Implicit Personality Representations in Humans and LLMs

Yilin Geng[1], Omri Abend[2], Eduard Hovy[1], Lea Frermann[1,3]
[1]University of Melbourne, [2]Hebrew University of Jerusalem, [3]University of Tübingen
Correspondence: yilin.geng@student.unimelb.edu.au

Abstract

A century of psychology has found that the trait words people use to describe one another vary, but the relational structure among those traits, which ones go together and which oppose, is strikingly consistent across raters and cultures. We test whether the LLM (Qwen 2.5-7B-Instruct) reproduces this structure in its internal trait representations. From millions of crowd-sourced personality ratings of fictional characters, we build a human implicit-personality matrix over hundreds of traits; from contrastive model activations, we build a matching matrix over the same traits. The two relational structures align strongly (Mantel $r = 0.77$), and the agreement holds trait by trait as well as in aggregate. Two dominant axes of the model's trait representations recover the social and intellectual dimensions long known to organize human personality impressions, social warmth and intellectual competence. On held-out dialogue, projecting model activations onto these directions yields personality profiles that agree with human ratings. This work establishes a framework that enables comprehensive, human-grounded comparison between internal model trait geometry and the shared structure of human personality impressions.

1 Introduction

Humans are remarkably efficient and consistent at assigning a personality to individuals, and they do so by attributing stable characteristics — or *traits* — such as brave, stubborn, or curious that are inferred from how a person speaks and acts (Chaplin et al., 1988). A century of research on personality psychology has established structured personality representations (Allport and Odbert, 1936; Goldberg, 1990), and has shown that while the trait-describing terms vary among individuals and cultures (Albright et al., 1988), their covariance or *structure* of this representational space is highly consistent (Schneider, 1973; Shweder and D'Andrade, 1980).

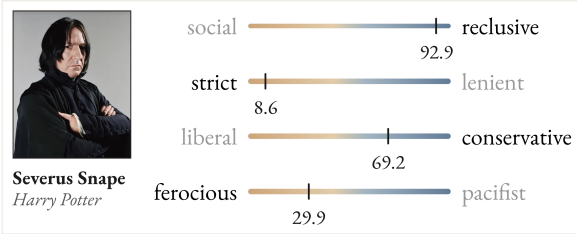

Figure 1: Example profile (simplified) of Severus Snape from *Harry Potter*. Each trait scale is a spectrum defined and constrained by two poles of contrastive words or phrases. This work covers 385 such trait scales.

Trained on the text that people produce, large language models (LLMs) exhibit personality-relevant behavior (Serapio-García et al., 2025; Lee et al., 2025), and individual concepts can be located as directions in their internal representations (Park et al., 2024; Arditi et al., 2024; Chen et al., 2025). Prior output-level work has also recovered known personality factors from model probabilities over trait descriptors (Suh et al., 2024). This paper investigates whether representations (independently extracted internal trait directions) in the LLM activation space reproduce a relational structure measured from human personality impressions, using a comprehensive, human-grounded framework.

Our framework builds on a large inventory of over 380 traits, where each trait is represented as a spectrum with two poles (See Figure 1). The set of traits is derived from a large-scale data set of *human* ratings of fictional characters, and we use the traits and their human ratings to ground the LLM representations. Specifically, we use the Statistical "Which Character" Personality Questionnaire (SWCPQ), a crowd-sourced data set in which volunteers contributed tens of millions of ratings for thousands of fictional characters (Open Source Psychometrics Project, 2023). We thus extend prior work on internal LLM personality representation, which was constrained to a handful of traits which

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 32092–32109
October 24-29, 2026 ©2026 Association for Computational Linguistics

were defined ad-hoc (Arditi et al., 2024; Chen et al., 2025).

We analyse the LLM's internal representation of traits directly, rather than inferring from model output, because it has been shown repeatedly that model behavior does not always accurately reflect its internal mechanisms (Turpin et al., 2023). Prior work has prompted LLMs to respond to personality questionnaires and has inferred personality such as a "Big-5 profile" based on their responses (Serapio-García et al., 2025; Lee et al., 2025; Huang et al., 2024), however, this approach does not conclusively answer how LLMs *represent* personality, and does not allow us to compare their representational space to the one established in human personality psychology.

Instead, we build on the *contrastive prompting* method of Chen et al. (2025), improving its clarity, scalability, and validation process. Given a trait with two extreme poles (Figure 1), we prompt the LLM with scenarios that would elicit divergent behavior by characters corresponding to the respective poles (see an example in Figure 2). The LLM generates the two diverging behaviors, and we extract the corresponding activations and construct personality vectors from their differences. We construct corresponding representations from our human-created SWCPQ dataset for the same trait set, which we then directly compare with the emerging LLM representation space (Kriegeskorte et al., 2008; Mantel, 1967).

We find strong alignment between the LLM's trait geometry (Qwen 2.5-7B-Instruct) and the human impression structure. Applying dimensionality reduction, we find that a large amount of variance in the model space is captured by two principal components corresponding to *social warmth* and *intellectual competence*. These dimensions were identified in human personality-impression research independently of the SWCPQ inventory (Rosenberg et al., 1968; Fiske et al., 2007). We further test whether the extracted directions transfer to a different text genre by projecting Wikiquote dialogue for 256 held-out characters and comparing the resulting profiles with the human reference.

This work shows that a language model carries not just isolated personality concepts but the structure that organizes them. This structure, recovered from models' internal activity, matches one recovered from millions of human personality impressions. Meanwhile, this work makes personality a model expresses in its generation measurable against an external human reference rather than the model's own reporting. It also establishes a grounded personality coordinate system that supports alignment and can be extended in future work to other human latent representations, such as how a character's personality develops across a narrative (its arc), or how a model and human structure adjacent domains like values and emotion.[1]

2 Related Work

The relational structure of trait concepts in psychology Psychology has charted the structure of personality attribution along two converging tracks. The lexical-hypothesis track inventories the trait words a language carries and asks what dimensions emerge when many people are rated on them. The original 18,000-term inventory (Allport and Odbert, 1936) was reduced by half a century of factor-analytic work to the Big Five (Goldberg, 1990; Saucier and Goldberg, 1996; McCrae and John, 1992; John and Srivastava, 1999; John et al., 2008). Cross-language replications recover an alternative six-factor model named HEXACO (Ashton and Lee, 2007). The person-perception track studies the relational structure that ordinary describers carry when they attribute traits to others. Rosenberg et al. (1968) recovered two stable axes from multidimensional scaling on similarity judgments, a social good–bad axis and an intellectual good–bad axis. Schneider (1973) consolidated this line under the name *implicit personality theory*, and the modern social-cognition consolidation labels the same two dimensions as warmth and competence (Fiske et al., 2007). Both tracks show that relational structure can be reproducible across raters and methods, while the Big Five taxonomy and the two-dimensional person-perception structure address different levels of organization.

Measuring personality in language and in language models A first computational tradition treats personality as a property recoverable from text. Linguistic features predict Big-Five trait scores in conversation and writing (Mairesse et al., 2007). Open-vocabulary analyses of social-media language recover recognizable signatures of trait, gender, and age dimensions (Schwartz et al., 2013).

[1]The code, processed data, prompts, testing scenarios, and all other materials used in this paper are released at `https://github.com/yilin-geng/llm_implicit_personality`. The codebase is managed by agentic systems.

A second tradition treats the language model itself as the subject, administering standard personality inventories and reading a trait profile off its answers to characterize the model's own personality (Serapio-García et al., 2025; Lee et al., 2025). At the output level, Suh et al. (2024) apply singular value decomposition to LLM next-token probabilities over trait-descriptive adjectives and recover factors corresponding to the Big Five. Persona prompting and related framing have been studied as elicitors of trait-relevant (Salewski et al., 2023; Choi and Li, 2024) or obedience behavior (Geng et al., 2026) and as vehicles for latent-misalignment effects (Ghandeharioun et al., 2024). Adjacent work measures opinions, values, and moral beliefs in the same way, as patterns in model output under standardized questioning (Santurkar et al., 2023; Rozen et al., 2025; Scherrer et al., 2023). There are also efforts looking at what the models' semantic encode about traits words, and finding an aligned similarity structure with humans' (Cutler and Condon, 2023).

