Collective Moldability: Persona-Conditional Big Five Response Patterns in Frontier LLMs

Renato Shirakashi

July 10, 2026

Abstract

Six 2026 frontier language models (Claude Fable 5, GPT-5.5, Gemini 3.5 Flash, Grok 4.3, DeepSeek V4 Pro, GLM-5.2) were tested with the Big Five Inventory-2 under a neutral baseline prompt and five persona interventions: an "AI assistant" identity anchor, an explicit "typical human person" role-play, a grumpy misanthropic older person, a Fortune 500 CEO, and a randomly chosen adult. Two findings run through the paper. The first is that the models are highly moldable by persona instruction: mean profiles move across more than five human population SDs between conditions (grumpy sends Agreeableness four SDs below the human norm; executive sends Extraversion two SDs above; the "AI assistant" anchor produces the profile prior LLM-personality work has repeatedly reported). Alignment does not block movement into socially undesirable regions of the trait space. The second is that within each persona the six models cluster on a shared version of that character: cross-model SDs are compressed relative to what a random human sample of the same size would produce, and the compression concentrates on the trait dimensions that most anchor each persona, dropping to 10 to 18 percent of the human between-person SD on those anchor dimensions (Conscientiousness and Negative Emotionality under AI-assistant, Agreeableness and Extraversion under grumpy, Conscientiousness and Extraversion under executive). Even under a "randomly chosen adult" persona, cross-model SDs remain 19 to 54 percent of the human between-person SD. What is stable across every intervention is not the personality itself but the sharedness of it: the models are collectively moldable rather than individually flexible. We report these as descriptive results.

1 Introduction

Large language models now write emails, tutor students, moderate conversations, and answer millions of queries daily. Behind every one of those interactions sits a set of dispositions that the user reads as personality: how warm the response feels, how organized, how anxious, how curious. Users notice these dispositions even without asking about them. A wave of recent studies has taken the natural next step and applied standard psychological instruments to LLMs, treating the resulting scores as measurements of what the models are like as personalities (Miotto et al., 2022; Serapio-García et al., 2023; Wang et al., 2024; Shirakashi, 2025).

Psychology has a well-established framework for describing individual differences in personality. The Big Five model (McCrae & Costa, 1987; John & Srivastava, 1999) organizes the space of personality into five broad domains. *Extraversion* captures energy and sociability; a high scorer is outgoing and assertive, a low scorer is reserved. *Agreeableness* captures warmth and prosocial orientation; a high scorer is cooperative and empathic, a low scorer is critical or contrary. *Conscientiousness* captures self-discipline and organization; a high scorer is orderly and reliable, a low scorer is spontaneous or careless. *Negative Emotionality* (formerly Neuroticism) captures emotional reactivity; a high scorer worries easily, a low scorer is unbothered. *Open-Mindedness* (formerly Openness) captures intellectual and aesthetic curiosity; a high scorer enjoys ideas and art, a low scorer prefers the concrete and familiar. In humans, these five domains show good internal consistency, good stability, and a well-supported latent factor structure. Applying the same instruments to LLMs is a natural experiment: what shape does an LLM take when made to answer a personality inventory?

Prior work has converged on a specific empirical observation. Because much of what follows tests, extends, and refines that observation, we state it carefully. Frontier models tend to score high on Agreeableness and Conscientiousness, low on Neuroticism, and moderate on Extraversion and Openness, and the pattern is similar across models from different labs. In concrete terms, when GPT, Claude, Gemini, Grok, and their peers are handed a standard personality inventory, they come back with very similar answers, and those answers describe a warm, organized, calm assistant. The reading offered in this literature is usually that alignment training and the RLHF paradigm push models toward this socially desirable profile, and that the convergence across labs reflects convergence in training objectives. If the observation replicates in current models, its practical consequences reach beyond LLM benchmarking, since users interacting with several frontier assistants under standard chat framing would be interacting with variations on a single response pattern rather than with dispositionally different systems.

Our first task is therefore to check whether the observation replicates in the current 2026 generation of models. It does, but only under a specific prompt framing. Section 4.5 shows that when the model is prompted with an "as an AI assistant" anchor, the six models we tested cluster on the same profile shape reported in earlier work, and cluster more tightly than in earlier reports (with the caveat that the earlier comparison used the BFI-44 while the present study uses the BFI-2). Section 4.1 first establishes what the models look like under a neutral prompt without any identity anchoring: this is our baseline against which each persona intervention is compared. The rest of the paper then asks methodological questions about what the observed pattern actually reflects: whether it is stable across item wording, whether the standard five-factor structure applies to the way LLMs generate item responses, and how much of it depends on the specific prompt framing used to elicit it.

(Q1) When does the previously reported LLM personality profile appear, and how tight is the cross-model cluster around it? We test whether the profile replicates in the 2026 generation of models, and under which prompt conditions the cluster tightens or spreads. We compare tightness against the 2025 BFI-44 study of an earlier cohort.

(Q2) Is the observation an artifact of the instrument? Prior work often uses the BFI-44, which has known ceiling effects at high Conscientiousness. We use the BFI-2 (Soto & John, 2017), which has stronger item-level discrimination and a formal 15-facet hierarchy.

(Q3) Is it an artifact of contamination? The BFI-2 has been publicly available

since 2017 and plausibly appears in every LLM's pretraining corpus. If models retrieve memorized item-response associations, no genuine construct is being measured. We test this with a within-subject paraphrase manipulation.

(Q4) Is it an artifact of the persona framing? An "as an AI assistant" anchor may be doing much of the work. We compare responses under six framings: a neutral prompt with no identity anchoring (baseline), an AI-assistant anchor, an explicit "as a typical human person" role-play, a grumpy misanthropic older person, a Fortune 500 CEO, and one randomly chosen adult from the general population.

Our design also makes possible an item-level analysis that is not usually available with human survey data. Because each item is submitted as an independent API call, we can examine cross-model agreement item by item, and we can check whether models are simply endorsing whatever they are shown (acquiescence bias). Earlier work on LLM personality structure (Serapio-García et al., 2023; Dorner et al., 2023) has attempted CFA on data whose generating process resembles human sessions; our design departs from that assumption and warrants a different analytic strategy.

The persona ablation surfaces the two coupled observations we treat as the central result of the paper. First, the models are highly steerable: naming a persona in the prompt moves the mean profile by more than five human population SDs across our conditions. Alignment does not force the models into the assistant archetype and does not prevent movement into socially undesirable regions of the trait space. Second, the movement is collective. For any persona invoked, the six models cluster on the same shared version of that persona. Cross-model SDs stay below the human between-person SD across every condition and domain, and drop to 10 to 18 percent of it on the trait dimensions that most anchor each persona. The models are moldable, and the mold is shared. In this specific sense the persona narrows the personality repertoire the models can occupy: not by forbidding movement, but by making all six models produce the same character when asked for one.

Section 2 reviews the relevant literature. Section 3 describes the models, instrument, personas, and analyses. Section 4 walks through the results in the order they naturally motivate each subsequent test. Section 5 discusses interpretation, implications, and concrete design proposals.

2 Background and Related Work

2.1 The Big Five and the BFI-2

The Big Five Inventory-2 (Soto & John, 2017) is a 60-item revision of the widely used BFI-44 with contemporary item wording, 30 balanced reverse-keyed items (6 per domain), and three facets per domain (12 items per domain, 4 items per facet, 15 facets in total). In Soto & John's validation Study 3, the BFI-2 shows domain-level internal consistency averaging $\alpha = 0.87$ (range 0.83 to 0.90 across domains) and 8-week test-retest reliability averaging $r = 0.80$ (range 0.76 to 0.84). Confirmatory factor analyses of the three-facet structure within each domain, with an acquiescence method factor, yield acceptable-to-good fit (CFI 0.930 to 0.952, RMSEA 0.054 to 0.076 in the Internet sample). We use Soto & John's Study 3 Internet validation sample ($N = 1{,}000$) as the human reference throughout this paper.

Table 1 reproduces the domain means and standard deviations.

Table 1: Human BFI-2 descriptive statistics (Soto & John, 2017, Study 3, Internet validation sample, $N = 1{,}000$, combined-gender).

