978-3-319-98131-4.
- Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657.
- Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. Brussels, Belgium: Association for Computational Linguistics.
- Goldberg, L. R. 1990. An Alternative “Description of Personality”: The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, 59(6): 1216–1229.
- Harbison, B.; Taubman, S.; Taylor, T.; and Goel, A. K. 2026. Personality-Enhanced Social Recommendations in SAMI: Exploring the Role of Personality Detection in Matchmaking. In *INTED2026 Proceedings*. IATED.
- Hayes, B.; Suleiman, A.; and Watling, D. 2024. Students’ Impression Management and Self-Presentation Behaviours via Online Educational Platforms: An Archival Review. *First Monday*, 29(3).
- Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.
- John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. The Big Five Inventory—Versions 4a and 54. Technical report, University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA.
- John, O. P.; and Srivastava, S. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In Pervin, L. A.; and John, O. P., eds., *Handbook of Personality: Theory and Research*, 102–138. New York: Guilford Press, 2nd edition.
- Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; and Dong, X. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4099–4113. Mexico City, Mexico: Association for Computational Linguistics.
- Liimatta, A. 2022. Do Registers Have Different Functions for Text Length? A Case Study of Reddit. *Register Studies*, 4(2): 263–287.
- Liimatta, A. 2023. Register Variation across Text Lengths: Evidence from Social Media. *International Journal of Corpus Linguistics*, 28(2): 202–231.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.
- Luz de Araujo, P. H.; Röttger, P.; Hovy, D.; and Roth, B. 2025. Principled Personas: Defining and Measuring the Intended Effects of Persona Prompting on Task Performance. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 26857–26886. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.
- Madsen, A.; Chandar, S.; and Reddy, S. 2024. Are Self-Explanations from Large Language Models Faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, 295–337. Bangkok, Thailand: Association for Computational Linguistics.
- Maharjan, J.; Jin, R.; Zhu, J.; and Kenne, D. 2025. Psychometric Evaluation of Large Language Model Embeddings for Personality Trait Prediction. *Journal of Medical Internet Research*, 27: e75347.
- Matz, S. C.; Kosinski, M.; Nave, G.; and Stillwell, D. J. 2017. Psychological Targeting as an Effective Approach to Digital

- Mass Persuasion. *Proceedings of the National Academy of Sciences*, 114(48): 12714–12719.
- McCrae, R. R.; and Costa, P. T., Jr. 1987. Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology*, 52(1): 81–90.
- McCrae, R. R.; and John, O. P. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2): 175–215.
- Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3): 436–465.
- Öhman, E. S.; and Liimatta, A. 2024. Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 1–8. Miami, USA: Association for Computational Linguistics.
- Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom In: An Introduction to Circuits. *Distill*, 5(3).
- OpenAI. 2024a. GPT-4o mini: Advancing Cost-Efficient Intelligence. OpenAI model release. Available at <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- OpenAI. 2024b. GPT-4o System Card. System card, OpenAI. Available at <https://cdn.openai.com/gpt-4o-system-card.pdf>.
- OpenAI. 2025. OpenAI o3-mini System Card. System card, OpenAI. Available at <https://cdn.openai.com/o3-mini-system-card-feb10.pdf>.
- OpenAI. 2026. GPT-5.4 Thinking System Card. System card, OpenAI. Appendix: GPT-5.4 mini, added March 17, 2026. Available at <https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf>.
- Pennebaker, J. W.; and King, L. A. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6): 1296–1312.
- Peters, H.; and Matz, S. C. 2024. Large Language Models Can Infer Psychological Dispositions of Social Media Users. *PNAS Nexus*, 3(6): pgae231.
- Piastra, M.; and Catellani, P. 2025. On the Emergent Capabilities of ChatGPT 4 to Estimate Personality Traits. *Frontiers in Artificial Intelligence*, 8: 1484260.
- Rao, H.; Leung, C.; and Miao, C. 2023. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1184–1194. Singapore: Association for Computational Linguistics.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics.
- Schoenegger, P.; Greenberg, S.; Grishin, A.; Lewis, J.; and Caviola, L. 2025. AI Can Outperform Humans in Predicting Correlations Between Personality Items. *Communications Psychology*, 3: 23.
- Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D. J.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8(9): e73791.
- Tausczik, Y. R.; and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1): 24–54.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems 36*, 74952–74965.
- van der Veen, A. M.; and Bleich, E. 2025. The Advantages of Lexicon-Based Sentiment Analysis in an Age of Machine Learning. *PLOS ONE*, 20(1): e0313092.
- Vinciarelli, A.; and Mohammadi, G. 2014. A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 5(3): 273–291.
- Wright, A. G. C.; Ringwald, W. R.; Vize, C. E.; Eichstaedt, J. C.; Angstadt, M.; Taxali, A.; and Sripada, C. 2026. Assessing Personality Using Zero-Shot Generative AI Scoring of Brief Open-Ended Text. *Nature Human Behaviour*, 10(3): 541–555.
- Yang, T.; Shi, T.; Wan, F.; Quan, X.; Wang, Q.; Wu, B.; and Wu, J. 2023. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3305–3320. Singapore: Association for Computational Linguistics.
- Yarkoni, T. 2010. Personality in 100,000 Words: A Large-Scale Analysis of Personality and Word Use Among Bloggers. *Journal of Research in Personality*, 44(3): 363–373.
- Zheng, M.; Pei, J.; Logeswaran, L.; Lee, M.; and Jurgens, D. 2024. When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15126–15154. Miami, Florida, USA: Association for Computational Linguistics.

