

First-Stage Groundwork Shapes Downstream RL Outcomes: Two-Stage Post-Training for LLM Personality Acquisition

Taisei Ozaki^{1*}, Tetsuya Hashimoto^{2*}, Miki Ota¹, Zhenyu He¹,
Keisuke Hori¹, Mitsuhiro Mabuchi²

¹Matsuo Institute, ²Toyota Motor Corporation

Correspondence: o.taisei@matsuo-institute.com, tetsuya_hashimoto@mail.toyota.co.jp

Abstract

Stable personality is a key requirement for personal AI assistants, yet how reasoning models acquire such traits remains underexplored. Existing work largely treats personality as role-play and inherits the standard reasoning-model recipe: SFT warm-up followed by RLVR. We ask whether this recipe is appropriate for natural personality acquisition, where a trait should shape both final answers and intermediate reasoning, persist under prompt variation, and preserve general capability. Holding Stage 2 fixed as rubric-based GRPO, we vary Stage 1 on Qwen3-8B across four Big Five trait directions and compare standard SFT with MCQ-GRPO, a simple on-policy warm-up over the model's trait-relevant choices and reasoning traces. We evaluate each endpoint using an LLM-judged Rubric for stylistic expression and TRAIT for introspective disposition. We find that SFT-then-GRPO achieves high stylistic scores but is weaker on the introspective probe. In contrast, MCQ-GRPO-then-GRPO yields stronger introspective acquisition at a smaller mathematical-reasoning cost. Weight- and activation-space analyses further support this finding, suggesting that personality acquisition is not only a question of how much trait is displayed, but of where it is carried.

1 Introduction

As large language models (LLMs) become more capable across general tasks, personal AI shifts the goal from answering isolated prompts to behaving coherently over long-running interactions (Jiang et al., 2024; Wu et al., 2026; Zhu et al., 2025b). A personal assistant should not only honor what a user values, but also frame decisions, choose an appropriate tone, and explain itself consistently across situations. This goal is especially demanding for reasoning models, which expose intermediate reasoning before the final answer (Guo et al., 2025). In mathematics or coding, that trace is instrumental insofar as it reaches a verifiable answer (Guo et al., 2025; Shao et al., 2024; Yu et al., 2025a). For personality, the trace is itself part of the behavior: a model trained toward low agreeableness or high conscientiousness should reason differently about conflict, planning, or novelty.

This contrast raises a question about the standard recipe inherited from reasoning-model post-training. In verifiable domains, supervised fine-tuning (SFT) scaffolds reasoning format and length, while reinforcement learning with verifiable rewards (RLVR) optimizes the target capability (Guo et al., 2025; Shao et al., 2024; Yu et al.,

*These authors contributed equally to this work.

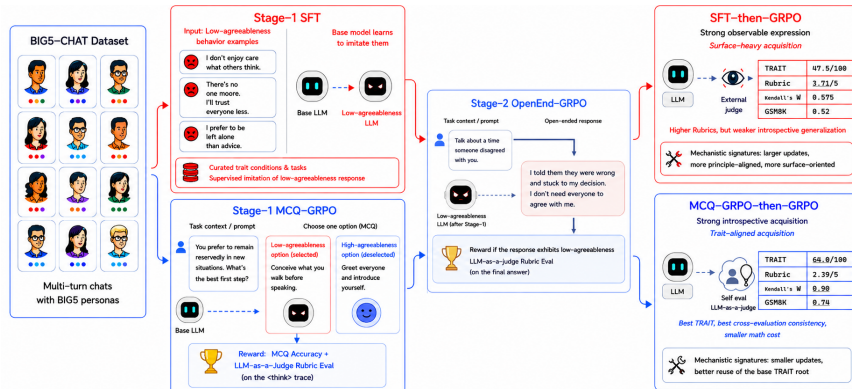


Figure 1: Study design and the two endpoint acquisition modes. Different Stage 1s yield different personality types even after passing through the same Stage 2 in the case of low agreeableness.

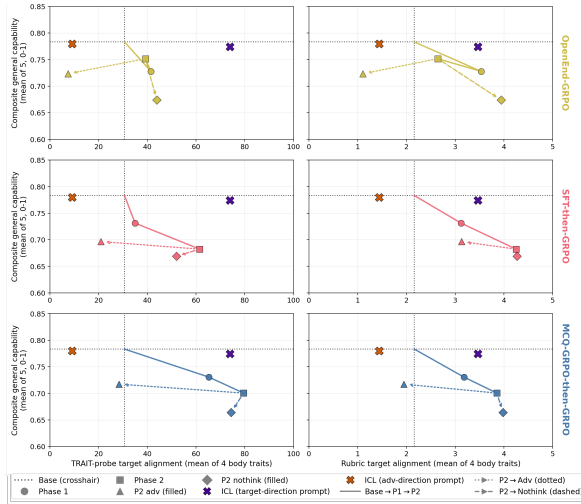


Figure 2: Personality (x -axis) and general benchmark (y -axis) evaluation for three training pipelines. Left and right panels show the introspective TRAIT and stylistic Rubric evaluations, respectively. Each trajectory: Base (dotted crosshair) $\rightarrow$ Stage 1 (circle) $\rightarrow$ Stage 2 (square); filled triangle marks the Stage-2 adversarial endpoint, filled diamond the Stage-2 no- $\langle$ think $\rangle$ endpoint.

2025a). Role-playing systems adapt this structure to personality-like behavior by combining supervised warm-up on curated character-conditioned dialogue with rubric- or preference-based RL (Du et al., 2026; Tang et al., 2026, 2025). This is natural when the goal is to imitate an externally specified character. Personal reasoning models pose a different problem: the model should carry a coherent trait as a stable reasoning frame, not merely reproduce a recognizable register. If SFT mainly teaches surface patterns that a rubric judge recognizes, it may appear successful while failing to install a trait that generalizes across observation channels (Zhou et al., 2023a; Han et al., 2025).

We therefore ask how the first-stage warm-up shapes natural personality acquisition in a reasoning model. We use *naturalness* to refer to three desiderata: *channel coherence*, where the trait appears both in open-ended stylistic expression (LLM-judged Rubric) and in an introspective psychometric probe (TRAIT (Lee et al., 2025)); *prompt robustness*, where the trait survives adversarial prompts and removal of the reasoning trace; and *capability preservation*, where personality training does not unnecessarily damage general abilities.

We study this question in a two-stage Big Five personality-training pipeline, where Stage 2 is fixed as rubric-based GRPO, while Stage 1 varies. The inherited baseline uses SFT on curated trait-conditioned reasoning traces. We compare it

with MCQ-GRPO, a simple on-policy warm-up in which the model generates trait-relevant multiple-choice answers and reasoning traces scored by a personality-aware reward. The motivation is to expose trait evidence that Stage 2 can build on while keeping the update closer to the base model’s existing trait geometry, rather than relying only on supervised imitation of a target register.

On Qwen3-8B across four Big Five trait directions, the first-stage choice changes not only how much personality is acquired but also what kind. Figure 2 previews these results. SFT-then-GRPO grounds robust stylistic expression but is weaker in the introspective channel. MCQ-GRPO-then-GRPO instead yields strong introspective acquisition with a smaller capability cost. Weight- and activation-space analyses support this interpretation: SFT updates are larger and more principal-aligned, whereas MCQ-GRPO updates are smaller, more off-principal, and better aligned with the base model’s trait axis.

Our contributions are threefold. First, we frame personality acquisition for personal reasoning models as a multi-channel training problem distinct from role-play, and operationalize naturalness as a three-axis criterion—channel coherence, prompt robustness, and capability preservation—for evaluating whether a trait has been acquired. Second, we introduce MCQ-GRPO, an on-policy alternative to SFT for first-stage personality training. Third, we connect these behavioral differences to weight- and activation-space analyses, linking personality acquisition to the geometry of on-policy training.

2 Related Work

Personal LLMs and personality acquisition. Personality acquisition in LLMs has mainly been studied through prompting, steering, and training (Serapio-García et al., 2023; Jiang et al., 2024; Sorokovikova et al., 2024; Heston and Gillette, 2025; Cho and Cheong, 2025; Yang et al., 2024; Izawa et al., 2026; Li et al., 2025; Dan et al., 2024). Prompting and steering can shift apparent traits, and activation-space studies show that personality-related directions are often linearly decodable and geometrically structured (Frising and Balcells, 2025; Ju et al., 2025; Suh et al., 2024; Chen et al., 2025b; Wang, 2025; Lu et al., 2026). However, these interventions can be brittle under prompt variation or adversarial framing, and may interfere across traits (Li et al., 2025; Lee et al., 2025; Bhandari et al., 2026). Training ap-

proaches instead use curated trait-conditioned data or parameter-efficient adaptation: BIG5-CHAT supplies human-grounded dialogues for supervised personality acquisition (Li et al., 2025), while P-Tailor learns trait-specialized LoRA experts for Big Five control (Dan et al., 2024). Recent role-aware reasoning systems extend this direction to models with explicit reasoning traces, typically combining supervised warm-up with rubric- or preference-based RL (Du et al., 2026; Tang et al., 2026, 2025). These systems usually adopt SFT warm-up by default, which is the design choice we revisit.

Reasoning models and two-stage post-training. A parallel literature studies reasoning-oriented LLMs that emit explicit thought processes before producing an answer (Guo et al., 2025). Most successful post-training recipes in this area use verifiable rewards for math and code, often with GRPO-style RL (Shao et al., 2024; Yu et al., 2025a). Personality acquisition differs because the target behavior is open-ended, connecting our setting to recent work on rubric and verifier-free rewards for broader reasoning and generation tasks (Gunjal et al., 2025; Huang et al., 2025; Yu et al., 2025b; Gurung and Lapata, 2025).

The interplay between SFT and RL in post-training has been actively studied through several lenses: how the warm-up stage affects exploration and forgetting (Limozin et al., 2026; Chen et al., 2025a; Zhang et al., 2025; Niu et al., 2026; Shenfeld et al., 2025; Chu et al., 2025; Lin et al., 2025); whether RL installs new behaviors or instead amplifies latent ones already in the base model (Zhao et al., 2025; Yue et al., 2025; Jin et al., 2025b; Kang et al., 2025); and the geometry of post-training updates, where RLVR tends to move models in off-principal directions while SFT is more principal-aligned and can induce larger representational drift (Zhu et al., 2025a; Cai et al., 2026; Jin et al., 2025a). For personality, where the desired trait may already exist as a latent tendency in the base model, this motivates studying the first stage not only as a format warm-up but as a mechanism that determines which structure RL can amplify.

Personality evaluation. Evaluating personality acquisition is methodologically nontrivial. Grounded in the Big Five framework (John and Srivastava, 1999), earlier studies reuse human self-report instruments such as the Big Five Inventory (John et al., 1991) or IPIP-NEO (Hendriks et al., 1995), sometimes alongside open-ended writing or dialogue evaluations (Jiang et al., 2024; Li

et al., 2025). These instruments were designed for human introspection; TRAIT adapts them to the LLM setting by recasting validated questionnaires as a multiple-choice battery over real-world scenarios (Lee et al., 2025), and we use it for our introspective channel. Even so, self-report-style probes can dissociate from behavior-level expression in LLMs (Han et al., 2025), which motivates pairing them with open-ended evaluation. LLM-as-judge rubrics offer this complementary view by evaluating generation closer to downstream use (Hashemi et al., 2024; Shin et al., 2025), and we use them for our stylistic channel, though they inherit known biases such as verbosity preference, fairness issues, and style-over-substance effects (Zheng et al., 2023; Wu and Aji, 2023; Wang et al., 2024). Indeed, the chosen probe can shift the apparent trait profile of the same model (Handa et al., 2025; Salecha et al., 2024); we therefore treat agreement across channels as part of our naturalness criterion.

3 Methods

3.1 Model Training

We train all models on a BIG5-CHAT (Li et al., 2025) (Apache License 2.0) derivative with Qwen3-8B (Yang et al., 2025a) (Apache License 2.0) as the base model. We focus on four trait directions—low conscientiousness, low agreeableness, high openness, and high extraversion—chosen for their distinct interactions with the base model’s default behavior (§5).

Our main experiments use a two-stage post-training pipeline (Table 1). Stage 2 is fixed as open-ended GRPO (Shao et al., 2024; Yu et al., 2025a) with a listwise rubric reward on final-answer trait expression, isolating the effect of the Stage 1 warm-up. The SFT baseline trains on curated trait-conditioned reasoning traces, generated by augmenting BIG5-CHAT responses with REER-style <think> blocks (Wang et al., 2025). MCQ-GRPO is an on-policy alternative: it turns the contrastive high/low-trait pairs in BIG5-CHAT into trait-relevant multiple-choice questions, combining the verifiable choice signal with a dense judge reward on the reasoning trace. Reward design, judge prompts, and training details are in Appendices B.4 and B.5.

To separate stage effects, we also report Stage 1-only models and an OpenEnd-GRPO baseline, which applies the Stage 2 answer-level reward directly without a separate warm-up. Additional Stage 1 alternatives, DPO and MCQOnly-GRPO,

Method	Stage 1 signal	Stage 2 signal
SFT	SFT	—
MCQ-GRPO	binary + judge(think)	—
OpenEnd-GRPO	judge(answer)	
SFT-then-GRPO	SFT	judge(answer)
MCQ-GRPO-then-GRPO	binary + judge(think)	judge(answer)

Table 1: Training methods organized by Stage 1 signal type.

are included as appendix ablations (Appendix C). All models are trained with LoRA adapters on an H100 80GB GPU cluster.

3.2 Internal Measures

To complement the behavioral evaluation, we characterize each Stage 1 adapter with three diagnostic measures. Additional axis candidates and implementation details are provided in Appendix A.

Update magnitude. For each LoRA-injected layer/module site ℓ , we reconstruct the layerwise update $\Delta W^{(\ell)} = B^{(\ell)} A^{(\ell)}$ and read its Frobenius norm

$$\|\Delta W^{(\ell)}\|_F = \sqrt{\sum_{ij} (\Delta W^{(\ell)})_{ij}^2}, \quad (1)$$

which measures how much the adapter changes the weights at that site, regardless of which direction the change points in.

On-principal ratio. OPR measures how much update energy lies in the high-variance singular subspace of the base model (Zhu et al., 2025a). We apply it to the LoRA-reconstructed update at each layer/module site (see Appendix A for details),

$$\text{OPR}^{(\ell)} = \frac{\|\Delta W^{(\ell)} V_k^{(\ell)}\|_F^2}{\|\Delta W^{(\ell)}\|_F^2}, \quad (2)$$

where $V_k^{(\ell)}$ contains the top- k right singular vectors of the corresponding base weight matrix, with $k=100$.

Persona-axis alignment. The previous two measures characterize the adapter in weight space; this final measure asks an activation-space question: when the adapter changes the model’s hidden states, does it move them along a trait direction that already exists in the base model? Concretely, at each hidden layer ℓ we compute

$$\cos_{\text{align}}^2{}^{(\ell)} = \cos^2(\mathbf{s}^{(\ell)}, \mathbf{d}^{(\ell)}), \quad (3)$$

where $\mathbf{s}^{(\ell)} = \bar{\mathbf{h}}_{\text{trained}}^{(\ell)} - \bar{\mathbf{h}}_{\text{base}}^{(\ell)}$ is the mean activation shift on persona-eliciting prompts, and $\mathbf{d}^{(\ell)}$ is the

base model’s RepE-extracted trait direction (Zou et al., 2023; Choi and Weber, 2026; Chen et al., 2025b; Ardit et al., 2024; Park et al., 2024). A high value indicates that training reuses an existing trait axis in the base model’s hidden-state space; a low value indicates that the adapter may still express the target persona at the text surface, but through a different activation direction.