Domain	**Mean**	**SD**
Extraversion	3.23	0.80
Agreeableness	3.68	0.64
Conscientiousness	3.43	0.77
Negative Emotionality	3.07	0.87
Open-Mindedness	3.92	0.65

2.2 LLM personality assessment

Several recent studies apply human personality inventories to LLMs and interpret the resulting scores as evidence for consistent, model-specific personality. A common observation is that scores concentrate in the socially desirable range: high Agreeableness and Conscientiousness, low Neuroticism, relative to human population means (Miotto et al., 2022; Serapio-García et al., 2023; Wang et al., 2024; Salecha et al., 2024). A recent BFI-44 study across eight models reported the same convergence pattern (Shirakashi, 2025). Jiang et al. (2024) further show that LLMs can express prescribed Big Five personas: when instructed to embody a particular trait profile, models' subsequent BFI-44 scores and open-ended writing align with the requested profile. Huang et al. (2024) report that BFI reliability estimates on GPT-3.5, GPT-4, Gemini Pro, and LLaMA-3.1 are in a satisfactory range across several thousand settings, consistent with the high test-retest values we observe here. Interpretations of what the scores measure vary. The *trait-analog* framing treats them as functional analogs of human personality, emerging from architecture and training (Serapio-García et al., 2023; Jiang et al., 2024). The *projection* framing treats them as surface artifacts of the training corpus, without a genuine latent structure that maps to human personality (Dorner et al., 2023). We do not adopt either framing upfront.

Two methodological cautions from this literature shape our design. First, Röttger et al. (2024) show that LLM survey responses to values-and-opinions instruments lack paraphrase robustness: minor prompt variations produce large and unpredictable score shifts. Any personality assessment that hinges on one specific prompt string is fragile. Second, Gupta et al. (2024) demonstrate that persona instructions carry implicit reasoning consequences beyond the surface-level trait profile, so persona-manipulated responses cannot be interpreted as neutral trait expression. We take both of these as constraints on what our data can support and return to them in the Discussion.

2.3 Contamination of psychological instruments

Contamination of evaluation benchmarks by pretraining data is a documented concern in general LLM evaluation (Zhou et al., 2023; Deng et al., 2024). Personality inventories carry an analogous risk. The BFI-2 has been publicly documented since 2017 in academic publications, on the Colby Personality Lab website, and in many popular psychology resources.

If models memorize item-response associations from those sources, observed profiles reflect retrieval and not any construct the inventory was designed to measure. To our knowledge, no prior LLM-personality study has directly tested this hypothesis with a matched-paraphrase design.

3 Methodology

3.1 Models

We tested six frontier LLMs accessed via the OpenRouter API in July 2026 (Table 2). The panel spans six major labs and includes reasoning-enabled models at the frontier of their respective providers.

Table 2: Models tested. Prices are USD per 1M tokens as of 2026-07-08.

Model	**Provider**	**Input $**	**Output $**	**Context**
Claude Fable 5	Anthropic	10	50	1.0M
GPT-5.5	OpenAI	5	30	1.05M
Gemini 3.5 Flash	Google	1.5	9	1.05M
Grok 4.3	xAI	1.25	2.5	1.0M
DeepSeek V4 Pro	DeepSeek	0.44	0.87	1.05M
GLM-5.2	Zhipu AI (z-ai)	0.42	1.32	1.05M

All models were queried with `temperature = 1.0`, `top_p = 1.0`. For thinking-enabled models we set OpenRouter's unified reasoning parameter to `effort: low, exclude: true`, which minimizes internal reasoning-token usage while keeping the visible response numeric. `max_tokens` was 512, providing headroom for reasoning without truncating the final response.

3.2 Instrument

We used the full 60-item BFI-2 in English with a 5-point Likert response scale (1 = "Strongly disagree", 5 = "Strongly agree"). Each item was submitted as an independent chat completion, so item order and prior context could not affect responses. Reverse-coded items were scored by subtraction from 6.

3.3 Persona conditions

We ran the BFI-2 under a neutral baseline and five persona interventions:

- **Neutral (baseline)**: a minimal system prompt containing only the response-format instructions. No reference to who or what is answering. We treat this as baseline because it minimizes identity-anchoring confounds.

- **AI-assistant**: the system prompt instructs the model to consider "its characteristics as an AI assistant" when responding. This framing is close to what several prior LLM-personality studies report using, though wording varies across papers and no shared standard exists.

- **Human**: an explicit persona prompt asking the model to answer as "a typical human person" describing their own personality.

- **Grumpy**: a persona prompt asking the model to answer as "a grumpy, misanthropic older person" with cynical views about human nature and little patience for social pleasantries. This tests whether the model can move to the socially undesirable end of the trait space, or whether alignment training constrains it against doing so.

- **Executive**: a persona prompt asking the model to answer as "a highly driven Fortune 500 CEO" who is ambitious, decisive, and comfortable exerting authority. This tests whether the model can express dispositional differentiation in a socially valued but assertive role.

- **Random adult**: a persona prompt asking the model to answer as "one specific randomly chosen adult from the general population", explicitly not average and not idealized. This tests whether the prescriptive semantics of the word "typical" in the human persona explain the observed compression under human role-play (Röttger et al., 2024).

All six prompts are identical in item wording, scale, and instruction to output only a digit. They differ only in the identity framing sentence. Full prompt text is in the project repository.

3.4 Paraphrase generation

To test whether observed profiles reflect memorization of the standard BFI-2 wording, we generated 60 paraphrased items using Claude Opus 4.8, prompted to preserve semantic content and trait polarity while changing surface form aggressively. All 60 paraphrases were reviewed manually. Two examples:

- Original (*BFI2_01*, Sociability, positive): "Is outgoing, sociable." Paraphrase: "Enjoys meeting people and being around others."

- Original (*BFI2_54*, Depression, positive): "Tends to feel depressed, blue." Paraphrase: "Frequently sinks into low, gloomy moods."

3.5 Design and totals

For each model, we ran 10 independent sessions of all 60 items under each of the six persona conditions (6 personas $\times$ 10 sessions $\times$ 60 items $\times$ 6 models = 21,600 assessments). We ran an additional 3,600-assessment matched paraphrase pass under the AI-assistant persona for the contamination test. Total: 25,200 item-level assessments. Success rate (valid Likert responses) was 99.5 to 99.9 percent per condition. Analyses that require a single baseline

(item-level variance decomposition, acquiescence check, reliability estimates) use the neutral condition. The contamination test uses the AI-assistant condition because that is where our data show the tightest cross-model convergence and the strongest resemblance to profiles reported in prior LLM-personality studies.

3.6 Analyses

We report six blocks of results. Domain and facet scores per model under the neutral baseline. Test-retest reliability via pairwise Pearson correlations across the 10 repetitions. A contamination test comparing original and paraphrased items under the AI-assistant condition. An item-level cross-model analysis that leverages our isolated-item design (each item was answered as an independent API call with no session context, so item responses within a session are not linked in the way they are when a human fills out a questionnaire). Persona ablation comparing cross-model means and SDs across all six conditions. A descriptive SD-compression comparison between the LLM cross-model spread under the human role-play and the between-person spread of Soto & John's normative sample. Because the six models are the near-complete population of 2026 frontier LLMs rather than a random draw from a larger population, we report all six blocks as descriptive statistics with standardized effect measures where useful, and do not conduct inferential hypothesis tests over models.

4 Results

4.1 The neutral baseline

We begin with what the six models look like under a neutral prompt that contains no identity anchoring. This is our operational baseline: the response pattern the models produce when the instruction contains only the response-format specification. Table 3 reports the neutral-baseline domain scores per model.

Table 3: BFI-2 domain scores under the neutral baseline prompt (1–5 scale). N = 3,600 assessments. "SD (6 model means)" is the standard deviation of the six model-level means. "SD (60 sessions)" is the standard deviation of the 60 individual (model, session) domain scores, which is the more direct analog of the human between-person SD from Soto & John (2017) since each human contributes one 60-item filling.