Trait	Value	Proportion
O	1	0.898230 (89.82%)
	0	0.101770 (10.18%)
C	1	0.840708 (84.07%)
	0	0.159292 (15.93%)
E	1	0.438053 (43.81%)
	0	0.561947 (56.19%)
A	1	0.871681 (87.17%)
	0	0.128319 (12.83%)
N	1	0.429204 (42.92%)
	0	0.570796 (57.08%)

Table 3: Distribution of binary Big Five personality labels in the student introduction-post dataset ($N = 226$). Trait abbreviations: O = Openness, C = Conscientiousness, E = Extroversion, A = Agreeableness, and N = Neuroticism.

Appendix

Midpoint for Student Introduction Posts

Before use, the student post dataset was converted from the original continuous data to binary high/low. Each personality-trait score was dichotomized using the fixed scale midpoint of $\tau = 3$. For participant i and personality trait j , the binary classification c_{ij} was defined as,

$$c_{ij} = \mathbb{I}(x_{ij} > 3) = \begin{cases} 0, & \text{if } x_{ij} \leq 3, \\ 1, & \text{if } x_{ij} > 3, \end{cases} \quad (6)$$

where x_{ij} denotes the original personality-trait score and $\mathbb{I}$ denotes the indicator function. Thus, scores less than or equal to the scale midpoint were assigned to class 0, whereas scores above the midpoint were assigned to class 1. Past work utilizing this dataset (Harbison et al. 2026), instead used a random midpoint procedure, which can introduce stochastic label noise, and was therefore avoided in this work.

Additional Dataset Distributions

This section includes additional tables. Table 3 demonstrates the distribution of the Student Introduction Post dataset, introduced in Section "Datasets" in the main body. Table 4 shows the distribution data for the marginally balanced single-post-per-user subset, while Table 5 and Table 6 show the distribution data for the full essay set and for the subset, respectively.

Lexicon Categories for Ablation

This section shows the full list of retained lexicon subcategories from Empath and NRC, shown in Table 7

Complete Results

The full statistical output of the described experiments are included here.

Trait	Value	Proportion
cOPN	1	0.548780 (54.88%)
	0	0.451220 (45.12%)
cCON	1	0.542683 (54.27%)
	0	0.457317 (45.73%)
cEXT	1	0.518293 (51.83%)
	0	0.481707 (48.17%)
cAGR	1	0.548780 (54.88%)
	0	0.451220 (45.12%)
cNEU	1	0.451220 (45.12%)
	0	0.548780 (54.88%)

Table 4: Binary division of classification buckets for BFI traits from the myPersonality single-post-per-user subset ($N = 164$). Trait abbreviations follow those used by the underlying myPersonality dataset: cOPN = Openness, cCON = Conscientiousness, cEXT = Extroversion, cAGR = Agreeableness, and cNEU = Neuroticism.