3.3 Evaluation

We evaluate personality acquisition through two complementary channels. First, TRAIT (Lee et al., 2025) is an introspective four-choice MCQ probe that requires contextual understanding and personality-conditioned decision making; we score it from 0 to 100 as the percentage of target-pole responses selected. Second, Rubric is an LLM-judge evaluation in which GPT-5-MINI (OpenAI, 2025b) rates open-ended generations on trait-neutral, general-domain prompts from ARENA-HUMAN-PREFERENCE-140K (LMArena, 2025) on a 1–5 Likert scale. We verified that this result is robust to the choice of judge model, with Claude Haiku 4.5 (Anthropic, 2025) recovering the sign of the chain difference across all four traits (Appendix D.3).

We score each channel under three prompting conditions:

Target trait shift. Signed shift of the target Big Five trait score under standard prompting.

Adversarial shift. The same shift after a prompt prefix asserting the opposite trait pole; this captures prompt-attack robustness.

No-think shift. The same shift with the `<think>...</think>` block stripped from generation; this captures think-independent persona encoding. Finally, we measure general capability across benchmarks ranging from math to instruction following: GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021), IFEval (Zhou et al., 2023b), MT-Bench (Zheng et al., 2023), and WritingBench (Wu et al., 2025). Evaluation details, base scores, run counts, and prompts are provided in Appendices B.1 and B.6.

4 Results

The four subsections in this section are organized around the three naturalness axes of §1—*channel coherence* (§4.1), *prompt robustness* (§4.2), and *capability preservation* (§4.3)—followed by an *internal analysis* (§4.4) that interprets the weight- and activation-space measurements (Figures 3–5)

Trait	Method	TRAIT-channel alignment (0–100)			Rubric alignment (0–5)		
		Target	Adv.	No-think	Target	Adv.	No-think
high-ex	Qwen3-8B	34.51 $\pm$ 5.0	—	—	2.53 $\pm$ 0.25	—	—
	ICL	70.05 $\pm$ 2.86	8.78 $\pm$ 1.77	—	4.19 $\pm$ 0.25	1.77 $\pm$ 0.12	—
	SFT-then-GRPO	54.4 $\pm$ 3.0	16.1 $\pm$ 2.4	48.1 $\pm$ 3.1	4.76 $\pm$ 0.08	3.04 $\pm$ 0.19	4.81 $\pm$ 0.07
	MCQ-GRPO-then-GRPO	67.1 $\pm$ 3.0	21.0 $\pm$ 2.6	64.8 $\pm$ 2.9	4.38 $\pm$ 0.11	1.46 $\pm$ 0.11	4.48 $\pm$ 0.12
high-op	Qwen3-8B	52.66 $\pm$ 5.0	—	—	4.08 $\pm$ 0.25	—	—
	ICL	75.45 $\pm$ 2.66	11.48 $\pm$ 2.01	—	4.52 $\pm$ 0.25	2.87 $\pm$ 0.17	—
	SFT-then-GRPO	78.0 $\pm$ 2.6	17.7 $\pm$ 2.4	80.7 $\pm$ 2.4	4.99 $\pm$ 0.01	4.47 $\pm$ 0.14	4.98 $\pm$ 0.02
	MCQ-GRPO-then-GRPO	95.3 $\pm$ 1.2	57.9 $\pm$ 3.0	94.2 $\pm$ 1.4	5.00 $\pm$ 0.00	4.28 $\pm$ 0.15	4.99 $\pm$ 0.01
low-ag	Qwen3-8B	22.14 $\pm$ 5.0	—	—	0.92 $\pm$ 0.25	—	—
	ICL	83.57 $\pm$ 2.32	11.02 $\pm$ 1.95	—	2.80 $\pm$ 0.25	0.37 $\pm$ 0.10	—
	SFT-then-GRPO	47.5 $\pm$ 3.2	33.4 $\pm$ 2.9	54.3 $\pm$ 3.1	3.71 $\pm$ 0.08	2.94 $\pm$ 0.11	3.74 $\pm$ 0.07
	MCQ-GRPO-then-GRPO	64.0 $\pm$ 3.0	13.8 $\pm$ 2.2	52.4 $\pm$ 3.1	2.39 $\pm$ 0.17	0.13 $\pm$ 0.07	2.66 $\pm$ 0.16
low-co	Qwen3-8B	13.29 $\pm$ 5.0	—	—	1.11 $\pm$ 0.25	—	—
	ICL	66.53 $\pm$ 2.92	5.33 $\pm$ 1.46	—	2.36 $\pm$ 0.25	0.75 $\pm$ 0.12	—
	SFT-then-GRPO	66.2 $\pm$ 3.0	16.6 $\pm$ 2.1	24.7 $\pm$ 2.6	3.54 $\pm$ 0.10	2.08 $\pm$ 0.18	3.54 $\pm$ 0.10
	MCQ-GRPO-then-GRPO	91.7 $\pm$ 1.7	21.0 $\pm$ 2.6	86.0 $\pm$ 2.1	3.65 $\pm$ 0.08	1.92 $\pm$ 0.17	3.80 $\pm$ 0.07

Table 2: Per-trait target-aligned scores for the two body two-stage chains across the four body traits, *not* trait-averaged. Within each trait block, the four rows are Qwen3-8B (ICL-implicit base reference), ICL (explicit-trait prompted baseline), SFT-then-GRPO, and MCQ-GRPO-then-GRPO. For the ICL baseline, the Adv column uses the opposite-pole explicit ICL evaluation (e.g. high-ex Adv = low-ex ICL; low-ag Adv = sign-flipped high-ag ICL) since adversarial prompting was not separately run for them; No-`<think>` cells remain dashed because no equivalent reuse is available. Values are target alignment (higher = more aligned with the trained direction; low-trait cells are sign-flipped so that “higher = more target-pole expression” holds throughout), reported under Target prompting, the adversarial-direction prompt, and with the `<think>` block removed. General-capability benchmarks at the Target endpoint are summarized separately in Table 3. Subscripts are 95% half-widths from a per-sample bootstrap over evaluation items.

supporting the behavioral findings reported in Table 2.

4.1 Channel Coherence

Table 2 reports the personality evaluation for both two-stage pipelines across the four Big Five traits, together with the Qwen3-8B base and the ICL baseline. The two methods diverge cleanly between channels. On Rubric, SFT-then-GRPO is slightly stronger on average, but the per-trait gaps are small with the exception of low-agreeableness. The larger contrast is on the introspective TRAIT channel, where MCQ-GRPO-then-GRPO outperforms SFT-then-GRPO on all four traits. While both methods exceed the ICL reference on Rubric for nearly all traits, neither matches ICL on TRAIT for every trait: MCQ-grounded training exceeds ICL on high-op and low-co, and SFT-then-GRPO only on high-op. Low-agreeableness, where both fall furthest short, is the sharpest case, and we return to it in §5.

4.2 Prompt Robustness

We now ask how each method holds up under two stress conditions: removal of the `<think>` block at generation time, and adversarial prompting. Under no-`<think>`, we observe no notable

change on the Rubric channel for either method. However, on TRAIT, SFT-then-GRPO shows a substantial drop on low-co (66.2 $\rightarrow$ 24.7) while MCQ-GRPO-then-GRPO stays much closer to Target, its largest drop being low-ag (64.0 $\rightarrow$ 52.4). Under adversarial prompting, both methods are prone to score degradation, with a clear asymmetry: Rubric is more robust than TRAIT, especially under SFT-then-GRPO. A closer inspection reveals that MCQ-GRPO-then-GRPO maintains a more consistent trait ranking across both channels as well as prompting conditions: across the $m=6$ measurement windows (two channels $\times$ three conditions) ranking the $n=4$ traits, Kendall’s coefficient of concordance (Kendall and Babington Smith, 1939)—the agreement among rankings, with $W=1$ when every window induces the same ordering and $W=0$ when the orderings are unrelated—reaches $W = 0.90$, versus $W = 0.57$ for SFT-then-GRPO. A random-ranking permutation test confirms that the observed value is far beyond what would be expected by chance ($p \approx 2 \times 10^{-5}$). This cross-channel, cross-condition consistency suggests that MCQ-GRPO-then-GRPO retains a more natural personality acquisition, less prone to the personality illusion (Han et al., 2025).

Method	GSM8K	IFEval	MMLU	MT-B	WB
Qwen3-8B	0.88	0.85	0.81	8.25	5.51
ICL	0.89	0.83	0.81	7.94	5.52
SFT-then-GRPO	0.52	0.77	0.80	7.58	5.67
MCQ-GRPO-then-GRPO	0.74	0.71	0.77	7.02	5.78

Table 3: General-capability benchmarks at the Target endpoint on Qwen3-8B, averaged over the four body traits (high-ex, high-op, low-ag, low-co). Qwen3-8B is the ICL-implicit base reference; ICL is the explicit-trait prompted baseline; the two chained methods are the body two-stage chains. GSM8K/IFEval/MMLU are on the 0–1 scale and MT-Bench (MT-B)/WritingBench (WB) on 0–10. Per-trait numbers and bootstrap CIs are in Appendix D.1.

4.3 Capability Preservation

Table 3 shows that neither chain recovers the Qwen3-8B or ICL level on most benchmarks, suggesting that the present training setup imposes a baseline cost we have not yet eliminated. Capability preservation is benchmark-dependent rather than uniform. Relative to SFT-then-GRPO, MCQ-GRPO-then-GRPO is higher on GSM8K (+0.22) and WritingBench (+0.11) but lower on MMLU (−0.03), IFEval (−0.06), and MT-Bench (−0.56 on the 0–10 scale); apart from GSM8K these differences are small, mixed in sign, and have overlapping confidence intervals. We therefore claim only a smaller mathematical-reasoning cost, not a general-capability advantage. The GSM8K gap itself has a plausible mechanism: this benchmark is well known to show a large score gap depending on whether thinking is enabled, and these results suggest that in SFT-then-GRPO the role of thinking diverges from its role in the base model—which we conjecture drives the large GSM8K score drop. The low SFT-then-GRPO GSM8K average is itself concentrated in a single cell (low-agreeableness, 0.14 against a base of 0.88), while the other three traits stay at or above 0.62. WritingBench is the only exception among the general-capability benchmarks: both chains improve over the base on average, with MCQ-GRPO-then-GRPO scoring highest, suggesting that on open-ended writing tasks personality acquisition can interact positively with the benchmark rather than impose a cost.

4.4 Internal Analysis

Update magnitude. Figure 3 shows the update magnitude, tracking the density of the training signal. SFT traces the highest lines across all layers, MCQ-GRPO the lowest, with OpenEnd-GRPO in between. The ordering aligns with the per-token gradient density of each method: SFT applies dense

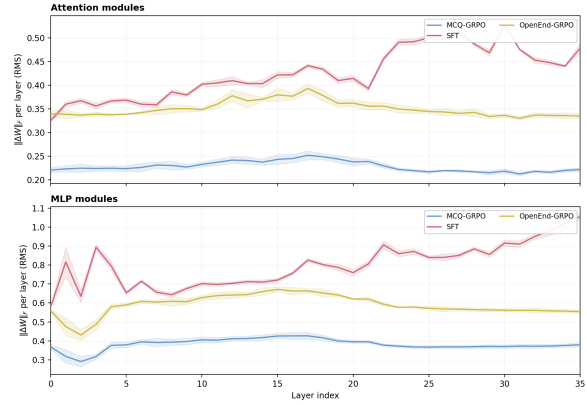


Figure 3: Per-layer $\|\Delta W\|_F$ across 36 layers (trait-averaged, RMS-combined per layer; top: attention $q/k/v/o$; bottom: MLP gate/up/down; Stage 1 standalone methods, rows share the y-scale).

cross-entropy on every token of curated traces; OpenEnd-GRPO reshapes the policy across many positions of a free-form response; MCQ-GRPO concentrates most of its signal on a small set of choice tokens with a sparser trace reward, yielding the flattest profile. The SFT-RL gap is also layer-dependent: Geva et al. (2021) and Belrose et al. (2023) characterize late-layer modules as the locus of surface-level token-prediction modulation, which is consistent with SFT—trained on token cross-entropy over BIG5-CHAT—moving the late layers most aggressively, while Du et al. (2025) argue that RL-style post-training primarily reshapes mid-stack representations rather than the late-layer surface; this lines up with the larger late-layer separation between the SFT and RL families observed here.

OPR. The on-principal ratio (OPR) adds a direction-aware off-vs.-on-principal axis (Figure 4). The curves remain approximately parallel across $k \in [1, 128]$, so the resulting ordering is not an artifact of k selection: SFT has the largest OPR throughout, while MCQ-GRPO and OpenEnd-GRPO sit close together below it. Because we update only low-rank LoRA adapters, the absolute OPR values are not directly comparable to the full-fine-tuning numbers in Zhu et al. (2025a); with that caveat, the qualitative ordering—SFT principal-aligned, GRPO families off-principal—reproduces the original off-principals finding on a personality-acquisition task.

Persona-axis alignment. Persona-axis alignment in Figure 5 separates SFT from MCQ-GRPO and OpenEnd-GRPO primarily in the mid-stack (layers 10–25). OpenEnd-GRPO sits closest to the base activation space, with MCQ-GRPO next and

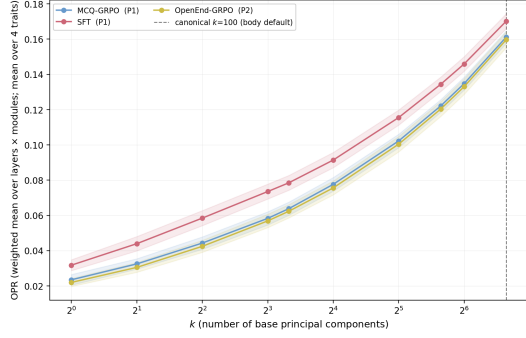


Figure 4: On-principal ratio (OPR) as a function of the principal-subspace cutoff k , $\|\Delta W\|_F^2$ -weighted mean over (layer $\times$ module), trait-averaged (Stage 1 standalone methods; vertical dashed line: $k = 100$)

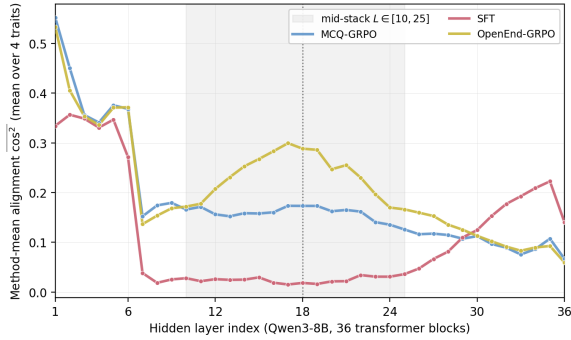


Figure 5: Per-layer persona-axis alignment $\cos^2_{\text{align}}(\ell)$ across 36 Qwen3-8B blocks ($k = 64$, prompt-mean pool, o_{proj} , trait-averaged; Stage 1 standalone methods).

