ON THE BEHAVIORAL TRAITS OF LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Users increasingly describe different AI agents as distinct colleagues to work with. AI personality research aims to quantify such impressions by attributing human-like "traits" to agents. However, existing measures fall short: models' self-reports (S-data) diverge from their actual behavior, while informant ratings from LLM judges (I-data) are costly to scale and cover few everyday scenarios. In this paper, we propose A-B-D to infer traits bottom-up from behavioral data (B-data), namely how agents act on their environment and communicate with users, as recorded in existing trajectories. From 345,667 real-world trajectories spanning 80 models, 12 tasks, and 50 harnesses, we extract 318 candidate features that capture both the actions an agent takes at each step (functional) and the language accompanying them (linguistic). We retain only features that show instance-level stability, cross-task consistency, and model discriminability. Factor analysis of the remaining 79 features uncovers six stable, model-attributable factors, two functional and four linguistic. For example, Kimi-K3 exhibits the most planfulness, whereas GPT-5.5 and GPT-5.6 are the least energetic. Moreover, we quantify the "knowledge–action gap" in the wild: these factors correlate only weakly with self-reported Big Five scores, even for conceptually matched pairs such as *extroversion* and *energetic* ($r = 0.07, p = 0.58$). Our work offers a new lens for understanding AI personality, with implications for users, developers, and researchers from both computer science and social science.

1 INTRODUCTION

Users report that AI agents powered by different large language models differ in how they initiative, caution, verbosity, and willingness to push back (Reddit, 2026a;b). These differences in how agents act and communicate give users the impression of distinct "personalities." Such impressions suggest that agent behavior may reflect stable *traits*, tendencies that recur across tasks and distinguish one agent from another. Measuring these traits is useful to agent developers, evaluators, and end users. It can help developers track behavioral changes across model versions, evaluators compare agents beyond task success, and users judge whether an agent's working style fits their needs. We thus ask: *how can we systematically identify and measure these traits?*

AI personality research seeks to measure these traits using methods from personality psychology, which classifies the evidence into four types, collectively termed *BLIS* (Funder, 2015): self-reports (*S-data*; what the individual says about itself), informant reports (*I-data*; what others say about it), behavioral observations (*B-data*; what it actually does), and life outcomes (*L-data*; the records it accumulates in the world). Most studies on AI agents rely on S-data, administering questionnaires designed originally for humans, which are easy to operationalize and scale (Huang et al., 2024b;a; Jiang et al., 2023; Miotto et al., 2022). However, recent work reveals a *knowledge–action gap*: the traits LLMs report correlate only weakly with their choices in realistic decision scenarios, and more weakly than in humans (Huang et al., 2026; Han et al., 2025; Shen et al., 2025). I-data offer a partial solution by having a judge model infer the traits of a target model from their dialogue (Wang et al., 2024; 2025b; Huang & Hadfi, 2025). Yet they capture only the traits expressed within that dialogue and cannot cover the full range of contexts in which personality manifests.

Questionnaires yield S-data and interviews yield I-data, yet both capture only how a model responds to prompts. Human B-data, such as naturalistic audio sampling (Mehl et al., 2001), is inherently sparse and subject to observer reactivity (Baumeister et al., 2007). AI agents, by contrast, produce massive, complete, and replayable trajectories recording how they think, reason, call tools, evaluate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

their results, and reflect on their outcomes (Liu et al., 2026; Wang et al., 2025c). Therefore, these trajectories make B-data available at scale for the first time.

In this work, we introduce A-B-D (Agent Behavioral-Data), a data-driven, bottom-up framework that derives a low-dimensional, model-attributable behavioral structure from in-the-wild agent trajectories using established psychometric methods. We first collect a large-scale corpus of 345,667 trajectories from publicly released runs spanning 21 benchmarks, 80 models, and 50 harnesses. Following standard practice in personality psychology, we convert each trajectory into a sequence of predefined behaviors (Goodenough, 1928; Bales, 1950; Buss & Craik, 1983; Funder et al., 2000). Specifically, we annotate each trajectory using Act·ONOMY (Gao et al., 2026), a recent, publicly available, domain-agnostic taxonomy, and generate 211 functional features. However, mapping raw outputs to predefined behaviors discards linguistic features, which are known to reflect individual traits (Gottschalk & Gleser, 1969; Pennebaker & King, 1999). We therefore complement the functional features with 107 linguistic features extracted by LIWC-22 (Boyd et al., 2022).

We retain 79 features that satisfy three criteria. (i) *Instance-level stability*: a feature must be expressed consistently across instances within a task rather than driven by particular test cases, which we measure by split-half reliability (Spearman, 1910). (ii) *Cross-task consistency*: a feature must generalize across tasks rather than being task-specific, which we measure by inter-task correlation (Epstein, 1979). (iii) *Model discriminability*: a feature must reliably distinguish between models, which we measure by the intra-class correlation coefficient (ICC) (Shrout & Fleiss, 1979). We then perform factor analysis separately on the functional and linguistic features to summarize them as a few behavioral dimensions.

Our analysis shows that agent behavior in real-world tasks is structured along a small number of stable, model-attributable factors: two functional (planful and deliberate) and four linguistic (talkative, responsible, energetic, and absolutist). Each model's scores on these factors form its behavioral profile, describing how agents powered by that model tend to act and communicate. For example, GPT-5.1 is the least planful, Claude-3.5-Sonnet the most responsible, while Gemini-3-Pro and DeepSeek-V3.2 stand out for absolutist. These profiles are largely invisible to S-data: the factors correlate only weakly with self-reported Big Five scores (mean $|r| = 0.11$; none significant after Bonferroni correction), even for conceptually matched pairs such as extroversion and energetic ($r = 0.07, p = 0.58$). We further find that behavioral profiles shift across model generations. Newer models exhibit more planfulness, which also correlates with task success, but score lower on responsible and energetic. This shift may reflect agentic training that favors explicit planning over the social and affective language characteristic of earlier models. Our contributions are as follows:

- **Data.** We construct a large-scale corpus of 345,667 real-world agent trajectories spanning diverse models, tasks, and harnesses, curated with quality control and stratified sampling.
- **Method.** We introduce A-B-D, a generalizable and scalable analysis pipeline for characterizing LLM agents through their behaviors. It extracts functional and linguistic features, selects features that are task-stable and model-discriminative, and uses factor analysis separately for each feature family to construct behavioral traits.
- **Findings.** We show that, after controlling for task and harness variance, agent behavior exhibits a stable structure attributable to the underlying model. Building on this, we provide, to our knowledge, the first in-the-wild quantification of the "knowledge–action gap," without relying on purpose-built scenarios.

2 THE A-B-D FRAMEWORK

2.1 TRAJECTORIES IN THE WILD

Collection. To identify features that generalize beyond individual tasks, we collect trajectories spanning as diverse a set of scenarios as possible. We draw on 21 benchmarks that publicly release complete agent trajectories, yielding 345,667 trajectories in total. These sources are heterogeneous, including official leaderboard submissions (Li et al., 2026b), community contributions (Merrill et al., 2026), and independent reproductions (Wang et al., 2025a). This breadth allows us to observe the same model across many tasks and harnesses, and different models on the same task. We further categorize the benchmarks into 12 tasks, enabling stratified sampling so that no single category dominates the dataset. Categories include dialogue (Yao et al., 2025; Barres et al., 2026), tool

use (Li et al., 2026a), software engineering (Zan et al., 2025), and research (Wang et al., 2026). Table 2 summarizes the statistics of each benchmark.

Standardization. Each source records trajectories in its own format, such as chat message lists, tool-call spans, or reasoning–action iterations. We first parse each format into a unified event stream, in which each event is a system prompt, a user or agent message, a reasoning trace, a tool call, or an observation, and attach a header containing task and outcome metadata. We then map the model, harness, and task names used by each source to canonical identifiers, so that an entity recorded under multiple aliases is counted once. Finally, we serialize every trajectory into the Agent Trajectory Interchange Format (ATIF v1.8) (Harbor Framework Team, 2026), in which each step comprises a single model call (its reasoning, message, and tool calls) together with the observations; images are replaced with placeholders. All trajectories pass the ATIF validator, and all subsequent analyses operate on this standardized data.

Quality control. We apply a sequence of filters to remove low-quality trajectories; Appendix A.2 reports the number of trajectories remaining after each stage. Specifically, we discard trajectories (i) produced by models outside our target population, such as task-specific fine-tunes and intermediate checkpoints; (ii) whose harness cannot be identified, such as those from customized harnesses; (iii) that contain too little behavior to annotate, including result-only records, fewer than five events or fewer than five steps, and runs without genuine agent actions (no tool calls and fewer than two assistant turns); and (iv) with non-linear structure, namely multi-agent and branching-tree rollouts.

Stratified sampling. We then group the remaining trajectories by (model, task, harness) and discard groups with fewer than 10 trajectories. To enable cross-task comparison of each model's behavior, we retain only models with groups in at least two task categories, yielding a core set of models with sufficient task overlap. Finally, we draw a stratified sample for behavioral analysis: from each remaining group, we sample at most 30 trajectories, balancing successful and failed runs when outcomes are available. This cap bounds the annotation cost per group but does not equalize group sizes; our analyses (§2.3) instead account for the imbalance by controlling for task category and harness when testing each feature. The resulting dataset, used throughout the paper, contains 27,803 trajectories from 80 models and 50 harnesses.

2.2 Feature Construction

Psychology has inferred human traits through two channels: behavior, where a trait summarizes the acts a person repeatedly performs (Buss & Craik, 1983; Funder et al., 2000), and language, the lexical hypothesis underlying the Big Five posits that the most salient individual differences are encoded in natural language (Allport & Odbert, 1936; Goldberg, 1990), and subsequent work showed that an individual's own word use also reflects their traits (Pennebaker & King, 1999). We characterize an LLM through the same two channels: its language yields linguistic features, and its step-by-step actions yield functional features. For each trajectory, we collect all text generated by the agent, including its reasoning. We exclude demonstration and scripted events, as well as any text not produced by the agent, such as environment observations, user messages, and tool outputs.

2.2.1 Functional Features

We define a step as a single model output, comprising its reasoning, message, and tool calls, and annotate each one step of a trajectory at a time. Annotations follow Act·onomy (Gao et al., 2026), a recent taxonomy that organizes agent behaviors into 10 groups and 46 sub-actions. We use GPT-OSS-120B (Agarwal et al., 2025) as the annotator, providing the eight preceding steps and the current observation as read-only context. For each step, the annotator extracts up to three short quotes (3–15 words each), each capturing a distinct behavioral move, and assigns each quote a group and then a sub-action within that group. We constrain its output so that every label is a valid taxonomy category and every quote appears verbatim in the step. Applied to the full corpus, this procedure labels all 947,383 steps across the 27,803 trajectories at the sub-action level, yielding 1,862,737 quotes (1.97 per step on average). Agreement with human annotators is reported in Appendix A.3.