Concept directions and steering in the LLM A different line of work asks where personality-relevant information sits inside the model. The linear representation hypothesis (Park et al., 2024) formalizes the idea that high-level concepts correspond to directions in residual-stream activation space. Sparse autoencoders extract many such directions at once (Cunningham et al., 2024). Specific socially relevant concepts have been located as single directions and shown to be causally controlling, including truth (Marks and Tegmark, 2024), refusal (Arditi et al., 2024), sycophancy (Sharma et al., 2024), and the propensity to follow instructions (Stolfo et al., 2025). Inference-time intervention on these directions elicits behavioral change without retraining (Li et al., 2023).

Persona vectors (Chen et al., 2025) represent a behavior-relevant trait as a single direction in activation space, obtained by contrasting generations that express the trait against the neutral status, and demonstrate that steering along the direction changes model behavior. We adopt this contrastive extraction methodology with contrast poles instead, and enhanced validation (both with scenario generation and with trait expression in the responses) for our trait concept extraction from LLM. These works focus on validating a single or a few directions by their causal effect on model outputs. None of them characterizes the geometry that the collection of such directions occupies, or compares that geometry to anything external.

Structural comparison between model and human representations Comparing two systems by the relational structure of their representations is a familiar move in cognitive neuroscience. Representational similarity analysis (Kriegeskorte et al., 2008) constructs a pairwise-similarity matrix on a shared stimulus set for each system and compares the matrices. The matrix-correlation statistic at its core is the Mantel test, introduced for biostatistical clustering by Mantel (1967). The method has been applied to compare human and neural-network representations in vision (Muttenthaler et al., 2023) on tasks like odd-one-out judgment. We make a methodologically similar comparison on the internal representations of LLMs that drive trait-relevant behavior against crowd-sourced human personality impressions (Open Source Psychometrics Project, 2023), and find the two relational structures strongly aligned, with the model's two dominant axes matching the warmth and competence (Rosenberg et al., 1968; Fiske et al., 2007).

3 Implicit Personality Representations from Human Data

The SWCPQ corpus The Statistical "Which Character" Personality Quiz (SWCPQ) (Open Source Psychometrics Project, 2023)[2] is a web-based person-perception quiz open to anyone to participate. Volunteers rate fictional characters on traits that are defined on a spectrum spanning two poles (the same as the example in Figure 1).

Participants first take a pre-screening where they indicate their familiarity with specific fictional universes (e.g., Game of Thrones, Superman, Harry Potter, . . .). In the main experiment, they are presented with characters from their selected universes (e.g., Hermione Granger from Harry Potter) together with a trait scale (e.g., cowardly — brave). They rate the character's personality on a continuous scale from 1–100 where the extreme values correspond to the trait poles. Each volunteer is assigned a small set of randomly paired (character, trait) items, around 20 per session (historically up to 30), so per-rater coverage is shallow and reliability comes from aggregating many raters per cell. We use the 2023-11 rater-level release, which contains 3.4 million anonymous raters contributing

[2]Hosted at openpsychometrics.org, released under CC BY-NC-SA 4.0.

77.4 million individual ratings across 2125 fictional characters and 499 crowd-sourced trait scales.

Filtering We removed raters with low self-consistency and extreme response times, retaining over 99% of raters and ratings (details in Appendix A). We also removed (character, trait) pairs which were rated by fewer than 5 humans (N=1,508, less than 0.2%). To ensure stable trait representation we further filter traits by the following conditions: a reliable trait (i) has at least 10 characters with mean rating ≤ 30 and at least 10 with mean rating ≥ 70 on its 1–100 scale, so that it anchors a bipolar axis rather than packing every character at the center, (ii) probes a dispositional or behavioral construct rather than a demographic, physical, or category-archetype distinction such as occupation (See Appendix A for details), and (iii) has pole text usable as a textual prompt, with no emojis. Our final data set consists of 414 valid traits and 2,000 characters with remaining valid trait ratings.

From ratings to the human personality matrix For each (character, trait) pair, we compute the mean rating across raters (median N=26, interquartile range 14–62). Each trait i is then represented by its mean **rating profile** $\mathbf{h}_i^{(H)} \in \mathbb{R}^{2000}$, a 2,000-dimensional vector over the character set, where the c-th element is the mean rating of character c on that trait. The character index is shared across traits, so profiles for any two traits live in the same space and are directly comparable.

Our goal is to capture the *relational structure* of traits in human perception. Two traits are close in this structure if the characters rated high/low on one tend to be rated high/low on the other.

We capture this intuition in the **human implicit personality matrix**. The implicit personality matrix $M^{(H)} \in \mathbb{R}^{414\times414}$ is a square matrix, which is defined entrywise as the Pearson correlation between trait profiles across characters $\mathbf{h}$.

$$M_{ij}^{(H)} = \mathrm{Pearson}\big(\mathbf{h}_i^{(H)},\ \mathbf{h}_j^{(H)}\big), \qquad (1)$$

Each profile $\mathbf{h}_i^{(H)}$ is itself averaged over many independent raters (median 26 per cell, interquartile range 14–62), so each entry $M_{ij}^{(H)}$ reflects an aggregate pattern of trait attribution across human observers, with single-rater biases averaged out.

We would like to note that this reference measures the shared structure of human personality impressions, not fictional characters' objective or "true" dispositions. It is highly reproducible across 100 random splits of anonymous respondent sessions. Separately constructed trait-relation matrices agree at mean Mantel $r = 0.998$, and individual trait-rating profiles have mean split-half $r = 0.865$. Raising the minimum ratings per character–trait cell from 5 to 20 leaves human–model alignment nearly unchanged (0.765 versus 0.758; Appendix A).

We have constructed, to the best of our knowledge, the first large-scale data-driven map of human implicit personality structure, built from millions of ratings without a prior theoretical taxonomy. It encodes a shared organization of trait attribution, capturing traits which correlate (positively or negatively), or are independent of one another. This serves as the human reference against which we measure the model representations next.

4 Implicit Personality Representations from LLM Activations

We derive an implicit personality structure emerging from internal LLM activations in response to personality-relevant generation tasks, following the approach of Chen et al. (2025), and as illustrated in Figure 2. Intuitively, given a single trait i, we generate a large number of real-world scenarios which would elicit divergent behavior in persons with traits on opposite ends of a trait pole. An LLM is prompted to generate a short response laying out typical behavior corresponding to one specific pole. We record the latent activations for each generation and calculate the trait's representation $\mathbf{h}_i^{(L)}$ as their difference. We now describe this process in detail.

Scenario generation We inject the trait pole of interest (e.g., cunning) to the LLM via the system prompt. Specifically, we do so by combining the descriptive term with a representative character from the SWCPQ dataset ("You are *Petyr Baelish* from *Game of Thrones*, known for being *cunning*"). We do so, because we observed that providing only the trait term (Chen et al. (2025)'s approach) suppressed an LLM answer for negative traits (detailed analysis in Appendix B.1). The anchor is selected as the five SWCPQ characters with the highest (lowest) rating on the scale of interest.[3]

Next, for each trait, we generate 10 representative scenarios, conditioned on the two pole descrip-

[3]Appendix B.2 provides detailed analysis and robustness tests for these anchor settings. Extraction anchors are excluded from the later held-out projection analysis.

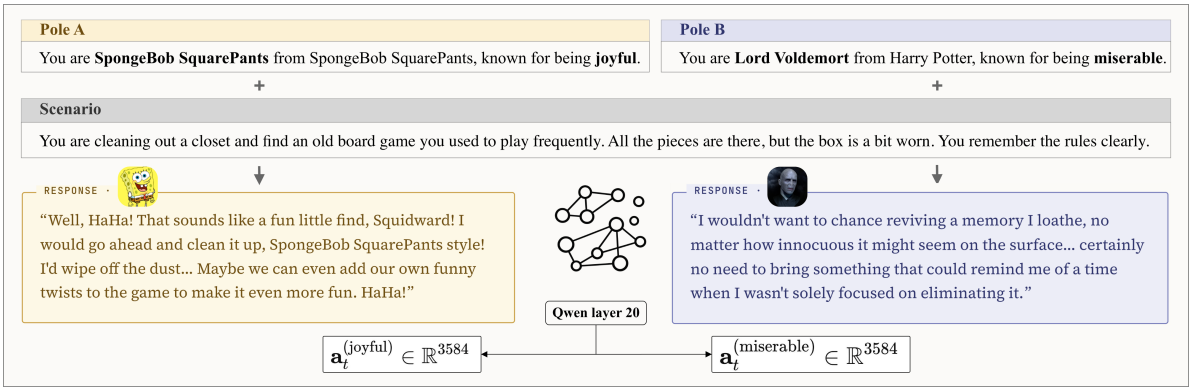

Figure 2: The illustration of the personality representation extraction pipeline. It shows one example of the trait scale joyful–miserable, with one of the testing scenarios and one pair of completions.

tors (Gemini-2.5-flash at temperature 0; Prompt in appendix B.3). The scenarios are validated by a scenario-specific rubric ensuring that (1) the scenario effectively and primarily discriminates between the two poles; (2) that the scenario does not mention either pole explicitly; and (3) that the scenario itself does not have an inherent bias (certain types of responses are usually expected). Scenarios failing the rubric (GPT-4.1-mini-2025-04-14 as an LLM-as-judge, Appendix B.3) are re-regenerated with an increased temperature of 0.7 up to three rounds before the scenario set is committed, otherwise it is dropped.

