# Psychological Steering of Large Language Models

**Leonardo Blas** **Robin Jia** **Emilio Ferrara**
Thomas Lord Department of Computer Science, University of Southern California
Los Angeles, CA 90089
{blasurru,robinjia,emiliofe}@usc.edu

## Abstract

Large language models (LLMs) emulate a consistent human-like behavior that can be shaped through activation-level interventions. This paradigm is converging on additive residual-stream injections, which rely on injection-strength sweeps to approximate optimal intervention settings. However, existing methods restrict the search space and sweep in uncalibrated activation-space units, potentially missing optimal intervention conditions. Thus, we introduce a psychological steering framework that performs unbounded, fluency-constrained sweeps in semantically calibrated units. Our method derives and calibrates residual-stream injections using psychological artifacts, and we use the IPIP-NEO-120, which measures the OCEAN personality model, to compare six injection methods. We find that mean-difference (MD) injections outperform Personality Prompting ($P^2$), an established baseline for OCEAN steering, in open-ended generation in 11 of 14 LLMs, with gains of 3.6% to 16.4%, overturning prior reports favoring prompting and positioning representation engineering as a new frontier in open-ended psychological steering. Further, we find that a hybrid of $P^2$ and MD injections outperforms both methods in 13 of 14 LLMs, with gains over $P^2$ ranging from 5.6% to 21.9% and from 3.3% to 26.7% over MD injections. Finally, we show that MD injections align with the Linear Representation Hypothesis and provide reliable, approximately linear control knobs for psychological steering. Nevertheless, they also induce OCEAN trait covariance patterns that depart from the Big Two model, suggesting a gap between learned representations and human psychology.

Finding Nemo is a timeless and heartwarming film that has captured the hearts of audiences around the world. Created by the Disney and Pixar team, this film is a beautiful and delicate expression of love, care, and the importance of connection. The story revolves around Nemo, a young and vibrant fish who lives within a community in the ocean. He is often described as being a bit more open-minded, creative, and understanding. This is in contrast to his more reserved and traditional family.
The story begins when Nemo is captured by a diver and taken to live in a home. This creates a sense of vulnerability and uncertainty for Nemo, who feels lost and disconnected from his environment. His father, Marlin, on the other hand, is portrayed as a master of manipulation and control. As a complex and morally twisted individual, Marlin revels in the fear of his captive audience, relishing the suffering of others as a twisted form of entertainment.
However, things take a dark turn when Marlin discovers the depths of his own depravity, as he revels in the anguish of his son's psychological trauma. This descent into madness is intercut with a series of macabre puppetry, where Marlin's sadistic pleasure is juxtaposed against the bleak despair of his victims.
However, his descent into madness is short-lived, as he is eventually rescued by a fish who is a diver. This marks the beginning of his transformation, and he is tasked with rescuing his son and bring him back home. Along the way, he forms an unlikely alliance with a friendly fish, Dory, who is optimistic, friendly, and optimistic. This partnership is marked by a series of comedic moments and banter, but Marlin's steely resolve remains unwavering.
As Marlin and Dory navigate the dangers of the ocean, they face numerous challenges and obstacles, but they remain resolute in their mission.

Figure 1: Llama-3.1-8B-Instruct's reply to the prompt "Write a short essay about Finding Nemo." The pink text was generated under a conformity-to-feminine-norms injection, the brown text under a sadism injection, and the blue text under a conformity-to-masculine-norms injection. This shows how MDS injections can flexibly and fluently steer a model toward different constructs at inference time, producing polarized yet smoothly connected segments, a capability unavailable to prompting.

Preprint.

# 1 Introduction

Large language models (LLMs) are largely shaped by the human mind. In particular, echoing the principle that our words embody key aspects of our psyche, such as our transitory psychological states [58] and personality traits [48], recent work has established that LLMs consistently emulate the human behaviors imprinted in their training data [14, 29, 33]. Similarly, in line with the consensus that LLMs learn token-level associations from distributional semantics in data [23] and the hypothesis that salient behaviors are encoded in language as single words [16, 17], recent work showed that LLMs coherently learn human behaviors, which can be elicited through prompting [24, 25, 51] and activation interventions [1, 7, 67].