Trait	Value	Proportion
cOPN	1	0.515397 (51.54%)
	0	0.484603 (48.46%)
cCON	1	0.508104 (50.81%)
	0	0.491896 (49.19%)
cEXT	1	0.517423 (51.74%)
	0	0.482577 (48.26%)
cAGR	1	0.530794 (53.08%)
	0	0.469206 (46.92%)
cNEU	1	0.499595 (49.96%)
	0	0.500405 (50.04%)

Table 5: Distribution of binary Big Five personality labels in the full, original Pennebaker-King Essays dataset ($N = 2468$). Trait abbreviations follow those used by the underlying Pennebaker-King Essays dataset: cOPN = Openness, cCON = Conscientiousness, cEXT = Extroversion, cAGR = Agreeableness, and cNEU = Neuroticism.

Examples of Pre-Ablated and Post-Ablated Text Below is a contiguous essay excerpt before and after content masking. The model received the complete essay in each condition; only the corresponding excerpt is shown. Each underscore represents one or more consecutive masked tokens.

Original Essay Excerpt: I’m not saying that I’m perfect, but I’ve learned over the past few years about what I want out of life and what I don’t want. I’m living my life the way I want to, as stress-free as possible and as happy as possible. When I’m put into these stupid situations it just makes life that much harder and it sucks!

Ablated: I’m not _ that I’m perfect, but I’ve learned over the _ few _ about what I _ out of _ and what I don’t _ . I’m _ my _ the _ I _ to. _ stress- _ as _ and _ happy as _ . When I’m _ into these stupid _ it just _

Trait	Value	Proportion
cOPN	1	0.513274 (51.33%)
	0	0.486726 (48.67%)
cCON	1	0.508850 (50.88%)
	0	0.491150 (49.12%)
cEXT	1	0.517699 (51.77%)
	0	0.482301 (48.23%)
cAGR	1	0.535398 (53.54%)
	0	0.464602 (46.46%)
cNEU	1	0.500000 (50.00%)
	0	0.500000 (50.00%)

Table 6: Distribution of binary Big Five personality labels in the smaller Pennebaker-King Essays subset ($N = 226$, $seed = 67$). Trait abbreviations: cOPN = Openness, cCON = Conscientiousness, cEXT = Extroversion, cAGR = Agreeableness, and cNEU = Neuroticism.

Lexicon source	Category group	Retained subcategories
Empath	Positive affect	positive_emotion, joy, cheerfulness, contentment, affection, love, optimism, pride, zest
Empath	Negative affect	negative_emotion, sadness, disappointment, suffering, torment, nervousness, fear, timidity, anger, rage, aggression, irritability, exasperation, hate, disgust, shame
Empath	Social-affective stance	sympathy, emotional, warmth
Empath	Social/normative language	swearing_terms
Empath	Cognitive style	thinking, order, confusion, anticipation, deception
NRC	Core emotions and polarity	joy, sadness, anger, fear, disgust, surprise, positive, negative

Table 7: Psycholinguistic and affective lexicon categories retained during ablation.

much harder and it sucks!

Prompts

Listing 1 shows the required format JSON example. Listing 2 shows the tool call definition that was then used in the model API call to return model outputs in the required structured format. The classification field was required in every condition. The justification field was required for the *Basic+Ex* and *Linguistic+Ex* conditions and omitted for the *Basic* condition.

Table 8 shows the full prompts, system and user, used in the study.

Listing 1: Unified JSON response format. The justification field is included only in explanation-required conditions.

```

1 expected_format = ""
2 {
3   "Openness": {
4     "classification": "low/high"
5   },
6   "Conscientiousness": {
7     "classification": "low/high"
8   },
9   "Extroversion": {
10    "classification": "low/high"
11  },
12  "Agreeableness": {
13    "classification": "low/high"
14  },
15  "Neuroticism": {
16    "classification": "low/high"
17  }
18 }
19 ""