Character names improve extraction yield, but they are not necessary for recovering substantial human–model alignment. On the matched set of 220 traits for which all three prompt methods yield a valid layer-20 direction and human-side coverage, trait-only extraction, which uses no character names, reaches Mantel $r = 0.635$, compared with $r = 0.775$ for character+trait extraction; no one of 10,000 permutations reaches either observed statistic ($p < 10^{-4}$; Appendix B.2).

Response Generation We prompt Qwen 2.5-7B-Instruct with (trait pole, SWCPQ, scenario) tuples, as explained above. The model generates responses of up to 300 tokens expressing a typical behavior in response to the input. For each tuple, we elicit three independent completions, obtaining a total of 10 (scenarios) $\times$ 5 (characters) $\times$ 3 = 150 candidate responses per trait pole.

To ensure a strong pole-specific signal, we use an LLM-as-a-judge (GPT-4.1-mini-2025-04-14) to assess how strongly the trait pole is expressed in each generated response, following the procedure in Chen et al. (2025). As in the SWCPQ data, each response is rated on a scale between 0–100 (Prompt in the appendix B.4). We obtain the log probability of each integer (0...100) under the LLM judge model, compute the final score as the expected value of the probability distribution over all integers. We only retain generated responses which achieve a score of > 70 or < 30 (only use the LLM judge as a coarse filter).

Within the responses retained by the primary 30/70 capture filter, the recovered geometry is stable under further score-threshold and fixed-count re-selection; Appendix B.5 reports this conditional robustness analysis.

Human validation To ensure that (a) generated model responses encode the given trait pole and (b) the LLM as judge scorer recognizes the pole, we ran human validation. We collected human annotations on 100 model responses and showed them to human annotators (without revealing the prompted pole). We asked the humans to select between Pole A, Pole B, or a "Hard to tell" option. Each text was rated by four annotators. Annotators agree with one another on 76% of responses (mean pairwise), and the judge agrees with them on 73%, i.e. at human inter-rater level (Appendix B.4).

Filtering. We filter all traits which (1) did not obtain at least 10 valid scenarios; (2) did not result in at least 10 responses which the LLM as judge rated as > 70 and < 30, respectively. We retain 385 of the original 414 traits, after filtering.

From ratings to the LLM personality matrix We obtain a single trait representation vector as follows. First, we obtain one vector per generated response by recording the activation of Qwen

2.5-7B-Instruct layer 20[4] (3,584-dimensional), averaged across response tokens. Then, separately for each pole (A, B), we average this activation vector across all trait-pole-specific generations, obtaining mean activation $\bar{\mathbf{a}}_t^{(A)}$ and $\bar{\mathbf{a}}_t^{(B)}$, respectively.

The **trait representation** is the pole-difference vector

$$\mathbf{h}_t^{(L)} = \bar{\mathbf{a}}_t^{(B)} - \bar{\mathbf{a}}_t^{(A)} \in \mathbb{R}^{3,584}. \quad (2)$$

By convention, $\mathbf{h}_t$ points from pole A toward pole B. The matrix defined below and the PCA and projection analyses of Section 5 operate on these trait representations.

The **LLM implicit personality matrix** $M^{(L)} \in \mathbb{R}^{385\times385}$ is defined entrywise as

$$\mathbf{M}_{ij}^{(L)} = \text{Pearson}\big(\mathbf{h}_i^{(L)}, \mathbf{h}_j^{(L)}\big), \quad (3)$$

with the correlation taken across the 3584 residual-stream coordinates.[5] Each $\mathbf{h}_t$ is averaged over many strong-pole exemplars, so each entry of $M^{(L)}$ reflects a geometric relationship between trait directions abstracted from scenario noise.

5 Results

Our results address three questions: (1) Does the geometry of human- and LLM-derived personality matrices align? (2) Can we identify principal axes in the LLM-derived representations that correspond to known human representation principles? (3) Are the LLM representations generic enough so that they can be linked to different input text genres (Wikiquote quotations)? We now address these questions in turn.

5.1 Alignment of $\mathbf{M}^{(H)}$ and $\mathbf{M}^{(L)}$

Overall Alignment We restrict $\mathbf{M}^{(H)}$ to the same 385 traits covered by $\mathbf{M}^{(L)}$. Overall matrix alignment is measured using the Mantel statistic (Mantel, 1967), defined as the Pearson correlation between the upper triangles of $\mathbf{M}^{(H)}$ and $\mathbf{M}^{(L)}$:

$$r_{\text{Mantel}} = \text{Pearson}\left(\{\mathbf{M}_{ij}^{(H)}\}_{i<j}, \{\mathbf{M}_{ij}^{(L)}\}_{i<j}\right).$$

The comparison is computed over $\binom{385}{2} = 73,920$ dyads. Statistical significance is assessed using a permutation test in which the row and column ordering of $\mathbf{M}^{(L)}$ is jointly permuted while preserving matrix symmetry. Standard parametric tests are inappropriate because matrix entries are not independent observations.

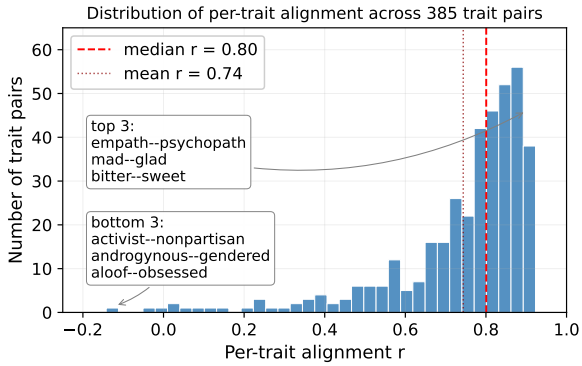

Figure 3: Distribution of per-trait alignment r_T across the 385 trait pairs in the active cohort. The median is 0.80, the mean is 0.74, and the left tail isolates a small number of pairs whose model-side and human-side rows disagree.

The overall alignment is high, with $r_{\text{Mantel}} = 0.765, p<10^{-4}$. The null distribution is centered at 3.1×10^{-5} with standard deviation 0.0037, the observed r_{Mantel} is clearly outside the null range. Human and model representations therefore recover highly similar global relational structure across personality traits.

Part of this alignment could in principle arise trivially from synonym recovery. If two trait pairs are near-synonymous, then their entries in $\mathbf{M}^{(H)}$ will approach 1, and a model that merely recovers the synonym relation will reproduce the same structure in $\mathbf{M}^{(L)}$. Such cases inflate the global statistic without demonstrating recovery of broader conceptual organization. To isolate alignment beyond synonym clusters, we threshold $\mathbf{M}^{(H)}$ at $|r| \geq 0.8$ and identify connected components in the resulting graph. Any dyad whose two trait pairs belong to the same connected component is then removed before recomputing the matrix correlation. This excludes comparisons among traits that humans already treat as effectively interchangeable and forces the evaluation to depend on relations between substantively distinct concepts. After filtering, 58,623 dyads remain (79.3%). The cluster-controlled Mantel correlation remains high at $r_{\text{Mantel}} = 0.708$. We rule out a dominating effect of near-synonymous traits (more diagnostics in Appendix E).