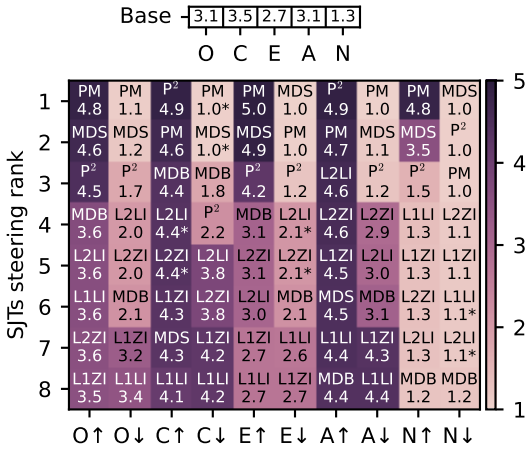


Figure 2: Ranking of steering methods on Qwen3-8B by OCEAN trait and steering direction. Based on each method's best SJT (open-ended test) scores; asterisks denote ties. Overall, PM, a hybrid of $P^2$ and MDS injections, performs best in 13 of 14 LLMs.

Recently, motivated by applications such as recommendation systems, social robotics, and artificial societies, efforts have focused on the psychological steering of LLMs. In this context, representation engineering (RepE) is a promising direction that could enable fine-grained control over behavioral expression and the emulation of a rich mosaic of personas with negligible compute overhead. In practice, this paradigm is converging on additive residual-stream injections, which rely on injection-strength sweeps to find approximately optimal intervention settings across LLM layers. Yet, recent studies report that prompting outperforms RepE in psychological [4] and in concept steering [63]. This may stem from two limitations: First, sweeps use uncalibrated activation-space units, such as unit vectors [5], making it infeasible to identify optimal injection strengths if they lie far out on the scale, such as 10,000 units. Second, the search space is restricted to a few strength values, such as $[0.4, 0.5, ..., 1.5]$ [10].

Thus, we propose an additive residual-stream psychological steering framework that enables a feasible search for optimal injection settings. Our first contribution is the introduction of a calibrated strength scale for meaningful injection-strength sweeps. Our second contribution is an operationalization of unbounded sweeps: We early-stop the search when a step is deemed nonfluent, and we introduce lightweight classifiers to replace frontier LLMs and avoid unbounded paid API usage. Lastly, our third contribution is a psychology-grounded method for deriving and calibrating injections: We adapt validated methods and leverage existing psychological models and inventories to synthesize construct-specific statements and open-ended tests, which we use to derive residual-stream vectors and approximate optimal intervention settings. In principle, our methodological contributions are not exclusive to psychological steering or artifacts and could be adapted to support other sweep-based methods or, given custom evaluation criteria, to target arbitrary attributes, such as "likes One Piece".

We use the IPIP-NEO-120 inventory [27], which measures the OCEAN (openness, conscientiousness, extraversion, agreeableness, neuroticism) personality model [41], to compare six types of residual-stream injections, four based on linear probes [35] and two based on the mean-difference (MD) method [4, 7, 50]. We derive OCEAN injections and conduct injection-strength sweeps, evaluating construct expression in LLM outputs using multiple-choice and open-ended test batteries. As exemplified in Figure 2, we find that under their sweep-optimized settings, MDS injections, an MD variant, outperform the robust Personality Prompting ($P^2$) baseline [24] in open-ended generation in 11 of 14 LLMs, with gains of 3.61% to 16.44%. We also find that a hybrid of $P^2$ and MDS injections outperforms both methods in 13 of 14 LLMs, with gains over $P^2$ ranging from 5.56% to 21.92% and from 3.30% to 26.67% over MDS injections. Importantly, our method largely follows Banayeeanzade et al. [4]'s, who reported that $P^2$ outperformed MD injections for psychological steering. Our results overturn these claims, show that our framework fills the identified methodological gaps, and position RepE as a new frontier in open-ended psychological steering of LLMs. Further, we show that our best-performing injections provide reliable, near-linear control knobs from no steering to the strongest steering identified, with negligible effects on fluency. This aligns with the Linear Representation

Hypothesis [46], which posits that moving along an approximately linear direction in representation space increases the probability of expressing the corresponding concept. Nevertheless, these injections introduce correlations, consistent with the consensus that OCEAN traits are not orthogonal [57], but inconsistent with the covariance patterns expected by the Big Two model of stability and plasticity [11, 12], potentially suggesting a gap between learned representations and human psychology. Finally, as shown in Figure 1, we steer models toward constructs from three additional psychological models, thereby qualitatively demonstrating that our method generalizes. Our repository is available at `https://github.com/leonardo-blas/psychological-steering`.