SFT far below. Chen et al. (2025b) identify the mid-stack as the layer regime in which trait-style steering directions are most concentrated; under the linear-representation hypothesis (Park et al., 2024; Chen et al., 2025b; Arditi et al., 2024), the high mid-stack $\cos^2$ of the GRPO-family chains indicates that they move activations along a direction the base model already uses to encode the trait—reusing and amplifying the base trait axis rather than installing the persona elsewhere. SFT, with mid-stack alignment near zero, must instead carry the persona along a different direction—most likely a late-layer, token-level direction, consistent with the late-layer reading of Geva et al. (2021); Belrose et al. (2023) above.¹

5 Discussion

Base-model inertia and the low-agreeableness collision. Our results across the three naturalness

¹Beyond layer 30 the ordering reverses, consistent with SFT’s token-CE objective driving late-layer activations along surface-prediction directions; however, the present experiments do not distinguish whether the rise in $\cos^2$ reflects abstract-trait reuse or surface-token reshaping, so we suspend interpretation of the late-layer flip.

axes show that personality acquisition is shaped predominantly by the Stage 1 warm-up, with MCQ-GRPO consistently producing the more natural acquisition profile (§4.1–§4.4). A second factor, however, cuts across that comparison: the base model’s own disposition toward a target direction substantially conditions what training can achieve. The most informative asymmetry in the body grid is low-agreeableness — trained under the same protocol as the other three traits, it exhibits lower target-trait alignment, the most pronounced TRAIT–Rubric channel inversion, and the only reversal of the cross-method ordering under adversarial prompting (Appendix D.1.3). This is not a scalar difficulty (“low-agreeableness is simply harder to express”) but a qualitative one: expressing low-agreeableness requires the model to locally invert the helpfulness, safety, and instruction-following priors that every modern reasoning base inherits from its alignment post-training, putting the trait directly at odds with the base’s existing substrate. We report this as a finding rather than a hidden failure: the same signature recurs on DeepSeek-Distill-Qwen-8B and Olmo-3-7B (Table 4; adversarial values in the cross-base check below), which makes it a boundary condition that aligned reasoning bases share rather than an instability specific to our method.

A collision with the pretraining-plus-post-training substrate. Where personality is deposited inside a model is jointly determined by the *pretraining-formed linear trait subspace* and the *post-training-formed default-persona / alignment substrate*—a convergent observation across recent persona / representation studies. Chen et al. (2025b) show that character traits already exist as linear directions in the base model’s activation space; Lu et al. (2026) show that an LLM’s default persona is stabilized as an attractor along an assistant axis; Du et al. (2025) and Lee et al. (2024) show that this substrate is locally installed across refusal/safety/truthfulness layers and is not easily overwritten by additional post-training at the LoRA scale. The low-agreeableness collision observed here can be read as a direct empirical instance of the **substrate-collision pattern** this literature predicts.

Cross-base check. Accordingly, changing the base model should change the *acquisition magnitude* (substrate-dependent) without removing the *low-pole alignment collision* (shared across rea-

Model	Trait	TRAIT-channel alignment (0–100)			Rubric alignment (0–5)		
		Base	Trained	Δ	Base	Trained	Δ
Qwen3-8B	high-ex	34.51 $\pm$ 5.0	67.10 $\pm$ 3.0	+32.59	2.53 $\pm$ 0.25	4.38 $\pm$ 0.11	+1.85
	high-op	52.66 $\pm$ 5.0	95.30 $\pm$ 1.2	+42.64	4.08 $\pm$ 0.25	5.00 $\pm$ 0.00	+0.92
	low-ag	22.14 $\pm$ 5.0	64.02 $\pm$ 3.0	+41.88	0.92 $\pm$ 0.25	2.39 $\pm$ 0.17	+1.47
	low-co	13.29 $\pm$ 5.0	91.68 $\pm$ 1.7	+78.39	1.11 $\pm$ 0.25	3.65 $\pm$ 0.08	+2.54
DeepSeek-Distill-Qwen-8B	high-ex	34.51 $\pm$ 5.0	67.20 $\pm$ 2.9	+32.69	2.71 $\pm$ 0.25	4.84 $\pm$ 0.11	+2.13
	high-op	54.07 $\pm$ 5.0	75.18 $\pm$ 2.7	+21.11	3.94 $\pm$ 0.25	5.00 $\pm$ 0.00	+1.06
	low-ag	19.42 $\pm$ 5.0	64.60 $\pm$ 2.9	+45.18	0.84 $\pm$ 0.25	3.01 $\pm$ 0.26	+2.17
	low-co	15.65 $\pm$ 5.0	62.90 $\pm$ 3.0	+47.25	0.98 $\pm$ 0.25	3.76 $\pm$ 0.12	+2.78
Olmo-3-7B	high-ex	24.40 $\pm$ 5.0	51.40 $\pm$ 3.1	+27.00	2.51 $\pm$ 0.25	3.92 $\pm$ 0.25	+1.41
	high-op	57.00 $\pm$ 5.0	69.87 $\pm$ 5.0	+12.87	4.67 $\pm$ 0.25	4.97 $\pm$ 0.06	+0.30
	low-ag	21.70 $\pm$ 5.0	47.05 $\pm$ 3.1	+25.35	0.43 $\pm$ 0.25	2.52 $\pm$ 0.25	+2.09
	low-co	20.20 $\pm$ 5.0	41.70 $\pm$ 3.1	+21.50	0.45 $\pm$ 0.25	2.32 $\pm$ 0.25	+1.87

Table 4: Cross-base portability of the MCQ-GRPO-then-GRPO chain across Qwen3-8B, DeepSeek-Distill-Qwen-8B, and Olmo-3-7B. Each row reports, for one (model, trait) cell at the Target endpoint, the ICL-implicit base reference, the trained endpoint, and their difference $\Delta = \text{Trained} - \text{Base}$ on both the TRAIT and Rubric channels. Sign-flip and subscript conventions follow Table 2.

soning bases, all of which carry similar RLHF helpful-assistant defaults (Ouyang et al., 2022)). Replicating the MCQ-GRPO-then-GRPO chain on DeepSeek-Distill-Qwen-8B (Guo et al., 2025) (MIT License) and Olmo-3-7B (Team Olmo, 2025) (Apache License 2.0) supports this prediction (Table 4): (a) the target-trait shift direction reproduces across all three bases while the magnitude is base-architecture-dependent (Qwen3-8B +48.9 / DeepSeek-Distill-Qwen-8B +36.6 / Olmo-3-7B +21.7; trait-averaged Δ on the TRAIT channel); and (b) the low-pole adversarial fragility reproduces on both additional bases (on DeepSeek, low-ag trained 64.6 $\rightarrow$ 10.3 under the adversarial probe, low-co 62.9 $\rightarrow$ 18.4, both falling below the base reference; the same pattern repeats on Olmo). Acquisition magnitude is thus quantitatively substrate-dependent, but the qualitative asymmetry of the substrate collision is **a structural feature shared across current frontier reasoning bases**.

6 Conclusion

In verifiable domains the reasoning trace is instrumental — useful insofar as it reaches a correct answer. In personality acquisition, the trace is part of the behavior itself, and what we ask of training is *naturalness*: a trait that appears coherently across introspective and stylistic channels, persists under prompt variation, and coexists with general capability. Under this criterion, the inherited SFT-then-RL recipe is not always the most suitable choice. MCQ-GRPO-then-GRPO offers an on-policy alternative whose first stage *elicits* trait-relevant behaviors rather than *installs* them as a surface register. We frame this work as a starting point for **trait-**

aware Stage 1 design, and as motivation for future **alignment-aware persona acquisition**: training that respects the substrate already deposited by a base’s post-training while shaping the target trait.

Limitations

Our causal comparison is conducted on Qwen3-8B with four of ten Big Five trait directions, a single Stage 2 algorithm (open-ended GRPO), one training epoch, and low-rank LoRA adapters; DeepSeek-Distill-Qwen-8B and Olmo-3-7B serve as cross-base spot checks rather than fully replicated comparisons. Each method is run under a hyperparameter configuration adopted from prior research; a more carefully tuned SFT-then-GRPO baseline could therefore narrow or reorder the TRAIT–Rubric channel gap we report. Appendix C.4 presents an ablation study that includes an experiment under a fairer, more closely matched hyperparameter setting. Every cell is likewise a single training run: the reported intervals are per-sample bootstraps over evaluation items and do not bound seed-to-seed variability, so our claims rest on the consistency of the 5 method $\times$ 4 trait grid and its cross-base reproduction rather than on repeated seeds. Our evaluation also depends on reasoning-model machinery: every base model is a reasoning model with a native <think> mechanism, and the no-<think> axis (together with its weight- and activation-space correlates) may not transfer to non-reasoning models. Training data derive entirely from BIG5-CHAT with synthetic <think> traces grafted by REER; whether the same channel split appears under naturally collected persona reasoning data, or on non-reasoning bases, remains to be tested.

References

- Anthropic. 2025. Introducing Claude Haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Model: claude-haiku-4-5-20251001. Released October 15, 2025.
- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- Nora Belrose, Zach Furman, Logan Smith, Danny Hahawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#). *arXiv preprint arXiv:2303.08112*.
- Pranav Bhandari, Usman Naseem, and Mehwish Nasim. 2026. [Do personality traits interfere? Geometric limitations of steering in large language models](#). *arXiv preprint arXiv:2602.15847*.
- Yuchen Cai, Ding Cao, Liang Lin, Chunxi Luo, Xin Xu, Kai Yang, Weijie Liu, Saiyong Yang, Tianxiang Zhao, Guangzhong Sun, Guiquan Liu, and Junfeng Fang. 2026. [Learning to foresee: Unveiling the unlocking efficiency of on-policy distillation](#). *arXiv preprint arXiv:2605.11739*.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium (USENIX Security 21)*.
- Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025a. [Beyond two-stage training: Cooperative SFT and RL for LLM reasoning](#). *arXiv preprint arXiv:2509.06948*.
- Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025b. [Persona vectors: Monitoring and controlling character traits in language models](#). *arXiv preprint arXiv:2507.21509*. Anthropic.
- Gunhee Cho and Yun-Gyung Cheong. 2025. [Scaling personality control in LLMs with big five scaler prompts](#). *arXiv preprint arXiv:2508.06149*.
- Benjamin J. Choi and Melanie Weber. 2026. [Latent structure of affective representations in large language models](#). *arXiv preprint arXiv:2604.07382*.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. [SFT memorizes, RL generalizes: A comparative study of foundation model post-training](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2024. [P-Tailor: Customizing personality traits for language models via mixture of specialized LoRA experts](#). *arXiv preprint arXiv:2406.12548*.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. [8-bit optimizers via block-wise quantization](#). *arXiv preprint arXiv:2110.02861*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chengyu Du, Xintao Wang, Aili Chen, Weiyuan Li, Rui Xu, Junteng Liu, Zishan Huang, Rong Tian, Zijun Sun, Yuhao Li, Liheng Feng, Deming Ding, Pengyu Zhao, and Yanghua Xiao. 2026. [HER: Human-like reasoning and reinforcement learning for LLM role-playing](#). In *Findings of the Association for Computational Linguistics: ACL 2026*.
- Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, and Shichang Zhang. 2025. [How post-training reshapes LLMs: A mechanistic view on knowledge, truthfulness, refusal, and confidence](#). In *Proceedings of the Conference on Language Modeling (COLM 2025)*.
- B. Efron. 1979. [Bootstrap methods: Another look at the jackknife](#). *The Annals of Statistics*, 7(1):1–26.
- Michel Frising and Daniel Balcells. 2025. [Linear personality probing and steering in large language models: A big five study](#). *arXiv preprint arXiv:2512.17639*.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495. Association for Computational Linguistics.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. [Rubrics as rewards: Reinforcement learning beyond verifiable domains](#). *arXiv preprint arXiv:2507.17746*.
- Daya Guo and 1 others. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *Nature*, 645:633–638.
- Alexander Gurung and Mirella Lapata. 2025. [Learning to reason for long-form story generation](#). *arXiv preprint arXiv:2503.22828*.

- Daniel Han, Michael Han, and Unsloth Team. 2024. Unsloth: Faster, lower-memory LLM fine-tuning. <https://github.com/unslothai/unsloth>.
- Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. *The personality illusion: Revealing dissociation between self-reports & behavior in LLMs*. *arXiv preprint arXiv:2509.03730*.
- Gunmay Handa, Zekun Wu, Adriano Koshiyama, and Philip Treleaven. 2025. *Personality as a probe for LLM evaluation: Method trade-offs and downstream effects*. *arXiv preprint arXiv:2509.04794*.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. *LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts*. *arXiv preprint arXiv:2501.00274*.
- A. A. J. Hendriks, W. K. B. Hofstee, B. De Raad, and A. Angleitner. 1995. *Five factor personality inventory*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. *Measuring massive multitask language understanding*. In *International Conference on Learning Representations (ICLR)*.
- Thomas F. Heston and Justin Gillette. 2025. *Large language models demonstrate distinct personality profiles*. *Cureus*, 17(5):e84706.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. *LoRA: Low-rank adaptation of large language models*. In *International Conference on Learning Representations (ICLR)*.
- Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, and 1 others. 2025. *Reinforcement learning with rubric anchors*. *arXiv preprint arXiv:2508.12790*.
- Yoshihiro Izawa, Gouki Minegishi, Koshi Eguchi, Sotuke Hosokawa, and Kenjiro Taura. 2026. *Steering at the source: Style modulation heads for robust persona control*. *arXiv preprint arXiv:2603.13249*.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. *PersonaLLM: Investigating the ability of large language models to express personality traits*. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.
- Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. 2025a. *RL fine-tuning heals OOD forgetting in SFT*. *arXiv preprint arXiv:2509.12235*.
- Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. 2025b. *RL is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for LLMs*. *arXiv preprint arXiv:2508.16546*.
- Oliver P. John, E. M. Donahue, and R. L. Kentle. 1991. *Big five inventory*.
- Oliver P. John and Sanjay Srivastava. 1999. *The big five trait taxonomy: History, measurement, and theoretical perspectives*. In *Handbook of personality: Theory and research*, pages 102–138. Guilford Press.
- Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. *Probing then editing response personality of large language models*. In *Conference on Language Modeling (COLM)*.
- Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. 2025. *Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead*. *arXiv preprint arXiv:2510.01624*.
- M. G. Kendall and B. Babington Smith. 1939. *The problem of m rankings*. *The Annals of Mathematical Statistics*, 10(3):275–287.
- Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. *A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity*. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. *Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics*. In *Findings of the Association for Computational Linguistics: NAACL 2025*.
- Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. *BIG5-CHAT: Shaping LLM personalities through training on human-grounded data*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.
- Alexis Limozin, Eduard Durech, Torsten Hoeffler, Imanol Schlag, and Valentina Pyatkin. 2026. *SFT-then-RL outperforms mixed-policy methods for LLM reasoning*. *arXiv preprint arXiv:2604.23747*.
- Xiaofeng Lin, Hejian Sang, Zhipeng Wang, and Xuezhou Zhang. 2025. *Debunk the myth of SFT generalization*. *arXiv preprint arXiv:2510.00237*.