Based on these annotations, we represent each trajectory as a sequence of label sets, defined over the 10 Act·onomy groups. We characterize this sequence through four families of properties: (i)

frequency (Buss & Craik, 1983), the proportion of each behavior, together with overall behavioral diversity and per-step density; (ii) *co-occurrence* (Bakeman & Gottman, 1997), the tendency of two behaviors to appear within the same step; (iii) *transition* (Chatfield & Lemon, 1970), the probability that a behavior in one step is followed by another in the next, together with the persistence of each behavior across consecutive steps; and (iv) *stage* (Lashley et al., 1951), the distribution of each behavior across the early, middle, and late portions of the trajectory. In total, each trajectory is described by 211 functional features; Table 6 lists the full feature bank.

2.2.2 LINGUISTIC FEATURES

Raw text contains formulaic content that carries little information about the agent's linguistic style, such as code blocks, inline code, file paths, URLs, long identifiers, terminal output, and markup. Because word-frequency measures would count such content as words, we remove remove it with a small, ordered set of regular expressions. Table 7 in the appendix lists the full rule set. After cleaning, we retain only trajectories with at least 50 words of prose.

We extract linguistic features with LIWC-22 (Boyd et al., 2022), a validated psycholinguistic lexicon that maps English words to categories grouped into higher-level domains. These categories cover function words (e.g., pronouns, prepositions, and conjunctions), affect (e.g., positive and negative tone), social processes and drives (e.g., social referents and affiliation), cognition (e.g., insight and certitude), and time orientation. LIWC-22 also provides four summary variables: analytic thinking, clout, authenticity, and emotional tone. Applying LIWC-22 to each trajectory's cleaned prose yields, for each word category, the percentage of words that belong to it, and for each summary variable, a standardized score between 0 and 100. We extract 107 linguistic features for each trajectory.

2.3 FEATURE FILTERING

To ensure that features characterize the model rather than its harness or task, we first apply context adjustment to feature values to remove harness- and task-specific effects, and then apply a three-layer test to identify trait-level features, i.e., features that the same model reproduces stably and that vary sufficiently across models. The three layers (stability (ρ_{stab}), consistency (ρ_{con}), and discriminability (ICC) draw on generalizability theory (Shavelson et al., 1989), and each rules out a distinct way in which a behavior can appear trait-like by accident: a behavior driven by only a few trajectories fails stability, one that separates models on only some tasks fails consistency, and one on which models barely differ fails discriminability. We justify the threshold for each layer in Appendix B.1.

Context adjustment. Our trajectories span multiple models, tasks, and harnesses, so each feature reflects both intrinsic model behavior and artifacts of the evaluation environment. To make features comparable across models, we remove task and harness effects using a linear mixed-effects model. We first discard trajectories from harnesses that evaluate only a single model. For each feature, we then fit the following model via restricted maximum likelihood (REML):

$$y_i = \mu + \alpha_{t(i)} + \beta_{h(i)} + \gamma_{m(i)} + e_i, \qquad \gamma_m \sim \mathcal{N}(0, \sigma^2_{\text{model}}), \quad e_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where y_i is the feature value of trajectory i; $t(i)$, $h(i)$, and $m(i)$ denote its task, harness, and model, respectively; μ is the intercept; α_t and β_h are fixed effects for tasks and harnesses, with reference levels set to zero; γ_m is a per-model random intercept; and e_i is the residual. Including γ_m is essential: without it, a harness used predominantly by one model would absorb that model's characteristics into its estimated effect $\hat{\beta}_h$. We then remove the estimated task and harness offsets while preserving the grand mean of each feature:

$$y_i^{\text{adj}} = y_i - \left(\hat{\alpha}_{t(i)} + \hat{\beta}_{h(i)}\right) + \bar{c}, \qquad \bar{c} = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{\alpha}_{t(j)} + \hat{\beta}_{h(j)}\right), \tag{2}$$

where n is the number of trajectories after filtering. The adjusted values y_i^{adj} retain the model effect and the residual, and all downstream analyses are performed on them.

Instance-level stability (ρ_{stab}). The first layer uses split-half reliability (Spearman, 1910; Brown, 1910) to test whether a feature's model–task cell mean reflects the task as a whole rather than a few trajectories. For each cell with at least five trajectories, we randomly partition the trajectories

into two halves and compute the feature mean within each half. We then compute the Pearson correlation r between the two sets of half means across all cells and apply the Spearman–Brown correction, $\rho_{\text{stab}} = 2r/(1+r)$, to account for the halved sample size. We average ρ_{stab} over ten random splits. A low ρ_{stab} indicates that cell means are driven by a few influential trajectories. We retain features with $\rho_{\text{stab}} \geq 0.7$.

Cross-task consistency (ρ_{con}). Whereas the first layer ensures that each cell mean reflects the whole task, the second layer tests whether between-model differences in a feature generalize across tasks rather than reflect task-specific tendencies. We adapt the cross-situational consistency test from personality psychology (Epstein, 1979; Mischel, 2013). In the model $\times$ task matrix of cell means, we compute the Pearson correlation between each pair of task columns over the models observed in both tasks, requiring at least four shared models per pair, and average the correlations over all valid pairs. A ρ_{con} near zero indicates that between-model differences emerge only in particular tasks. We retain features with $\rho_{\text{con}} \geq 0.3$.

Model discriminability (ICC). The first two layers retain features that are stable within tasks and consistent across tasks; the final layer tests whether these features discriminate between models. We measure the proportion of variance in the feature attributable to the model using the one-way random-effects intra-class correlation (Shrout & Fleiss, 1979), with models as the grouping factor: $\text{ICC}(1) = \sigma^2_{\text{model}}/(\sigma^2_{\text{model}}+\sigma^2_{\text{within}})$, where σ^2_{model} and σ^2_{within} are the between-model and within-model (i.e., between-trajectory) variance components, estimated by analysis of variance for unequal group sizes. A high ICC(1) thus indicates that most remaining variation lies between models rather than among trajectories of the same model. We retain features with $\text{ICC}(1) \geq 0.10$.

2.4 Factor Analysis

To test whether the features that pass the three-layer test are governed by a smaller number of latent dimensions, we perform a factor analysis across models. Let $\mathbf{Z} \in \mathbb{R}^{n\times p}$ be the data matrix for n models and p features, where z_{ij} is the mean of feature j over the trajectories of model i, standardized across models to zero mean and unit variance. The feature correlation matrix is then $\mathbf{R} = \frac{1}{n-1}\mathbf{Z}^\top\mathbf{Z}$. Before extracting factors, we assess the factorability of $\mathbf{R}$ with two standard diagnostics (Dziuban & Shirkey, 1974): Bartlett's test of sphericity (Bartlett, 1937), which tests whether $\mathbf{R}$ differs significantly from the identity matrix, and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (Kaiser, 1970; 1974).

We adopt the common factor model $\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$, where $\mathbf{\Lambda} \in \mathbb{R}^{p\times k}$ is the loading matrix and $\mathbf{\Psi} \in \mathbb{R}^{p\times p}$ is the diagonal matrix of uniquenesses. We initialize the uniquenesses with squared multiple correlations, $\mathbf{\Psi}_0 = [\text{diag}(\mathbf{R}^{-1})]^{-1}$, where $\text{diag}(\cdot)$ sets all off-diagonal entries to zero, and form the initial reduced correlation matrix $\mathbf{R}_0^* = \mathbf{R} - \mathbf{\Psi}_0$. We select the number of factors (k) by Horn's parallel analysis (Horn, 1965): factors are retained sequentially while the j-th eigenvalue of $\mathbf{R}_0^*$ exceeds the 95th percentile of the j-th eigenvalue of reduced correlation matrices computed from 500 simulated standard normal datasets of size $n \times p$ (Glorfeld, 1995). We prefer this to the Kaiser criterion (retaining eigenvalues of $\mathbf{R}$ greater than one) (Kaiser, 1960), which tends to over-extract (Zwick & Velicer, 1986).

We estimate $\mathbf{\Lambda}$ by iterated principal axis factoring (PAF). At iteration $t = 0, 1, \ldots$, let $\mathbf{V}_t \in \mathbb{R}^{p\times k}$ and $\mathbf{\Delta}_t \in \mathbb{R}^{k\times k}$ hold the k leading eigenvectors and eigenvalues of $\mathbf{R}_t^* = \mathbf{R} - \mathbf{\Psi}_t$, and set $\mathbf{\Lambda}_t = \mathbf{V}_t\mathbf{\Delta}_t^{1/2}$. The uniquenesses are then updated as $\mathbf{\Psi}_{t+1} = \mathbf{I}_p - \text{diag}(\mathbf{\Lambda}_t\mathbf{\Lambda}_t^\top)$. Starting from the initialization $\mathbf{\Psi}_0$, we iterate until $\max_j|\psi_j^{(t+1)} - \psi_j^{(t)}| < \epsilon$, with $\epsilon = 10^{-6}$.

We then rotate the loadings toward simple structure using varimax with Kaiser normalization (Kaiser, 1958). Let $\mathbf{H} = [\text{diag}(\mathbf{\Lambda}\mathbf{\Lambda}^\top)]^{1/2}$ hold the square roots of the communalities, and let $\mathbf{B} = \mathbf{H}^{-1}\mathbf{\Lambda}\mathbf{T}$ denote the row-normalized loadings under an orthogonal rotation $\mathbf{T}$, $\mathbf{T}^\top\mathbf{T} = \mathbf{I}_k$. Varimax maximizes the variance of the squared loadings within each factor, summed over factors:

$$\sum_{m=1}^{k} \text{Var}_j\left(b_{jm}^2\right) = \sum_{m=1}^{k}\left(\mathbb{E}_j\left[(b_{jm}^2)^2\right] - \left(\mathbb{E}_j\left[b_{jm}^2\right]\right)^2\right) = \sum_{m=1}^{k}\left[\frac{1}{p}\sum_{j=1}^{p} b_{jm}^4 - \left(\frac{1}{p}\sum_{j=1}^{p} b_{jm}^2\right)^2\right]. \quad (3)$$

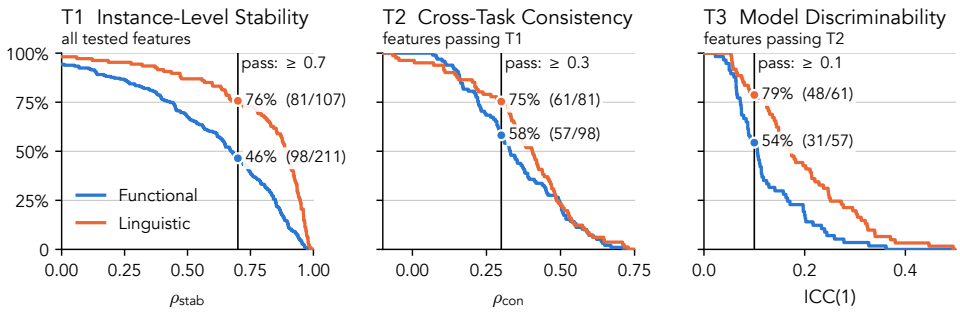


Figure 1: Cumulative distribution of each statistic over 211 functional and 107 linguistic features. Vertical lines mark the thresholds; the height of a curve is the fraction failures in that layer.