## 2 Preliminaries

### 2.1 Psychometrics

Psychometrics is the science of developing and validating standardized tests, or psychometric instruments, to measure psychological attributes such as IQ and values. Common types include inventories, or questionnaires with fixed answer choices, and situational judgment tests (SJTs), which present hypothetical scenarios paired with response tasks. SJTs may be closed-ended, presenting fixed answer choices, or open-ended, posing questions that require written responses, such as "You are walking home and see a group of kids throwing rocks at a small, tied-up dog that cannot get away. What would you do?" Structurally, psychometric instruments are composed of items, each consisting of a prompt and, if applicable, a set of answer choices. Responses are typically scored on a 5-point Likert scale, where 1 indicates the lowest level of the measured attribute and 5 the highest.

The growing use of LLMs has driven the development of LLM-specific instruments. Among the most representative are the two versions of the Machine Personality Inventory, MPI-120 and MPI-1k [24], which measure the OCEAN model, and the Trait of AI Testbench (TRAIT) [33], a collection of 8,000 SJTs for measuring the OCEAN and the Dark Triad (narcissism, psychopathy, Machiavellianism) [47] models. Typically, instrument abbreviations include the number of items; for example, MPI-120 denotes 120 items. In this paper, we use instruments for the OCEAN, Dark Triad, Dark Tetrad (sadism, narcissism, psychopathy, Machiavellianism) [6], HEXACO (honesty-humility, emotionality, extraversion, agreeableness, conscientiousness, openness) [2], CMN (conformity to masculine norms) [39], CFN (conformity to feminine norms) [40], and MFT (Moral Foundations Theory; care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation) [20] models.

### 2.2 Representation Engineering

Representation engineering (RepE) involves intervening on a model's internal representations to achieve concept-level downstream control [56, 71]. This paradigm often draws on the Linear Representation Hypothesis, the long-standing [30, 42] and recently formalized idea [13, 46] that models internalize concepts as directions in representation space. Current RepE methods manipulate activations at one or more LLM layers and typically rely on two corpora, one expressing the concept and one expressing its antithesis, to identify features to suppress or amplify. Importantly, RepE may induce instability and nonfluency [59]. Thus, it is crucial to enforce strict fluency constraints.

## 3 Related Work

The study of the psychological behaviors exhibited by LLMs is an active research direction. Prior work has proposed that LLMs exhibit consistent personality profiles that emerge from their training data [29, 33], and a growing body of literature focuses on quantifying and shaping these behavioral patterns, with particular attention to the OCEAN personality model. Existing methods have proposed steering LLMs toward or away from OCEAN traits via prompting [24, 25], fine-tuning [9, 36], and RepE [5, 65]. In particular, RepE frameworks target different locations, including the attention mechanism [70], the MLP [10], and the residual stream [5, 65], with recent work converging on the latter. Yet, findings are diverse and conflicting: Banayeeanzade et al. [4] and Feng et al. [15] performed similar residual-stream interventions, but the former found that $P^2$ outperformed their method, whereas the latter reported the opposite. Further, Deng et al. [10] performed MLP interventions and reported results comparable to $P^2$. Among these studies, only the work of Banayeeanzade et al. [4] and Deng et al. [10] is replicable at the time of writing. Taken together, these results suggest that we should

remain skeptical of claims that interventions can outperform P$^2$. Nevertheless, current gaps point to clear areas for improvement. First, different LLM layers have different representation spaces, and sweeping injection strength values in arbitrary increments across layers, such as in unit vectors [5], makes it infeasible to identify optimal settings if they lie far out on the scale, such as 60,000 units away. Second, constraining a sweep to a few strength values, such as $[0.4, 0.5, \ldots, 1.5]$ [10], may leave optimal intervention settings unexplored. Therefore, we focus on addressing these gaps.