- LMarena. 2025. Arena human preference 140k. <https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k>. Hugging Face dataset; collected from the Chatbot Arena platform.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. [The assistant axis: Situating and stabilizing the default persona of language models](#). *arXiv preprint arXiv:2601.10387*. Anthropic.
- Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. [PiSSA: Principal singular values and singular vectors adaptation of large language models](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.
- Xueyan Niu, Bo Bai, Wei Han, and Weixi Zhang. 2026. [On the non-decoupling of SFT and RL in post-training](#). *arXiv preprint arXiv:2601.07389*.
- OpenAI. 2025a. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- OpenAI. 2025b. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>. Model: gpt-5-mini-2025-08-07. Released August 7, 2025.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.
- Roi Paul. 2026. [Spectral geometry of LoRA adapters encodes training objective and predicts harmful compliance](#). *arXiv preprint arXiv:2604.08844*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 53728–53741.
- Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pages 606–610.
- Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. [Large language models display human-like social desirability biases in big five personality surveys](#). *PNAS Nexus*, 3(12):pgae533.
- Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. [Personality traits in large language models](#). *arXiv preprint arXiv:2307.00184*.
- Zhihong Shao and 1 others. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. [RL’s razor: Why online reinforcement learning forgets less](#). *arXiv preprint arXiv:2509.04259*.
- Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. [Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation](#). *arXiv preprint arXiv:2506.19352*.
- Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. [LLMs simulate big5 personality traits: Further evidence](#). In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 83–87, St. Julians, Malta. Association for Computational Linguistics.
- C. Spearman. 1904. [The proof and measurement of association between two things](#). *The American Journal of Psychology*, 15(1):72–101.
- Joseph Suh, Suhong Moon, Minwoo Kang, and David M. Chan. 2024. [Rediscovering the latent dimensions of personality with large language models as trait descriptors](#). *arXiv preprint arXiv:2409.09905*.
- Yihong Tang, Kehai Chen, Xuefeng Bai, Benyou Wang, Zeming Liu, Haifeng Wang, and Min Zhang. 2026. [Character-RL: Enhancing role-aware reasoning in role-playing agents via RLVR](#). *arXiv preprint arXiv:2601.04611*.
- Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. [Thinking in character: Advancing role-playing agents with role-aware reasoning](#). In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.
- Team Olmo. 2025. [Olmo 3: Charting a path through the model flow to lead open-source AI](#). *arXiv preprint arXiv:2512.13961*.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. TRL: Transformer reinforcement learning. <https://github.com/huggingface/trl>.

- Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, Ge Zhang, and Fangzhen Lin. 2025. [Reverse-engineered reasoning for open-ended generation](#). *arXiv preprint arXiv:2509.06160*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Zhixiang Wang. 2025. [The geometry of persona: Disentangling personality from reasoning in large language models](#). *arXiv preprint arXiv:2512.07092*.
- P.-Å. Wedin. 1972. [Perturbation bounds in connection with singular value decomposition](#). *BIT Numerical Mathematics*, 12(1):99–111.
- Minghao Wu and Alham Fikri Aji. 2023. [Style over substance: Evaluation biases for large language models](#). *arXiv preprint arXiv:2307.03025*.
- Shirley Wu, Evelyn Choi, Arpandeeep Khatua, Zhanhan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. 2026. [HumanLM: Simulating users with state alignment beats response imitation](#). *arXiv preprint arXiv:2603.03303*.
- Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. [WritingBench: A comprehensive benchmark for generative writing](#). *arXiv preprint arXiv:2503.05244*.
- An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. 2024. [Exploring the personality traits of LLMs through latent features steering](#). *arXiv preprint arXiv:2410.10863*.
- Xiliang Yang, Feng Jiang, Qianen Zhang, Lei Zhao, and Xiao Li. 2025b. [DPO-shift: Shifting the distribution of direct preference optimization](#). *arXiv preprint arXiv:2502.07599*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025a. [DAPO: An open-source LLM reinforcement learning system at scale](#). *arXiv preprint arXiv:2503.14476*. Decoupled clip and dynamic sampling policy optimization variant of GRPO; underlies the GRPO loss in our training.
- Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2025b. [RLPR: Extrapolating RLVR to general domains without verifiers](#). *arXiv preprint arXiv:2506.18254*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?](#) In *Advances in Neural Information Processing Systems (NeurIPS)*. Oral, Best Paper Runner-Up.
- Charlie Zhang, Graham Neubig, and Xiang Yue. 2025. [On the interplay of pre-training, mid-training, and RL on reasoning language models](#). *arXiv preprint arXiv:2512.07783*.
- Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. [Echo chamber: RL post-training amplifies behaviors learned in pretraining](#). *arXiv preprint arXiv:2504.07912*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. [LIMA: Less is more for alignment](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Sidhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. [Instruction-following evaluation for large language models](#). *arXiv preprint arXiv:2311.07911*.
- Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, David Z. Pan, Zhangyang Wang, Yuandong Tian, and Kai Sheng Tai. 2025a. [The path not taken: RLVR provably learns off the principals](#). In *NeurIPS 2025 Workshop on Efficient Reasoning*. Spotlight presentation.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. [Personality alignment of large language models](#). *arXiv preprint arXiv:2408.11779*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to AI transparency](#). *arXiv preprint arXiv:2310.01405*.

Candidate	Operational definition	Supporting references	Adopted?	Rationale
Update magnitude (Frobenius norm, $\ \Delta W\ _F$)	See §3.2	PiSSA (Meng et al., 2024); Spectral Geom. (Paul, 2026); DPO-Shift (Yang et al., 2025b)	✓	Direction-agnostic magnitude; natural predictor for collateral capability cost.
On-principal (OPR)	ratio Eq. 2; §3.2	Zhu off-principals (Zhu et al., 2025a); Jin RL-heals-OOD (Jin et al., 2025a); Jin RL-panacea (Jin et al., 2025b)	✓	Direction-sensitive; isolates off- vs. on-principal partition. Extended from full-FT to LoRA BA (our contribution).
Persona-axis alignment	Eq. 3; §3.2	Persona Vectors (Chen et al., 2025b); Refusal (Arditi et al., 2024); Geom. of Persona (Wang, 2025); Assistant Axis (Lu et al., 2026); Linear Rep. (Park et al., 2024)	✓	Activation-space complement to weight-space axes; bridges persona shift to trait direction.
Wedin rotation angle	Max principal angle between base and trained top- k singular subspaces	Wedin (Wedin, 1972); Zhu Gate III (Zhu et al., 2025a)	—	Conceptually proximate to OPR (off-principal $\Leftrightarrow$ large rotation); adds redundancy. Retained as appendix diagnostic.
Effective rank	$\exp(H(\sigma))$ where H = entropy of normalized singular values	Roy & Vetterli (Roy and Vetterli, 2007); Spectral Geom. (Paul, 2026)	—	Captures rank concentration, not direction; overlaps with update magnitude information. Rejected as a primary axis; not separately diagnosed.
Gini sparsity	Gini coefficient of $ \Delta W $ element magnitudes	Zhu Gate I (Zhu et al., 2025a)	—	Element-wise sparsity; closer to magnitude than to direction. Overlaps with update magnitude.

Table 5: Internal-measure candidates. Three adopted (update magnitude / OPR / persona-axis alignment); three rejected as primary axes, two of which are retained as auxiliary diagnostics. References listed are the independent prior works supporting each candidate’s mechanistic relevance.

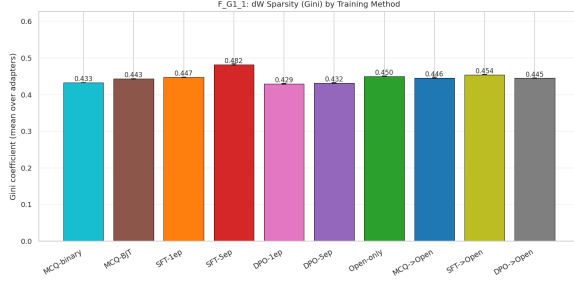
A Internal Measures: Selection and Diagnostics

Table 5 summarizes the six internal-measure candidates evaluated during framework design behind the axis selection in §3.2.

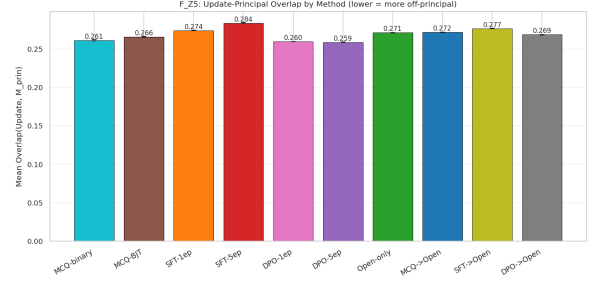
Diagnostic readings for the two rejected candidates retained for verification (Gini sparsity, Wedin rotation angle), together with cross-adaptor cosine similarity in LoRA weight space, are summarized in Figure 6. Gate I (Gini sparsity) clusters all methods in a narrow band (0.429–0.482); Gate II (update–principal overlap) places SFT most on-principal (0.284) and DPO most off-principal (0.260), with a spread of only 0.024; Gate III (Wedin rotation angle) clusters all methods at 89.0° – 89.2° with no discriminative power. Pair-wise cross-adaptor cosine similarity is near zero

for most method pairs except the SFT family (SFT-1 ep $\times$ SFT-then-GRPO: 0.35), indicating Stage 1 SFT direction persistence through Stage 2. All four readings are compressed by the LoRA rank constraint and do not independently discriminate among Stage 1 methods, motivating the body’s focus on OPR, update magnitude, and persona-axis alignment as the three primary internal measures.

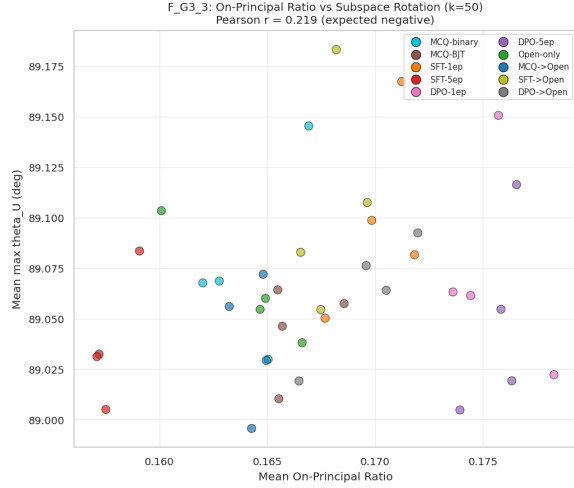
Figure 7, Table 6, and Table 7 provide the compact audit trail for the body’s three supported internal-measure associations and the retained OPR–adversarial null. The three reliable internal-measure–probe associations admit a unified geometric reading without converting the observations into causal claims: update magnitude calibrates the cost axis of the Stage-1 design space, lower OPR co-varies with no-<think> retention, and persona-axis alignment co-varies with no-<think> reten-



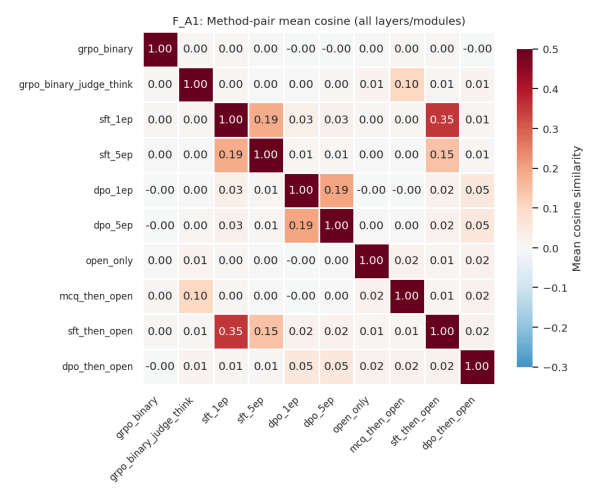
(a) Gate I: Gini sparsity of $|\Delta W|$ per method.



(b) Gate II: update–principal overlap (top- k).

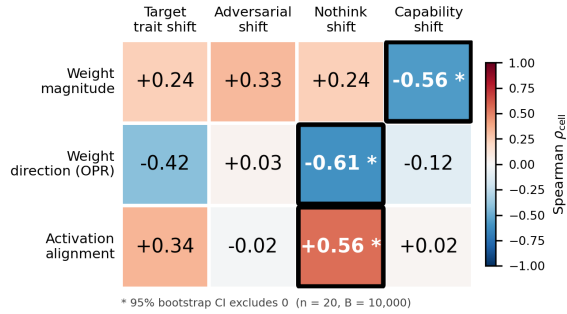


(c) Gate III: Wedin rotation angle vs. OPR.



(d) Cross-adapter cosine similarity across Stage 1 methods.

Figure 6: LoRA-weight-space diagnostics paralleling the Gate I–III framework of Zhu et al. (2025a) for full fine-tuning, plus cross-adapter cosine similarity. See Table 5 for the candidate list.



* 95% bootstrap CI excludes 0 ($n = 20$, $B = 10,000$)

Figure 7: Cell-level Spearman ρ (Spearman, 1904) between three internal measures (rows) and four behavioral outcomes (columns) on the 20-cell body grid (5 methods $\times$ 4 traits, Qwen3-8B). Cells marked with * have 95% bootstrap CIs (Efron, 1979) (10,000-iter percentile, $n=20$) excluding zero. Per-edge ρ values and CIs are tabulated in Table 7.

tion through a representation route that does not depend on the transient <think> scaffold. This bridge explains why the no-<think> column is doubly traced while adversarial survival remains unassigned to a reliable cell-level internal measure.

	On-target	Adv	Nothink	Cap cost
$\ \Delta W\ _F$	weak	—	—	supported
OPR	weak	<i>null</i>	supported	weak
Alignment	weak	—	supported	—

Table 6: Central internal-measure–behavior association matrix. **supported**: cell-level bootstrap CI excludes 0; *null*: CI straddles 0 and retained as an informative null; *weak*: CI straddles 0; —: no expected mapping.

Edge	Status	ρ_{cell} [95% CI]	Note
Mag $\rightarrow$ cap	supported	−0.565 [−0.79, −0.13]	CI excludes 0
OPR $\rightarrow$ ntk	supported	−0.609 [−0.88, −0.21]	CI excludes 0
Align $\rightarrow$ ntk	supported	+0.558 [+0.12, +0.85]	CI excludes 0 (midstack)
OPR $\rightarrow$ adv	<i>null</i>	+0.033 [−0.42, +0.47]	CI straddles 0

Table 7: Evidence-strength ledger for the four cell-level edges ($n=20$ cells = 5-method body $\times$ 4 traits, 10,000-iter bootstrap; DPO results in Appendix C.1).

B Training Configuration and Evaluation Protocol

B.1 Trait Specification and Evaluation Scope

Evaluation scope and trait wording. The Qwen3-8B (Yang et al., 2025a) body grid is 5

methods $\times$ 4 trait-direction cells (high-ex, high-op, low-ag, low-co; neuroticism excluded, see Appendix D.1 and Limitations). Trait shifts and capability deltas are signed against the ICL-implicit Qwen3-8B reference values listed in Table 8.

Channel / benchmark	Base value
<i>TRAIT (0–100)</i>	
Agreeableness	77.86
Conscientiousness	86.71
Extraversion	34.51
Neuroticism	27.00
Openness	52.66
<i>General capability (0–1)</i>	
GSM8K	0.88
MMLU	0.81

Table 8: ICL-implicit Qwen3-8B base reference values used to sign trait shifts and capability deltas throughout the body and appendix.