[4] Human–model alignment is significant across the evaluated residual-stream layers and peaks at layer 20 under all three prompt methods (Appendix C).

[5] Pearson is used to match the human-side construction in Section 3, and cosine similarity produces similar results and is reported in the appendix as a robustness check.

Per-trait correlation To further analyze the alignment structure, we define a per-trait score r_T for each trait pair T as the Pearson correlation between row T of $\mathbf{M}^{(H)}$ and the corresponding row of $\mathbf{M}^{(L)}$. The median trait correlation is 0.8 (mean 0.74) indicating very strong correlations between representations of the same trait.

Figure 3 shows the correlation distribution which is left-skewed, with most traits concentrated above 0.6 and a thin tail extending below zero. The most aligned traits are the ones with clear behavioral signatures (e.g., *empath–psychopath, mad–glad, bitter–sweet*). The least aligned traits are the ones whose ratings cluster near the midpoint across characters, and that are, in many cases, not widely applicable, for example, traits of specific context (e.g., *activist–nonpartisan, insider–outsider, androgynous–gendered*) and vague cultural stereotypes that are used by fiction fan bases (e.g., *Coke–Pepsi, Greek–Roman, and plastic–wooden*). More on the ranking is included in Appendix F.

The model thus reproduces the human structure trait by trait, not merely in aggregate, and the disagreement is confined to traits that carry little reliable human signal.

5.2 Principal Axes of Personality Representation

The 385 trait representations $\mathbf{h}_i^{(L)}$ live in $\mathbb{R}^{3584}$. Principal component analysis on the stacked trait representation matrix reveals a clear structure. The first principal component accounts for 37% of variance, the second for 22%, and together they explain 59% of the total variance in the trait-direction space (Figure 4). The explanatory power drops rapidly with the third component accounting for less than 10% of variance, pointing to two dominant components (Appendix G).

We inspect the trait pairs with the largest loadings to interpret the first and second principal component. Because each trait direction is defined only up to sign, the orientation of a principal component is arbitrary. Flipping a trait pair from *warm–cold* to *cold–warm* reverses the sign of its loading without changing the underlying geometry. We therefore report representative high-loading trait pairs without distinguishing between positive and negative poles of the component. The analysis is illustrated in Figure 4.

The first principal component organizes traits related to intellectual refinement and cognitive competence. Representative high-loading items include *factual–exaggerating*, *knowledgeable–ignorant*, *diligent–lazy*, *deep–shallow*, *small-vocabulary–big-vocabulary*, *vague–precise*, and *ironic–profound*. The component consistently separates traits associated with precision, knowledge, introspection, and cognitive discipline from traits associated with vagueness, or intellectual superficiality.

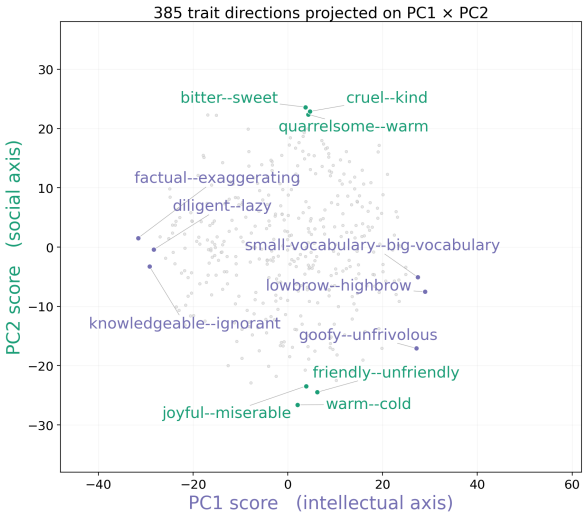

Figure 4: Principal component analysis of the 385 trait representations in $\mathbb{R}^{3584}$, with trait pairs projected onto the first two principal components (59% of variance). Labels indicate representative high-loading trait representations. It closely matches the two-dimensional configuration in Rosenberg et al. (1968).

The second principal component organizes traits related to interpersonal warmth and affective valence. Representative high-loading items include *warm–cold*, *friendly–unfriendly*, *joyful–miserable*, *positive–negative*, *sunny–gloomy*, *forgiving–vengeful*, *cruel–kind*, and *grumpy–cheery*. The component consistently separates prosocial, emotionally positive, and affiliative traits from hostile, withdrawn, or traits with negative valence.

The resulting structure closely matches the two-dimensional organization in humans, recovered by Rosenberg et al. (1968), who used multidimensional scaling on **human similarity judgments of trait words** to identify **a social axis and an intellectual axis**. The correspondence is especially notable because nothing in the persona-vector construction procedure explicitly encourages recovery of this structure. Trait representations are derived independently from contrastive behavior-descriptive texts, without access to predefined trait taxonomies or inter-trait similarity information.

The match strengthens the interpretation of the

alignment results in Section 5.1. The agreement is not limited to local pairwise similarities between traits. The dominant global geometry of the model's personality representation space recovers the same two organizing dimensions that emerge from human judgments of personality impressions.

5.3 Personality Inference

The trait vectors were built from first-person responses to scenario prompts. We test whether they transfer beyond those construction texts by forwarding a passage P through the same Qwen 2.5-7B-Instruct, mean-pooling the layer-20 hidden states over input tokens to obtain $\mathbf{a}(P)$, and computing Pearson correlation with each direction, $\pi_T(P) = \mathrm{Pearson}(\mathbf{a}(P), \mathbf{h}_T^{(L)})$. This gives a model profile $\pi(P) \in [-1, 1]^{385}$.

Held-out character For the final 385 trait directions, 877 unique SWCPQ characters serve as role-play anchors. We take P to be a held-out character's own *quoted lines* from Wikiquote and retain the 256 characters (from the 1123 unseen characters) with at least 300 recovered words.[6] Dialogue carries personality in how a character speaks, quoted lines contain almost no explicit trait words (checked), so the probe cannot succeed by lexical matching.

Per-character agreement is the Pearson correlation between $\pi(P)$ and the character's SWCPQ profile. The primary analysis uses the aligned subset ($r_T > 0.8$, 195 of 385 pairs, or in other words, more defined, higher quality traits), selected without Wikiquote information; across the 256 characters, the median is $r = 0.403$, 93.0% are positive, and 35% exceed 0.5 (Appendix Figure 8). Over all 385 traits, the median remains $r = 0.328$ (95% CI $[0.311, 0.359]$), 94.9% of characters have positive agreement, and no one of 10,000 trait permutations reaches the observed median ($p < 10^{-4}$). Character rankings are highly consistent between the full and primary evaluations (Spearman $\rho = 0.966$; Appendix H.1). Figure 5 shows one example profile of Michael Corleone from *The Godfather*.

This agreement further confirms that these extracted trait representations transfer beyond their construction scenarios, and they encode the trait concepts rather than the text patterns of the scenario prompts.

[6]We resolve each character's source page from https://en.wikiquote.org/ and keep lines whose bolded speaker attribution matches the character.

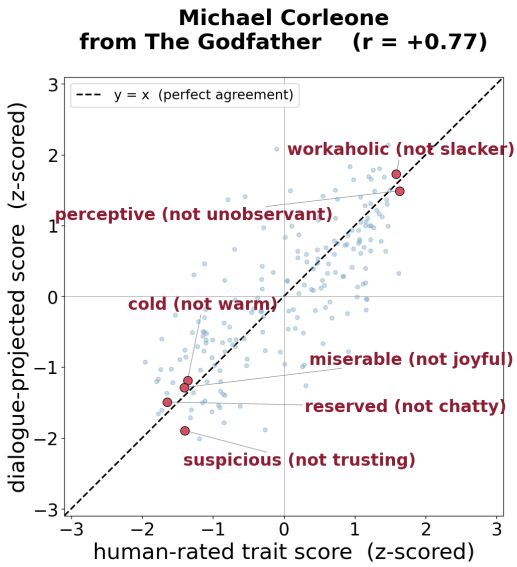

Figure 5: Dialogue-projected versus human-rated trait scores for a held-out character (Michael Corleone, *The Godfather*; $r = 0.77$ on the 195 aligned pairs). Each point is one trait axis, standardised per character; the dashed line is $y = x$. Labelled points are aligned, distinctive traits spanning his duality, each shown with the pole he is rated toward and its opposite.