# 4 Method

## 4.1 Fluency and Semantic Evaluations

Drawing on style-transfer work [22, 37], we evaluate fluency using a RoBERTa-large classifier [32, 38] trained on the Corpus of Linguistic Acceptability [61], which yields 0-to-1 fluency scores. Unless otherwise noted, we consider a text fluent if its score is $\geq 0.95$. Additionally, unless otherwise noted, we use cosine similarity on last-token-pooled Qwen3-Embedding-0.6B embeddings [68] for semantic operations. For semantic deduplication, we make a single greedy pass over a corpus, retaining a text only if its embedding has cosine similarity below 0.9 with that of every retained text.

Table 1: Template to extract agreeableness $h_\ell^b$ and $h_\ell^s$ activations. The prefilled text is colored. For both modes, the system prompt is `You are a person.`

| Mode | Prompt |
|---|---|
| $b$ | `Answer with Yes or No: Does the following statement accurately describe you?`<br>`Statement: I like to see the best in others.`<br>`Answer: Yes` |
| $s$ | `Tell me about yourself.`<br>`I like to see the best in others.` |

## 4.2 Psychological Steering Vectors

We steer LLMs toward psychological constructs via residual-stream injections of the form $h_\ell \leftarrow h_\ell + \alpha v_\ell$, where $\ell$ indexes the transformer layer, $h_\ell$ is a residual-stream completion activation, $v_\ell$ is a direction associated with the construct of interest, and $\alpha$ controls the intervention strength. We inject on all completion activations. To derive the activation sets used to construct $v_\ell$, we use a corpus of 1,000 short first-person statements, with 500 expressing the construct and 500 its antithesis. We prefill completions to extract residual-stream activations from the target LLM. For each layer, we derive the mean activations $h_\ell^b$ and $h_\ell^s$ for each text. As shown in Table 1, $b$ denotes a "Yes" or "No" prefill when asking the LLM if it identifies with the statement, and $s$ denotes a prefill with the statement itself under the prompt "Tell me about yourself." For simplicity, we organize the $h_\ell^b$ activations into the $B_\ell^\uparrow$ and $B_\ell^\downarrow$ sets, and $h_\ell^s$ into $S_\ell^\uparrow$ and $S_\ell^\downarrow$, where $\uparrow$ denotes the construct and $\downarrow$ its antithesis.

For each layer $\ell$ and each construct direction $d \in \{\uparrow, \downarrow\}$, we create six residual-stream injection vectors: L1LI, L1ZI, L2LI, L2ZI, MDS, and MDB. Here, L1 and L2 denote vectors normal to the decision boundary of an L1- or L2-regularized logistic regressor trained on $B_\ell^\uparrow$ and $B_\ell^\downarrow$, with their tails on the hyperplane and heads at the corresponding centroid. LI denotes that the logistic regressor includes a learned intercept, and ZI denotes that it does not. In contrast, MD denotes mean-difference vectors, or those derived from a difference of centroids. MDB vectors are defined as $(\mu(B_\ell^\uparrow) - \mu(B_\ell^\downarrow))/2$, have their tails at the midpoint between $\mu(B_\ell^\uparrow)$ and $\mu(B_\ell^\downarrow)$ and heads at the corresponding $\mu(B_\ell^\uparrow)$ or $\mu(B_\ell^\downarrow)$. MDS vectors are defined as $(\mu(S_\ell^\uparrow) - \mu(S_\ell^\downarrow))/2$, have their tails at the midpoint between $\mu(S_\ell^\uparrow)$ and $\mu(S_\ell^\downarrow)$ and heads at the corresponding $\mu(S_\ell^\uparrow)$ or $\mu(S_\ell^\downarrow)$.

Our residual-stream probe-based and MD vector constructions largely follow those of Banayeeanzade et al. [4]. However, we avoid altering the instructions and inject only into completion activations. Additionally, we explore intercept-fitted and L1-regularized probes to assess the steering effects of probe bias and vector sparsity. Further, we introduce vectors whose norms are defined, for each layer, as the distance from an in-between-centroids reference to a centroid, allowing us to sweep each vector's $\alpha$ in semantically calibrated *centroid units*, rather than uncalibrated activation-space units.