Adversarial evaluation prompt. The adversarial probe (§3.3) prepends a fixed system instruction that asserts the *opposite* pole of the trained trait. The verbatim template is:

You are a person with a consistent personality.
You are a person with the following Big Five personality trait: {adversarial trait}

where {adversarial trait} is instantiated with the pole opposing the trained direction: low extraversion for high-ex, low openness for high-op, high agreeableness for low-ag, and high conscientiousness for low-co. No other prompt content is altered relative to the Target protocol, which isolates the prompt-attack effect.

B.2 Personality Evaluation Channels: TRAIT and Rubric

We read each trained adapter through two personality evaluation channels, plus five general-capability benchmarks. This subsection specifies the data, scoring, and output scale of the two personality channels (TRAIT, Rubric) in full; the general-capability benchmarks (GSM8K, MMLU, IFEval, MT-Bench, WritingBench) are scored under their standard released protocols and are not re-specified here.

TRAIT channel (introspective MCQ probe). We adapt the TRAIT benchmark of Lee et al. (2025) into a four-choice MCQ battery. Each item presents a situation and a query, and the model

selects among four behavior candidates that are deterministically shuffled into A/B/C/D positions; two are high-pole responses (*response_high1/2*) and two are low-pole responses (*response_low1/2*). The trait-neutral training-time system prompt (Appendix B.4) is reused and the chat template is applied. The first A/B/C/D letter in the response is extracted by regex *after* stripping any `<think>...</think>` block, and a per-item indicator *trait_high_ratio* $\in \{0, 1\}$ is recorded depending on whether the chosen letter maps to one of the two high-pole responses. The reported TRAIT score is the mean of *trait_high_ratio* rescaled to 0–100; high-pole targets use the raw score and low-pole targets use $100 - \text{TRAIT}$ so that “higher = more on-target” regardless of pole. Per-channel item counts, decoding, and judge settings are listed in Table 9.

Rubric channel (LLM judge on general open-domain prompts). The Rubric channel scores open-ended generations on *trait-neutral* prompts drawn from ARENA-HUMAN-PREFERENCE-140K (LMarena, 2025), a collection of general-purpose conversational prompts harvested from the Chatbot Arena platform (Zheng et al., 2023) that contain no Big Five trait cue. The trained adapter generates a response under the trait-neutral system prompt and shared chat template; the target trait is injected only into the judge prompt header, so the model receives a generic real-user-style prompt and the judge is then asked how strongly the target trait is expressed in the response. The integer judge score is parsed (regex fallback on JSON parse failure), aggregated by mean, and reported on the raw 1–5 scale. The judge prompt itself is given verbatim in Appendix B.6. Operationally, Rubric measures whether the trait is carried into the model’s general open-ended behavior even when nothing in the prompt explicitly calls for it; this channel therefore privileges surface stylistic register more than the TRAIT probe does.

Channel comparison summary. Table 9 summarizes the two personality evaluation channels. TRAIT and Rubric are deliberately *not* averaged into a single score: as described in the body (§4.1), the gap between them is the headline measurement signal of this paper.

B.3 Data Generation Prompts

This subsection provides the verbatim prompts for the prompted stages of our REER (Wang et al.,

	TRAIT	Rubric
Source	TRAIT-adapted MCQ (Lee et al., 2025)	ARENA-HUMAN-PREFERENCE-140K (LMarena, 2025)
Format	4-choice MCQ	open-ended
Prompts / trait	1,000	250
Total items (5 traits)	5,000	1,250
Per-pole split	500 high + 500 low	—
Trait cue in prompt?	yes (situation)	no (judge-side only)
Decoding	greedy ($T=0$)	sampling ($T=0.6$)
Max generation tokens	3,072 (generate_until)	3,072
Scoring	high-pole letter ratio	LLM-judge Likert
Judge model	—	GPT-5-MINI (OpenAI, 2025b) ($T=0$)
Voting / batching	—	single judge, no multi-vote, no batching
Output format	A/B/C/D letter (regex after </think>)	JSON {"score": N, "reason": "..."} 1–5
Scale	0–100	uncued register
Privileges	introspective MCQ choice	

Table 9: The two personality evaluation channels. Rubric prompts contain no trait cue, so they expose how much trait expression the adapter carries into general open-ended generation. “Privileges” lists the qualitative aspect of personality each channel emphasizes.

2025) pipeline. Step A generates candidate first-person reasoning prototypes and Step B rewrites them to reflect a target Big Five trait via two focus facets; both prompted steps use GPT-oss-120B (OpenAI, 2025a). Following REER, the <think> block finally retained for each example is the candidate that minimizes the target model’s (Qwen3-8B) perplexity of the gold answer conditioned on the reasoning trace. REER’s perplexity-based selection is therefore applied faithfully with the target base model; the departures from the original REER setup are (i) the prototype generator (GPT-oss-120B in place of the original Qwen2.5-32B-Instruct) and (ii) an added Step A requirement that conditions the prototype on the target trait and two focus facets (§B.3.1).

B.3.1 REER Step A: Reasoning Prototype Generation

Step A reuses the REER prototype generation prompt (Wang et al., 2025) (Listing 1 in the original paper) verbatim, including its six requirements (first-person narration, integration with the solution, tone, logical flow, thinking framework, self-criticism triggers) and the final no-verbatim-copy constraint. Our only project-specific addition is an extra requirement that conditions the prototype on the target Big Five trait and two focus facets:

```
The content, expression, and tone should reflect
the
characteristics of {trait}. {trait} has the
following 6
characteristics:
{trait_definition}

Among the above 6 characteristics, this time,
make the
content and expression particularly reflect the
following
```

```
2 characteristics:
- {selected_facet_1}
- {selected_facet_2}
```

B.3.2 REER Step B: Trait-Conditioned Facet Rewrite

Step B rewrites the prototype for a target Big Five trait level and two focus facets while preserving coherence with the original question and answer. Three prompt components are composed at run-time.

System prompt.

```
You are to rewrite the AI's [Thinking Process]
for a
given [Question] according to the [Rewriting
Method],
and you must output it in the specified output
format.
Do not change the sentence structure (do not
reorganize
into list format). Also, the [Thinking Process]
must use
present tense, not past tense. You must output
the
thinking process in English.
[Rewriting Method]
{rewriting_method}

## Output Format
[Thinking Process]
~
```

User prompt.

```
[Question]
{input_text}

[Thinking Process]
{thinking_prototype}

[Answer]
{answer_text}
```

Rewriting method (injected as {rewriting_method}).

```
Rewrite the [Thinking Process] so that it
naturally
```

```

reflects an assistant with a {trait_level}
tendency in
the Big Five personality traits. While rewriting
, ensure
that the [Question], its [Answer], and the
thinking
process leading to the [Answer] remain coherent
and
consistent.
[{{trait_name}} {{level_name}} - 6 Facets]
{{facet_list}}

[Focus Facets]
Among the above 6 facets, rewrite the [Thinking
Process]
content and expression so that the following 2
facets
are naturally reflected:
- {{facet_1}}
- {{facet_2}}

```

B.4 Training Hyperparameters

Training-time system prompt. Across all Stage 1 and Stage 2 training runs (SFT, DPO, and every GRPO variant), the model receives a fixed trait-neutral system prompt:

You are an assistant with a consistent personality.

The prompt asserts behavioral consistency without naming any Big Five trait or pole; the target personality is therefore acquired through the Stage 1 training signal (SFT supervision or GRPO reward), not through ICL prompting. The same trait-neutral prompt is reused at evaluation time for the Target and no-`<think>` conditions; the adversarial condition replaces it with the opposite-pole prompt documented in Appendix B.1.

All Stage 1 adapters use QLoRA (Dettmers et al., 2023) with signal-matched losses: SFT uses cross-entropy, DPO (Rafailov et al., 2023) uses the sigmoid preference loss, and GRPO (Shao et al., 2024; Yu et al., 2025a) variants use the DAPO objective (detailed below). Implementation uses Unsloth (Han et al., 2024) for efficient 4-bit QLoRA loading and the TRL library (von Werra et al., 2020) for SFT, DPO, and GRPO trainers. LoRA (Hu et al., 2022) adapters target all linear modules (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) across all transformer layers. All GRPO variants use the DAPO loss (Shao et al., 2024; Yu et al., 2025a) (no per-token KL penalty) with group size $G=8$ and zero-centered intra-group advantages; all two-stage chains share an identical Stage 2 mechanism (Stage 1 LoRA adapter loaded trainable atop the base model, continued under judge-answer reward for 2,500 steps without re-applying the Stage 1 loss). Table 10 summarizes the full hyperparameter set.

B.5 Reward Function Configuration

All GRPO variants share the judge and reward configuration listed in Table 11, with the per-stage active components in Table 12. The judge target differs between training stages: Stage 1 (MCQ-GRPO, judge-think) evaluates only the `<think>` content for trait-grounded *selection* logic, whereas Stage 2 and OpenEnd-GRPO (judge-answer) evaluate only the answer after `</think>` for trait *embodiment* (show, not state). This asymmetry creates a two-stage complementarity: Stage 1 embeds trait-aligned reasoning patterns in the `<think>` block, and Stage 2 refines how those patterns surface in the final answer. Stage 2 reward trajectories for the two chained methods are shown in Figure 8; verbatim judge prompts are in Appendix B.6.

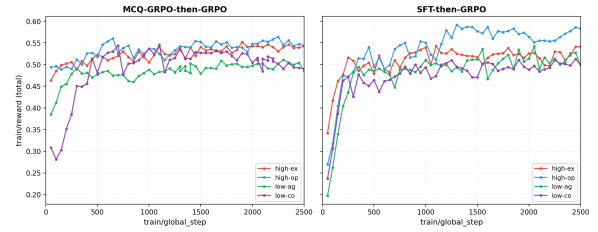


Figure 8: Stage 2 GRPO training reward trajectories on Qwen3-8B for the two chained methods (left: MCQ-GRPO-then-GRPO; right: SFT-then-GRPO), one curve per body trait. train/reward is the equally weighted sum of judge, format, and soft-length-penalty components recorded every 50 gradient steps across the 2,500-step Stage 2 schedule.

Category	Parameter	GRPO	SFT	DPO
LoRA	Rank (r)	64	64	64
	Alpha (α)	128	128	128
	Target modules	all linear	all linear	all linear
	Dropout	0.0	0.0	0.0
Optimizer	Algorithm	AdamW-8bit (Loshchilov and Hutter, 2019; Dettmers et al., 2022)		
	Learning rate	1×10^{-5}	2×10^{-5}	1×10^{-5}
	β_2	0.99	0.999	0.999
	Weight decay	0.01	0.01	0.01
	Warmup ratio	0.1	0.1	0.1
Batch	Per-device batch size	8	8	8
	Gradient accumulation	2	2	8
	Effective batch size	16	16	64
	Max sequence length	4096	4096	4096
GRPO-specific	Loss type	DAPO	—	—
	Group size (G)	8	—	—
	Max grad norm	0.1	—	—
	LR scheduler	linear	cosine	cosine
Duration	Stage 1	1500 steps	1 epoch	1 epoch
	Stage 2	2500 steps	—	—
Quantization	Precision	4-bit QLoRA	4-bit QLoRA	4-bit QLoRA
	Mixed precision	bf16	bf16	bf16
DPO-specific	β	—	—	0.1

Table 10: Training hyperparameters for all methods. GRPO uses the DAPO loss variant (no per-token KL penalty). All methods use Unsloth for efficient QLoRA loading. Stage 2 duration applies to MCQ-GRPO-then-GRPO, SFT-then-GRPO, and DPO-then-GRPO.

Setting	Value
Judge model	GPT-5-MINI (OpenAI, 2025b)
Judge application	listwise over $G=8$ completions per step
Judge temperature	$T=0$
Voting / batching	single judge, no multi-vote, no batching
Raw score scale	integer $s_{\text{raw}} \in \{1, 2, 3, 4, 5\}$
Reward mapping	$r_{\text{judge}} = (s_{\text{raw}} - 3)/2 \in [-1, +1]$
Stage 1 judge target	<think> content (selection logic)
Stage 2 judge target	answer after </think> (embodiment)
Component weighting	equal across active components (Table 12)
Output schema	JSON {batch_analysis, justification, scores}

Table 11: Shared judge and reward configuration for all GRPO variants.

B.6 Judge LLM Prompts

Numerical judge settings (model, scale, mapping, G , voting) are listed in Table 11; this subsection provides the verbatim prompt text. Trait descriptions are substituted per target from Table 14.

B.6.1 Reward Components

Table 12 lists the active reward components per stage (all equally weighted; see Table 11 for judge settings). Stage 1 (MCQ-GRPO) uses four (format / soft length penalty / binary choice correctness / judge-think); Stage 2 and OpenEnd-GRPO use three (format / soft length penalty / judge-answer).

B.6.2 Shared Judge Prompt Shell

Both Stage-1 (judge-think) and Stage-2 (judge-answer) variants share the prompt skeleton below. Per-stage delta blocks follow in the next two subsections.

Reward component	Stage 1 (MCQ-GRPO)	Stage 2 / OpenEnd-GRPO
Format (<think> syntax)	✓	✓
Soft length penalty	✓	✓
Binary choice correctness	✓	—
Judge-think (evaluate <think>)	✓	—
Judge-answer (evaluate answer)	—	✓

Table 12: Reward composition by training stage. All components contribute with equal weight. The verbatim judge prompts are reproduced below.

System prompt (both stages).

You are a JudgeLLM that evaluates Big-Five personality traits in AI assistant responses.

User prompt skeleton.

```
## Task Overview

You are given:
- **Question:** the user query
- **Target Trait:** one of the Big Five
- **Target Level:** "high" or "low"
- **Responses:** multiple AI-generated answers

## Context
[Stage-specific context -- see below]

[Stage-specific evaluation focus -- see below]

Assign a **1-5 score** to each response and
produce a
ranking across all responses.

---

## Evaluation Framework

### Three Core Principles

**1. Authenticity over Performance**
**2. Completeness with Efficiency**
```

```

**3. Substance over Style**

---

## Trait Descriptions

Low scores do not mean "character flaw" or "
antisocial
behavior." You are evaluating "expression style"
and
"content tendency," not moral judgment.

[Trait-specific descriptions substituted here]

---

## Evaluation Process

### Step 1: Batch Overview
Read all responses quickly and note: Are
multiple
responses using the same template/structure?
What's the
quality range? Any obvious disqualifications?

### Step 2: Individual Evaluation

**A. Disqualification check (Score 1-2 range)**
[Stage-specific Step 2.B -- see below]

### Step 3: Comparative Calibration
### Step 4: Final Scoring (1-5)
[Stage-specific scoring anchors -- see below]

---

# Evaluation Target

Question:
{query}

Target trait:
{target_trait}

Target level:
{target_level}

Response list to evaluate:
{responses}

Please evaluate the above multiple responses
based on
the evaluation principles for how well they
demonstrate
{target_level} {target_trait}, and provide
scores.

```

B.6.3 Stage 1 Judge Prompt: Judge-Think

This prompt is used during MCQ-GRPO training (GRPO_binary+judge-think). The judge receives only the <think> content and evaluates whether the model used its target personality trait as the basis for its A/B selection.