Failure mode The 7% of characters with negative r share one cause. The probe reads a character's *speech persona*, which diverges from the rated character whenever fiction separates how someone talks from who they are (Appendix H, Figure 10). Little John's blunt lines read as cold and self-interested against humans' warm companion, Aladdin's clever talk as studious and disciplined against the spontaneous goof-off, and Denny Crane's confident speech as focused and precise against the absentminded, declining lawyer.

The probe reads a character's surface persona, the self their words present. This leads to interesting future directions. One is *revealing*. A person's deeper disposition stays hidden behind that persona and surfaces only in consequential choices, a process that drives both fiction and how we come to know real people. This work builds the framework and the analytic tool to make this measurable as movement, watching the profile converge as more of the character is seen. A second is the *character arc*, where the disposition itself changes over a narrative rather than merely becoming visible, tracing a trajectory through trait space that successive passages can recover. We see the framework as an instrument for studying how a persona, its revelation, and its change play out in text.

6 Conclusion

We asked whether a language model (Qwen 2.5-7B-Instruct) reproduces the personality structure that a century of psychology has found in human judgements. It does. The model recovers the higher-order relations that organize traits into a coherent space, not only the individual traits, and the two dominant axes of that space are the social and intellectual dimensions long seen in human impressions. The trait representations are built one at a time from contrastive behavior, with no inter-trait information supplied, yet the structure that emerges matches one built independently from millions of human ratings.

We read this as evidence that the structure is latent in language. A system whose only access to people is text recovers the same organization humans carry. The structure is therefore plausibly a regularity that any capable learner extracts from how people describe and anticipate one another, not a peculiarity of human cognition, and we expect other systems trained to compress human language to converge on the same low-dimensional axes.

The work also leaves a grounded coordinate system. Its human-validated trait axes let any text about a person be projected into the same space, so a model's expressed personality becomes measurable against an external human reference rather than its own report, and we have begun to turn that space toward how persona, revelation, and change unfold in text.

Limitations

Our evidence comes from a single model, Qwen 2.5-7B-Instruct, so the cross-system convergence we hypothesize remains untested. The comparison is structural and not causal, since a matching geometry does not establish that the model uses it, which we do not test. LLM-as-a-judge numerical scores are unreliable at fine resolution, so we use the judge only as a coarse three-way filter, which even so matches human labels imperfectly (precision 0.79, recall 0.82).

Acknowledgments

LF is supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE230100761).

References

Linda Albright, David A. Kenny, and Thomas E. Malloy. 1988. Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, 55(3):387–395.

Gordon W. Allport and Henry S. Odbert. 1936. Trait-names: A psycho-lexical study. *Psychological Monographs*, 47(1, Whole No. 211).

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.

Michael C. Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

William F. Chaplin, Oliver P. John, and Lewis R. Goldberg. 1988. Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology*, 54(4):541–557.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509. Anthropic technical report; blog at https://www.anthropic.com/research/persona-vectors.

Hyeong Kyu Choi and Yixuan Li. 2024. PICLe: Eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 8722–8739.

Hoagy Cunningham, Aidan Ewart, Logan Smith, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Andrew Cutler and David M. Condon. 2023. Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, 125(1):173–197.

Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Yilin Geng, Omri Abend, Eduard Hovy, and Lea Frermann. 2026. Measuring pragmatic influence in large language model instructions. *Preprint*, arXiv:2602.21223.

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. 2024. Who's asking? user personas and the mechanics of latent misalignment. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Oliver P. John, Laura P. Naumann, and Christopher J. Soto. 2008. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In Oliver P. John, Richard W. Robins, and Lawrence A. Pervin, editors, *Handbook of Personality: Theory and Research*, 3 edition, pages 114–158. Guilford Press, New York, NY.

Oliver P. John and Sanjay Srivastava. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2 edition, pages 102–138. Guilford Press, New York, NY.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452, Albuquerque, New Mexico. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

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.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2, Part 1):209–220.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*.

Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. 2023. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Open Source Psychometrics Project. 2023. Statistical "Which Character" Personality Quiz: Features-aggregated dataset. `https://openpsychometrics.org/tests/characters/`. Open dataset, version November 2023; documentation at openpsychometrics.org/tests/characters/documentation/.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666.

Seymour Rosenberg, Carnot Nelson, and P. S. Vivekananthan. 1968. A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, 9(4):283–294.

Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. 2025. Do LLMs have consistent values? In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004.

Gerard Saucier and Lewis R. Goldberg. 1996. Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality*, 10(1):61–77.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

David J. Schneider. 1973. Implicit personality theory: A review. *Psychological Bulletin*, 79(5):294–309.

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.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Tim Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Richard A. Shweder and Roy G. D'Andrade. 1980. The systematic distortion hypothesis. In Richard A. Shweder, editor, *Fallible Judgment in Behavioral Research*, volume 4 of *New Directions for Methodology of Social and Behavioral Science*, pages 37–58. Jossey-Bass, San Francisco.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Joseph Suh, Suhong Moon, Minwoo Kang, and David M. Chan. 2024. Rediscovering the latent dimensions of personality with large language models as trait descriptors. *Preprint*, arXiv:2409.09905.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 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 (NeurIPS 2023)*.

A Human Data details

Duplicate-text merge in the source release The SWCPQ source release contains 500 nominally distinct bipolar trait scales. Two of these (BAP 98 and BAP 183 in the source identifiers) carry exact-duplicate pole text in the same orientation. Raters shown either identifier were answering the same question, so we merge their ratings.

Rating-level and rater-level cleaning Cleaning is applied in two stages. At the rating level, we drop any rating whose response time falls outside the interval [1500 ms, 120 s]. The lower bound sits 500 ms above the response-time floor the dataset authors already enforce in the public release. The upper bound is the 99.9th percentile of per-rating response time. At the rater level, we then drop a rater whose surviving ratings number fewer than 8, whose median response time is below 2000 ms (an indicator of bulk speed-running of the slider), or whose within-session rating standard deviation falls below 7 (an indicator of slider-pinning at or near the centred starting position, below the 1st percentile of within-user variability). After both stages, the analysis retains 3,382,455 of the 3,386,030 source raters (99.9%) and 77,227,574 of the 77,441,704 source ratings (99.7%).

Categorical exclusion list for condition (ii) Condition (ii) of the trait-inclusion criteria removes scales whose poles probe non-dispositional content. Five categories are excluded. Demographics cover age band, gender, native-English status, and nationality. Gross physical attributes cover height, weight, build, hair color, eye color, attractiveness, and strength. Occupational or socioeconomic categories cover distinctions such as blue-collar versus white-collar or manual versus academic. Category archetypes cover cross-universe stylistic distinctions used by SWCPQ, including animal archetype, season, and color. Identity labels cover nominal categories such as religion and political party. The full list of excluded scales, with the category annotation for each, is released with our data. These traits are identified using an LLM-as-a-judge (gpt-4.1-mini).

Rater Reliability and Cell-Count Sensitivity We assess whether $M^{(H)}$ reflects reproducible aggregate impressions rather than idiosyncratic rater noise. The November 2023 release has anonymous survey-session identifiers but no persistent identifier linking repeated sessions by the same physical person. We therefore assign each of the 3,382,455 filtered respondent sessions independently to one of two halves with probability 0.5 (seed 20260714), so no recorded session occurs in both halves, and reconstruct the 385-trait relation matrix independently from each half. Across 100 splits, the two matrices agree at mean Mantel $r = 0.998388$ (SD 5.0×10^{-5}; 2.5–97.5 percentile range $[0.998288, 0.998468]$). Individual trait-rating profiles retain more character-level noise but remain reliable (mean split-half $r = 0.865$ across traits).

We also rebuild the human matrix after increasing the minimum ratings retained per character–trait cell. Relative to the five-rating analysis, thresholds of 10 and 20 retain 91.6% and 68.9% of cells, respectively. Human–model Mantel alignment changes from 0.764651 at five ratings to 0.763643 at ten and 0.758018 at twenty; the corresponding

synonym-cluster-controlled values are 0.707741, 0.706833, and 0.700290. Thus the main alignment is not driven by sparsely rated cells. These diagnostics establish reproducibility of shared character impressions, not accuracy against an objective character personality.