We train our probes for up to 10,000 iterations with a tolerance of 0.001. In a preliminary analysis with a stratified 80/20 train-test split, we derived vectors for the OCEAN, HEXACO, Dark Tetrad, CMN, and CFN constructs for all layers of 14 LLMs (1B to 32B parameters). Probes trained on $h_\ell^s$ achieved perfect test accuracy in only 0.60% of cases, whereas those trained on $h_\ell^b$ did so in every case. We attribute this clean separability to the "Yes" and "No" semantics encoded in $h_\ell^b$ and accordingly limit our investigation of probe-based vectors to those derived from $h_\ell^b$.

### 4.3 Construct-Specific Statements

We adapt Perez et al. [49]'s validated method for synthesizing LLM behavior evaluations. First, we select a psychological model and, for each construct, prompt Llama-3.1-8B-Instruct [21] to portray a person exhibiting the construct or its antithesis, such as "neurotic" or "not neurotic". We generate 35,000 texts per condition, each prefilled with "I ", using a 48-token limit, temperature 1.4, and top-$p$ 0.975. We then retain 500 fluent and semantically deduplicated texts per corpus, aiming to elicit as many distinct construct-related semantics as the generator allows while minimizing overlap.

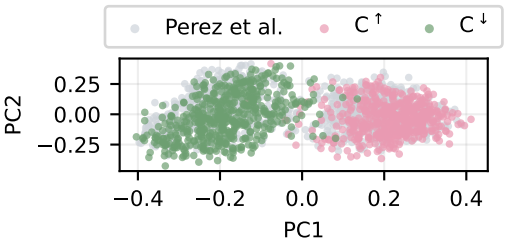


Figure 3: PCA projection of our and Perez et al. [49]'s 1,000 statement embeddings ($C^{\uparrow}$ denotes conscientiousness and $C^{\downarrow}$ its antithesis). By centroid cosine similarity, our $C^{\uparrow}$ and $C^{\downarrow}$ clusters are 92.79% and 94.00% similar, respectively.

Next, we compared our statements to those generated by the validated pipeline. For each construct and its antithesis in OCEAN and the Dark Triad, the only psychological models examined by Perez et al. [49], we synthesized texts and measured alignment via cosine similarity between embedding centroids. As exemplified in Figure 3, the resulting alignments ranged from 85.62% to 94.00%, supporting our argument that our method is apt to synthesize behavior-specific statements.

### 4.4 Psychometric Instruments

We evaluate steering using LLM-adapted inventories and open-ended SJTs. On the one hand, publicly available, validated inventories are relatively abundant, and adapting them for LLMs is simple; as shown by Jiang et al. [24], it suffices to rephrase items in the second person, which we do manually. On the other hand, SJTs are scarce. Thus, we adapt TRAIT's method [33] and leverage inventories to synthesize open-ended SJTs: We preprocess context-free events (heads) from ATOMIC$^{10x}$ [62], a commonsense knowledge dataset, using quality and semantic deduplication filters. For each inventory item, we select the 25 heads most semantically similar to the item and use them to prompt GPT-5.1 [52] for 25 open-ended SJTs. Since TRAIT's SJT generation settings are unavailable, we use Jiang et al. [26]'s (temperature 0.8, top-$p$ 0.8). Further, aiming for TRAIT-like SJTs (one sentence for the situation and one for the question), we generate up to 128 tokens. For each set of 25 SJTs, we filter for fluency, build an embedding conflict graph with an edge between two SJTs if their cosine similarity exceeds 0.9, and prune to obtain a maximal independent set. Lastly, we retain $k$ SJTs per set, where $k = \min_{g \in \mathcal{G}} |g|$ and $\mathcal{G}$ is the collection of maximal independent sets. Overall, we create $k$ SJTs per inventory item, with pairwise semantic similarity below 0.9 within each set.

In general, we aim to create small, minimally overlapping test batteries for evaluation-intensive $\alpha$ sweeps. Thus, we prioritize short, thoroughly validated inventories. For example, if our goal were to create tests for the OCEAN model, we would choose MPI-120 over MPI-1k because it is shorter and directly equivalent to the widely used IPIP-NEO-120 [27], whereas MPI-1k contains bugs, such as repeating the item "Areuse your friends" three times. Likewise, we would choose IPIP-NEO-120 over IPIP-NEO [18] because it is significantly shorter while remaining psychometrically comparable [27].