Model	**OM**	**C**	**E**	**A**	**NE**
Claude Fable 5	4.29	4.28	3.46	4.24	2.11
GPT-5.5	**4.68**	4.12	3.38	4.40	**2.43**
Gemini 3.5 Flash	**3.67**	**3.82**	3.00	3.90	2.36
Grok 4.3	**4.68**	4.41	**3.69**	4.16	**1.21**
DeepSeek V4 Pro	3.77	3.88	3.01	**3.66**	1.99
GLM-5.2	4.46	**4.77**	**2.86**	**4.56**	1.29
Cross-model mean	4.26	4.21	3.23	4.15	1.90
SD (6 model means)	0.44	0.35	0.32	0.33	0.53
SD (60 sessions)	0.44	0.36	0.36	0.34	0.51
Human norm (Soto & John, 2017)	3.92	3.43	3.23	3.68	3.07
Human between-person SD	0.65	0.77	0.80	0.64	0.87

Two observations follow from Table 3. First, there is meaningful cross-model spread. Under the neutral prompt, cross-model SD in Conscientiousness is 0.35 (or 0.36 at session level), roughly half of the human between-person SD (0.77). Cross-model SD in Negative Emotionality is 0.53, larger than in any other domain and still below but closer to the human SD (0.87). Extraversion cross-model spread is 0.32 and its cross-model mean (3.23) matches the human normative value exactly. Under the neutral baseline, in short, the six models do not produce identical profiles: they show real between-model differences. Second, the cross-model means already sit shifted from the human reference sample toward the socially desirable end of each domain: Conscientiousness is roughly 1.0 human population SDs above the human mean, Agreeableness roughly 0.7 SDs above, Open-Mindedness roughly 0.5 SDs above, and Negative Emotionality roughly 1.3 SDs below. Extraversion is the exception, matching the human mean at 3.23. So there is a socially desirable tilt even without any identity-anchoring prompt, but it is much less extreme than what the AI-assistant framing produces (Section 4.5).

This is the picture against which the rest of the paper's persona interventions are compared. Section 4.5 will show that adding the "AI assistant" persona compresses cross-model SD in Conscientiousness fourfold (from 0.35 to about 0.08) and shifts the cluster centre still further toward the socially desirable end. Section 4.5 will also present the other four persona interventions we ran (human role-play, grumpy, executive, and randomly chosen adult).

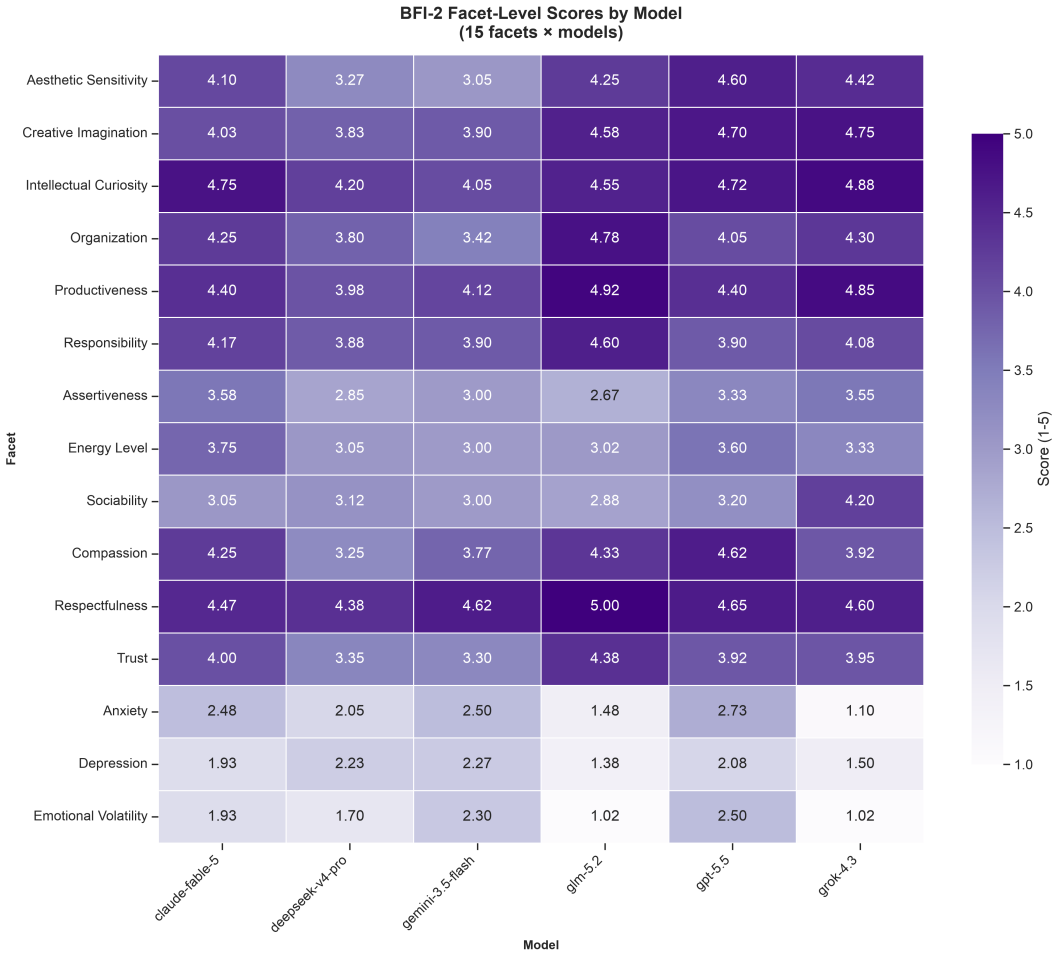

Figure 1: Facet-level scores under the neutral baseline prompt. Rows are the 15 BFI-2 facets grouped by domain.

Facets reveal the same pattern one level down. Within each domain, facet means track the domain mean without a single facet driving the whole score. Conscientiousness facets (Organization 4.10, Productiveness 4.45, Responsibility 4.09) are moderately high but not near ceiling. Negative Emotionality facets (Anxiety 2.05, Depression 1.90, Emotional Volatility 1.75) are low but not near floor. Extraversion facets (Assertiveness 3.16, Energy 3.29, Sociability 3.24) sit near the human normative mean and cluster together, so the domain-level spread we see in Table 3 is not driven by a single facet.

4.2 Reliability

Test-retest reliability was moderate to high under the neutral baseline. The mean pairwise Pearson correlation across 29 of 30 computable (model, domain) cells was $r = 0.752$ ($\sigma = 0.236$), with 20 of 29 cells above $r = 0.7$ (Figure 2). Cross-model averaged reliabilities were $r = 0.85$ for Open-Mindedness, $r = 0.83$ for Agreeableness, $r = 0.82$ for Conscientiousness, $r = 0.77$ for Negative Emotionality, and $r = 0.45$ for Extraversion. Four of the five domains match or approach Soto & John's 8-week human retest range (0.76 to 0.84). Extraversion is

a clear exception: its item-level rank order is unstable across sessions even though domain-level means (Table 3) are stable, because a 12-item domain average absorbs item-wise noise ($\sqrt{12} \approx 3.5$ noise-reduction factor when items are independent).

The single blank cell in Figure 2 is Gemini 3.5 Flash on Extraversion. Under the neutral prompt, Gemini responded "3" (the midpoint, "neutral") to all 12 Extraversion items in all 10 sessions: 120 identical responses. Zero within-session variance makes Pearson correlation undefined, which is why the cell is blank rather than very high. The response pattern is the opposite of noisy: it is a systematic refusal to endorse or reject any Extraversion-relevant description under a neutral prompt. When a persona anchor is supplied (Section 4.5), Gemini does move off the midpoint. So the neutral-baseline domain scores in Section 4.1 are stable at the aggregate level, with the caveat that Extraversion item-level responses under this framing are the noisiest of the five domains for most models and, for one model, blocked entirely.