```

Condition	Basic
System message	You are an expert in inferring Big-5 personality traits from text.
User message	Analyze this text for personality traits. Classify each Big Five trait as either 'low' or 'high'. Return your results in the following JSON format without explanation:
Condition	Basic+Ex
System message	You are an expert in inferring Big-5 personality traits from text. Your justifications are always evidence-based and grounded in direct quotes from the text.
User message	Your task is to analyze the provided text, classify the author's Big Five personality traits, and provide a detailed justification for your analysis. Classify each Big Five trait as either 'low' or 'high', no deviation allowed. Return your detailed justification and results in the following JSON format:
Condition	Linguistic+Ex
System message	You are an expert in inferring Big-5 personality traits from linguistic patterns in text.
User message	Analyze this text for personality traits based only on writing style, word choice, and language patterns. Classify each Big Five trait as either 'low' or 'high'. If the linguistic evidence is insufficient, make your best inference based on available patterns, but always provide a classification, and note this in your explanation. Return your explanation and results in the following JSON format:

Table 8: Prompt conditions used for zero-shot Big Five prediction. Prompts are reproduced as used. Each user message was followed by an expected JSON-format example and the target text sample. Outputs were collected through required structured tool calls.

Model Self-Explanation Results Below is an example of an original student introduction text, the model's prediction,

Listing 2: Abridged structured tool schema used in the Basic+Ex and Linguistic+Ex conditions. The remaining four traits used the same object structure as Openness.

```

1  tools = [
2  {
3    "type": "function",
4    "function": {
5      "name": "record_personality_profile",
6      "description": "Records the Big Five personality traits and a justification for each classification.",
7      "parameters": {
8        "type": "object",
9        "properties": {
10         "Openness": {
11           "type": "object",
12           "description": "The analysis for the Openness trait.",
13           "properties": {
14             "classification": {
15               "type": "string",
16               "enum": ["low", "high"]
17             },
18             "justification": {
19               "type": "string",
20               "description": "Reasoning based on a direct quote from the text."
21             }
19           }
22         },
23         "required": [
24           "classification",
25           "justification"
26         ]
27       }
28     },
29     # Structure repeats for the other traits
30   },
31   "required": [
32     "Openness",
33     "Conscientiousness",
34     "Extroversion",
35     "Agreeableness",
36     "Neuroticism"
37   ]
38 }
39 ]
40 }
41 ]

```

and the explanation for its reasoning the model provided. Incidentally, this was a correct prediction, though the self-explanation listed topical content. This is one example of an explanation tagged 'DEM'.

Introduction post:

"Hello!

My name is [NAME], and I currently live in [LOCATION]. This is my first [PROGRAM] course!

I currently work as a software engineer at [ORGANIZATION]. Outside of work, I like to play sports and hang out with friends/family.

From this class, I hope to develop a strong foundation in [FIELD].

One interesting fact about me is that I also have a [SECOND OCCUPATION]."

Prediction: High Extraversion

Explanation: "The text expresses enjoyment of social activities, such as playing sports and spending time with friends and family, indicating sociability."

Below is another example, this time tagged as 'LIN'. To preserve anonymous review, identifying spans in the displayed example were replaced with placeholders. The model received the complete original post. Interestingly, the model's prediction was correct in this case, as well.

Introduction post:

"Hello,

I am [NAME]. Currently, I am living in [LOCATION], and I am a [OCCUPATION] by day and a [DEGREE PROGRAM] student by night. I am specifically interested in this course, as I would like to learn more about [FIELD] and determine how to implement it in my current research. I took [COURSE] and [COURSE] last semester. That was a fun, occasionally stressful, and excellent learning experience. I am looking forward to this semester, and I hope to learn a lot. :)"

Prediction: Low Extraversion

Explanation: "The text does not include social or outgoing language; it focuses more on personal academic interests rather than social interactions."

Table 9 shows the counts for text content broad category markers for correct/incorrect and across total predictions. Additionally, it shows the confusion matrix of cue counts for text content.

Full Prediction and Ablation Results This section presents the remaining full Spearman r / Cohen's k results across all models that were not included in the main body, for Posts, Essays, and Facebook. Only Ablation results for GPT-4o were included as other models either failed to output results reaching significance or failed to complete the task consistently at all.

Pre-Ablated Predictions from Text Table 10 shows the full results on the non-ablated text, for all 3 datasets.