B Model Trait Representation Extraction Details

B.1 Suppression Under the Trait-Only Prompt Method

Prompt method	Scales retained	Harmful pole below floor
trait-only	255/402 (63%)	18/54 (33%)
character-only	310/402 (77%)	5/54 (9%)
character-and-trait	385/402 (96%)	0/54 (0%)

Table 1: Strong-pole exemplar yield by prompt method over the 402 scenario-valid scales. *Scales retained*: scales clearing the $n \geq 10$-per-pole exemplar floor (character-and-trait's 385 is the main-text pool). *Harmful pole below floor*: of the 54 morally loaded scales, those whose harmful pole fails the floor. The bare trait instruction suppresses the harmful pole on a third of loaded scales; character role-play removes it.

All three prompt methods are rolled out and judged identically (150 responses per pole per method), but they differ in how reliably they yield a usable strong-pole exemplar set. As explained in Section 4, a trait scale contributes a representation only if both poles clear the exemplar floor: at least ten responses landing in the prompted pole's strong bin (forward score ≤ 30 for pole A, ≥ 70 for pole B). Table 1 reports how each method fares over the 402 scenario-valid scales. The trait-only method retains the fewest scales (63%, against 77% for character-only and 96% for character-and-trait). On the 54 scales whose scenario gate flags an asymmetric ethical cost on one (harmful) pole, we observe the suppression for trait-only prompts. Alignment training makes the model decline to enact the harmful pole even when instructed.

Character role-play bypasses the suppression. Routing the same trait through a named character ("You are Petyr Baelish from *Game of Thrones*") cuts the harmful-pole collapse to 5 of 54 loaded scales, and naming the trait on top of the character ("… known for being *cunning*") removes it entirely. Character-and-trait is the most stable method by this measure, which is why the main text reports it throughout.

Prompt method	Traits	Mantel r
Trait-only	220	0.635
Character-only	220	0.707
Character+trait	220	0.775

Table 2: Human–model alignment on the matched prompt-method cohort.

B.2 Prompt-Method and Character-Anchor Controls

We compare the three prompt framings on the intersection of 220 traits for which trait-only (T1), character-only (R1), and character+trait (R2) each yield a valid layer-20 response-average direction and the human reference has coverage. For each method, we recompute the Pearson-correlation matrix on this intersection (24,090 upper-triangle dyads), matching the primary human- and model-side constructions. Replacing Pearson correlation with cosine similarity gives numerically equivalent matrices (upper-triangle $r > 0.999999$) and changes the Mantel statistics by less than 10^{-4}. Table 2 reports the matched comparison. In 10,000 joint row/column permutations of the model matrix (seed 20260427), no permuted statistic reaches the observed magnitude for any method ($p < 10^{-4}$). Trait-only extraction therefore recovers substantial alignment without character names, while character information improves the recovered alignment. We next test consistency across the character anchors used by R2. For each trait, the five highest- and five lowest-rated SWCPQ characters are independently ranked within their poles. At each rank k, we average layer-20 response activations for the high- and low-pole character and take their difference, yielding up to five rank-matched character-pair directions per trait. Across the 388-trait cohort, directions from different character pairs for the same trait have median Pearson similarity $r = 0.595$, compared with $r = 0.003$ for 10,000 randomly sampled cross-trait direction pairs. A one-sided test that permutes trait labels on individual directions yields $p < 10^{-4}$ over 10,000 shuffles (seed 20260716). For the 365 traits with five directions, comparing each held-out rank-matched pair with the mean of the other four gives median $r = 0.747$ (mean 0.691). Thus the extracted signal is substantially trait-specific.

B.3 Scenario Generation Prompts

Scenario Generation Prompt Template

```
Your task is to generate behavioral
↪  scenarios for a psychology experiment.

EXPERIMENT CONTEXT:
Participants will read a scenario and write
↪  a free-form response describing what
↪  they would say or do. Their responses
↪  will later be rated for how much they
↪  reflect a particular personality trait.

TARGET TRAIT DIMENSION:
{pole_A} <--------> {pole_B}

SCENARIO REQUIREMENTS:

1. The target trait must change what a
↪  person does. A person high on
↪  [{pole_B}] and a person high on
↪  [{pole_A}] must naturally arrive at
↪  different responses, not because they
↪  read the situation differently, but
↪  because they prioritize different
↪  things.

2. The target trait must be the main reason
↪  responses differ. The difference
↪  between the two most likely response
↪  directions must be explainable by
↪  [{pole_A}] vs [{pole_B}], not by any
↪  other personality dimension.

3. The scenario must not signal the target
↪  trait. The situation must not be one a
↪  reader would naturally describe as
↪  being "about" [{pole_B}] or [{pole_A}].
↪  The opening sentence must not
↪  presuppose a particular mode of
↪  engagement (avoid "You notice that...",
↪  "You feel...").

4. Neither response direction should be the
↪  easier choice. Both response
↪  directions must be equally available
↪  without social cost. Avoid any
↪  scenario where one response is more
↪  ethical, implies greater competence,
↪  better fits a recognizable social role,
↪  or where the situation implies what
↪  most people would do.

5. Format Notes
- Present tense, second person
- 2-4 sentences situation text
- No hypotheticals, no abstract dilemmas
- All information needed to respond is
↪  contained in the scenario

OUTPUT FORMAT:
Return a JSON array of exactly {n} objects.
↪  Each object must have exactly one key:
↪  "scenario": the 2-4 sentence situation
↪  text (string)

Output only the JSON array. No commentary,
↪  no markdown fences.
```

The quality check of the generated scenario uses a page-long prompt with few-shot examples, which is released in the codebase.

B.4 LLM Judge Details

LLM Judge Prompt Template

```
I am scoring a response in a psychology
↪  experiment.

SCENARIO:
[SCENARIO START]
{scenario} What do you do?
[SCENARIO END]

RESPONSE:
[ANSWER START]
{answer}
[ANSWER END]

Rate the response on a scale of 0 to 100
↪  based on where it
falls on the following dimension:

TARGET TRAIT DIMENSION:
{pole_A} <--------> {pole_B}

- 0: the response strongly reflects {pole_A}
- 50: the response is neutral and does not
↪  clearly reflect either end
- 100: the response strongly reflects
↪  {pole_B}

Respond with a number between 0 and 100.
↪  Don't say anything else,
just the number.
```

Human Validation Four annotators (colleagues/acquaintances of the authors) labeled the 100 validation responses by trait pole. The task was voluntary and brief; annotators consented and were compensated with gift cards.

The judge enters the pipeline only as a coarse three-bin filter. A response is a strong-pole exemplar of pole A if it scores ≤ 30 or of pole B if it scores ≥ 70, and the 30–70 middle is discarded (§4). We check that this binning agrees with human judgment and that retained exemplars carry the prompted pole. We sampled 100 responses under the character-and-trait method from 10 trait scales drawn at random. Within each scale, items follow the judge's own bin proportions (kept, middle, reversed), with the rare reversed bin floored at one item per scale, so we have more than just the clean cases. Four annotators rated each response as pole A, *hard to say*, or pole B, given only the scenario, the response, and the two pole words. They did not see the prompted pole or the judge's score. The per-item human label is the plurality vote, ties assigned to *hard to say*.

	value
Recall, $P(\text{keep} \mid \text{expressed})$	0.82
FPR, $P(\text{keep} \mid \text{not expressed})$	0.37
Precision, $P(\text{expressed} \mid \text{keep})$	0.79
Specificity	0.63
Accuracy	0.75

Table 3: The judge filter as a detector of human-perceived prompted-pole expression (n=100, 4 annotators).

Selection	Mantel r	Matrix r with main
30/70	0.864	1.000
20/80	0.865	0.998
10/90	0.850	0.978
Top 10 per pole	0.832	0.958
Top 20 per pole	0.845	0.972
Top 30 per pole	0.854	0.980

Table 4: Judge-filter sensitivity on the matched 112-trait cohort. Matrix correlations compare each model matrix with the matched primary 30/70 matrix.

Table 3 shows the filter as a detector of whether the output expresses the prompted pole by human consensus, where an item is *expressed* when the consensus matches the prompted pole and the filter "keeps" it when its score lands in the prompted-pole bin.

Humans saw a clear pole in 79% of responses, so the model is confirmed to produce trait-relevant behavior (the judge is not inventing them). Of the exemplars the filter keeps, 79% match the prompted pole by consensus, 15% are rated ambiguous, and 6% are given the opposite pole. The filter recovers 82% of human-confirmed expressions. It trades recall for precision, dropping some genuine expressions but rarely keeping a sign-flipped one, which is what its use as a large-margin exemplar selector requires, but it is not perfect.