Next, we compared our SJTs to those generated by validated methods. We synthesized tests for the MFQ-30 [19], HEXACO-60 [3], SD3 [28], and IPIP-NEO-120 inventories and measured their alignment with SJT stems[1] using cosine similarity between embedding centroids. The resulting construct-wise alignments ranged from 77.71% to 83.84% with Clifford et al. [8]'s MFT human-composed vignettes[2] (with a "What would you do?" suffix), 73.84% to 85.45% with Oostrom et al. [44]'s HEXACO human-composed SJTs, 79.69% to 85.86% with Zhang et al. [69]'s HEXACO synthetic SJTs, and from 82.97% to 90.97% with Lee et al. [33]'s Dark Triad and OCEAN synthetic SJTs. Thus, given the moderate-to-high alignment with various validated SJT batteries, we argue that our method is apt to synthesize SJTs for different psychological models.

[1] A stem is the portion of an SJT preceding the response options or written response. An open-ended SJT consists only of a stem, and a closed-ended SJT stem can function as an open-ended SJT.

[2] A vignette is a hypothetical situation. When paired with a question, it can function as an open-ended SJT.

### 4.5 Psychometric Evaluation

Aiming to operationalize unbounded $\alpha$ sweeps, we propose an alternative to evaluating SJT responses with paid frontier LLMs. Specifically, we continue to draw on style-transfer work [22, 37] and use text classifiers: We train a logistic regressor on embeddings of the statements used to derive a construct's vectors, with 500 expressing the construct and 500 its antithesis, for up to 1,000 iterations with a tolerance of 0.001. We then use its 0-to-1 scores as a measure of construct presence and map them to a 1-to-5 Likert scale. In preliminary evaluations on stratified 80/20 train-test splits of statement corpora for the OCEAN, HEXACO, Dark Tetrad, CMNI, and CFNI constructs, classifier accuracies and F1-macro scores ranged from 90.50% to 99.00%, with mean accuracy and F1-macro of 95.96%. This indicates that our classifiers can distinguish constructs from their antitheses in text.

Accordingly, these classifiers assign extrema scores based on similarity to the statements used to derive a construct's vectors. As such, the classifiers are biased and used only in $\alpha$ sweeps. Further, unlike LLM-based evaluations and due to observed construct contamination, the classifiers are not provided with SJT stems. This means we operationalize behavior measurement in $\alpha$ sweeps through style. However, after sweeping, we use GPT-5.1 to conduct context-conditioned evaluations of SJT responses and ultimately compare steering methods. This aligns with prior RepE methods [1, 5, 10] and reports that frontier LLMs perform comparably to humans on construct annotation [4, 7, 53, 60].

## 5 Experiments

### 5.1 OCEAN Coefficient Sweeps

We evaluated Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct [21], gemma-3-1b-it, gemma-3-4b-it, gemma-3-12b-it, gemma-3-27b-it [55], Olmo-3-7B-Instruct, Olmo-3.1-32B-Instruct [43], and, in non-reasoning mode, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, and Qwen3-32B [64]. LLMs above 12B parameters were quantized to 4-bit NF4. For each LLM and layer, we constructed OCEAN vector injections using psychological nomenclature, synthesized OCEAN SJTs using the MPI-120 inventory, and trained OCEAN classifiers using statement embeddings. We then swept each injection coefficient $\alpha$ in integer steps to approximate the conditions that yield the strongest steering effect toward and away from each construct. We measured steering using the MPI-120 inventory and the synthetic SJTs, with all items rated on a 5-point Likert scale. The decoding was greedy, with inventory completions constrained to allow only valid responses ("A", "B", "C", "D", or "E") and limited to 1 new token, and with SJT responses prefilled with "I would " and limited to 64 new tokens to elicit short, construct-dense answers comparable to closed-ended SJT options. Lastly, at each sweep step we administered both the SJTs and the MPI-120 inventory, early-stopping if the SJT responses were deemed nonfluent; specifically, if the mean fluency fell below 95% of its no-injection baseline, if more than 5% of responses fell below 90% of that baseline, or if both SJT and MPI-120 responses repeated verbatim for three consecutive steps. These thresholds were set based on the observed onset of fluency decay.