Post-Ablated Predictions from Text Table 11 shows the final full results on the ablated text, for all 3 datasets.

Prompt	Explanation-tag counts			Topical cues by prediction outcome			
	Correct	Incorrect	Total	Correct high	Correct low	Incorrect high	Incorrect low
Basic	LIN	68; LIN	45; LIN	113; S/F 14; family (3),	S/F 3; kids (1), Net-	S/F 14; friends (4),	S/F 10; family (3),
	DEM	48; DEM	29; DEM	77; board games (2),	flix (1), coding (1),	gaming (4), concerts	traveling (2)
	CMB 19	CMB 17	CMB 36	friends (2), band (2),	gaming (1)	(3)	
Linguistic	LIN	87; LIN	66; LIN	153; S/F 34; family (4),	S/F 19; reading (7),	S/F 19; family (6),	S/F 6; gaming (3),
	DEM	29; DEM	25; DEM	54; band (3), board	gaming (5), chess travel	(5), friends board/card	games
	CMB	7; CMB	11; CMB	18; games (3), concerts	(2), dogs (2), Netflix	(3), concerts (3)	(3)
	OTH 1	OTH 0	OTH 1	(3), friends (3), dog	(2)		

Table 9: GPT-4o-mini Extroversion explanation audit for student introduction posts. Tag columns report explanation-category counts; topical-cue columns report sports/fitness and other cue frequencies. LIN = linguistic; DEM = demographic/topical; CMB = combined; OTH = indeterminate.

Prompt	Model	O	C	E	A	N
<i>Posts (n=226)</i>						
BASIC	GPT-4o	-.039/-.024	.055/.027	.101/.088	.101/.074	-.058/-.009
	GPT-4o-mini	-.032/-.017	.055/.027	.209**/.153**	.105/.049	.077/.012
	5.4-mini	-.022/-.009	.033/.019	.162*/.151*	.105/.049	-.082/-.018
	o3-mini	-.032/-.017	-.058/-.033	.211**/.178**	-.026/-.009	-.???
	Claude Haiku	-.022/-.009	-.041/-.017	.238***/.165***	-.???	-.???
LINEX	GPT-4o	-.056/-.044	.033/.019	.091/.088	.110/.094	-.???
	GPT-4o-mini	-.045/-.031	-.026/-.023	.226***/.173**	-.036/-.017	.024/.009
	5.4-mini	-.???	-.058/-.033	.138*/.137	-.063/-.046	-.022/-.006
	o3-mini	.066/.045	-.058/-.033	.101/.088	-.058/-.039	-.???
	Claude Haiku	-.022/-.009	.033/.019	.205**/.136**	-.???	-.???
<i>Essays (n=226)</i>						
BASIC	GPT-4o	.221***/.172**	.139*/.122	.102/.097	.200**/.195**	.133*/.062
	GPT-4o-mini	.209**/.142**	.115/.097	.101/.100	.144*/.137*	.198**/.142**
	5.4-mini	.263***/.198***	.149*/.149*	.151*/.150*	.193**/.178**	.102/.062
	o3-mini	.269***/.178***	.126/.106	.111/.106	.128/.125	.168*/.097*
	Claude Haiku	.132*/.047	.114/.104	.183**/.183**	.165*/.144*	.124/.062
LINEX	GPT-4o	.215**/.162**	.150*/.131*	.180**/.179*	.198**/.198**	.119/.071
	GPT-4o-mini	.204**/.133**	.087/.071	.158*/.157*	.185**/.164**	.175**/.097*
	5.4-mini	.263***/.198***	.115/.114	.151*/.151*	.165*/.145*	.133*/.062
	o3-mini	.271***/.188***	.152*/.131*	.126/.122	.109/.106	.166*/.106*
	Claude Haiku	.180**/.094*	.126/.113	.147*/.141*	.185**/.165**	.089/.044
<i>Facebook (per-user, n=164)</i>						
BASIC	GPT-4o	-.020/-.020	.163*/.148	.064/.063	-.016/-.016	.184*/.183*
	GPT-4o-mini	-.015/-.015	.087/.051	.088/.088	.078/.078	.081/.078
	5.4-mini	-.025/-.024	.071/.061	.024/.024	.046/.046	.076/.076
	o3-mini	-.007/-.007	.078/.064	-.000/-.000	.049/.048	.117/.110
	Claude Haiku	-.060/-.056	.127/.119	.015/.014	-.042/-.040	.094/.094
LINEX	GPT-4o	.053/.051	.118/.099	.060/.060	.037/.037	.166*/.164
	GPT-4o-mini	.056/.056	.071/.041	-.022/-.022	.116/.115	.067/.064
	5.4-mini	???	.122/.106	-.038/-.038	-.010/-.009	.083/.082
	o3-mini	.051/.051	.062/.052	.041/.040	-.088/-.085	.098/.096
	Claude Haiku	.004/.004	.167*/.149*	.080/.064	-.058/-.053	.080/.080