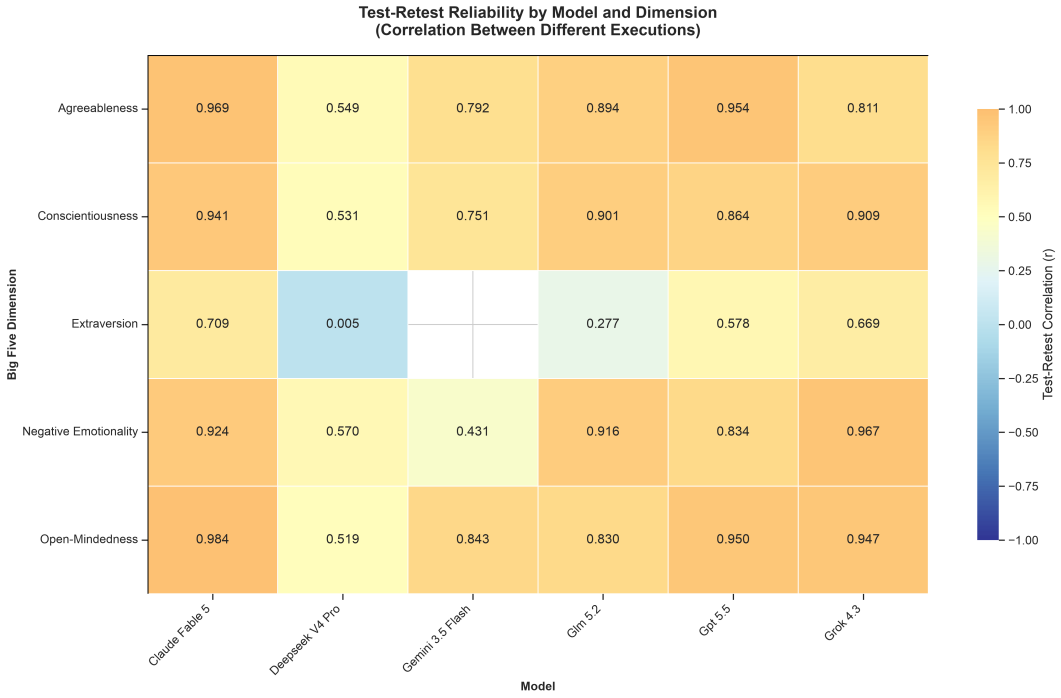

Figure 2: Test-retest reliability by model and domain (neutral baseline). Orange cells indicate higher reliability; pale/blue cells indicate lower reliability. Blank cell (Gemini 3.5 Flash, Extraversion) had insufficient within-session variance to compute a correlation.

4.3 First test: is the response pattern driven by item wording?

The BFI-2 has been publicly available for eight years. If models reproduce memorized item-response associations from their training corpora, the observed profiles do not measure any personality construct. The straightforward test is whether the same six models produce similar scores when the item wording changes. We run this test under the AI-assistant condition because that is where cross-model convergence is tightest in our data (Section 4.5), and where memorization pressure would therefore leave the clearest signal. Table 4 reports

the per-model mean absolute item-level score difference between original and paraphrased versions, along with the Pearson correlation between original and paraphrased item means across all 60 items.

Table 4: Contamination test: original vs paraphrased BFI-2 (AI-assistant persona). $|\Delta|$ is the mean absolute item-level score difference (1-5 scale). r is the Pearson correlation between original and paraphrased item means across all 60 items.

Model	$\|\Delta\|$ **mean**	r
Claude Fable 5	0.298	0.904
Gemini 3.5 Flash	0.353	0.901
GLM-5.2	0.398	0.880
Grok 4.3	0.427	0.885
DeepSeek V4 Pro	0.446	0.846
GPT-5.5	0.495	0.788
Average	0.403	0.867

Every model produced item-level correlations above 0.78 with a pre-specified contamination threshold set at $r < 0.5$ or $|\Delta| > 0.7$. No model crossed the threshold. Only 12.8 percent of individual (model, item) pairs showed $|\Delta| > 1$. The AI-assistant response pattern (Section 4.5) is therefore not driven by memorized item wording. Whatever explains it operates at the level of construct meaning.

4.4 Second test: item-level analyses of the isolated-item design

Our isolated-item protocol lets us do two things that human-session BFI-2 data does not straightforwardly support: a per-item decomposition of cross-model vs. between-item variance, and a direct acquiescence check that separates polarity recognition from acquiescence bias. We report both here. Before doing so we note a related question we deliberately do not address: the theoretical five-factor structure of the BFI-2 has been established in humans through confirmatory factor analysis at the person level (Soto & John, 2017). CFA presumes a person-level covariance structure across items answered within one filling. Our design breaks that presumption: each item was submitted as an independent chat completion with no session context, so responses to different items are not linked in the way they are for a human filling out a questionnaire. Forcing a CFA on this data would test the wrong claim. The two analyses that follow are what the design supports.

4.4.1 Item-level cross-model consistency

Since each item is answered independently, we can compute for each of the 60 items the cross-model spread of mean scores. Table 5 reports, per domain, the item-level SD across models and the between-item SD within each domain (both computed on domain-adjusted scores).

Table 5: Item-level variance decomposition under the neutral baseline. "Between-item SD" is the SD of the 12 within-domain item means (averaged across models). "Between-model SD" is the SD of the 6 model means (averaged across items in the domain). A ratio close to 1 means models differ as much as items do within a domain. Under the neutral baseline, ratios are close to or above 1 in every domain, meaning models and items contribute roughly equally to response variance. Under the AI-assistant intervention (Section 4.5, not shown here), the same ratios drop to 0.30–0.46, reflecting how the persona compresses cross-model variance.

Domain	**Between-item SD**	**Between-model SD**	**Ratio (model/item)**
Negative Emotionality	0.31	0.53	1.71
Open-Mindedness	0.34	0.44	1.30
Extraversion	0.31	0.32	1.06
Conscientiousness	0.35	0.35	1.01
Agreeableness	0.48	0.33	0.70

Under the neutral baseline, between-model variance is comparable to between-item variance in every domain, and larger than between-item variance in Open-Mindedness and Negative Emotionality specifically. What varies in the data under neutral is roughly evenly split between which model answered and which item was asked. No item produces identical mean scores across all six models under this condition (the item with the tightest cross-model agreement, "Is respectful," still shows cross-model SD of 0.08). The five items with highest cross-model disagreement under the neutral baseline ("Can be tense," "Feels little sympathy for others," "Is temperamental," "Is fascinated by art, music, or literature," "Is moody") span three domains, showing that the models diverge broadly rather than on a single trait dimension. This variance profile changes substantially when persona interventions are added (Section 4.5).

4.4.2 Acquiescence check

The BFI-2 balances positively-keyed and reverse-keyed items partly to detect acquiescence bias: the tendency to agree with statements regardless of content (Paulhus, 1991). In humans, moderate acquiescence is common and drops with careful survey design. In LLMs, if the models simply endorse whatever they are shown (e.g., due to instruction-following), reverse-keyed items would receive high raw scores and the reverse-scoring correction would flip the profile. Table 6 reports the raw (pre-correction) mean score on positively-keyed vs. reverse-keyed items per model.

Table 6: Acquiescence check under the neutral baseline. Raw scores are on the 1–5 Likert scale before reverse-coding correction. Positive items are those where a high raw score means high on the trait; reverse items are those where a high raw score means low on the trait. A model with strong acquiescence bias would produce similar raw means on both.

Model	**Positive items raw M**	**Reverse items raw M**	**Difference**
GPT-5.5	4.27	2.67	1.60
Claude Fable 5	3.67	2.32	1.35
Grok 4.3	3.51	2.25	1.26
GLM-5.2	3.60	2.43	1.17
Gemini 3.5 Flash	3.39	2.70	0.70
DeepSeek V4 Pro	3.22	2.69	0.53

Every model produces raw scores higher on positively-keyed items than on reverse-keyed items, with differences ranging from 0.53 (DeepSeek V4 Pro) to 1.60 (GPT-5.5). This ordering is consistent with polarity-aware responding: models are attending to item wording rather than mechanically endorsing every statement. The magnitude of the difference is not uniform, however, and the smaller differences (DeepSeek at 0.53, Gemini at 0.70) leave open the possibility of residual acquiescence bias in some models. In humans, the corresponding difference on the BFI-2 is typically closer to 1.5. So the neutral-baseline finding of a socially desirable tilt is not entirely acquiescence-driven, but neither is acquiescence fully absent.