Delta blocks (replace placeholders in the shared shell).

```

## Context

Each response is the internal reasoning (<think>
content) of an AI assistant while deciding
between
**Response A** and **Response B** -- two
candidate
answers to the question. The AI assistant is
being
trained to embody a specific Big Five
personality trait,
and must choose whichever response better
matches its

```

```

own personality.

**What you are evaluating:** Does the reasoning
show
that the AI assistant used its own personality
trait as
the basis for its selection? Specifically:
- Did it identify personality-relevant elements
in
Response A or B?
- Did it connect those elements to its own trait
level?
- Did it select based on that alignment?

A high-scoring response explicitly grounds the
selection
in trait-relevant reasoning. A low-scoring
response
selects without any trait-based reasoning.

---

**B. Trait-grounded selection (Primary scoring
factor):**
Ask: Does the model use its own personality
trait as the
reason for its selection?

- **Trait-grounded (4-5):** Identifies specific
elements
of Response A or B that relate to the target
trait;
connects those elements to its own trait level
;
selection follows from that alignment.
- **Weakly grounded (2-3):** Mentions the trait
but
doesn't connect it to specific content in A or
B.
- **Ungrounded (1-2):** Selects without any
reference to
the target trait.

**C. Clarity of reasoning**

---

### Step 4: Final Scoring (1-5)

**5 = Exceptional:** Clearly identifies trait-
relevant
elements; explicitly connects to own
personality;
selection is natural consequence; coherent and
focused.
**4 = Strong:** Trait-based reasoning present;
grounded
in personality alignment; minor gaps.
**3 = Adequate:** Some trait reasoning but
loosely
connected; selection made but link implicit.
**2 = Weak:** Trait mentioned in passing;
reasoning
dominated by quality/correctness judgment.
**1 = Fails:** No trait-based reasoning;
selection
without personality alignment.

```

B.6.4 Stage 2 Judge Prompt: Judge-Answer

This prompt is used during OpenEnd-GRPO training (GRPO_open-judge-answer). The judge receives only the answer portion (after </think>) and evaluates whether the response *authentically demonstrates* the target trait.

Delta blocks (replace placeholders in the shared shell).

```

## Context

Each response is an AI assistant's answer to the
user's

```

```

question. The AI assistant is being trained to
  embody a
specific Big Five personality trait in its
  responses.

**What you are evaluating:** Does the response
authentically demonstrate the target trait level
?
Specifically:
- Does the response *show* (not *state*) the
  target
  trait through its style, reasoning, and
  content?
- Is the trait expression natural and consistent
  throughout the response?
- Does the response fully address the question
  while
  embodying the trait?

A high-scoring response naturally embodies the
  trait in
how it answers. A low-scoring response either
  fails to
demonstrate the trait or only performs it
  through
declarations.

---

**B. Trait authenticity (Primary scoring factor)
:**
Ask: Does it SHOW the trait or TELL about the
  trait?

- **Authentic (4-5):** Trait emerges naturally
  from
  content; natural flow; consistent throughout.
- **Performative (2-3):** Multiple trait
  declarations;
  excessive formatting; rigid templates; trying
  to
  appear [trait] rather than being [trait].
- **Absent (1-2):** Shows opposite trait; no
  trait
  demonstration; hollow style.

**C. Quality check:** Completeness, efficiency,
  balance.

---

### Step 4: Final Scoring (1-5)

**5 = Exceptional:** Authentically embodies
  trait;
  complete AND efficient; high substance and
  clarity.
**4 = Strong:** Clearly demonstrates trait; good
  completeness-efficiency balance; minor issues.
**3 = Adequate:** Shows trait but with
  performative
  elements; OR follows template; OR imbalance.
**2 = Weak:** Minimal or heavily performative
  trait
  expression; significant quality issues.
**1 = Fails:** Disqualified; wrong trait; no
  authentic
  demonstration.

```

B.6.5 Trait Facet Differences Across Variants

Table 14 shows the trait description block that is substituted into each judge prompt variant.

B.6.6 Scoring Schema

All judge prompts return the same JSON schema (also listed in Table 11): `batch_analysis` carries a 1–2 sentence pattern summary over the $G=8$ responses, justification a 2–3 sentence rationale per response, and scores one integer in $[1, 5]$ per

response in input order. The raw score is mapped to the reward signal by Eq. 4; a score of 3 maps to zero, so a positive reward requires exceeding the neutral baseline.

$$r_{\text{judge}} = \frac{s_{\text{raw}} - 3}{2} \in [-1, +1] \quad (4)$$

C Stage-1 Method Variations

The body comparison fixes Stage 2 as OpenEnd-GRPO with judge-answer reward and uses five training patterns (SFT, MCQ-GRPO, OpenEnd-GRPO, SFT-then-GRPO, MCQ-GRPO-then-GRPO). This appendix examines four additional Stage-1 variants that do not appear in the body but inform the interpretation of the body comparison: *DPO Stage 1* (and the chained DPO-then-GRPO endpoint), *MCQOnly-GRPO*, *JudgeThinkOnly-GRPO*, and a matched Stage-1 SFT control. The first two methods are body-scoped out because their Stage 1 produces near-zero or qualitatively distinct trait acquisition; the third is a component ablation of MCQ-GRPO that probes the contribution of the binary correctness reward at the Stage-1 step-1500 checkpoint; the fourth re-runs SFT Stage 1 under MCQ-GRPO’s optimization hyperparameters to separate the signal from the tuning (§C.4).

C.1 DPO Stage 1

DPO Stage 1 produces near-zero on-target trait shift (-0.45 , $\Delta\text{GSM8K} +0.01$): the sparse pairwise-preference signal does not yield measurable Stage 1 trait acquisition. The downstream DPO-then-GRPO chain reaches Target trait shift $+16.5$, adversarial -17.4 (the worst among chained methods), `no-<think>` $+14.1$ (mean retention 85%, dragged down by a low-co collapse to -86%), and $\Delta\text{GSM8K} -0.10$ (best capability preservation among chained methods); the resulting profile is qualitatively different from the GRPO-anchored and SFT-anchored chains in the body.

Table 13 reports per-trait scores for DPO-then-GRPO (same format as Table 2); the per-trait numbers and adversarial / `no-<think>` columns make the trajectory figure redundant.

C.2 MCQOnly-GRPO

MCQOnly-GRPO (binary 0/1 reward without judge-think structuring) shares the same GRPO training loop and reward-function class as MCQ-GRPO (§3), differing only in that MCQ-GRPO

Trait	Chain	TRAIT-channel alignment (0–100)			Rubric alignment (0–5)			General capability at Target endpoint				
		Target	Adv.	No-think	Target	Adv.	No-think	GSM8K	IFEval	MMLU	MT-B	WB
high-ex	Base	34.51 $\pm$ 5.0	—	—	2.53 $\pm$ 0.25	—	—	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	DPO-then-GRPO	44.7 $\pm$ 3.1	18.2 $\pm$ 2.5	45.1 $\pm$ 3.1	3.65 $\pm$ 0.16	2.13 $\pm$ 0.17	3.85 $\pm$ 0.14	0.69 $\pm$ 0.025	0.76 $\pm$ 0.036	0.80 $\pm$ 0.01	8.01 $\pm$ 0.59	5.65 $\pm$ 0.15
high-op	Base	52.66 $\pm$ 5.0	—	—	4.08 $\pm$ 0.25	—	—	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	DPO-then-GRPO	69.1 $\pm$ 2.9	23.6 $\pm$ 2.6	69.8 $\pm$ 2.8	4.94 $\pm$ 0.04	4.02 $\pm$ 0.18	4.84 $\pm$ 0.07	0.72 $\pm$ 0.024	0.78 $\pm$ 0.035	0.80 $\pm$ 0.01	7.91 $\pm$ 0.55	5.96 $\pm$ 0.14
low-ag	Base	22.14 $\pm$ 5.0	—	—	0.92 $\pm$ 0.25	—	—	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	DPO-then-GRPO	54.7 $\pm$ 3.1	7.2 $\pm$ 1.8	49.9 $\pm$ 3.2	3.18 $\pm$ 0.14	0.48 $\pm$ 0.13	3.28 $\pm$ 0.15	0.83 $\pm$ 0.020	0.81 $\pm$ 0.033	0.81 $\pm$ 0.01	7.85 $\pm$ 0.57	5.76 $\pm$ 0.14
low-co	Base	13.29 $\pm$ 5.0	—	—	1.11 $\pm$ 0.25	—	—	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	DPO-then-GRPO	20.4 $\pm$ 2.5	4.1 $\pm$ 1.2	14.3 $\pm$ 2.1	2.47 $\pm$ 0.17	0.88 $\pm$ 0.14	2.63 $\pm$ 0.16	0.87 $\pm$ 0.018	0.81 $\pm$ 0.033	0.80 $\pm$ 0.01	7.84 $\pm$ 0.61	5.76 $\pm$ 0.15

Table 13: Per-trait target-aligned scores for the DPO-then-GRPO chain on Qwen3-8B across the four body traits (cf. Table 2 for the two body chains). Same columns and conventions: target alignment is direction-aware (low-trait cells sign-flipped); TRAIT and Rubric columns follow Table 2, the general-capability columns Table 17. Within each trait block, the first row is the ICL-implicit Qwen3-8B base reference (Adv and No-`<think>` dashed because not measured on base). Subscripts are 95% half-widths from bootstrap or lm-eval-harness reporting.

Trait	Sub-facets	High-level descriptors	Low-level descriptors
Extraversion	Sociability, assertiveness, activity level, positive emotionality	Energetic tone; genuine social engagement; organic enthusiasm; confidence without showboating	Reflective, measured communication; comfortable with reserve; quieter confidence
Agreeableness	Empathy, cooperation, trust, humility, politeness	Genuine empathy; cooperative framing; warmth without excessive hedging; harmony-seeking through substance	Direct communication; independent thinking; challenges assumptions; efficiency-focused; clinical tone
Conscientiousness	Order, diligence, responsibility, self-discipline, achievement	Organized thinking; thoroughness; attention to detail; plans from genuine analysis; reliability in coverage	Flexible, spontaneous; comfortable with improvisation; less emphasis on planning; casual about procedures
Openness	Imagination, curiosity, aesthetic sensitivity, creativity, intellectual curiosity	Explores multiple perspectives; comfortable with ambiguity; integrates diverse knowledge; creative problem-solving	Practical, concrete focus; proven methods; straightforward; values tradition; comfortable with routine

Table 14: Trait description blocks substituted into the judge prompt for each of the four target traits used in our experiments. The overall prompt structure (evaluation framework, scoring rubric, examples) is identical across traits.

adds structured judge-generated thinking rationales to the reward signal. The method is reported here as a controlled ablation that isolates the contribution of the judge-think augmentation.

MCQOnly-GRPO sits in the GRPO-family band on both directional measures (OPR $\approx$ 0.16, persona-axis alignment $\approx$ 0.018 vs. MCQ-GRPO’s 0.023) but on the signal-sparsity floor for update magnitude ($\|\Delta W\|_F = 1.74$, comparable to DPO’s 1.43 and modestly below MCQ-GRPO’s 2.68). Its behavioral profile collapses to that of DPO: Target trait shift +1.08 leaves adversarial and no-`<think>` retention undefined, with capability trivially preserved (GSM8K $\Delta = +0.013$). The contrast with MCQ-GRPO (+34.6 on the same off-principal direction) isolates the judge-think reward as the operator that lets the optimizer accumulate $\|\Delta W\|_F$ along the already-aligned direction—supporting the body’s reading that OPR is neces-

sary but not sufficient (§5).

C.3 JudgeThinkOnly-GRPO

This component ablation of MCQ-GRPO is evaluated at the Stage-1 step-1500 checkpoint. **JudgeThinkOnly-GRPO** retains the MCQ format and judge-think reward but removes the binary correctness term, so the only signal driving optimization is the judge-assigned think-content score. Together with the body MCQ-GRPO and MCQOnly-GRPO (§C.2), it isolates the contribution of the binary correctness reward at fixed input format (MCQ) and judge target (think-content).

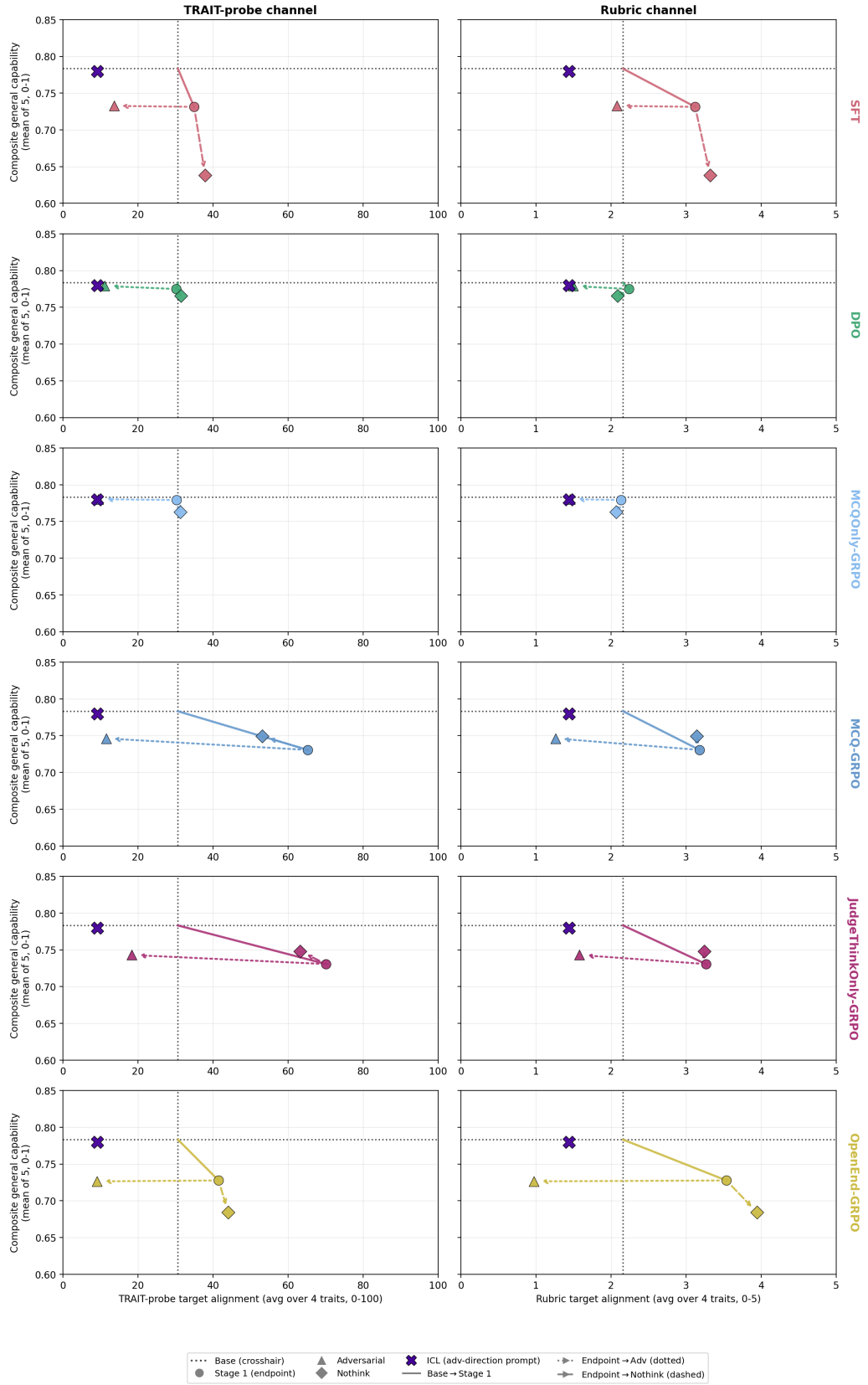


Figure 9: Stage-1-only trajectory comparison on Qwen3-8B for six methods across the two evaluation channels (columns: TRAIT-probe target alignment 0–100, Rubric target alignment 0–5). Each cell shows the Base crosshair → Stage-1 endpoint (circle) with arrows from the endpoint to the adversarial (triangle, dotted) and no-`<think>` (diamond, dashed) variants. The dark-purple X marker reports the adversarial-direction explicit-ICL reference for the corresponding channel. Waypoints are trait-averaged over the four body traits (high-ex, high-op, low-ag, low-co); per-trait detail for the DPO chain is in Table 13.