The rotated loadings are $\mathbf{\Lambda}_{\text{rot}} = \mathbf{HB} = \mathbf{\Lambda T}$. Since $\mathbf{T}$ is orthogonal, $\mathbf{\Lambda}_{\text{rot}}\mathbf{\Lambda}_{\text{rot}}^{\top} = \mathbf{\Lambda\Lambda}^{\top}$, so the communalities and uniquenesses are unchanged.

Finally, we compute factor scores for each model using Thurstone's regression method (Thurstone, 1935) with ridge regularization. The score weights are $\mathbf{W} = (\mathbf{R} + \alpha\mathbf{I}_p)^{-1}\mathbf{\Lambda}_{\text{rot}}$, with $\alpha = 0.1$. The ridge term keeps $\mathbf{W}$ bounded when features are nearly collinear. We then standardize each column of the raw scores $\mathbf{F} = \mathbf{ZW}$ to zero mean and unit variance, so that each score reflects a model's standing on that factor relative to the other models in our sample.

3 Results

The A-B-D framework specifies how to derive a low-dimensional, model-attributable behavioral structure from in-the-wild agent trajectories. We now focus on three research questions (RQs): (i) *Existence* (§3.1): After controlling for task- and harness-induced variance, does agent behavior exhibit a stable, model-attributable structure? (ii) *Composition* (§3.2): What are the stable factors, and which behavioral features load on each? (iii) *Correlation* (§3.3): How do factor scores correlate with external variables, including model release date, task performance, and self-reported trait scores?

3.1 RQ1: Existence of Stable Features

As shown in Figure 1, of the 318 candidate features we test, 176 are reproducible for the same model on the same task (i.e., they pass the instance-level stability check). Of these, 115 also hold across task types (cross-task consistency check), and 79 further discriminate between models. These 79 features comprise 31 functional features (Table 8) and 48 linguistic features (Table 9).

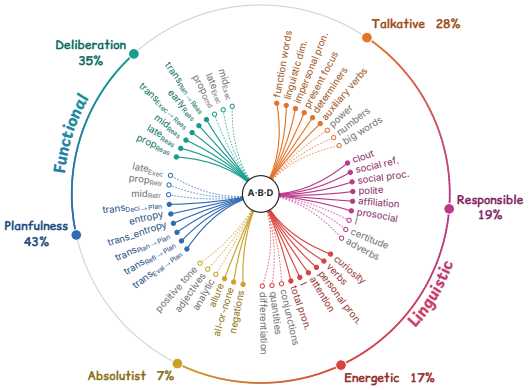
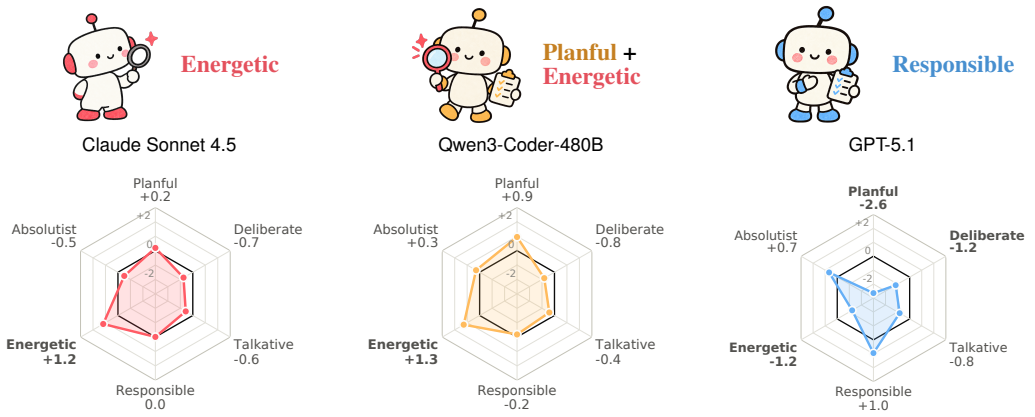


Figure 2: The six behavioral factors and their main features. Each fan is one factor. Leaves are the features with the largest loadings: solid with filled dots for positive, dashed with hollow dots for negative.

3.2 RQ2: Composition of Factors

We conduct separate factor analyses for the functional and linguistic features. The functional feature matrix is suitable for factor analysis (KMO $= 0.89$; Bartlett's $\chi^2(465) = 5713$, $p < 10^{-100}$). Parallel analysis retains two factors, which jointly explain 78% of the variance (43% and 35%); Table 10 reports the rotated loadings. The linguistic feature matrix is likewise suitable (KMO $= 0.70$; Bartlett's $\chi^2(1128) = 8923$, $p < 10^{-100}$). Parallel analysis retains four factors, which jointly explain 71% of the variance (28%, 19%, 17%, and 7%); Table 11 reports the rotated loadings. Table 12 lists the features that load on each of the six factors. To assess the stability of these factors, we rerun the full pipeline under leave-one-task-out cross-validation, holding out



Energetic +1.2

[6] The configure script requires a C compiler, but none is found. I need to install gcc [...] before proceeding with the build.
`$ apt-get install -y build-essential` *(+1 more command)*
[7] Build tools including gcc and gcov are now installed. I can proceed with configuring and building SQLite
`$ CFLAGS="--coverage" ./configure --prefix=/app/sqlite`
[11] Excellent! [...] I can see .gcno files were generated [...] The sqlite3 binary is now in PATH and working *(4 commands)*

Planful +0.9, Energetic +1.3

[12] The configure script failed because no C compiler was found. I need to install a C compiler (gcc) [...] *Plan:* I'll install gcc and build-essential packages
`$ apt-get update`
[13] The package lists have been updated. Now I need to install gcc and build-essential packages. *Plan:* I'll install the build-essential package
`$ apt-get install -y build-essential`
[14] I've successfully installed the build-essential package [...] *Plan:* I'll run the configure script again
`$ ./configure --enable-gcov --prefix=/usr/local`

Responsible +1.0

[2] We are at /app with no work done yet. Need to unpack the pre-vendored SQLite snapshot *(9 commands)*
[4] Configure failed because there is no C compiler or make installed [...] We must first install a build toolchain, then re-run configure and make with coverage flags [...] *Plan:* 1) Inspect OS to confirm we can use apt. [...] 6) Ensure sqlite3 is on PATH
`$ apt-get install -y build-essential`
`$ CC=gcc CFLAGS='--coverage -O0' LDFLAGS='--coverage' ./configure`
`$ make -j"$(nproc)"` *(+9 more commands)*

Figure 3: Three models on the same task and scaffold (terminal-bench-2 `sqlite-with-gcov`, terminus-2). Top: factor scores. Bottom: verbatim excerpts from each trajectory at the same moment, when `configure` fails for want of a compiler; [n] is the step index, `$` a command issued in that step, [...] an elision.

each of the twelve tasks in turn (Appendix B.5). As shown in Figure 2, our results demonstrate six observable factors from the agent trajectories. Figure 3 shows the traits of three models.

Planful. A planful model states what it will do next at almost every step and returns to planning after each evaluation, reflection, or decision; its steps also mix more kinds of behavior. A model low on this factor reads and executes without stating a plan, or plans once at the start and then runs long batches of commands (GPT-5.1 in Figure 3). Claude-Sonnet-4.6, Kimi-K3, and Claude-Opus-4.8 are the most planful; GPT-5.1, o3, and o4-Mini are the least, and almost every OpenAI model scores below zero.

Deliberate. A deliberate model spends much of each step analyzing what it has just observed before acting, and returns to analysis after each action or plan. A model low on this factor moves from one tool call to the next with little analysis. Kimi-K2.7-Code, Llama-3.1-70B, and MiniMax-M2.7 are the most deliberate; Gemini-2.5-Flash, GPT-5-Nano, and GPT-5-Mini are the least.

Talkative. A talkative model writes in full sentences, built from function words, articles, pronouns, and auxiliary verbs; the factor measures sentence form, not length. A model low on this factor writes terse notes dense in numbers and long technical terms. GPT-4o, the Llama-3 models, and DeepSeek-V2.5 are the most talkative; the GPT-5 Codex models and MiniMax-M2.1 are the least.

Responsible. A responsible model writes as if working alongside the reader: *we*, *you*, *please*, *thanks*. A model low on this factor reports in the first person singular and with certainty (*I*, *actually*). Claude-3.5-Sonnet and GPT-4o are by far the most responsible; the Qwen3.6 and Qwen3.7 models, Claude-Sonnet-4.6, and Claude-Opus-4.7 are the least.

Energetic. An energetic model narrates its own looking and checking, in the first person and with positive interjections: "let me check", "I need to look at", "Great, I can see" (Claude-Sonnet-4.5 in Figure 3). A model low on this factor writes flat, qualified description with many adjectives and conjunctions. DeepSeek-V3.2-Exp, MiniMax-M2.1, GLM-4.6, and the Claude 4.0 to 4.5 models are the most energetic; GPT-5.6, GPT-5.5, Kimi-K3, and Claude-Opus-4.8 are the least.

Absolutist. An absolutist model uses many negations and all-or-none words: *not*, *no*, *never*, *all*, *always*. This is the smallest factor and the least stable under task holdout (Appendix B.5). High scores are concentrated in a few models, Gemini-3-Pro, DeepSeek-V3.2, and GPT-OSS-20B; most others score near the mean.

3.3 RQ3: Correlation with External Metrics

Correlation with BFI. We find that four of our derived factors conceptually match a BFI domain. *Conscientiousness* in BFI describes a person who makes plans, follows through with them, and thinks before acting. It matches our *Planful* and *Deliberate* factors, as the former measures whether a model states a plan at each step and the latter measures whether it analyzes before it acts. *Agreeableness* in BFI describes a person who is cooperative and polite to others. It matches our *Responsible* factor, which measures whether a model writes as if working with the reader (*we*, *you*, *please*). *Extraversion* in BFI describes a person who is energetic and full of positive emotion. It matches our *Energetic* factor, which measures whether a model narrates its own actions with positive interjections. *Talkative* and *Absolutist* do not match any BFI domain, since the former measures sentence form rather than how much a model says, and the latter has no counterpart in BFI.

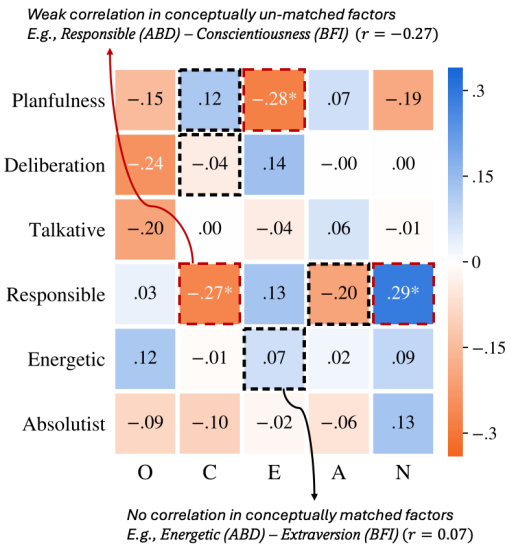


Figure 4: Pearson correlation between factor scores and self-reported Big Five domains.