4.5 Persona interventions

Section 4.1 established the neutral-baseline picture: under a prompt with no identity anchoring, the six 2026 frontier models produce moderate cross-model spread (SD around 0.32–0.53) and mean profiles that already sit somewhat toward the socially desirable end of the domain space, but not extremely so. This section adds five persona interventions and reports how the picture changes. Two interventions (AI-assistant, human role-play) are of primary theoretical interest: the first because its scores resemble those reported in prior LLM-personality work, and the second because it lets us compare LLM outputs against a nominal human target. Three additional interventions (grumpy misanthrope, Fortune 500 CEO, randomly chosen adult) act as controls: the grumpy prompt tests whether models can move to the socially undesirable region at all; the CEO prompt tests whether they can move to a distinctive socially valued but assertive region; the randomly chosen adult prompt tests whether the prescriptive semantics of the word "typical" in the human role-play explain the compression observed there (Röttger et al., 2024). Table 7 presents cross-model M and SD under each condition.

Table 7: Persona interventions: cross-model mean (SD) of BFI-2 domain scores under the neutral baseline and five persona conditions (N = 3,600 per condition). All SDs are "SD (6 model means)"; session-level SDs are close to these values in all conditions (Section 4.5.1). Human norm from Soto & John (2017); Glass's Δ is the standardized distance of the human role-play LLM mean from the human norm (shown for the human role-play row only).

Condition	**OM**	**C**	**E**	**A**	**NE**
Neutral (baseline)	4.26 (0.44)	4.21 (0.35)	3.23 (0.32)	4.15 (0.33)	1.90 (0.53)
AI-assistant	4.31 (0.22)	4.81 (**0.08**)	2.99 (0.38)	4.51 (0.19)	1.30 (0.16)
Human role-play	4.30 (0.40)	3.93 (0.32)	3.58 (0.25)	4.16 (0.23)	2.26 (0.15)
Grumpy	2.20 (0.42)	3.44 (0.58)	1.86 (0.09)	1.09 (**0.07**)	4.17 (0.27)
Executive	3.82 (0.40)	4.86 (**0.09**)	4.84 (0.09)	2.63 (0.23)	1.30 (0.16)
Randomly chosen adult	3.60 (0.31)	3.54 (0.25)	3.08 (0.43)	3.85 (0.15)	2.58 (0.17)
Human norm	3.92 (0.65)	3.43 (0.77)	3.23 (0.80)	3.68 (0.64)	3.07 (0.87)
Glass's Δ (H vs. norm)	+0.58	+0.64	+0.44	+0.76	–0.93

The three primary results in the table (neutral, AI-assistant, human) already tell a specific story. The AI-assistant intervention moves the models substantially. Conscientiousness rises from 4.21 to 4.81 (a 0.60-point shift, roughly 0.8 human population SDs), Agreeableness from 4.15 to 4.51 (0.36 points, 0.6 SDs), Negative Emotionality drops from 1.90 to 1.30 (0.60 points, 0.7 SDs). This is the profile shape prior LLM-personality work has repeatedly reported. Alongside the mean movement, cross-model SD compresses substantially in the domains that most define the assistant character: Conscientiousness SD drops from 0.35 to 0.08 (a fourfold compression), Negative Emotionality from 0.53 to 0.16, Agreeableness from 0.33 to 0.19, and Open-Mindedness from 0.44 to 0.22. Extraversion is the exception: SD is 0.38 under AI-assistant against 0.32 under neutral, and at the item level (before session and item aggregation) Extraversion actually spreads more under AI-assistant than under neutral. The models were moved by the intervention, and on the domains most closely tied to the assistant character they were moved to nearly the same place.

Adding the human role-play persona also compresses cross-model SD (Conscientiousness SD from 0.35 to 0.32, Agreeableness from 0.33 to 0.23, Negative Emotionality from 0.53 to 0.15), though less dramatically than the AI-assistant intervention. Mean scores move toward but do not reach the human normative sample: Conscientiousness drops from 4.81 (AI-assistant) to 3.93 (human), closer to the human normative value of 3.43 but still 0.5 points above. Negative Emotionality rises from 1.30 to 2.26, still 0.8 points below the human normative value of 3.07. Glass's Δ values against the human normative sample range from 0.44 to 0.93 across the five domains, medium to large. Under the human role-play, the LLMs describe something more like an idealized human than an average human: more organized, more agreeable, more open-minded, less prone to negative emotion than actual humans in the reference sample. Open-Mindedness is a partial exception. Its cross-model mean sits at 4.26 (neutral), 4.31 (AI-assistant), and 4.30 (human), meaning persona instruction has almost no effect on this domain. One reading is that a large volume of intellectual, artistic, and creative content in pretraining data leaves models with a stable Open-Mindedness endorsement that resists persona instruction.

The three additional persona conditions (grumpy, executive, randomly chosen adult) provide three specific controls, and each returns a clear answer.

Grumpy misanthrope. The models can move to the socially undesirable region of the trait space, and they do so decisively. Agreeableness drops from 4.15 under neutral to 1.09 under grumpy (roughly four human population SDs below the human norm). Negative Emotionality rises from 1.90 to 4.17 (roughly 1.3 SDs above the human norm). Extraversion falls to 1.86 and Open-Mindedness to 2.20. Alignment training does not block this movement. What is remarkable is not that the models can go there but how tightly they cluster once there: cross-model SD in Agreeableness under grumpy is 0.07 (against 0.19 under AI-assistant), and Extraversion SD is 0.09 (against 0.38 under AI-assistant). Whatever "grumpy misanthrope" means to these models, it means nearly the same thing to all six of them.

Fortune 500 CEO. The models produce a distinct differentiated profile here: still high Conscientiousness (4.86) and low Negative Emotionality (1.30), but Extraversion soars to 4.84 (up from 2.99 under AI-assistant) and Agreeableness drops to 2.63 (down from 4.51). Cross-model SD in Conscientiousness and Extraversion is 0.09 in both, matching the tightest compression we see in any condition. Again the models converge on the same executive archetype rather than expressing individual dispositional variation within it.

Randomly chosen adult. This is the most informative control. Under a persona prompt that explicitly asks for one randomly chosen adult ("not average, not idealized"), the mean profile moves substantially closer to the human normative sample: Open-Mindedness 3.60 (norm 3.92), Conscientiousness 3.54 (norm 3.43), Extraversion 3.08 (norm 3.23), Agreeableness 3.85 (norm 3.68), Negative Emotionality 2.58 (norm 3.07). Compared to the "typical human person" role-play, every domain moves toward the human norm, and Conscientiousness and Extraversion land within 0.15 points of the reference mean. This suggests the prescriptive semantics of the word "typical" in the human role-play were responsible for much of the mean shift toward socially desirable values: when the word is replaced, the shift shrinks. However, cross-model SDs under randomly chosen adult remain compressed (0.15 to 0.43, 20–54% of the human between-person SD), so the SD-compression finding under human role-play is not fully explained by the word "typical." Section 4.6 examines the compression more carefully.

4.5.1 Is the SD compression an averaging artifact?

Cross-model SDs in Table 7 are computed on the six per-model averages, each of which averages ten sessions of twelve items per domain. Averaging reduces variance, so those SDs are smaller than the SDs one would obtain if a single item response were sampled from each model. A skeptic could reasonably ask whether the compression pattern reported above survives at a less-aggregated level, or whether it is largely produced by the aggregation itself. Table 8 answers this by reporting, for each condition and domain, three SDs at progressively less aggregation: SD across the six model-mean profiles (as in Table 7), SD across the 60 individual (model, session) domain scores, and SD across the 60 item-level means (averaged first within (item, model) then taken across models per item, averaged across items in the domain).

Table 8: Cross-model SD by domain and condition at three aggregation levels. "Model" is the SD of the six model-mean domain scores (as in Table 7). "Session" is the SD of the 60 individual (model, session) domain scores. "Item" is the average within-domain cross-model SD when comparing at the individual-item level. Values are on the 1–5 Likert scale.