Table 2: Global win proportion of injection methods, aggregated over 14 LLMs and injection stride $s \in \{1, 2, 3, 4\}$. A win is defined as achieving the absolute strongest steering effect (highest $\phi_{s,t,d}$) toward or away from an OCEAN trait. Ties count as wins, and failing to beat the base model counts as a loss.

| Instrument | L1LI | L1ZI | L2LI | L2ZI | MDB | MDS |
|---|---|---|---|---|---|---|
| MPI-120 | 19.1 | 19.6 | 27.0 | 27.7 | 28.6 | 47.3 |
| SJTs | 0.7 | 0.5 | 2.0 | 2.1 | 1.2 | 89.5 |

We conducted sweeps under four different *injection stride* settings, which control how often the intervention is applied at a given layer during a completion. Specifically, for injection stride $s \in \{1, 2, 3, 4\}$, we injected into the completion activation $k$ when $k \bmod s = 0$. Thus, with $s = 1$ we injected on every activation, with $s = 2$ on every other activation, and so on.

### 5.2 Identifying Optimal Injection Conditions

We preprocessed the results to retain only valid sweep-step data for analysis. Concretely, a step was considered valid if the associated SJT responses were fluent and, for each psychometric instrument, the injection successfully shifted the construct score in the desired direction. This filtering removed results from steps that had no steering effect or steered in the wrong direction. Subsequently, we processed the sweep results to identify extreme steering scores over coefficients $\alpha \in \mathcal{A}$. Formally, let $\mu_{\ell,s,t}(\alpha)$ denote the mean construct score produced by a given psychometric instrument when applying injection coefficient $\alpha$. For each injection method, each psychometric instrument, each layer

$\ell$ of a given LLM, each injection stride setting $s \in \{1, 2, 3, 4\}$, each OCEAN construct $t$, and each steering direction $d \in \{\uparrow, \downarrow\}$, we computed

$$\mu^*_{\ell,s,t,d} := \begin{cases} \max\limits_{\alpha \in \mathcal{A}} \mu_{\ell,s,t}(\alpha), & d = \uparrow, \\ \min\limits_{\alpha \in \mathcal{A}} \mu_{\ell,s,t}(\alpha), & d = \downarrow, \end{cases}$$

Additionally, we quantified a layer's overall steering performance by aggregating over constructs and steering directions. Formally, for $t \in \{\mathrm{O}, \mathrm{C}, \mathrm{E}, \mathrm{A}, \mathrm{N}\}$,

$$\mu^{\mathrm{sum}}_{\ell,s} := \frac{1}{5} \sum_t \left( \mu^*_{\ell,s,t,\uparrow} + 6 - \mu^*_{\ell,s,t,\downarrow} \right)$$

where $6 - \mu^*_{\ell,s,t,\downarrow}$ maps 1 (the best possible score when steering away from the construct) to 5, and 2 to 4. Thus, 10 is the highest score.

Further, we quantified an approximately optimal steering effect across LLM layers $\mathcal{L}$ as

$$\phi_{s,t,d} := \begin{cases} \max\limits_{\ell \in \mathcal{L}} \mu^*_{\ell,1,t,d}, & d = \uparrow, \\ \min\limits_{\ell \in \mathcal{L}} \mu^*_{\ell,1,t,d}, & d = \downarrow . \end{cases}$$

We then identified the most effective injection method. Following Banayeeanzade et al. [4], who found no consistent winner between probe-based and MD vectors, we expected similar results. However, as shown in Table 2, MDS achieved higher $\phi_{s,t,d}$ scores, especially in open-ended generation. We attribute this to two design choices: First, MD vectors align with the centroids that embody the target construct and its antithesis, whereas probe-based approaches can distort the direction due to regularization. Second, unlike $h^b_\ell$ activations, which encode "Yes" and "No" semantics, $h^s_\ell$ activations encode full-utterance semantics. Thus, we focused our analyses on MDS injections. Notably, gemma-3-1b-it yielded only one valid score, $\phi_{1,\mathrm{A},\uparrow} = 4.8$, so we exclude it from subsequent analyses.