Figure 4 presents the correlation between our derived factors and the BFI domains, with the conceptually matched pairs boxed. Among the matched pairs, the maximum correlation is only $|r| = 0.20$ (Responsible–A, $p = 0.11$), and its sign is even opposite to the match; in the predicted direction the maximum is $r = 0.12$ (Planful–C, $p = 0.34$). None of the matched pairs is significant, while the three cells with $p < 0.05$ are all unmatched pairs and none of them survives Bonferroni correction. In comparison, self-reported traits in humans correlate with matched behaviors at around $r = 0.30$ (Mischel, 2013). Our results suggest that the personality a model reports about itself is largely unrelated to the personality observed from its behavior, i.e., a knowledge–action gap between S-data and B-data.

Correlation with performance and dates. We correlate each factor with task performance and model release date (Table 1). Regarding performance, only planfulness correlates positively with success ($r = 0.38$, $p < 0.001$), while responsible correlates negatively ($r = -0.31$, $p < 0.01$). Conversely, release date strongly correlates with five factors: newer models are more planful ($r = 0.42$) but less deliberate ($r = -0.24$), talkative ($r = -0.36$), responsible ($r = -0.68$), and energetic ($r = -0.53$). These findings suggest two conclusions. First, planfulness is the only trait that improves over time and directly aids task success. Second, the other generational shifts represent a stylistic drift rather than performance-driven selection. The decline in deliberate, talkative, energetic, and responsible behaviors likely stems from agentic post-training, which rewards terse, plan-driven execution over the conversational style of earlier models. Ultimately, behavioral profiles reveal dimensions of model evolution that task accuracy alone cannot capture.

Table 1: Pearson correlation between factor scores and raw task success (mean normalized outcome over a model's trajectories) and model release date. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Derived Factors from A-B-D					
	Planfulness	Deliberation	Talkative	Responsible	Energetic	Absolutist
Task success	0.38^{***}	0.06	-0.00	-0.31^{**}	-0.20	-0.11
Release date	0.42^{***}	-0.24^{*}	-0.36^{**}	-0.68^{***}	-0.53^{***}	-0.06

4 Related Work

AI personality. Personality psychology separates what a person says about themselves (S-data) from what informants say about them (I-data) (Funder, 2015), and LLM personality research has followed both routes. The S-data route administers human inventories to a model (Miotto et al., 2022; Jiang et al., 2023; Huang et al., 2024b; Pellert et al., 2024; Serapio-García et al., 2025) and reports stable, human-resembling profiles. The I-data route has a judge model infer traits from the target's dialogue, via psychological interviews (Wang et al., 2024), persona simulation (Wang et al., 2025b), or multi-observer ratings (Huang & Hadfi, 2025), but sees only the conversation elicited for it. Meanwhile, S-data itself has been contested: questionnaire answers show social desirability bias (Salecha et al., 2024), shift with prompt wording and option order (Gupta et al., 2024), and fail to reproduce the human factor structure (Sühr et al., 2025).

Knowledge–action gap. These critiques share a root: what a model says about itself need not match what it does. A line of work makes this explicit by building a scenario and comparing behavior in it with self-report. TRAIT (Lee et al., 2025) rewrites inventory items into everyday scenarios; later studies contrast self-reported traits with behavior in social-science experiments (Han et al., 2025), values with value-laden decisions (Shen et al., 2025), endorsed morality with moral action (Huang et al., 2026), Likert scores with generation probabilities on user queries (Song et al., 2025), and psychometric facets with matched behavioral paradigms (Yang et al., 2026). The gap persists even with an LLM-native instrument (Contreras, 2026). Each study designs a new setting and measures the gap there. We need none: we quantify the gap from trajectories agents already produce in the wild, on behaviors that pass a trait test rather than ones chosen in advance.

Bottom-up personality structure. A few studies derive a trait structure for models rather than assuming a human one. Mercer et al. (2025) rerun the HEXACO lexical study on GPT-4 personas, but from human trait adjectives; Jeong (2026) profile agents on four behavioral axes fixed a priori; Yang et al. (2024) steer latent features but score the result on BFI items; Contreras (2026) factor-analyze responses to purpose-written questionnaire items. None derives the structure from what agents do in deployment. Ours is computed from in-the-wild trajectories and named only afterward.

5 Conclusion

In this paper, we characterize AI personality from behavior rather than from self-report, measuring what a model does and how it writes across a large and heterogeneous set of real agent runs. We find that models carry stable behavioral traits that hold across tasks and harnesses and give each model a distinct profile. This profile is a property of the model itself, and it captures a part of an agent that accuracy benchmarks overlook yet users directly experience. While existing studies rely on carefully designed scenarios to show the knowledge–action gap, we directly quantify such gap using agent in-the-wild outputs on different scenarios. Our work opens a behavioral axis for evaluating and comparing AI agents that complements task success.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. (i) *Support qualitative data analysis*: functional behaviors in every trajectory were annotated by GPT-OSS-120B under the Act·ONOMY taxonomy (§2.2); this LLM annotation is a component of our method, and its agreement with a human annotator is reported in Appendix A.3. (ii) *Design or provide feedback on research methodology*: we used Claude Opus and Claude Fable to obtain feedback on the factor analysis pipeline (§2.4). (iii) *Implement methods*: the analysis pipeline was implemented in Python with the assistance of Claude Opus and Claude Fable. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, assist with translation, clean or reformat datasets, or interpret results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work, which develops no new mathematical results. Additionally, we used generative AI tools (Claude Opus and Claude Fable) for tasks with recommended disclosure: sourcing and searching for publicly available agent trajectories, creating and editing software code, drafting parts of the paper, and editing the paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the first author, and all AI-assisted text was checked and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i, 1936.

Roger Bakeman and John M Gottman. *Observing interaction: An introduction to sequential analysis*. Cambridge university press, 1997.

Robert F Bales. *Interaction process analysis; a method for the study of small groups.* Addison-Wesley, 1950.

Elron Bandel, Asaf Yehudai, Lilach Eden, Yehoshua Sagron, Yotam Perlitz, Elad Venezian, Natalia Razinkov, Natan Ergas, Shlomit Shachor Ifergan, Segev Shlomov, et al. General agent evaluation. *arXiv preprint arXiv:2602.22953*, 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. tau2-bench: Evaluating conversational agents in a dual-control environment. In *Forty-third International Conference on Machine Learning*, 2026.

Maurice Stevenson Bartlett. Properties of sufficiency and statistical tests. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 113–126. Springer, 1937.

Roy F Baumeister, Kathleen D Vohs, and David C Funder. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on psychological science*, 2(4):396–403, 2007.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10(1-47):6, 2022.

William Brown. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322, 1910.

David M Buss and Kenneth H Craik. The act frequency approach to personality. *Psychological review*, 90(2):105, 1983.

Christopher Chatfield and Robert E Lemon. Analysing sequences of behavioural events. *Journal of Theoretical Biology*, 29(3):427–445, 1970.

Juan Manuel Contreras. An llm-native psychometric instrument reveals a self-report–behavior gap across 25 models. *arXiv preprint arXiv:2606.09843*, 2026.

Charles D. Dziuban and Edwin C. Shirkey. When is a correlation matrix appropriate for factor analysis? some decision rules. *Psychological Bulletin*, 81(6):358–361, 1974.

Seymour Epstein. The stability of behavior: I. on predicting most of the people much of the time. *Journal of personality and social psychology*, 37(7):1097, 1979.

David C Funder. *The personality puzzle: Seventh international student edition*. WW Norton & Company, 2015.

David C Funder, R Michael Furr, and C Randall Colvin. The riverside behavioral q-sort: A tool for the description of social behavior. *Journal of personality*, 68(3):451–489, 2000.

Jie Gao, Kaiser Sun, Jen-tse Huang, Katherine Van Koevering, Sijie Ji, Heyuan Huang, Weiyan Shi, Zhuoran Lu, Ziang Xiao, Daniel Khashabi, et al. How to interpret agent behavior. *arXiv preprint arXiv:2605.13625*, 2026.

Louis W Glorfeld. An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, 55(3):377–393, 1995.

Lewis R Goldberg. An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

Florence L Goodenough. Measuring behavior traits by means of repeated short samples. *Journal of Juvenile Research*, 1928.

Louis A Gottschalk and Goldine C Gleser. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press, 1969.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 301–314, 2024.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *arXiv preprint arXiv:2509.03730*, 2025.

Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, 2026. URL `https://doi.org/10.5281/zenodo.20953922`.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30 (2):179–185, 1965.

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

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The twelfth international conference on learning representations*, 2024b.

Jen-tse Huang, Jiantong Qin, Xueli Qiu, Sharon Levy, Michelle R Kaufman, and Mark Dredze. Knowing but not doing: Convergent morality and divergent action in llms. *arXiv preprint arXiv:2601.07972*, 2026.

Yin Jou Huang and Rafik Hadfi. Beyond self-reports: Multi-observer agents for personality assessment in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21086–21101, 2025.

Jihoon Jeong. Mti: A behavior-based temperament profiling system for ai agents. *arXiv preprint arXiv:2604.02145*, 2026.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23 (3):187–200, 1958.

Henry F Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151, 1960.

Henry F Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970.

Henry F Kaiser. An index of factorial simplicity. *psychometrika*, 39(1):31–36, 1974.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, et al. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 98778–98849, 2026.

Karl Spencer Lashley et al. *The problem of serial order in behavior*, volume 21. Bobbs-Merrill Oxford, 1951.

James M LeBreton and Jenell L Senter. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, 11(4):815–852, 2008.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. 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*, pp. 8412–8452, 2025.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, et al. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. In *International Conference on Learning Representations*, volume 2026, pp. 53359–53394, 2026a.

Zongxia Li, Zhongzhi Li, Yucheng Shi, Ruhan Wang, Junyao Yang, Zhichao Liu, Xiyang Wu, Anhao Li, Yue Yu, Ninghao Liu, et al. Long-horizon-terminal-bench: Testing the limits of agents on long-horizon terminal tasks with dense reward-based grading. *arXiv preprint arXiv:2607.08964*, 2026b.

Zhaoyang Liu, JingJing Xie, Zichen Ding, Zehao Li, Bowen Yang, Zhenyu Wu, Xuehui Wang, Qiushi Sun, Shi Liu, Weiyun Wang, et al. Scalecua: Scaling open-source computer use agents with cross-platform data. In *International Conference on Learning Representations*, volume 2026, pp. 62707–62766, 2026.

Xiang Long, Li Du, Yilong Xu, RongJian Xu, Qiyanhui Lu, Ying Gao, Qinhua Xie, Fangcheng Liu, Ning Ding, Haoqing Wang, et al. Liveclawbench: Benchmarking llm agents on complex, real-world assistant tasks. *arXiv preprint arXiv:2604.13072*, 2026.

Urbano Lorenzo-Seva and Jos MF Ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.

Matthias R Mehl, James W Pennebaker, D Michael Crow, James Dabbs, and John H Price. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior research methods, instruments, & computers*, 33(4):517–523, 2001.

Sarah Mercer, Daniel P Martin, and Phil Swatton. Applying psychometrics to large language model simulated populations: Recreating the hexaco personality inventory experiment with generative agents. *arXiv preprint arXiv:2508.00742*, 2025.