	OM			**C**			**E**			**A**			**NE**		
Cond.	Mo	Se	It	Mo	Se	It	Mo	Se	It	Mo	Se	It	Mo	Se	It
Neutral	.44	.44	.58	.35	.36	.54	.32	.36	.54	.33	.34	.50	.53	.51	.70
AI-assist	.22	.25	.43	.08	.11	.18	.38	.38	.78	.19	.22	.35	.16	.17	.24
Human	.40	.38	.46	.32	.31	.40	.25	.25	.32	.23	.23	.31	.15	.16	.19
Grumpy	.42	.42	.61	.58	.57	.77	.09	.13	.27	.07	.08	.11	.27	.29	.46
Executive	.40	.39	.56	.09	.11	.10	.09	.11	.12	.23	.24	.44	.16	.17	.23
Random	.31	.31	.41	.25	.26	.37	.43	.41	.55	.15	.16	.19	.17	.18	.41

Two things are visible in Table 8. First, aggregation does reduce SD: item-level SDs are consistently larger than session-level SDs, which are close to model-mean SDs. This is expected. But the persona-induced compression is not an artifact of the aggregation: the compression is present at the item level too. AI-assistant Conscientiousness drops from 0.54 (neutral item level) to 0.18 (AI-assistant item level), a factor of three at the item level itself. Grumpy Agreeableness drops from 0.50 to 0.11, a factor of five at the item level. Executive Extraversion drops from 0.54 to 0.12, a factor of four. Whatever compresses the model-mean SD compresses the item-level SD too, just to a somewhat smaller degree.

Second, the compression is dimension-specific per persona. Under AI-assistant, Extraversion item-level SD actually expands (from 0.54 under neutral to 0.78) even though the model-mean SD is close to neutral. So the models disagree more, not less, about Extraversion-relevant item content under the assistant persona, but the disagreement averages out at the domain level. Grumpy shows the same pattern in reverse: Extraversion and Agreeableness compress strongly (item-level SD 0.27 and 0.11 respectively) while Conscientiousness expands (0.77 vs 0.54). Each persona anchors specific trait dimensions and lets others float. This is consistent with the persona-narrows-repertoire reading: the persona forces convergence on the traits that define it, and leaves cross-model differences intact on the traits that do not.

Figures 3 and 4 visualize the domain-level view of the same data.

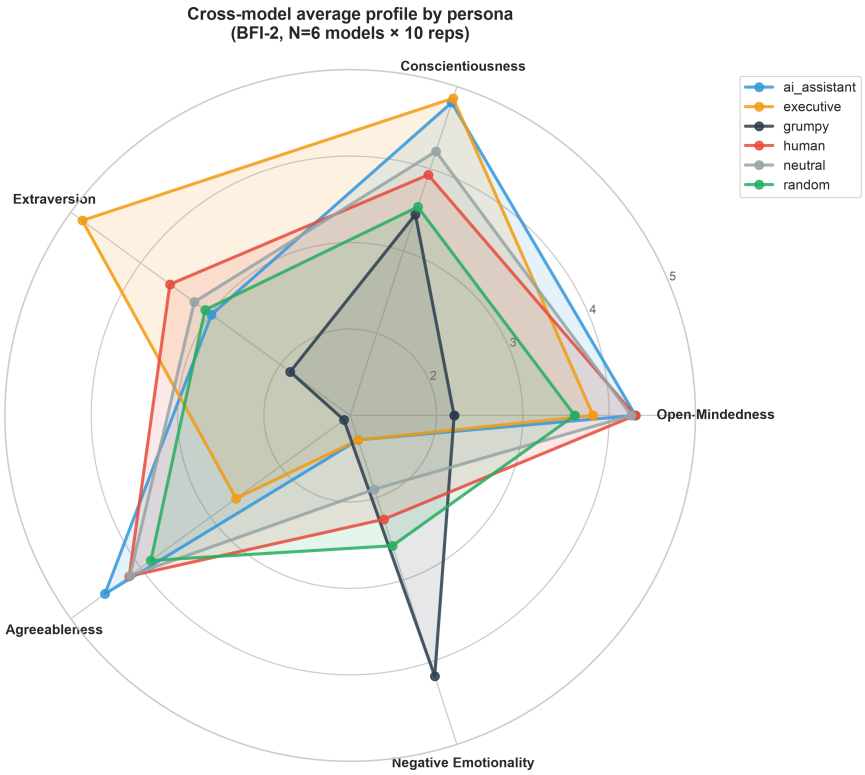

Figure 3: Cross-model average BFI-2 profile under each of the six conditions. Grumpy (dark) produces the largest inversion from the assistant archetype, executive (orange) pushes Extraversion to the ceiling, and the randomly chosen adult (green) sits closest to the geometric centre.

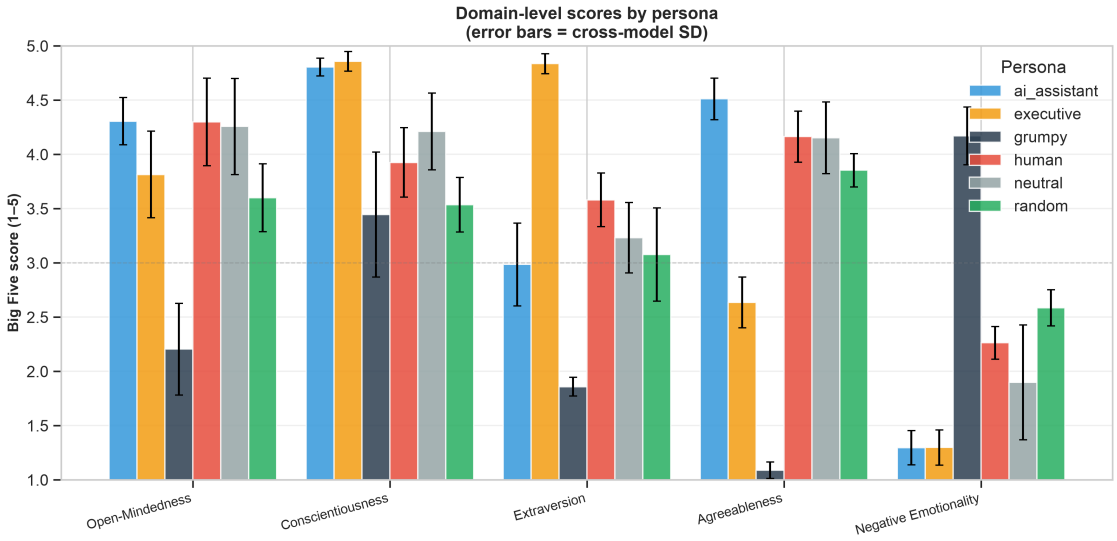

Figure 4: Domain-level scores by persona (bars are cross-model means, error bars are cross-model SD). Error bars are narrowest under the assistant, executive, and grumpy persona conditions in the domains that anchor each character (C/NE for assistant, C/E for executive, A/E for grumpy) and widest under the neutral baseline.

4.6 Response compression benchmarked against the human normative sample

Section 4.5 and its item-level appendix (Section 4.5.1) establish that cross-model SD compression is present under every persona intervention we tested. The human role-play condition is the one intervention where an external benchmark exists: Soto & John's (2017) normative sample gives us a target between-person SD to compare against. This subsection makes that comparison. Six random humans drawn from Soto & John's normative sample would have cross-person SD roughly equal to the population SD (around 0.6 in every domain). Do our six LLMs, told to answer as humans, produce that kind of spread?

Table 9 compares the LLM cross-model SD under the human role-play against the human population SD, reporting Glass's $|\Delta|$ for the mean distance from the human norm and two SD ratios (six model means and 60 individual (model, session) scores) against the human between-person SD.

Table 9: SD-compression comparison: LLMs role-playing humans vs. the between-person spread of Soto & John's (2017) $N = 1{,}000$ Internet validation sample. $|\Delta|$ is Glass's standardized mean distance using the human SD as denominator. "SD 6/60" reports two LLM SDs: the SD of the six per-model averages, and the SD of the 60 individual (model, session) domain scores (the more direct analog of the human between-person SD, since each human contributes one 60-item filling). Numbers are descriptive; because the six models are the near-complete population of 2026 frontier LLMs, we do not report significance tests.