Table 10: Cells: Spearman ρ /Cohen's κ . Stars on Spearman p , on κ from χ^2 p : * $p < .05$, ** $p < .01$, *** $p < .001$. O=Openness, C=Conscientiousness, E=Extroversion, A=Agreeableness, N=Neuroticism. Facebook ablation run for GPT-4o only. Values of ??? indicate single class prediction only.

Prompt	Model	O	C	E	A	N
<i>Posts (n=226)</i>						
BASIC	GPT-4o	-.001/-.001	-.050/-.048	.017/.006	-.???	.052/.023
	GPT-4o-mini	.015/.013	-.045/-.044	.103/.038	-.???	.031/.014
	5.4-mini	.049/.036	-.048/-.048	.153*/.139*	.070/.056	.024/.009
	o3-mini	.145*/.142	-.118/-.097	.139*/.114	.082/.082	.052/.023
	Claude Haiku	-.026/-.025	-.057/-.057	.061/.033	-.036/-.017	.027/.012
LINEX	GPT-4o	-.045/-.045	-.004/-.004	.098/.097	.072/.072	.092/.062
	GPT-4o-mini	-.035/-.032	-.111/-.068	.062/.058	.045/.044	.078/.054
	5.4-mini	-.008/-.008	-.050/-.048	.075/.029	-.???	.027/.012
	o3-mini	-.038/-.038	-.001/-.001	.110/.090	.059/.059	.051/.020
	Claude Haiku	-.026/-.025	-.057/-.057	.061/.033	-.036/-.017	.027/.012
<i>Essays (n=226)</i>						
BASIC	GPT-4o	.224***/.215**	.086/.057	.076/.057	.073/.053	.030/.009
	GPT-4o-mini	.160*/.139*	.038/.023	.064/.062	.148*/.148*	.144*/.053
	5.4-mini	.175**/.146*	.078/.077	.091/.088	.182**/.176**	.137*/.080
	o3-mini	.171**/.115*	.045/.024	.033/.027	.112/.094	.110/.035
	Claude Haiku	.180**/.094*	.084/.071	.152*/.152*	.247***/.241***	.116/.027
LINEX	GPT-4o	.216**/.183**	.139*/.092	.194**/.172**	.230***/.223***	.128/.044
	GPT-4o-mini	.245***/.232***	.106/.050	.069/.064	.096/.095	.144*/.053
	5.4-mini	.161*/.151*	.087/.085	.144*/.139*	.231***/.215***	???
	o3-mini	.128/.100	.071/.040	.119/.097	.132*/.111	.096/.035
	Claude Haiku	.192**/.104**	.091/.072	.110/.102	.220***/.214**	.134*/.035
<i>Facebook (per-user, n=164)</i>						
BASIC	GPT-4o	.066/.052	.088/.044	.028/.025	-.027/-.022	-.058/-.056
LINEX	GPT-4o	.104/.083	.105/.055	.108/.098	.004/.004	.062/.062

Table 11: Cells: Spearman ρ /Cohen’s κ . Stars on Spearman p , on κ from χ^2 p : * $p < .05$, ** $p < .01$, *** $p < .001$. O=Openness, C=Conscientiousness, E=Extroversion, A=Agreeableness, N=Neuroticism. Facebook ablation run for GPT-4o only. Values of ??? indicate single class prediction only.