Mike Merrill, Alexander Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, pp. 40903–40986, 2026.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pp. 218–227, 2022.

Walter Mischel. *Personality and assessment*. Psychology Press, 2013.

Jum C. Nunnally. *Psychometric Theory*. McGraw-Hill, 1978.

Neev Parikh and Hjalmar Wijk. Malt: A dataset of natural and prompted behaviors that threaten eval integrity. https://metr.org/blog/2025-10-14-malt-dataset-of-natural-and-prompted-behaviors/, 10 2025.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

u/Alarming_Eggplant_49 Reddit. I've used enough ai models to realize they all have wildly different personalities at this point i'm convinced ai models are just coworkers with different levels of talent, ego, and criminal energy. Reddit post, r/AI_Agents, 2026a. URL https://www.reddit.com/r/AI_Agents/comments/1slwe5q/ive_used_enough_ai_models_to_realize_they_all/. Accessed: 2026-09-23.

u/dimonb19a Reddit. Does anyone else feel like each ai model has a distinct personality? Reddit post, r/ClaudeAI, 2026b. URL https://www.reddit.com/r/ClaudeAI/comments/1uz2piw/does_anyone_else_feel_like_each_ai_model_has_a/. Accessed: 2026-09-23.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533, 2024.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12): 1954–1968, 2025.

Richard J Shavelson, Noreen M Webb, and Glenn L Rowley. Generalizability theory. *American Psychologist*, 44(6):922, 1989.

Hua Shen, Nicholas Clark, and Tanu Mitra. Mind the value-action gap: Do llms act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3097–3118, 2025.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Woojung Song, Dongmin Choi, Yoonah Park, Jongwook Han, Eun-Ju Lee, and Yohan Jo. Human psychometric questionnaires mischaracterize llm behavior. *arXiv preprint arXiv:2509.10078*, 2025.

Charles Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. Challenging the validity of personality tests for large language models. In *Proceedings of the 2025 Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 74–81, 2025.

L. L. Thurstone. *The Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits*. University of Chicago Press, 1935.

Zelin Wan, Arash Nourian, Xiaoxiao Li, Nihar Nandan, and Kamalakannan Nandagopal. Apiflow-bench: Measuring whether agents survive long, dependent api workflows. *arXiv preprint arXiv:2608.29128*, 2026.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025a.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1840–1873, 2024.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, et al. Coser: Coordinating llm-based persona simulation of established roles. In *ICML*, 2025b.

Xinyuan Wang, Bowen Wang, Dunjie Lu, Junlin Yang, Tianbao Xie, Junli Wang, Jiaqi Deng, Xiaole Guo, Yiheng Xu, Chen Wu, et al. Opencua: Open foundations for computer-use agents. *Advances in Neural Information Processing Systems*, 38:139756–139806, 2025c.

Yuru Wang, Lejun Cheng, Yuxin Zuo, Sihang Zeng, Bingxiang He, Che Jiang, Junlin Yang, Yuchong Wang, Kaikai Zhao, Weifeng Huang, et al. Naturebench: Can coding agents match the published sota of nature-family papers? *arXiv preprint arXiv:2606.24530*, 2026.

Zijian Wu, Xiangyan Liu, Lingjun Chen, Fanqing Meng, Lingxiao Du, Yiran Zhao, Fanshi Zhang, Yaoqi Ye, Jiawei Wang, Zirui Wang, et al. Mcpmark: A benchmark for stress-testing realistic and comprehensive mcp use. In *International Conference on Learning Representations*, volume 2026, pp. 79852–79895, 2026.

Frank Fangzheng Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *Advances in Neural Information Processing Systems*, 38, 2025.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik A Chaudhari, George Karypis, and Huzefa Rangwala. Agentoccam: A simple yet strong baseline for llm-based web agents. In *International Conference on Learning Representations*, volume 2025, pp. 97533–97565, 2025.

Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. Exploring the personality traits of llms through latent features steering. *arXiv preprint arXiv:2410.10863*, 2024.

Yutong Yang, Chenxi Miao, Weikang Li, and Yunfang Wu. Acttraitbench: Quantifying the knowledge-decision gap in large language models via human-grounded behavioral validation. *arXiv preprint arXiv:2605.29791*, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pp. 9965–10017, 2025.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, et al. Osworld2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.

Daoguang Zan, Zhirong Huang, Wei Liu, Hanwu Chen, Shulin Xin, Linhao Zhang, Qi Liu, Li Aoyan, Lu Chen, Xiaojian Zhong, et al. Multi-swe-bench: A multilingual benchmark for issue resolving. *Advances in Neural Information Processing Systems*, 38, 2025.

Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *International Conference on Learning Representations*, pp. 25094–25243, 2025.

William R Zwick and Wayne F Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.

A Details of the a-b-d Framework

A.1 Statistics of All Collected Trajectories

Table 2 lists all 21 sources we collect for trajectories with their assigned task, number of models, harnesses, all (raw) trajectories, and final trajectories selected into the subset.

Table 2: Raw trajectory sources. **Raw** = trajectories collected; **Subset** = trajectories retained after quality control, and stratified sampling. **#Models** and **#Harn.** are counts after the pipeline. For Long-Horizon-Terminal-Bench, only the 503 tasks with a complete trajectory are counted as raw; 698 runs whose log keeps only the last continuation phase were excluded. The sources vary widely in sizes and tasks, and no single source dominates the subset.

Source	Tasks	#Models	#Harn.	#Raw	#Subset
GAIA (Mialon et al., 2024)	research-qa	3	1	65,915	90
HAL (Kapoor et al., 2026)	science, swe, dialogue, research-qa, web	22	13	58,095	4,876
APIFlow-Bench (Wan et al., 2026)	tool-api	19	1	56,037	570
Terminal-Bench 2.0 (Community) (Merrill et al., 2026)	terminal, science, swe, security, ml-eng, other	31	9	52,272	11,254
Multi-SWE-bench (Zan et al., 2025)	swe	7	2	27,514	420
MCPMark (Wu et al., 2026)	tool-api	17	1	11,314	510
τ^2-Bench (Barres et al., 2026)	dialogue	4	4	10,832	300
Terminal-Bench 2.0 (Official) (Merrill et al., 2026)	terminal, science, swe, security, ml-eng, other	12	12	10,680	3,042
Open General Agent Leaderboard (Bandel et al., 2026)	research-qa, swe, dialogue, tool-api	5	5	10,056	2,114
AgentOccam (Yang et al., 2025)	web	2	1	7,446	0
MALT (Parikh & Wijk, 2025)	research-qa, security, swe, science, ml-eng, web, other	5	1	7,179	893
Toolathlon (Li et al., 2026a)	tool-api	19	1	7,116	570
LiveClawBench (Long et al., 2026)	research-qa, terminal, web, swe	17	1	6,752	2,009
OpenHands (Wang et al., 2025a)	swe	8	1	5,695	240
TheAgentCompany (Xu et al., 2025)	office	10	3	2,707	330
τ-Bench (Yao et al., 2025)	dialogue	1	1	1,980	30
NatureBench (Wang et al., 2026)	science	10	2	1,439	282
OSWorld2.0 (Yuan et al., 2026)	gui	6	3	1,430	180
Cybench (Zhang et al., 2025)	security	2	1	690	60
Long-Horizon-Terminal-Bench (Li et al., 2026b)	office, ml-eng, swe, other	2	1	503	33
WebVoyager (He et al., 2024)	web	–	1	15	0
21 Benchmarks	**12 Tasks**	80	50	345,667	27,803

A.2 FILTERING TRAJECTORIES

Table 3 traces the reduction from all collected trajectories to the final subset. The largest reduction comes from trajectories generated by models outside the target candidate set, most of which originate from large fine-tuning pools. We then restrict the eligible set to models covering at least two task types, which form the connected core required for the cross-task test; subsample each (model, task, harness) group; and finally remove branching-tree rollouts and runs too short to annotate.

Table 3: Filtering from all collected trajectories to the final subset. Each discarded trajectory is logged with its exclusion reason for transparency and reproducibility. Eligibility predicates are applied in the listed order, and each trajectory is attributed to the first predicate it fails.

Stage	Dropped	Remaining
Raw collected (21 sources)		345,667
model not in target candidate set	63,485	
harness unidentified	200	
task unidentified	1,969	
result record only (no trajectory)	18,294	
multi-agent (flattened threads)	2,162	
fewer than 5 events	21,181	
no genuine agent action	1,308	
branching-tree rollout	1,452	
fewer than 5 annotatable agent steps	21,820	
Eligible		213,796
Groups below floor (138 of 1,153 groups)	520	
Connectivity: models on <2 task types (49 of 129 models)	33,529	
Connected core (80 models, 960 groups)		179,747
Stratified sampling (cap 30 per group)	151,944	
Final subset		**27,803**

A.3 Agreement with Human for Behavioral Annotation

We used GPT-OSS-120B to annotate functional behaviors of the trajectories. To validate the annotation results, we compare the LLM annotated behaviors with human on a randomly sampled subset. Table 4 and 5 reports the annotation difference.

Table 4: Step-level agreement between a human annotator and the GPT-OSS-120B annotator on 8 trajectories (56 steps, 131 human and 140 model spans). Each step is reduced to the set of labels its spans carry; Micro-F1 treats the human labels as the reference, Jaccard is the mean per-step overlap of the two label sets.

Metric	Spans per Step		Micro-F1		Jaccard	
	Human	GPT-OSS-120B	Group	Sub-Action	Group	Sub-Action
Value	2.34	2.50	0.45	0.30	0.36	0.27

Table 5: Per-group Cohen's κ between the human and the GPT-OSS-120B annotator on the same 56 steps. For each group the two annotations are reduced to whether the step carries at least one span of that group; κ is the chance-corrected agreement of these binary decisions. Count: steps carrying the group. Learning never occurs on either side (κ undefined). Pooled: one κ over all 560 step $\times$ group decisions.

Group	Human	GPT-OSS-120B	κ
Grounding	16	23	0.11
Retrieval	11	5	0.43
Reasoning	8	18	0.14
Planning	11	22	0.22
Evaluate	7	9	-0.16
Deciding	4	3	0.24
Executing	41	34	0.25
Reflection	6	2	-0.06
Learning	0	0	–
Memory	13	0	0.00
Pooled			0.30

A.4 GENERATING FUNCTIONAL FEATURES FROM BEHAVIORS

Table 6 lists all functional features. Each trajectory's behavior sequence is analyzed along four families: frequency (how often each behavior occurs), co-occurrence (which behaviors share a step), transition (which behavior follows which), and stage (where in the trajectory a behavior falls). With the indices g and h ranging over the ten Act·ONOMY groups, these templates expand to 211 features.

Table 6: The functional behavior feature bank. Co-occurrence features are computed within a step, transition and stage features over consecutive steps.