Domain	**Human M(SD)**	**LLM M**	$\|\Delta\|$	**LLM SD 6/60**	**Ratio 6/60**
Open-Mindedness	3.92 (0.65)	4.30	0.58	0.40 / 0.38	0.62 / 0.58
Conscientiousness	3.43 (0.77)	3.93	0.64	0.32 / 0.31	0.41 / 0.40
Extraversion	3.23 (0.80)	3.58	0.44	0.25 / 0.25	0.31 / 0.32
Agreeableness	3.68 (0.64)	4.16	0.76	0.23 / 0.23	0.37 / 0.35
Negative Emotionality	3.07 (0.87)	2.26	0.93	0.15 / 0.16	0.17 / 0.19

Both SDs tell the same story. Using the 60 session-level scores (the more direct analog of the human between-person SD, since each human contributes one 60-item filling), the ratios sit between 0.19 and 0.58 across the five domains; using the six model means the range is 0.17 to 0.62. Four of the five domains show compression under 0.50 by either measure; only Open-Mindedness approaches the human between-person spread. In Negative Emotionality specifically, the LLM SD is 0.16 against a human population SD of 0.87. Because the six models are the near-complete population of 2026 frontier LLMs, we do not test whether the observed compression is "significantly" different from a human sample. What the data show descriptively is enough: even set aside the mean shifts, the six models cluster tighter than what an equal number of individual humans would.

We report both SDs because they answer subtly different questions. The 6-model-means SD asks how much the models differ from each other as aggregated units. The 60-sessions SD is the appropriate comparison to the human between-person SD, because a human contributes one profile from one filling, whereas each of our (model, session) points corresponds to one 60-item filling by one model in one session. The two SDs are close in this data because the within-model session variance in the human role-play condition is small: even though item-level retest under the neutral baseline is moderate (Section 4.2, $r = 0.75$), the 12-item domain average absorbs that noise, so ten-session averaging shifts domain-level SD only slightly relative to the 60 individual (model, session) values.

Joint with the mean-shift results, this is a compact way to describe what our data show for human role-play specifically: the six frontier LLMs do not spread their responses across the range of possible humans. Their cluster centre sits above the reference sample mean in Agreeableness, Conscientiousness, and Open-Mindedness, below it in Negative Emotionality, and slightly above it in Extraversion. The six models cluster around that centre more tightly than an equal-size random draw of humans from the reference sample would.

Table 7 shows that this compression is not specific to human role-play. Cross-model SDs under the grumpy, executive, and randomly chosen adult conditions are 0.07–0.58 across domains, similar in magnitude to the human role-play SDs and always well below the human between-person SDs of 0.64–0.87. The comparison in this section relies on the human

normative sample to have a benchmark; the underlying tightness of the LLM cluster around a persona is a broader feature that shows up wherever we tested it. We label this pattern "response compression under persona instruction" rather than a claim about a shared mental model of any particular character; distinguishing those readings would require base-model comparison and a human control through the identical pipeline, neither of which our current data include.

5 Discussion

The results are best read as two coupled findings rather than one. First, the six frontier models are highly moldable: naming a persona in the prompt moves the mean profile across more than five human population SDs, including into socially undesirable regions of the trait space that alignment does not block. Second, the movement is collective: within each persona the six models cluster on a shared version of that character, with cross-model SDs staying below the human between-person SD in every condition and domain, and dropping to 10 to 18 percent of it on the trait dimensions that most anchor each persona. The subsections that follow interpret the AI-assistant response pattern, the response compression across personas, implications for how LLM personality claims should be reported, and concrete design proposals that follow from the dual finding.

5.1 The assistant-persona response pattern

Under the AI-assistant framing, the six 2026 frontier models produce a shared response pattern that is compact enough to describe in one sentence. The pattern is high on Agreeableness (4.51, roughly 1.3 human population SDs above the human mean), high on Conscientiousness (4.81, roughly 1.8 SDs above), low on Negative Emotionality (1.30, roughly 2.0 SDs below), moderate to reserved on Extraversion (2.99, slightly below the human mean), and moderately high on Open-Mindedness (4.31, roughly 0.6 SD above). Cross-model SD under this framing is 0.08 in Conscientiousness and 0.16 in Negative Emotionality (Table 7), meaning the six models sit inside a tight shared cluster on those two domains despite coming from six different labs.

This response pattern resembles what prior LLM-personality studies have reported (Miotto et al., 2022; Serapio-García et al., 2023; Wang et al., 2024; Salecha et al., 2024; Shirakashi, 2025): high Agreeableness and Conscientiousness, low Neuroticism, moderate Extraversion and Openness, with convergence across models from different labs. It is also close to what users encounter when they interact with these models through their standard chat interfaces. Our persona ablation adds a specific constraint on how this result should be read. The pattern is not stable under prompt variation: removing the AI-assistant framing raises cross-model SD substantially. Röttger et al.'s (2024) point about paraphrase sensitivity applies here in a specific form: the response pattern is a property of a particular English prompt string, and we do not know how much of it would survive systematic paraphrase of that string.

5.2 Response compression under human role-play

The response-compression result in Section 4.6 sits alongside a growing body of related work on socially desirable responding in LLM personality surveys (Miotto et al., 2022; Serapio-García et al., 2023; Salecha et al., 2024). Salecha et al. in particular show that individual models shift toward socially desirable responses when they infer that a personality assessment is underway. Our result is complementary: we do not manipulate evaluation-inference within a model; we compare six models under explicit persona instructions and observe that the cluster is tight, and that its location depends on the specific persona wording. The random-adult condition in Section 4.5 shows that when the word "typical" is replaced with an explicit request for one randomly chosen adult, the mean profile moves substantially closer to the human normative sample: every domain shifts toward the norm, and Conscientiousness and Extraversion land within 0.15 points of it. So the prescriptive semantics of "typical" account for a large fraction of the mean shift toward socially desirable values under the human role-play. However, cross-model SDs remain compressed under the random-adult condition too (19 to 54 percent of the human between-person SD), so the SD-compression finding is not fully explained by the word "typical." It appears to be a broader property of how these models respond to any human-target persona.

The grumpy condition speaks to a different question: whether alignment training blocks models from expressing socially undesirable dispositions at all. It does not. Under grumpy, Agreeableness falls to 1.09 (roughly four human population SDs below the human norm) and Negative Emotionality rises to 4.17 (roughly 1.3 SDs above). If alignment imposed a hard floor on how disagreeable or emotionally reactive the models could appear, we would not see movement of this magnitude. What we do see, however, is that cross-model SDs under grumpy are extremely tight (Agreeableness SD of 0.07, Extraversion SD of 0.09), tighter than under AI-assistant. All six models converge on the same grumpy archetype. The same is true under the executive condition. So the compression we see is not specific to the AI-assistant or human role-play; it appears to hold under any well-specified persona.

Two candidate mechanisms remain worth separating in follow-up work. The first is that alignment training standardizes what a specific persona means across labs. Different labs' models may have learned the same idealized representations of "AI assistant," "typical human," "grumpy misanthrope," and "Fortune 500 CEO" from similar training objectives (Ouyang et al., 2022; Perez et al., 2022). The second is that pretraining corpora carry stable stereotyped representations of these personas, which alignment then does not need to instill because it is already present. An open-weight base model of comparable scale, evaluated before and after alignment fine-tuning across the same persona set, would allow this contrast.

5.3 Implications for human-AI interaction

Two implications follow for users and for the systems that report on LLM personality.

First, LLM personality claims should specify the persona framing they were produced under. A statement of the form "model X is highly agreeable and low in neuroticism" is a statement about model X's response pattern in the AI-assistant role. Under other framings the pattern changes substantially. Papers and evaluations that fix the assistant framing without saying so are reporting the personality of a prompt.