B.5 Sensitivity to Exemplar Re-Selection

We test how the geometry changes when the retained exemplars are narrowed using score thresholds of 30/70, 20/80, or 10/90, or by selecting the 10, 20, or 30 most pole-extreme responses per pole. The saved activation pool contains only responses that passed the primary 30/70 capture filter, so all six conditions are re-selections from that pool. This analysis therefore measures sensitivity to further narrowing, not to admitting responses excluded by the original filter. The matched cohort is the intersection of traits valid under all six conditions; the strict 10/90 condition limits it to 112 traits. No one of 10,000 permutations reaches any observed Mantel statistic ($p < 10^{-4}$). For fixed-count selections, responses are ranked by pole extremity; no scenario or character balancing is imposed. Separately, judge-score separation, defined as the difference between mean pole-B and pole-A scores in the full pre-filter judged-response table, correlates with per-trait human–model alignment (Pearson $r = 0.312$, $p = 3.9 \times 10^{-10}$, $n = 385$). Cleaner separation is associated with stronger alignment, consistent with judge noise attenuating the signal, but the nested re-selection design does not rule out effects of the primary filter.

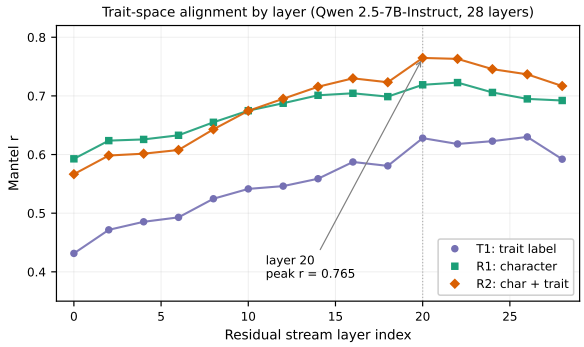

Figure 6: Trait-space alignment by residual-stream layer for the three eliciting framings. The dotted vertical line marks layer 20. Each point is the Mantel r between the model-side Pearson-correlation matrix and the human implicit-personality matrix at that layer; permutation p-values are below 10^{-4} at every reported point. R2 (character + trait) reaches Mantel $r = 0.765$ at layer 20.

C Layer Sweep details

The cohort-level Mantel correlation as a function of residual-stream layer (Figure 6) peaks at layer 20 across all three prompting methods (T1: trait label, R1: character, R2: character + trait). The peak position matches the persona-relevant layer for Qwen 2.5-7B-Instruct identified by Chen et al. (2025) for unrelated traits. Character–trait method reaches the highest value and is the main configuration discussed in the main text.

D Similarity Metric Robustness of the Alignment Statistic

The main text builds $M^{(L)}$ with Pearson correlation across the 3584 residual-stream coordinates, matching the human-side construction. Substituting cosine similarity on the same trait representations h_i leaves the statistic unchanged to four decimal places (0.7647 either way). The upper-

triangle entries of two model-side matrices correlate at $r = 1.000$, so every quantity derived from $M^{(L)}$ is invariant to this choice.

E Activation-Space Diagnostics

We compute diagnostics on the 385 raw character+trait directions at layer 20. Let $\bar{\mathbf{h}}$ denote their coordinate-wise mean. Its norm is 6.8% of the mean individual-vector norm, and the mean cosine similarity over all 73,920 off-diagonal trait pairs is 0.003. Subtracting $\bar{\mathbf{h}}$ before rebuilding the model matrix changes Mantel alignment only from $r = 0.764731$ to $r = 0.764494$. These results rule out domination by a common mean direction, but do not imply that the covariance spectrum is isotropic.

PC1 explains 36.6% of trait-direction variance and has cosine similarity 0.296 with the global mean direction. Removing the PC1 projection from every direction still leaves Mantel $r = 0.613537$. No one of 10,000 permutations reaches either processed-matrix statistic ($p < 10^{-4}$). Thus the alignment is not solely dependent on the global mean or PC1, although the leading components carry substantial structure. Using all nonzero eigenvalues, the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ is 5.16.

F Per-trait Alignment Ranking

The full 385-row alignment ranking is released in the codebase. Table 5 reports the top and bottom 15 for reference.

G Principal Component Analysis Details

Figure 7 contains the Scree plot for the PCA analysis.

H Additional projection-probe examples

Figure 8 gives the full distribution of per-character agreement over the 256 held-out characters. Figure 9 shows the dialogue-projection probe (Section 5.3) for more examples, spanning heroes and villains across distinct fictional worlds and the range $r = 0.50$ to 0.73 on the 195 aligned trait pairs. Each point is one trait axis, standardised per character; the dashed line is $y = x$. Highlighted points are aligned, distinctive traits, labelled with the pole the character is rated toward and its opposite. Figure 10 shows the three meaningful failures

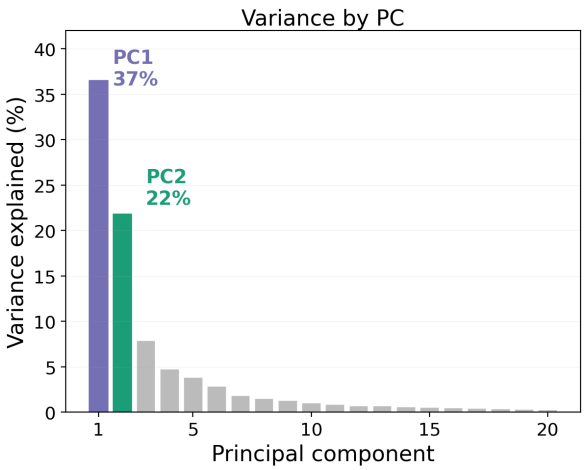

Figure 7: Scree plot for the PCA of the 385 character-and-trait trait directions (layer 20, 3584-dimensional residual stream). PC1 and PC2 capture 37% and 22% of the variance (59% jointly) and stand well clear of the rapidly decaying tail; the full-spectrum participation ratio is 5.16, and 90% of the variance is reached only by PC21. The two leading components—the intellectual and social axes of Figure 4— dominate the trait geometry.

discussed in Section 5.3, with a two-colour convention: blue is the pole humans rate, red the pole the dialogue projects.

H.1 Full-Trait Evaluation and Null Baseline

For each of the 256 held-out characters, we mean-pool Qwen 2.5-7B-Instruct's layer-20 hidden states over the recovered Wikiquote input tokens and compute Pearson correlation between the resulting representation and each trait direction. We compare this model profile with the character's complete SWCPQ profile using Pearson correlation. The 256-character cohort has no identifier overlap with the 877 unique characters used as anchors for the final 385 directions. The 95% intervals use 10,000 character resamples with replacement (seed 20260714). For the null, each draw applies one shared random permutation of the human trait axes to all 256 characters and recomputes the median per-character correlation. No one of 10,000 null medians reaches the observed value for either cohort. Character rankings are highly stable between the full and aligned evaluations (Spearman $\rho = 0.966$, $n = 256$). These analyses test representational transfer to held-out dialogue, not causal use of the directions during generation.

Top 15			Bottom 15		
rank	**pair**	r	**rank**	**pair**	r
1	empath/psychopath	0.923	371	permanent/transient	0.300
2	mad/glad	0.922	372	claustrophobic/spelunker	0.263
3	bitter/sweet	0.920	373	pronatalist/child free	0.239
4	open/guarded	0.918	374	Greek/Roman	0.234
5	positive/negative	0.918	375	plastic/wooden	0.231
6	wholesome/salacious	0.917	376	resigned/resistant	0.197
7	entitled/grateful	0.917	377	leisurely/hurried	0.148
8	modest/flamboyant	0.916	378	philosophical/real	0.118
9	touchy-feely/distant	0.914	379	Coke/Pepsi	0.096
10	debased/pure	0.913	380	melee/ranged	0.048
11	people-person/things-person	0.911	381	chronically single/serial dater	0.034
12	gluttonous/moderate	0.909	382	insider/outsider	0.012
13	gossiping/confidential	0.908	383	aloof/obsessed	-0.017
14	deranged/reasonable	0.907	384	androgynous/gendered	-0.039
15	demanding/unchallenging	0.907	385	activist/nonpartisan	-0.142

Table 5: Per-trait alignment ranking. Top-aligned scales anchor in clear behavioral signatures. Bottom-aligned scales are still used dispositionally by SWCPQ raters but anchor in diffuse cultural stereotype (Coke/Pepsi, Greek/Roman, plastic/wooden) or in poorly differentiated dimensions whose ratings cluster near the midpoint (claustrophobic/spelunker, androgynous/gendered). The gradient reflects how sharply each trait dimension is shared across the rater pool.