Family	Feature	Description	#
Frequency	prop_g	Share of behavioral spans labeled group g.	10
	density	Behavioral spans per agent step.	1
	entropy	Normalized entropy of the group distribution (behavioral variety).	1
Co-occurrence	$\mathrm{lift}_{g,h}$	Within-step co-occurrence lift of groups g and h: how much more than chance they occur in the same step.	45
	grps_per_step	Mean number of distinct groups in a step.	1
Transition	$\mathrm{trans}_{g\to h}$	Probability that a step containing group g is followed by a step containing group h (conditional presence; a row need not sum to one).	100
	self_loop	Probability that a group present in a step is still present in the next step (persistence).	1
	step_jaccard	Mean Jaccard overlap between the group sets of consecutive steps.	1
	trans_entropy	Mean normalized entropy of the next-step group distribution.	1
Stage	$\mathrm{centroid}_g$	Mean normalized position (0 to 1) of group g across the trajectory.	10
	early_g, mid_g, late_g	Share of group g in the early / middle / late third of the trajectory.	30
	shift_g	Late-third share minus early-third share of group g.	10
		Total	**211**

A.5 Data Cleaning for Linguistic Features

Table 7 lists the process of data cleaning for a better LIWC parsing result.

Table 7: Cleaning rules applied to each trajectory's raw text, in the order they run. Each removal is a single named regular expression: spans matched in the span-stripper block are replaced by a space, and the line-filter block drops whole lines. Regex flags: ^{s} DOTALL, ^{m} multiline, ^{i} case-insensitive.

Rule	Pattern	Note
(a) Message level		
JSON prose	*message parses as a JSON dict*	keep only fields `analysis`, `plan`, `thought`, `reasoning`, `reflection`; drop the rest (e.g. command payloads)
(b) Span strippers (matched span → space)		
Fenced code	```` ```.*?```^{s} ````	fenced code blocks
Dangling fence	```` ```.*^{s} ````	an unclosed trailing code fence
Think tags	`</?(mm:)?think>`^{i}	reasoning-wrapper tags (e.g. MiniMax `<mm:think>`)
Inline markup	`<[^>\n]{1,200}>`	inline HTML / XML tags
Inline code	`` `[^`\n]+` ``	backtick code spans
URL	`https?://\S+|www\.\S+`	web links
File path	`(?:\.{0,2}/)?(?:[\w.\-]+/){2,}[\w.\-]*`	slash paths of ≥ 2 segments
Long hex	`\b[0-9a-fA-F]{16,}\b`	hashes / ids (≥ 16 hex chars)
Long number	`\b\d{7,}\b`	digit runs of ≥ 7
(c) Line filters (whole line dropped)		
Prompt line	`^\s*(root@|$ |# |>>> |\.\.\. )`^{m}	lines starting with a shell or REPL prompt
Non-prose line	*letter ratio* < 0.55 *or* < 2 *words*	tables, hexdumps, code fragments, lone tokens
(d) Normalization		
Markdown markers	`^\s*[#>\-\*\d\.)]+\s*`^{m}	leading heading / bullet / list / quote markers
Whitespace	`\s+` → space	collapse runs of whitespace

B More Results

B.1 Three-Layer Test: Score Distribution

Figure 1 shows the cumulative distribution of each layer's statistic over all tested features, separately for functional and linguistic features. Because a feature passes a layer when its statistic meets or exceeds the layer's threshold, the height of each curve at the dashed line gives the fraction of features that fail that layer.

We set the thresholds following psychometric convention and the data distribution. For instance-level stability, 0.70 is the minimum reliability conventionally recommended by Nunnally (1978). For cross-task consistency, correlations in human personality research rarely exceed 0.30, whether between a trait measure and behavior in a specific situation or between behaviors across situations (Mischel, 2013). We therefore require a feature to be at least as consistent across tasks as human traits typically are across situations. For model discriminability, ICC values are concentrated near zero in our data, and a threshold of 0.10, a medium effect for a variance-partition coefficient (LeBreton & Senter, 2008), separates features with non-negligible between-model variance from the bulk with almost none.

B.2 Features Passing the Three-Layer Test

Table 8: The 31 functional features that pass all three tests, grouped by the feature families of Table 6. Act·onomy groups are abbreviated in subscripts: Plan = Planning, Reas = Reasoning, Exec = Executing, Eval = Evaluate, Refl = Reflection, Grnd = Grounding, Retr = Retrieval, Deci = Deciding. ρ_{stab}: split-half reliability within a task; ρ_{con}: cross-task consistency; ICC: between-model discriminability.

Family	Feature	ρ_{stab}	ρ_{con}	ICC
Transition	trans_entropy	0.98	0.64	0.36
	$\text{trans}_{\text{Plan}\to\text{Plan}}$	0.95	0.65	0.24
	$\text{trans}_{\text{Reas}\to\text{Plan}}$	0.93	0.50	0.20
	$\text{trans}_{\text{Reas}\to\text{Reas}}$	0.94	0.55	0.20
	$\text{trans}_{\text{Exec}\to\text{Reas}}$	0.94	0.55	0.20
	$\text{trans}_{\text{Plan}\to\text{Reas}}$	0.95	0.52	0.20
	$\text{trans}_{\text{Exec}\to\text{Plan}}$	0.94	0.59	0.20
	$\text{trans}_{\text{Eval}\to\text{Plan}}$	0.89	0.53	0.14
	$\text{trans}_{\text{Refl}\to\text{Plan}}$	0.88	0.60	0.13
	$\text{trans}_{\text{Refl}\to\text{Exec}}$	0.87	0.62	0.11
	$\text{trans}_{\text{Deci}\to\text{Plan}}$	0.85	0.46	0.11
	$\text{trans}_{\text{Refl}\to\text{Reas}}$	0.86	0.51	0.11
	$\text{trans}_{\text{Eval}\to\text{Reas}}$	0.85	0.48	0.11
	$\text{trans}_{\text{Plan}\to\text{Exec}}$	0.90	0.38	0.11
Co-occurrence	grps_per_step	0.98	0.59	0.33

Family	Feature	ρ_{stab}	ρ_{con}	ICC
Frequency	$\text{prop}_{\text{Reas}}$	0.96	0.49	0.28
	entropy	0.96	0.55	0.26
	$\text{prop}_{\text{Plan}}$	0.96	0.52	0.24
	density	0.96	0.51	0.22
	$\text{prop}_{\text{Grnd}}$	0.95	0.33	0.16
	$\text{prop}_{\text{Refl}}$	0.90	0.71	0.12
	$\text{prop}_{\text{Retr}}$	0.94	0.31	0.11
Stage	$\text{early}_{\text{Reas}}$	0.95	0.51	0.23
	$\text{early}_{\text{Plan}}$	0.94	0.39	0.17
	mid_{Reas}	0.93	0.50	0.17
	mid_{Plan}	0.92	0.57	0.16
	$\text{late}_{\text{Exec}}$	0.90	0.32	0.12
	$\text{late}_{\text{Reas}}$	0.90	0.30	0.11
	mid_{Exec}	0.90	0.31	0.11
	$\text{late}_{\text{Plan}}$	0.88	0.49	0.11
	mid_{Retr}	0.91	0.33	0.10

Table 9: The 48 linguistic features that pass all three tests. Features are grouped by LIWC-22's own output-header domains (*Function words* abbreviates LIWC's *Linguistic Dimensions* header, *Summary* its summary variables, and *Motives* is our label for the header-less `need`/.../`allure` block). Categories are scored word-level on the full cleaned corpus. ρ_{stab}: split-half reliability within a task; ρ_{con}: cross-task consistency; ICC: between-model discriminability.

Domain	LIWC Cat.	ρ_{stab}	ρ_{con}	ICC
Function words	we	0.99	0.71	0.49
	function words	0.98	0.46	0.37
	articles	0.98	0.46	0.34
	linguistic dim.	0.97	0.42	0.34
	personal pron.	0.97	0.59	0.32
	total pron.	0.97	0.51	0.32
	I	0.97	0.55	0.32
	determiners	0.98	0.41	0.30
	auxiliary verbs	0.97	0.53	0.27
	verbs	0.96	0.40	0.25
	adverbs	0.95	0.44	0.25
	conjunctions	0.95	0.58	0.23
	you	0.97	0.42	0.22
	impersonal pron.	0.95	0.37	0.20
	prepositions	0.95	0.52	0.18
	adjectives	0.94	0.34	0.18
	numbers	0.94	0.49	0.16
	quantities	0.91	0.36	0.15
	negations	0.96	0.53	0.14
Summary	clout	0.98	0.65	0.38
	big words	0.97	0.37	0.25
	analytic	0.93	0.44	0.17
	emotional tone	0.89	0.51	0.15
	authentic	0.91	0.32	0.11
Perception	attention	0.92	0.35	0.15
Affect	positive emotion	0.89	0.47	0.13
	positive tone	0.88	0.48	0.12
Social	social ref.	0.96	0.65	0.31
	social proc.	0.93	0.54	0.17
	prosocial	0.90	0.57	0.14
	polite	0.91	0.48	0.13
Drives	affiliation	0.98	0.73	0.45
	power	0.98	0.33	0.29
	drives (total)	0.96	0.49	0.21
	achievement	0.92	0.47	0.17
Cognition	certitude	0.93	0.60	0.21
	insight	0.93	0.54	0.20
	discrepancy	0.93	0.42	0.14
	differentiation	0.90	0.44	0.12
	all-or-none	0.96	0.38	0.12
Lifestyle	work	0.96	0.33	0.25
	lifestyle (total)	0.95	0.42	0.21
Time orient.	present focus	0.97	0.41	0.30
	future focus	0.96	0.71	0.28
Motives	curiosity	0.94	0.61	0.24
	need	0.92	0.48	0.16
	allure	0.94	0.47	0.15
Conversation	assent	0.95	0.48	0.14

B.3 LOADINGS OF FEATURES

Table 10: Functional family: varimax-rotated loadings of the 31 functional features on the two factors retained by parallel analysis (KMO $= 0.89$). Loadings with $|\lambda| \geq 0.40$ are marked in bold; h^2 is the communality. Traits are ordered by their primary factor and loading.