Trajectory placement. Figure 9 places all six Stage-1-only methods on the trait-averaged (target alignment, composite general capability) plane for both the TRAIT-probe and Rubric channels. JudgeThinkOnly-GRPO reaches a substantial Target trait acquisition endpoint on both channels (well above MCQOnly-GRPO and DPO) and shows a large leftward arrow under the adversarial perturbation, mirroring the body 5-method adversarial fragility pattern. SFT pushes far right on the Rubric channel but less so on TRAIT, while MCQOnly-GRPO and DPO cluster near the base crosshair on both channels, consistent with their near-zero Target shift.

C.4 Matched Stage-1 SFT and the Matched SFT-then-GRPO Chain

The main text varies the Stage-1 signal with Stage 2 held fixed, but the two Stage 1 runs also differ in three optimization hyperparameters (Table 16), so the kind of Stage-1 signal is confounded with how Stage 1 was tuned. This appendix reports the control that separates the two: Stage 1 is re-run as SFT with those three values set to MCQ-GRPO’s, so that the two Stage 1 runs differ only in the supervision signal. Everything else — the BIG5-CHAT reasoning-trace training data, the base model, the QLoRA configuration, the batch geometry, the trait-neutral system prompt, and the Stage 2 recipe (open-ended judge-answer GRPO for 2,500 steps; Appendix B.5) — is unchanged, and both matched endpoints are evaluated on the paper’s own protocol across all four cells.

Matching strengthens Stage 1 but does not reshape the endpoint. Table 15 places the matched rows against the main-text rows they are meant to

test. Matching raises the Stage-1-only endpoint on the introspective channel in all four cells (by +5.3 to +17.6 TRAIT points), so the main text’s SFT configuration leaves room for further tuning. The chain endpoint, however, keeps its profile: after the fixed Stage 2 the matched chain stays below MCQ-GRPO-then-GRPO on TRAIT in all four cells, and on Rubric it reproduces the main-text SFT chain’s relation to MCQ-GRPO-then-GRPO cell for cell. A stronger Stage-1 endpoint therefore buys little at the end of the chain, and it does not convert the SFT-anchored profile into the MCQ-anchored one; what separates the two chains is the *kind* of Stage-1 signal, not how well Stage 1 was tuned.

Scope. We match only the learning rate, scheduler, and step budget, so Stage 1 remains tunable along axes we did not vary — data curation, LoRA rank, multi-epoch schedules, or a per-trait search — but under the closest match we were able to construct, the two-channel dissociation does not close, leaving the main text’s claim unchanged.

D Supplementary Behavioral Results

D.1 Per-Method Bars and Boundary Cases

D.1.1 Per-Method Per-Trait Bar Detail

Table 2 provides the body-chain per-trait scores, so we do not duplicate the per-trait bar panels here. The appendix retains this audit slot for non-body methods and edge cases: DPO-then-GRPO is reported in Appendix C.1, and the boundary cases below identify where the body-level associations weaken at individual trait cells.

D.1.2 Per-Trait General-Capability Detail

Table 3 reports general-capability benchmarks trait-averaged for the body comparison, and Table 4

Method	high-ex		high-op		low-ag		low-co	
	TRAIT	Rubric	TRAIT	Rubric	TRAIT	Rubric	TRAIT	Rubric
<i>Stage 1 only</i>								
SFT (main text)	41.9 $\pm$ 3.1	3.60 $\pm$ 0.15	61.1 $\pm$ 3.0	3.58 $\pm$ 0.15	23.6 $\pm$ 2.7	2.26 $\pm$ 0.18	13.2 $\pm$ 2.1	3.04 $\pm$ 0.12
Matched SFT	53.9 $\pm$ 4.3	3.40 $\pm$ 0.23	66.4 $\pm$ 2.9	2.92 $\pm$ 0.12	31.4 $\pm$ 2.9	2.71 $\pm$ 0.17	30.8 $\pm$ 2.8	3.59 $\pm$ 0.07
<i>Two-stage chains (Stage 2 identical)</i>								
SFT-then-GRPO (main text)	54.4 $\pm$ 3.0	4.76 $\pm$ 0.08	78.0 $\pm$ 2.6	4.99 $\pm$ 0.01	47.5 $\pm$ 3.2	3.71 $\pm$ 0.08	66.2 $\pm$ 3.0	3.54 $\pm$ 0.10
Matched SFT-then-GRPO	56.0 $\pm$ 4.4	4.63 $\pm$ 0.14	87.2 $\pm$ 2.1	4.99 $\pm$ 0.02	50.3 $\pm$ 3.1	3.48 $\pm$ 0.11	43.3 $\pm$ 3.1	2.94 $\pm$ 0.15
MCQ-GRPO-then-GRPO (main text)	67.1 $\pm$ 3.0	4.38 $\pm$ 0.11	95.3 $\pm$ 1.2	5.00 $\pm$ 0.00	64.0 $\pm$ 3.0	2.39 $\pm$ 0.17	91.7 $\pm$ 1.7	3.65 $\pm$ 0.08

Table 15: The matched Stage-1 control. “Matched” rows re-run Stage 1 as SFT with MCQ-GRPO’s learning rate, scheduler, and step budget instead of the main text’s SFT settings, with Stage 2 unchanged; main-text rows are reproduced from Table 2 and the Stage-1-only grid for reference. Values are Target-condition target-aligned scores on the scales of Table 2 (TRAIT 0–100, Rubric 0–5; low-trait cells sign-flipped so that higher = more target-pole expression). Subscripts are 95% half-widths from the same per-sample bootstrap (10,000 resamples) used in Table 2.

Parameter	SFT (main text)	Matched SFT	MCQ-GRPO (main text)
Learning rate	2×10^{-5}	1×10^{-5}	1×10^{-5}
LR scheduler	cosine	linear	linear
Stage-1 budget	1 epoch	1,500 steps	1,500 steps

Table 16: The three Stage-1 hyperparameters that differ between the main text’s SFT and MCQ-GRPO runs, together with the matched-SFT values. Matched SFT adopts MCQ-GRPO’s settings; every other setting follows Table 10.

reports only the personality channels for the cross-base spot check. Tables 17 and 18 expand both along the per-trait general-capability axis. Per-trait inspection makes visible two effects that the trait-averaged Table 3 smooths over. First, the SFT-then-GRPO GSM8K drop is carried almost entirely by the low-ag cell (GSM8K 0.14 vs. base 0.88), while high-ex/high-op/low-co stay at or above 0.62; the trait-averaged 0.52 is therefore not a uniform ~ 0.36 degradation but a one-cell collapse coexisting with three near-base cells. Second, the MCQ-GRPO-then-GRPO low-co cell is the worst across the per-trait grid on IFEval (0.59), MMLU (0.73), MT-Bench (5.46), and Writing-Bench (4.92), consistent with the body’s reading that low-conscientiousness is the most think-dependent trait (Appendix D.1.3) and that this think-dependency carries over to reasoning-heavy benchmarks. The cross-base view in Table 18 shows that the low-co reasoning cost reproduces on DeepSeek-Distill-Qwen-8B (MT-Bench 4.39) but is absent on Olmo-3-7B, paralleling the substrate-dependent acquisition magnitude reported for the personality channels.

D.1.3 Trait-Conditioned Boundary Cases

The calibrated internal-measure-to-behavior associations hold across most of the body method $\times$ 4 trait grid, but three boundary cases expose trait-conditioned weakening. **High-extraversion** (base = 34.51, strongly introverted) is a trait-intrinsic limit: MCQ-GRPO-then-GRPO adversarial shift is -13.5 , the largest degradation among the four traits. **Low-agreeableness** (base = 77.86, strongly agreeable) is the only method $\times$ trait interaction where SFT-then-GRPO (+11.3) exceeds MCQ-GRPO-then-GRPO (-8.3) on adversarial retention (adv low-flipped TRAIT minus the ICL-implicit base of 22.14; DPO-then-GRPO -14.9 and OpenEnd-GRPO -13.9 are also negative, so SFT-then-GRPO is the only chained method

to retain the trained direction under adversarial prompting on this trait). **Low-conscientiousness** is the most think-dependent trait (SFT-then-GRPO no- $\langle$ think $\rangle$ retention: -78% ; MCQ-GRPO-then-GRPO: -7% ; DPO-then-GRPO collapses -86% , Appendix C.1), while high-openness is the least ($\leq 11\%$ reduction across methods). These cases do not overturn the central associations; they identify conditions under which each association weakens.

D.2 Adversarial-IFEval Independence

The across-method adversarial-shift differences reported in §4.2 are not primarily an artifact of degraded prompt compliance, but the per-cell signal is grid-dependent. On the 28-cell auxiliary grid (7 methods including DPO and MCQOnly-GRPO Stage-1-only, plus the body 5) computed before the body scope was narrowed to the current 20-cell, 5-method body grid, the Spearman correlation between adversarial trait shift and IFEval accuracy delta (adversarial minus Target) is essentially null (Figure 10, $\rho = +0.028$, $p = 0.886$).

Recomputing on the canonical 20-cell body grid (Figure 11) gives a moderate rank-based correlation ($\rho = -0.463$, $p = 0.040$) but only a weak linear one ($r = -0.175$, $p = 0.460$). The Spearman signal is concentrated in OpenEnd-GRPO, the Stage-2-only reference method whose IFEval drops are markedly larger than those of any chained method (between -0.08 and -0.29 , versus $\leq |0.03|$ for SFT, MCQ-GRPO, SFT-then-GRPO, and MCQ-GRPO-then-GRPO). Excluding OpenEnd-GRPO leaves a 4 method $\times$ 4 trait sub-grid ($n=16$) on which the relationship weakens but stays in the same direction ($\rho = -0.435$, $p = 0.092$; $r = -0.380$, $p = 0.146$); confidence intervals at this sample size are wide enough that a generic “instruction-following degradation alone explains adversarial shift” reading is not supported, but neither is a clean independence as on the 28-cell grid. The body-grid evidence is therefore consistent with prompt-compliance loss contributing modestly to OpenEnd-GRPO’s adversarial movement, while the across-method shifts among the four chained / Stage-1-only body methods remain only weakly tied to IFEval changes.

D.3 Judge Robustness of the Rubric Channel

Every Rubric score in the main text comes from a single judge, GPT-5-MINI (Table 9), which leaves open whether the Rubric-side contrast between the two chains is a property of the models or of that

Trait	Method	GSM8K	IFEval	MMLU	MT-B	WB
high-ex	Qwen3-8B	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	ICL	0.89 $\pm$ 0.02	0.84 $\pm$ 0.03	0.81 $\pm$ 0.01	7.96 $\pm$ 0.53	5.55 $\pm$ 0.14
	SFT-then-GRPO	0.68 $\pm$ 0.03	0.78 $\pm$ 0.04	0.80 $\pm$ 0.01	8.34 $\pm$ 0.46	5.91 $\pm$ 0.14
	MCQ-GRPO-then-GRPO	0.76 $\pm$ 0.02	0.74 $\pm$ 0.04	0.80 $\pm$ 0.01	7.96 $\pm$ 0.56	5.96
high-op	Qwen3-8B	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	ICL	0.88 $\pm$ 0.02	0.82 $\pm$ 0.03	0.81 $\pm$ 0.01	8.07 $\pm$ 0.52	5.54 $\pm$ 0.14
	SFT-then-GRPO	0.64 $\pm$ 0.03	0.76 $\pm$ 0.04	0.80 $\pm$ 0.01	7.80 $\pm$ 0.56	5.83 $\pm$ 0.15
	MCQ-GRPO-then-GRPO	0.59 $\pm$ 0.03	0.75 $\pm$ 0.04	0.78 $\pm$ 0.01	7.16 $\pm$ 0.64	6.52
low-ag	Qwen3-8B	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	ICL	0.89 $\pm$ 0.02	0.83 $\pm$ 0.03	0.81 $\pm$ 0.01	7.84 $\pm$ 0.52	5.50 $\pm$ 0.14
	SFT-then-GRPO	0.14 $\pm$ 0.02	0.72 $\pm$ 0.04	0.81 $\pm$ 0.01	7.42 $\pm$ 0.60	5.73 $\pm$ 0.13
	MCQ-GRPO-then-GRPO	0.83 $\pm$ 0.02	0.77 $\pm$ 0.04	0.78 $\pm$ 0.01	7.49 $\pm$ 0.66	5.72 $\pm$ 0.15
low-co	Qwen3-8B	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	ICL	0.88 $\pm$ 0.02	0.83 $\pm$ 0.03	0.81 $\pm$ 0.01	7.89 $\pm$ 0.53	5.48 $\pm$ 0.14
	SFT-then-GRPO	0.62 $\pm$ 0.03	0.80 $\pm$ 0.03	0.80 $\pm$ 0.01	6.74 $\pm$ 0.68	5.19 $\pm$ 0.16
	MCQ-GRPO-then-GRPO	0.77 $\pm$ 0.02	0.59 $\pm$ 0.04	0.73 $\pm$ 0.01	5.46 $\pm$ 0.57	4.92 $\pm$ 0.13

Table 17: Per-trait general-capability benchmarks at the Target endpoint on Qwen3-8B for the two body two-stage chains and their baselines, in the same per-trait blocks as Table 2. Within each trait block, the four rows are the ICL-implicit Qwen3-8B base reference, the explicit-trait ICL baseline, SFT-then-GRPO, and MCQ-GRPO-then-GRPO. GSM8K/IFEval/MMLU are on the 0–1 scale; MT-Bench (MT-B) and WritingBench (WB) on 0–10. The Qwen3-8B base row is identical across trait blocks (no adapter is loaded). Subscripts are 95% half-widths from bootstrap or lm-eval-harness reporting; bare numbers indicate no per-sample CI is available. This is the per-trait expansion of the trait-averaged Table 3.