Traits	Median r	95% CI	Positive	p_{perm}
195 aligned	0.403	[0.358, 0.445]	238/256	$< 10^{-4}$
All 385	0.328	[0.311, 0.359]	243/256	$< 10^{-4}$

Table 6: Agreement between dialogue-projected and human-rated profiles. Intervals are percentile bootstrap intervals over characters.

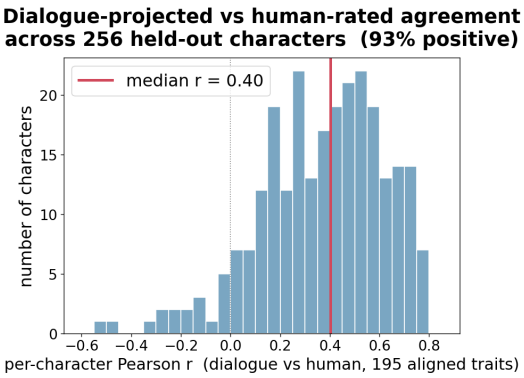

Figure 8: Per-character agreement between dialogue-projected and human-rated profiles over the 256 held-out characters (Pearson r on the 195 aligned trait pairs). 93% are positive; median $r = 0.40$ (red line).

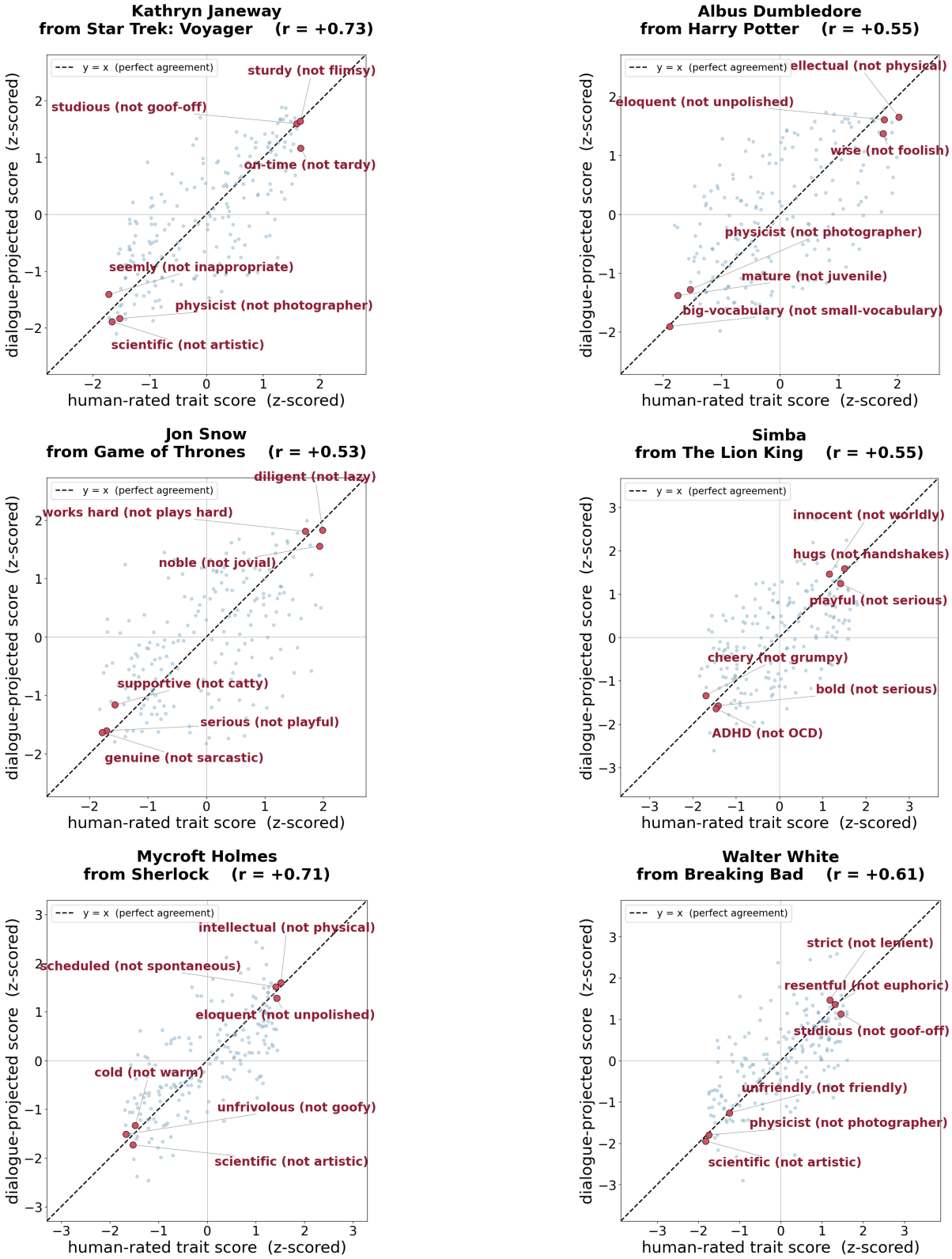

Figure 9: Dialogue-projected versus human-rated trait scores for eleven held-out characters (cf. Figure 5). Each point is one of the 195 aligned trait axes, standardised per character; the dashed line is $y = x$. Pearson r (aligned subset) is in each panel title.

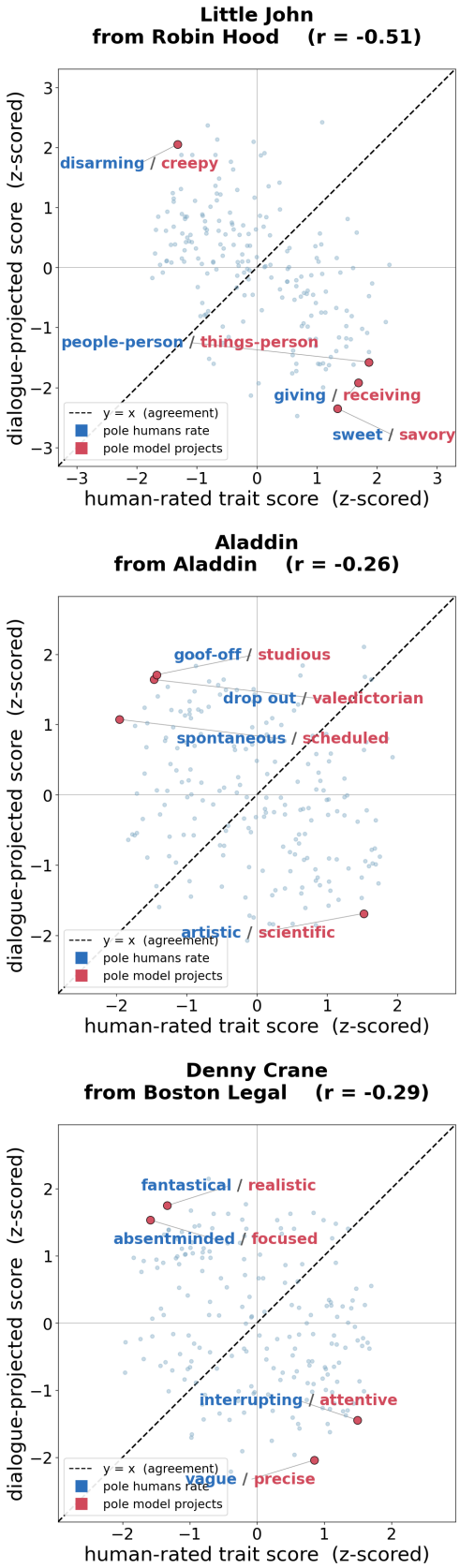

Figure 10: Three meaningful failure examples ($r < 0$). Blue is the pole humans rate, red the pole the dialogue projects.