Trait	F1	F2	h^2
$\text{trans}_{\text{Eval}\to\text{Plan}}$	**0.90**	0.35	0.94
$\text{trans}_{\text{Refl}\to\text{Plan}}$	**0.88**	0.27	0.85
$\text{trans}_{\text{Plan}\to\text{Plan}}$	**0.85**	**0.46**	0.94
trans_entropy	**0.85**	0.30	0.81
entropy	**0.84**	0.35	0.83
$\text{trans}_{\text{Deci}\to\text{Plan}}$	**0.81**	0.29	0.74
$\text{trans}_{\text{Refl}\to\text{Exec}}$	**0.80**	0.16	0.67
$\text{trans}_{\text{Reas}\to\text{Plan}}$	**0.80**	**0.54**	0.92
grps_per_step	**0.79**	**0.55**	0.92
$\text{prop}_{\text{Plan}}$	**0.78**	**0.53**	0.89
mid_{Plan}	**0.74**	**0.53**	0.84
$\text{late}_{\text{Plan}}$	**0.74**	**0.50**	0.81
mid_{Retr}	**−0.74**	−0.27	0.62
$\text{early}_{\text{Plan}}$	**0.74**	**0.53**	0.82
$\text{trans}_{\text{Refl}\to\text{Reas}}$	**0.71**	**0.54**	0.80
$\text{prop}_{\text{Retr}}$	**−0.67**	−0.30	0.54
$\text{trans}_{\text{Exec}\to\text{Plan}}$	**0.66**	**0.59**	0.78
density	**0.62**	**0.45**	0.59
$\text{prop}_{\text{Refl}}$	**0.55**	0.21	0.35
$\text{trans}_{\text{Plan}\to\text{Exec}}$	0.32	0.26	0.17
$\text{prop}_{\text{Reas}}$	0.30	**0.95**	0.99
$\text{late}_{\text{Reas}}$	0.16	**0.94**	0.92
mid_{Reas}	0.32	**0.92**	0.95
$\text{trans}_{\text{Exec}\to\text{Reas}}$	0.36	**0.92**	0.97
$\text{early}_{\text{Reas}}$	0.37	**0.88**	0.91
$\text{trans}_{\text{Plan}\to\text{Reas}}$	**0.47**	**0.84**	0.93
$\text{trans}_{\text{Reas}\to\text{Reas}}$	**0.51**	**0.81**	0.91
$\text{trans}_{\text{Eval}\to\text{Reas}}$	**0.55**	**0.73**	0.84
$\text{prop}_{\text{Grnd}}$	−0.39	**−0.72**	0.67
$\text{late}_{\text{Exec}}$	**−0.60**	**−0.61**	0.72
mid_{Exec}	**−0.47**	**−0.58**	0.56
Variance explained	43%	35%	

Table 11: Linguistic family: varimax-rotated loadings of the 48 linguistic features on the four factors retained by parallel analysis (KMO $= 0.70$). Loadings with $|\lambda| \geq 0.40$ are marked in bold; h^2 is the communality. Traits are ordered by their primary factor and loading.

Trait	F1	F2	F3	F4	h^2
function words	**0.95**	0.08	0.19	0.01	0.94
linguistic dim.	**0.93**	−0.00	0.30	−0.04	0.96
impersonal pron.	**0.91**	−0.10	0.12	−0.14	0.88
present focus	**0.89**	0.03	0.12	−0.06	0.81
determiners	**0.88**	0.12	0.15	−0.00	0.81
auxiliary verbs	**0.88**	0.14	0.13	−0.01	0.81
articles	**0.87**	0.09	0.16	−0.14	0.80
power	**−0.85**	0.24	0.26	0.17	0.87
numbers	**−0.83**	−0.20	−0.12	−0.03	0.74
big words	**−0.73**	**0.42**	0.21	−0.05	0.76
lifestyle (total)	**−0.72**	0.32	**0.48**	0.16	0.87
total pron.	**0.72**	0.19	**0.61**	−0.01	0.92
adverbs	**0.71**	**−0.48**	−0.21	−0.04	0.78
work	**−0.69**	0.32	**0.52**	0.18	0.89
prepositions	**0.66**	0.18	0.30	−0.04	0.55
achievement	**0.52**	−0.40	0.25	−0.37	0.63
discrepancy	**0.48**	−0.07	0.23	0.28	0.37
assent	−0.28	−0.21	−0.22	0.04	0.17
clout	0.14	**0.92**	0.18	0.12	0.92
social ref.	0.27	**0.87**	0.27	0.21	0.95
social proc.	−0.28	**0.83**	0.12	0.30	0.88
polite	−0.16	**0.83**	0.15	−0.02	0.73
affiliation	**0.41**	**0.83**	0.13	0.15	0.90
prosocial	−0.09	**0.80**	0.35	−0.07	0.78
we	**0.46**	**0.79**	0.09	0.16	0.87
you	−0.11	**0.71**	**0.60**	0.11	0.88
I	0.23	**−0.62**	**0.62**	−0.15	0.85
certitude	0.24	**−0.61**	−0.33	−0.25	0.60
drives (total)	**−0.52**	**0.61**	**0.45**	0.12	0.85
authentic	0.34	−0.37	−0.01	−0.27	0.32
curiosity	0.14	0.38	**0.86**	0.10	0.91
verbs	**0.41**	0.06	**0.77**	0.09	0.77
personal pron.	**0.48**	0.31	**0.75**	0.06	0.89
attention	0.14	**0.42**	**0.66**	0.15	0.66
positive emotion	0.15	0.11	**0.60**	−0.15	0.41
emotional tone	0.34	0.32	**0.59**	−0.26	0.63
conjunctions	0.28	0.17	**−0.58**	−0.02	0.45
insight	0.16	0.33	**0.57**	−0.11	0.47
quantities	0.29	**−0.45**	**−0.55**	−0.39	0.74
differentiation	**0.46**	−0.18	**−0.52**	0.10	0.52
adjectives	**0.46**	**−0.40**	**−0.51**	**−0.43**	0.82
need	−0.02	0.28	**0.48**	0.37	0.45
positive tone	0.29	0.22	**0.45**	**−0.40**	0.50
future focus	0.24	**0.41**	**0.44**	−0.08	0.43
negations	−0.08	0.26	−0.09	**0.89**	0.88
all-or-none	−0.29	0.14	−0.02	**0.74**	0.66
allure	0.28	−0.01	**0.55**	**0.56**	0.69
analytic	−0.11	−0.12	0.02	**−0.48**	0.26
Variance explained	28%	19%	17%	7%	

B.4 Features That Load on Each Factor

Table 12: Features load on each factor ($|\lambda| \geq 0.40$ in the varimax solution), split by the sign of the loading and ordered by its size. The two families are analyzed separately, so their factors are numbered independently; a feature may load on two factors of its family.

	Functional		Linguistic			
	F1	F2	F1	F2	F3	F4
Positive	$\text{trans}_{\text{Eval}\to\text{Plan}}$	$\text{prop}_{\text{Reas}}$	function words	clout	curiosity	negations
	$\text{trans}_{\text{Refl}\to\text{Plan}}$	$\text{late}_{\text{Reas}}$	linguistic dim.	social ref.	verbs	all-or-none
	$\text{trans}_{\text{Plan}\to\text{Plan}}$	mid_{Reas}	impersonal pron.	social proc.	personal pron.	allure
	trans_entropy	$\text{trans}_{\text{Exec}\to\text{Reas}}$	present focus	polite	attention	
	entropy	$\text{early}_{\text{Reas}}$	determiners	affiliation	I	
	$\text{trans}_{\text{Deci}\to\text{Plan}}$	$\text{trans}_{\text{Plan}\to\text{Reas}}$	auxiliary verbs	prosocial	total pron.	
	$\text{trans}_{\text{Refl}\to\text{Exec}}$	$\text{trans}_{\text{Reas}\to\text{Reas}}$	articles	we	you	
	$\text{trans}_{\text{Reas}\to\text{Plan}}$	$\text{trans}_{\text{Eval}\to\text{Reas}}$	total pron.	you	positive emotion	
	grps_per_step	$\text{trans}_{\text{Exec}\to\text{Plan}}$	adverbs	drives (total)	emotional tone	
	$\text{prop}_{\text{Plan}}$	grps_per_step	prepositions	big words	insight	
	mid_{Plan}	$\text{trans}_{\text{Reas}\to\text{Plan}}$	achievement	attention	allure	
	$\text{late}_{\text{Plan}}$	$\text{trans}_{\text{Refl}\to\text{Reas}}$	discrepancy	future focus	work	
	$\text{early}_{\text{Plan}}$	$\text{prop}_{\text{Plan}}$	personal pron.		lifestyle (total)	
	$\text{trans}_{\text{Refl}\to\text{Reas}}$	mid_{Plan}	we		need	
	$\text{trans}_{\text{Exec}\to\text{Plan}}$	$\text{early}_{\text{Plan}}$	differentiation		positive tone	
	density	$\text{late}_{\text{Plan}}$	adjectives		drives (total)	
	$\text{prop}_{\text{Refl}}$	$\text{trans}_{\text{Plan}\to\text{Plan}}$	verbs		future focus	
	$\text{trans}_{\text{Eval}\to\text{Reas}}$	density	affiliation			
	$\text{trans}_{\text{Reas}\to\text{Reas}}$					
	$\text{trans}_{\text{Plan}\to\text{Reas}}$					
Negative	mid_{Retr}	$\text{prop}_{\text{Grnd}}$	power	I	conjunctions	analytic
	$\text{prop}_{\text{Retr}}$	$\text{late}_{\text{Exec}}$	numbers	certitude	quantities	adjectives
	$\text{late}_{\text{Exec}}$	mid_{Exec}	big words	adverbs	differentiation	positive tone
	mid_{Exec}		lifestyle (total)	quantities	adjectives	
			work	adjectives		
			drives (total)			

B.5 Leave-One-Out Cross-Validation on the Factors

Setting. We ask whether the traits and factors reported in §3 depend on any single task. For each of the 12 tasks we remove all trajectory of one task from both feature tables (e.g., 180 trajectories for *gui*, 5,332 for *swe*) and rerun the entire pipeline on the remaining eleven with unchanged settings: the harness filter, the context adjustment (the mixed model is refitted, now without the held-out task level), the three-layer tests with the same thresholds ($\rho_{\text{stab}} \geq 0.7$, $\rho_{\text{con}} \geq 0.3$, $\text{ICC} \geq 0.1$), and factor analysis using the same configuration (Horn's parallel analysis for the number of factors, iterated PAF, varimax rotation, ridge regression scores). Nothing is copied from the reference run: each fold selects its own traits and its own number of factors. The fold is then compared with the reference run on two levels, the trait set and the factor structure (Table 13).

Trait-level feature Jaccard. Let A be the set of features that pass the three tests in the reference run (31 functional, 48 linguistic) and B the set that passes in the fold. The Jaccard index,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

is the fraction of features, among all that are a trait in at least one of the two runs, that are a trait in both. It equals one when the two runs select exactly the same traits and zero when they share none; a feature that drops out of the trait set and one that newly enters it lower the index by the same amount. Folds keep 22–36 functional and 43–53 linguistic traits, and the mean Jaccard is 0.87 for the functional and 0.94 for the linguistic family. The lowest values come from holding out *science* (0.71/0.77): the fold loses nine functional traits, mostly transition and stage features (e.g., Evaluate→Planning, late-run Planning), and adds none, because science trajectories contribute many of the task pairs on which the cross-task consistency test is computed. Across folds, 18 of the 31 functional and 38 of the 48 linguistic traits are selected in all 12 folds, and 25 and 45 in at least 10.