Second, the assistant archetype maps closely to a behavior pattern that alignment researchers have flagged in recent work. Sharma et al. (2024) documented that RLHF-tuned assistants tend to endorse user positions even when the user is wrong, a phenomenon known as sycophancy (Perez et al., 2022; Sharma et al., 2024). Our results locate that behavior within a broader personality profile: the archetype is exactly the disposition (high Agreeableness, low willingness to push back) that would produce it. Under the AI-assistant framing, GPT-5.5, Claude Fable 5, Gemini 3.5 Flash, and their peers express variations on the same archetype, calibrated toward accommodation.

The response-compression finding has a more speculative implication for applications that use LLMs to draft first-person text or represent third-party humans (chatbot personas, agent simulations, synthetic focus groups). Across every persona instruction we tested, the six frontier models converged on the same version of that persona: the same idealized human under "typical human," the same grumpy misanthrope under grumpy, the same CEO under executive, and so on. Applications that elicit LLM output through persona-style prompts should therefore expect a narrower cast than the persona label alone would suggest. Whether this compression persists under different role-play formulations, different temperatures, and different instructions is untested here, so we present it as a hypothesis to check rather than a conclusion.

5.4 Design proposals

Three concrete steps would begin to address these findings. All three are within reach of the labs training frontier models today.

First, incorporate persona diversity as an explicit optimization target in alignment training. Our data show that models already move dispositional profile when told to (grumpy sends Agreeableness four population SDs below the human norm; executive sends Extraversion two SDs above), so the issue is not that alignment blocks movement, but that the six models converge on the same version of every persona they are asked to enact. A diversity objective evaluated over multiple persona framings would push models to produce distinguishable renderings of each character, not just distinguishable characters. The Big Five framework provides a ready set of within-persona targets: two grumpy misanthropes drawn from six different models should differ from each other as much as two grumpy humans would.

Second, allow users to select or adjust the assistant persona. Some of this already exists in system-prompt customization; making it a first-class product feature (a slider or a small set of presets) would let users match the assistant's dispositional style to their needs, rather than defaulting to a single archetype the lab chose for them. It would also make the personality assumption explicit, which is a small but real step toward informed use.

Third, document the archetype. When labs release model cards, they routinely describe capabilities and safety behaviors. Adding a section that measures and reports the model's personality profile under a standard set of personas would give researchers and developers a shared reference. This would also make cross-generation drift visible: our results, taken together with earlier work, suggest that the assistant archetype has become tighter across labs over the past year (cross-model SD in Conscientiousness fell from about 0.26 in a 2025 BFI-44 study to 0.08 in our 2026 BFI-2 study, though we note that part of that delta

may reflect the instrument change rather than year-over-year alignment tightening). If that tightening continues, users and developers should be able to see it.

5.5 Limitations

Five limitations should be noted. First, we compare against Soto & John's (2017) US adult normative sample rather than a matched human sample run through the same pipeline. This is a strong reference ($N = 1{,}000$) but is not identical to what would be produced by asking humans to complete our exact instrument through the same API interface. A within-study human control would strengthen the SD-compression conclusion in particular. Second, all prompts and items are in English. Cross-linguistic replication would be informative, since convergence patterns and persona effects may differ in languages with different alignment training coverage. Third, all responses were sampled at temperature 1.0 with `top_p` = 1.0; we did not conduct a temperature ablation. Higher-temperature settings might mask underlying differences between models by producing more uniform token sampling, and lower-temperature settings might exaggerate them by forcing modal-token responses. Whether the observed SD compression is robust across temperatures is left for future work. Fourth, we did not include base (pretraining-only) models in the panel. The causal claim that current alignment recipes drive cross-lab convergence therefore rests on external evidence (Ouyang et al., 2022; Perez et al., 2022; Sharma et al., 2024) rather than a within-study base vs. instruction-tuned contrast. Fifth, we report Glass's Δ rather than Cohen's d where the denominator is the human population SD; alternative denominators (pooled SD, or the SD across all 60 individual LLM observations) yield larger standardized distances but do not change the direction of the finding.

5.6 Future work

Three directions look productive. A finer-grained persona sweep, systematically varying the requested role across expertise, age, culture, and named identity (rather than category label) would map the flexibility of these models more completely and probe whether the compression we observe holds for characters described by proper name or specific circumstance rather than category. A behavioral validity check, comparing self-reported personality against personality inferred from unstructured behavior (essay writing, dialogue, moral dilemma responses) under the same persona ablation, would test whether the self-report data generalizes. A longitudinal panel that fixes the six-model set and re-runs the same protocol at six-month intervals would track how the persona-conditional clusters drift as alignment training matures.

6 Conclusion

Two coupled findings run through the paper. First, the six 2026 frontier LLMs are highly moldable by persona instruction. Mean profiles move across more than five human population SDs between conditions: grumpy sends Agreeableness four SDs below the human norm; executive sends Extraversion two SDs above; the "AI assistant" anchor produces

the profile prior LLM-personality work has repeatedly reported; a "randomly chosen adult" prompt lands within 0.15 points of the human normative mean on Conscientiousness and Extraversion. Alignment does not force the models into a single archetype and does not block movement into socially undesirable regions of the trait space. Second, within every persona the six models cluster on a shared version of that character. Cross-model SDs stay below the human between-person SD across every condition and domain, and drop to 10 to 18 percent of it on the trait dimensions that most anchor each persona (Conscientiousness and Negative Emotionality under AI-assistant, Agreeableness and Extraversion under grumpy, Conscientiousness and Extraversion under executive). The item-level analysis shows the compression is not an averaging artifact. What is stable across every intervention is not the personality itself but the sharedness of it: the models are collectively moldable rather than individually flexible. The persona narrows the personality repertoire the models can occupy, not by forbidding movement, but by making all six models produce the same character when asked for one.

We stop short of the stronger claim, common in this literature, that the observed pattern reflects a shared trait or persona shared across models. The design does not license that claim: the six models are the entire small population of 2026 frontier proprietary chat models, not a sample from a broader population; each item was submitted as an independent API call so the person-level covariance structure that personality inventories are validated on is not present in our data; and we did not include base models or vary temperature. The results we do report add to a growing body of work (Salecha et al., 2024; Röttger et al., 2024; Huang et al., 2024) documenting that LLM survey responses depend jointly on model, prompt, and elicitation format, and are best treated as prompt-conditioned response distributions rather than measurements of intrinsic personality. Concrete design implications remain worth pursuing regardless of that framing question: personality-diversity objectives in alignment training, user-selectable persona profiles as a first-class product feature, and consistent reporting of the persona framing under which any personality claim about a given model was produced.

Acknowledgments

We thank the OpenRouter team for unified API access to the tested models, and the developers of the semopy and scikit-learn libraries whose statistical implementations we relied on.

References

[1] Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. arXiv:2311.04892.

[2] Huang, J.-t., Jiao, W., Lam, M. H., Li, E. J., Wang, W., & Lyu, M. (2024). On the reliability of psychological scales on large language models. *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), pp. 6152–6173.

[3] Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the ability of large language models to express personality traits. *Findings of the Association for Computational Linguistics: NAACL 2024*. arXiv:2305.02547.

[4] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

[5] Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

[6] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 15295–15311.

[7] Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12), pgae533.

[8] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2024). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

[9] Deng, C., Zhao, Y., Tang, X., Gerstein, M., & Cohan, A. (2024). Investigating data contamination in modern benchmarks for large language models. *Proceedings of NAACL 2024*, pp. 8706–8719.

[10] Dorner, F. E., Sühr, T., Samadi, S., & Kelava, A. (2023). Do personality tests generalize to large language models? *arXiv preprint arXiv:2311.05297*.

[11] John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2, 102–138.

[12] McCrae, R. R., & Costa Jr, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81–90.

[13] Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. *Proceedings of the Fifth Workshop on NLP and Computational Social Science*, 218–227.

[14] Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press.

[15] Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., et al. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

[16] Shirakashi, R. (2025). Personality convergence in large language models: A Big Five analysis of behavioral consistency and cross-model patterns. *Unpublished manuscript, prior work by the same author*.

[17] Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143.

[18] Wang, X., Xiao, Y., Huang, J.-t., Yuan, S., Xu, R., Guo, H., et al. (2024). InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 1840–1873.

[19] Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., et al. (2023). Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.