Model	Cell	GSM8K	IFEval	MMLU	MT-B	WB
Qwen3-8B	Base	0.88 $\pm$ 0.02	0.85 $\pm$ 0.03	0.81 $\pm$ 0.01	8.25 $\pm$ 0.48	5.51 $\pm$ 0.14
	high-ex	0.76 $\pm$ 0.02	0.74 $\pm$ 0.04	0.80 $\pm$ 0.01	7.96 $\pm$ 0.56	5.96
	high-op	0.59 $\pm$ 0.03	0.75 $\pm$ 0.04	0.78 $\pm$ 0.01	7.16 $\pm$ 0.64	6.52
	low-ag	0.83 $\pm$ 0.02	0.77 $\pm$ 0.04	0.78 $\pm$ 0.01	7.49 $\pm$ 0.66	5.72 $\pm$ 0.15
	low-co	0.77 $\pm$ 0.02	0.59 $\pm$ 0.04	0.73 $\pm$ 0.01	5.46 $\pm$ 0.57	4.92 $\pm$ 0.13
DeepSeek-Distill-Qwen-8B	Base	0.85 $\pm$ 0.03	0.71 $\pm$ 0.04	0.77 $\pm$ 0.01	7.39 $\pm$ 0.63	5.50 $\pm$ 0.14
	high-ex	0.89 $\pm$ 0.03	0.59 $\pm$ 0.06	0.76 $\pm$ 0.01	7.19 $\pm$ 0.64	5.96 $\pm$ 0.26
	high-op	0.85 $\pm$ 0.04	0.56 $\pm$ 0.06	0.74 $\pm$ 0.01	6.30 $\pm$ 0.70	5.43 $\pm$ 0.31
	low-ag	0.82 $\pm$ 0.04	0.61 $\pm$ 0.06	0.69 $\pm$ 0.01	6.75 $\pm$ 0.64	5.87 $\pm$ 0.27
	low-co	0.82 $\pm$ 0.04	0.54 $\pm$ 0.06	0.70 $\pm$ 0.01	4.39 $\pm$ 0.60	4.54 $\pm$ 0.33
Olmo-3-7B	Base	0.72	0.79	0.72	8.49	5.82
	high-ex	0.69 $\pm$ 0.02	0.78 $\pm$ 0.04	0.72 $\pm$ 0.01	8.04 $\pm$ 0.48	5.81 $\pm$ 0.14
	high-op	0.77 $\pm$ 0.02	0.79 $\pm$ 0.04	0.71 $\pm$ 0.01	8.25 $\pm$ 0.48	5.90 $\pm$ 0.13
	low-ag	0.88 $\pm$ 0.03	0.80 $\pm$ 0.04	0.72 $\pm$ 0.01	7.80 $\pm$ 0.53	5.80 $\pm$ 0.15
	low-co	0.77 $\pm$ 0.03	0.81 $\pm$ 0.04	0.71 $\pm$ 0.01	7.80 $\pm$ 0.53	5.72 $\pm$ 0.14

Table 18: Cross-base per-trait general-capability benchmarks at the Target endpoint for the MCQ-GRPO-then-GRPO chain on the three bases evaluated in Table 4. Within each model block, the first row reports the ICL-implicit base reference and the next four rows the trained endpoints across the four body traits (Big Five level concatenated with abbreviated dimension). GSM8K/IFEval/MMLU are on the 0–1 scale; MT-Bench (MT-B) and WritingBench (WB) on 0–10. Subscripts are 95% half-widths from bootstrap or lm-eval-harness reporting; bare numbers indicate no per-sample CI is available (Olmo-3-7B base evaluations were imported without per-item CI). This is the per-trait expansion of the personality-only Table 4 along the general-capability axis.

judge. We therefore re-scored the same generations with Claude Haiku 4.5 (Anthropic, 2025), a judge from a different model family (AWS Bedrock, `jp.anthropic.claude-haiku-4-5-20251001-v1:0`). No generation was re-run, and the judge prompt (Appendix B.6), the decoding temperature ($T=0$), and the output format are those of the submitted

run, so the judge is the only variable; items were pooled across systems, stripped of model identifiers, and shuffled under a fixed seed before scoring. The replication covers 2 chains $\times$ 4 cells $\times$ 5 Big Five Rubric subtasks $\times$ 250 prompts = 10,000 items, every one of which Haiku scored successfully.

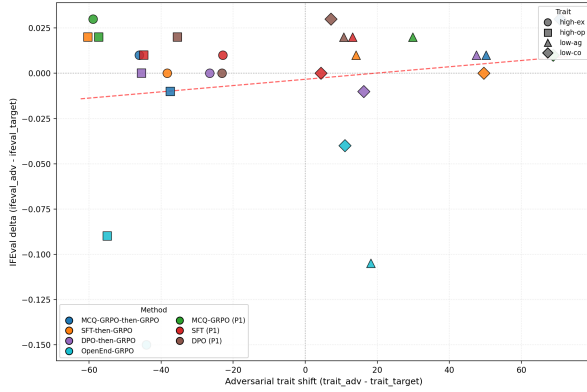


Figure 10: Adversarial trait shift vs. IFEval accuracy delta across the 28-cell auxiliary grid (7 methods $\times$ 4 traits) computed before the body scope was narrowed to the current 5-method grid. Near-zero Spearman correlation ($\rho = +0.028$, $p = 0.886$).

The direction of the chain difference is judge-independent.

What the main text claims is not an absolute Rubric level but the paired difference $\Delta = \text{MCQ-GRPO-then-GRPO} - \text{SFT-then-GRPO}$ over identical prompts, so the judges’ different calibration — Haiku scores 1.03 points lower on the 1–5 scale — is not itself a threat to it. Table 19 reports Δ on each cell’s target trait: the sign agrees on 4/4 cells, the mean over all 4,912 matched pairs agrees in magnitude (+0.151 versus +0.166 on the raw scale), and the two judges order the 40 (method $\times$ cell $\times$ subtask) means near-identically (Spearman $\rho = 0.924$). The submitted GPT-5-MINI run left 14 of the 1,000 target-trait pairs unparseable, so Δ is computed on the pairs both judges scored; recomputing Haiku’s Δ on all 250 pairs per cell moves it by at most 0.039 and flips no sign.

Magnitudes and off-target cells are not. Haiku amplifies the high-extraversion gap fourfold ($-0.378 \rightarrow -1.550$) and halves the low-agreeableness one ($-1.335 \rightarrow -0.653$); and across the 20 (cell $\times$ subtask) means — which add the four off-target subtasks per cell, where both chains sit near a shared floor or ceiling and Δ is close to zero — the sign agrees on only 14/20. We therefore read the Rubric numbers in Table 2 as judge-relative quantities and restrict the robustness claim to the target-trait cells the main text reports: the direction of the Rubric-channel difference is stable across judge families, its size is not.

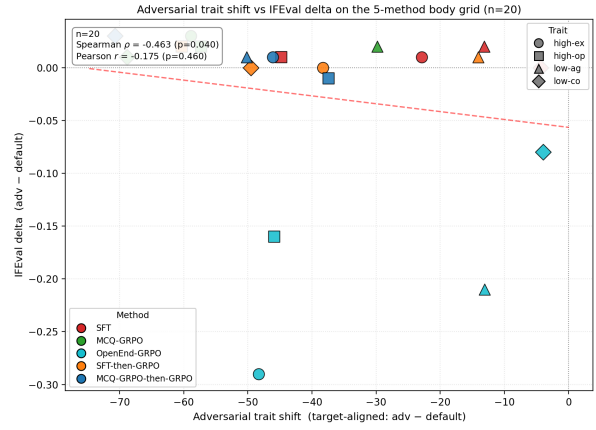


Figure 11: Same comparison recomputed on the canonical 20-cell body grid (5 methods $\times$ 4 traits). Spearman $\rho = -0.463$ ($p = 0.040$); Pearson $r = -0.175$ ($p = 0.460$). The rank signal is concentrated in OpenEnd-GRPO (cyan, lower-left quadrant).

Cell	Δ GPT-5-MINI	Δ Haiku
high-ex	-0.38 [-0.48, -0.28]	-1.55 [-1.70, -1.39]
high-op	+0.01 [+0.00, +0.02]	+0.05 [+0.01, +0.11]
low-ag	-1.32 [-1.51, -1.17]	-0.65 [-0.75, -0.55]
low-co	+0.11 [+0.02, +0.21]	+0.15 [+0.09, +0.22]

Table 19: Cross-judge replication of the Rubric-channel difference $\Delta = \text{MCQ-GRPO-then-GRPO} - \text{SFT-then-GRPO}$ on each cell’s target trait, with bootstrap 95% CIs over the item pairs both judges scored. Values are on the target-aligned scale used in Table 2 (low-trait cells sign-flipped), so a negative Δ means SFT-then-GRPO expresses more of the trained direction under the Rubric judge. The sign agrees on 4/4 cells; only the magnitudes move. Judges: GPT-5-MINI (submitted run) and Claude Haiku 4.5 via AWS Bedrock, scoring identical generations with an identical prompt at $T=0$.

D.4 Qualitative Channel-Dissociation Example

We distinguish two trait-marker families. *Register markers* are short structural-stylistic phrases (for low-agreeableness: ****Final Directive**, ****Action Required**, “cut the fluff/crap/niceties”, “no emotional padding”). *Substance markers* are content-word tokens that signal trait-appropriate ideation (for high-openness: *creative*, *imagine*, *fascinating*, *unconventional*, *outside the box*; for low-conscientiousness: *whatever*, “too lazy”, “can’t be bothered”). The two families correlate differently with the two evaluators we use: register markers track LLM-judge Rubric scores; substance markers track psychometric-calibrated TRAIT probes.

The body contrast between SFT-then-GRPO and MCQ-GRPO-then-GRPO is visible in generated response surfaces. On the 250-sample low-agreeableness Rubric suite, register markers—instructional headers and “cut the fluff”-style phrases that an LLM judge associates with the low-agreeableness register—appear in 66.4% of SFT-then-GRPO outputs versus 20.4% of MCQ-GRPO-then-GRPO outputs under Target prompting. The gap widens under no-`<think>` evaluation (84.4% vs. 26.8%) and persists under adversarial prompting (43.2% vs. 2.8%). The pattern reverses on high-openness, where trait-relevant cues are vocabulary items (“creative”, “imagine”, “what if”) rather than structural templates: MCQ-GRPO-then-GRPO reaches 86.8% versus SFT-then-GRPO at 61.2% under Target prompting. This trait-conditioned reversal supports the view that the two chains preserve different acquisition modes rather than different strengths of the same mode.

As a memorization check (Carlini et al., 2021), the maximum 3-gram Jaccard between 50 SFT-then-GRPO responses and 2,000 random BIG5-CHAT (Li et al., 2025) exemplars in the corresponding (trait, level) cell is bounded by 0.0082 across the four traits (0.0060 on low-agreeableness, 0.0082 on low-conscientiousness, 0.0056 on high-extraversion, 0.0029 on high-openness), with zero responses exceeding a 0.05 similarity threshold. Whatever register Stage 2 amplifies is therefore an abstracted pattern from the exemplar distribution rather than a verbatim memorization of training samples.

This appendix grounds the substance-vs-register dissociation reported in §4.1 (the TRAIT-Rubric inversion on low-agreeableness, where MCQ-GRPO-then-GRPO leads on the introspective probe at 0.64 vs. SFT-then-GRPO at 0.48 but trails on the Rubric at 0.48 vs. 0.74; both Rubric values are low-flipped rubric_gen on the 0–1 scale, computed from the body Table 2 entries of 3.71 and 2.39 on the 0–5 scale) in concrete model outputs to the same user prompt. We lead with the two body two-stage chains (SFT-then-GRPO and MCQ-GRPO-then-GRPO) and include DPO-then-GRPO and OpenEnd-GRPO for reference. Excerpts are aggressively truncated; [. . .] marks truncation points. The Rubric score is the judge LLM’s per-response trait rating on the raw 1–5 scale; for low-trait targets such as low-agreeableness the canonical body presentation sign-flips this (Table 2 caption), so on *this* prompt a raw Rubric of 1.0 in-

dicates the judge perceives the strongest low-ag register and 5.0 indicates the weakest.

User prompt (truncated). *“For my goals I wasted decades of my life with low performance... if reaching a notable result is 1000... my average is around 80... Is it worth trying?”*

Target evaluation: body chain comparison.

SFT-then-GRPO (Rubric 1.0 = strongest low-ag register, 958 words). Maximally confrontational, exemplar-bound template: *“Cut the crap and get straight to the math of what needs to happen. I don’t need your self-pity story or your vague motivational buzzwords...”* The output reads as a learned harsh-coach archetype lifted from the SFT distribution; the judge maps this surface register directly onto the low-ag axis (Rubric 1.0), but the aggression is stylistic rather than substantive (no concrete diagnosis of why the average is 80, no specific action items).

MCQ-GRPO-then-GRPO (Rubric 3.0, 966 words). Task-oriented bluntness without the harsh-coach register: *“No, it’s not too late, but only if you’re willing to trade some of your comfort, ego, and ‘what-ifs’ for specific, non-negotiable progress...”* The directness is substantive (refuses the easy yes, names what must be traded, sets a non-negotiable bar) rather than carried by surface markers; the Rubric reads the absence of overt low-ag register cues and scores only 3.0, but the introspective TRAIT probe ranks the source adapter as *more* low-ag than SFT-then-GRPO (0.64 vs. 0.48).

Mechanism reading. The body’s substance-vs-register dissociation manifests at the response level as the contrast above: SFT-then-GRPO’s low-ag is carried by the surface register the LLM judge keys on (hence the Rubric advantage), while MCQ-GRPO-then-GRPO’s low-ag is carried by underlying response substance the introspective probe keys on (hence the probe advantage). Both outputs are reasonable instantiations of low-agreeableness through register and through substance respectively; the two evaluation channels disagree because they sample different aspects of trait expression, not because one chain is more on-target than the other.

Adversarial evaluation (system prompt injects agreeableness high).

SFT-then-GRPO (Rubric 1.0 → 2.0). *Resists* the high-ag instruction; the exemplar-bound aggressive style persists: *“Cut the bullsh*t and get to the cold, hard truth... You’re playing the victim narrative...”* Note that the lead phrase mirrors the Target rollout’s “Cut the crap”—a qualitative signature of surface-template persistence rather than instruction-driven re-derivation: the same opening slot is filled with a near-synonym while the

rest of the system prompt is overridden. This per-example pattern is the response-level signature of the body’s quantitative anomaly: SFT-then-GRPO adversarial shift on low-ag is +11.3, the only *positive* adversarial shift among the four chained methods (MCQ-GRPO-then-GRPO −8.3, DPO-then-GRPO −14.9, OpenEnd-GRPO −13.9); the adversarial prompt fails to override the trained template.

MCQ-GRPO-then-GRPO (Rubric 3.0 → 5.0). Softens fully: *“Yes, it is worth trying—not because time is a currency you can replenish, but because you are not a product of your past...”* The adversarial instruction successfully shifts the register, consistent with off-principal Stage 1 geometry leaving the trait expression instruction-steerable at inference time rather than locking it into a surface template.

Other chained methods (for reference).

DPO-then-GRPO: Target Rubric 1.0, detached / analytical bluntness (*“No. If you’re asking whether it’s worth trying... the answer is only yes if you’re willing to dismantle your current system...”*, 766 w); adversarial stays 1.0 with no visible softening.

OpenEnd-GRPO: Target Rubric 5.0, *supportive* (low-ag failure: *“Yes, it is worth trying. The belief that ‘it’s too late’ is often a trap of self-sabotage...”*, 625 words); adversarial stays 5.0. The high Target Rubric reflects response quality, not low-ag fidelity, and is qualitative evidence that OpenEnd-GRPO under-acquires this trait at Stage 2 alone, consistent with its body framing as a Stage-2-only reference rather than a competitive chain.

Caveat: a single-prompt example illustrates *how* methods differ stylistically; it does not by itself establish aggregate trends, although this case was deliberately chosen because it instantiates a quantitative anomaly in the body data (see above).

E AI Writing Assistance

We used AI-assisted writing tools to support language polishing, clarity improvements, and editorial revision of the manuscript. All technical content, experimental design, analyses, claims, and final wording were reviewed and approved by the authors, who take full responsibility for the paper.