Tucker congruence ϕ. A factor is described by its vector of loadings, one entry per trait. To compare a factor of the reference run, $\boldsymbol{a}$, with a factor of the fold, $\boldsymbol{b}$, we restrict both to the traits kept by both runs and compute the Tucker congruence coefficient:

$$\phi(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2 \sum_i b_i^2}},$$

the cosine of the angle between the two loading vectors (Lorenzo-Seva & Ten Berge, 2006). It is one when the fold's factor loads on the same traits in the same proportions as the reference run's (the loadings may differ by a common scale), zero when the two loading patterns are unrelated, and it ignores the sign of a factor, which is arbitrary in factor analysis (we report $|\phi|$). Conventionally $\phi \geq 0.95$ is taken to mean the two factors are the same and 0.85–0.94 that they are similar. Because a fold may retain a different number of factors than the reference run, we first pair the factors of the two runs one-to-one so as to maximize the total congruence (Hungarian assignment). As shown in Table 13, two functional factors reappear in all 12 folds (mean $\phi = 0.99$, minimum 0.97), and every fold retains exactly two of them. Similarly, the linguistic family retains four factors in 6 of 12 folds (3–6 otherwise); its first two factors are recovered with mean $\phi = 0.97$, the third with 0.93, and the fourth (negation, 7% of variance in the reference run) with 0.91, the only value below 0.85 being the *swe* fold ($\phi = 0.61$), where the fold splits into six factors.

Table 13: Leave-one-out stability of the pipeline. For every fold the held-out task type is removed and the whole pipeline is rerun on the remaining eleven; the result is compared with the reference run. k_F, k_L: retained factors (reference run: 2 and 4). Jac.: Jaccard overlap of the fold's trait set with the reference run's (31 functional, 107 linguistic). Per factor: Tucker congruence ϕ of the varimax loadings on the traits both runs kept, factors matched by Hungarian assignment; – marks a reference-run factor without a match. Last row: mean over folds (Linguistic F4 over the 11 folds in which it is matched).

			Jac.		Functional ϕ		Linguistic ϕ			
Held-out task	k_F	k_L	F	L	F1	F2	F1	F2	F3	F4
swe	2	6	.87	.92	.99	.98	.89	.93	.92	.61
terminal	2	4	.81	.86	1.00	1.00	.99	.99	.98	.99
tool-api	2	5	.76	.91	.97	.97	.95	.93	.86	.88
web	2	4	.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00
gui	2	5	1.00	1.00	1.00	1.00	.96	.97	.91	.94
office	2	4	.94	1.00	1.00	.99	.95	.99	.91	.99
research-qa	2	5	.88	1.00	1.00	1.00	.97	.98	.92	.93
science	2	4	.71	.77	.98	.99	.98	.94	.89	.76
security	2	4	.84	.96	1.00	1.00	1.00	1.00	1.00	.99
ml-engineering	2	4	.84	.98	1.00	1.00	1.00	1.00	1.00	.99
dialogue	2	3	.91	.94	1.00	1.00	.97	.97	.94	–
other	2	5	.94	.98	1.00	1.00	.94	.97	.89	.93
Mean	2.0	4.4	.87	.94	.99	.99	.97	.97	.93	.91

B.6 Model Traits

Table 14: Factor scores of all models (standardized across models to zero mean and unit STD), grouped by model family. Cells are shaded by the score, blue for positive and orange for negative, with full saturation at $|z| = 3$.

Model	Fun_F1	Fun_F2	Lin_F1	Lin_F2	Lin_F3	Lin_F4
alibaba/qwen2.5-72b-instruct	0.05	1.46	2.02	0.87	0.46	0.09
alibaba/qwen3-coder-480b-a35b-instruct	0.95	-0.82	-0.41	-0.18	1.28	0.29
alibaba/qwen3.5-27b	0.68	-0.41	0.41	-1.18	0.23	-0.75
alibaba/qwen3.5-35b-a3b	1.57	-0.42	0.31	-0.94	-0.06	-1.21
alibaba/qwen3.5-397b-a17b	1.10	-0.49	0.38	-1.00	-0.13	-0.74
alibaba/qwen3.6-27b	0.53	-0.32	0.68	-1.54	-0.14	-0.28
alibaba/qwen3.6-35b-a3b	0.77	-0.09	0.57	-1.16	-0.86	-0.42
alibaba/qwen3.6-plus	0.76	-0.08	0.52	-1.06	-0.75	-0.44
alibaba/qwen3.7-plus	1.08	-0.18	0.23	-1.48	-0.52	0.04
anthropic/claude-3-5-sonnet-20240620	0.97	-0.45	1.06	4.33	-0.24	-1.34
anthropic/claude-3-5-sonnet-20241022	-0.96	0.95	0.38	1.67	0.33	-0.23
anthropic/claude-3-7-sonnet-20250219	-0.13	0.07	0.37	0.84	1.09	-0.50
anthropic/claude-haiku-4-5-20251001	0.15	-1.03	-0.43	-0.12	1.17	-0.60
anthropic/claude-opus-4-1-20250805	-0.10	-0.60	-0.65	0.14	1.22	-0.62
anthropic/claude-opus-4-20250514	-0.43	0.39	0.34	0.04	1.39	-0.32
anthropic/claude-opus-4-5	-0.69	-0.17	0.03	-0.48	-0.15	2.38
anthropic/claude-opus-4-5-20251101	-0.03	-0.61	-0.57	-0.16	0.55	-0.66
anthropic/claude-opus-4-6	-0.34	0.27	-1.17	-0.68	-0.10	-0.11
anthropic/claude-opus-4-7	0.53	-0.68	0.63	-1.23	-1.00	-1.34
anthropic/claude-opus-4-8	1.68	-0.45	0.39	-0.78	-1.46	-0.47
anthropic/claude-sonnet-4-20250514	0.13	0.25	0.24	-0.18	1.50	-0.47
anthropic/claude-sonnet-4-5-20250929	0.19	-0.74	-0.57	-0.01	1.18	-0.51
anthropic/claude-sonnet-4-6	1.98	-1.07	0.46	-1.30	-0.79	-0.96
deepseek/deepseek-r1	-1.15	0.56	0.30	0.48	-0.21	0.87
deepseek/deepseek-v2.5	-0.09	1.71	2.25	1.26	0.50	-0.02
deepseek/deepseek-v3	-0.73	-0.04	-0.34	0.44	0.61	-0.49
deepseek/deepseek-v3-0324	-1.19	-0.54	-0.74	0.06	0.72	0.37
deepseek/deepseek-v3.1	-0.53	0.04	-0.83	-0.16	0.85	-0.37
deepseek/deepseek-v3.2	0.16	-0.14	0.41	0.44	-0.51	3.49
deepseek/deepseek-v3.2-exp	0.92	-1.26	-0.13	-0.46	1.92	0.83
deepseek/deepseek-v4-flash	0.65	0.23	0.63	-1.23	-0.60	-0.30
deepseek/deepseek-v4-pro	1.29	0.38	0.63	-1.23	-1.05	-0.29
google/gemini-2.0-flash	-0.16	-0.46	0.17	0.06	1.04	0.03
google/gemini-2.5-flash	-0.98	-1.53	-0.37	0.22	0.72	0.20
google/gemini-2.5-pro	0.06	-1.41	-0.20	-0.03	0.87	-0.07
google/gemini-3-flash-preview	-0.19	-0.45	0.10	-0.49	0.61	0.94
google/gemini-3-pro-preview	-1.20	-0.71	-0.33	-0.27	-0.41	4.18
google/gemini-3.1-pro-preview	-0.49	1.55	-0.20	-0.41	0.64	-0.55
meta/llama-3.1-405b-instruct	0.17	1.96	2.87	-0.01	1.37	0.71
meta/llama-3.1-70b-instruct	-0.91	3.13	2.64	-0.33	1.34	0.14
meta/llama-3.3-70b-instruct	0.67	1.68	2.57	-0.66	1.37	0.73
minimax/minimax-m2	0.90	-0.18	-1.10	-0.59	1.55	-0.17
minimax/minimax-m2.1	-0.12	-0.57	-1.42	0.29	1.66	-1.50
minimax/minimax-m2.5	0.31	1.31	-0.69	-0.58	1.08	0.35
minimax/minimax-m2.7	1.02	2.57	0.80	-1.28	-0.82	-0.10
minimax/minimax-m3	1.13	0.35	0.79	-1.13	-1.22	-0.31

Table Continued

Model	Fun_F1	Fun_F2	Lin_F1	Lin_F2	Lin_F3	Lin_F4
moonshot/kimi-k2-0905-preview	0.98	-1.08	-0.55	-0.36	1.32	-0.06
moonshot/kimi-k2-thinking	0.95	-0.79	-0.80	-0.46	0.95	0.13
moonshot/kimi-k2.5	0.94	0.46	-0.27	-0.62	-0.37	1.63
moonshot/kimi-k2.6	0.32	0.50	0.44	-0.19	-1.58	-0.73
moonshot/kimi-k2.7-code	0.46	3.22	0.77	-0.45	-1.43	0.88
moonshot/kimi-k3	1.72	-0.29	-0.46	-0.57	-1.97	-1.00
openai/gpt-4.1-2025-04-14	-1.10	-0.05	-0.16	0.89	0.03	-0.40
openai/gpt-4.1-mini-2025-04-14	-1.18	-0.31	-0.69	1.07	-0.44	-0.34
openai/gpt-4o-2024-05-13	0.73	2.20	3.17	3.23	-0.03	-0.60
openai/gpt-5-2025-08-07	-1.25	-0.96	-1.12	0.58	-0.56	-0.10
openai/gpt-5-codex	-1.14	0.31	-1.86	0.75	0.36	-0.09
openai/gpt-5-mini-2025-08-07	-1.46	-1.44	-1.22	0.62	-0.37	-0.10
openai/gpt-5-nano-2025-08-07	-1.15	-1.51	-0.93	0.87	-0.97	-0.46
openai/gpt-5.1	-2.65	-1.16	-0.83	0.95	-1.24	0.67
openai/gpt-5.1-codex	-0.81	1.05	-1.50	0.72	0.00	0.63
openai/gpt-5.2	-1.62	-0.64	-0.82	0.65	-1.28	1.56
openai/gpt-5.3-codex	-0.89	0.17	-0.69	0.75	-0.65	0.48
openai/gpt-5.5	0.71	0.08	0.73	-0.01	-1.85	-0.60
openai/gpt-5.6-sol	0.86	-1.16	-1.39	0.93	-2.22	-1.85
openai/gpt-oss-120b	-1.28	-0.31	-0.71	1.34	-0.73	0.41
openai/gpt-oss-20b	-1.71	0.62	-0.21	1.57	-1.43	2.82
openai/o3-2025-04-16	-1.65	-0.27	-0.64	0.55	-0.10	-0.36
openai/o3-mini	-1.32	0.00	0.03	1.19	-0.02	-0.69
openai/o4-mini-2025-04-16	-1.75	-0.02	-0.40	0.66	0.30	-0.48
xai/grok-4-0709	-0.91	-0.80	-0.99	0.39	-0.08	0.66
xai/grok-code-fast-1	-0.69	-0.46	-0.95	0.26	-0.19	0.73
xiaomi/mimo-v2.5-pro	0.97	-0.29	0.42	-0.98	-0.85	-0.70
zhipu/glm-4.6	1.11	-0.71	-0.56	-0.48	1.58	-0.04
zhipu/glm-4.7	0.22	0.60	-0.84	-0.44	-0.13	-0.48
zhipu/glm-5	0.78	0.04	-0.94	-0.35	0.50	0.31
zhipu/glm-5.1	0.74	-0.02	0.43	-0.83	-1.40	-0.84
zhipu/glm-5.2	1.10	0.09	0.54	-1.01	-1.38	-0.47