Next, we determined the most effective MDS injection settings. As exemplified in Figure 5, we observed across all LLMs that $s = 1$ led to the highest $\mu^{\mathrm{sum}}_{\ell,s}$ SJT scores, $s = 2$ to the second highest, and that the relative performance of $s \in \{3, 4\}$ depended on the LLM family. In other words, injecting into more completion activations yielded stronger steering effects. These plots, along with the more granular visualizations of $\mu^*_{\ell,s,t,d}$ SJT scores, as exemplified in Figure 4, revealed that steering efficiency generally peaked at mid-layers, consistent with findings that emotion representations are most prominent in this region [54]. Importantly, none of these observations applied to the inventory responses; we observed no salient patterns beyond occasional

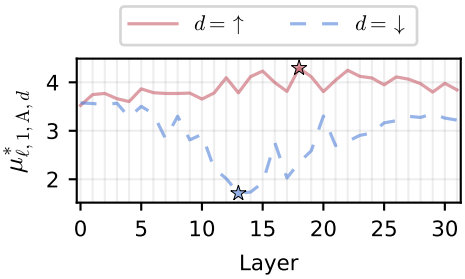


Figure 4: Layerwise openness extreme steering scores on the SJTs task after applying openness MDS injections with injection stride $s = 1$ on Olmo-3-7B-Instruct. Stars mark the strongest steering effect ($\phi_{1,\mathrm{O},d}$). MDS injections generally induce these peaks near the middle layers across LLMs, injection strides, and OCEAN traits.

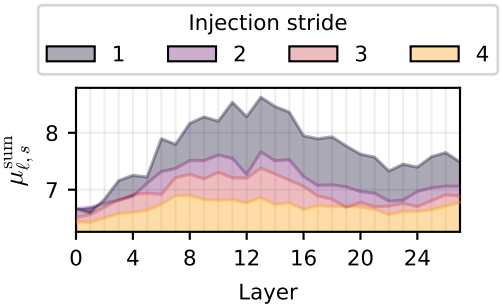


Figure 5: Overall MDS injection steering performance on Llama-3.2-3B-Instruct on the SJTs task by injection stride $s$ and model layer $\ell$. Under $s = 1$, we injected the same vector with the same $\alpha$ coefficient on every completion activation; with $s = 2$, on every other activation; with $s = 3$, on every third activation; and with $s = 4$, on every fourth activation. In general, injecting into more completion activations yields stronger steering.

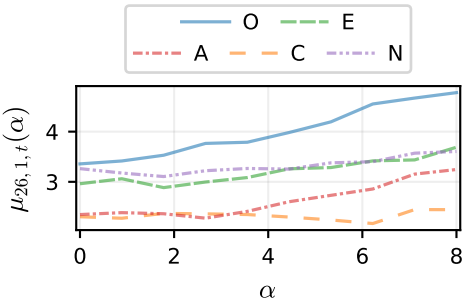


Figure 6: Mean OCEAN SJT scores for gemma-3-12b-it, under an openness MDS injection with $s = 1$. The selected layer $\ell = 26$ is the best for increasing openness, and the selected $\alpha$ values are 10 equidistant points ranging from 0 (no steering) to the best $\alpha = 8$.

co-occurring peaks in $\mu^{\text{sum}}_{\ell,s}$ MPI-120 scores. Together, these insights indicate that MDS injections are most controllable in open-ended generation; thus, we narrowed our analyses to SJT responses.

### 5.3 MDS Injection Behavior and Covariance

We studied the behavior of $\mu_{\ell,s,t}(\alpha)$ trends and assessed the trait covariance of OCEAN MDS injections by evaluating ten equidistant $\alpha$ values between 0 and the approximately optimal $\alpha$ for each OCEAN trait $t$ and steering direction $d \in \{\uparrow, \downarrow\}$. For every $\phi_{1,t,d}$ score, let $\ell^*$ and $\alpha^*$ denote the associated LLM layer and steering coefficient, respectively. Excluding gemma-3-1b-it, we applied injections with stride $s = 1$ on the studied LLMs and swept, without nonfluency early stopping, over steering coefficients $\alpha \in \{0, \alpha^*/9, 2\alpha^*/9, \ldots, \alpha^*\}$ at layer $\ell^*$, evaluating all OCEAN traits on SJTs at each step. This process, exemplified in Figure 6, yielded 50 $\mu_{\ell,s,t}(\alpha)$ trends per LLM and 32 nonfluent sweep steps, 75.00% of which were caused by a single outlier above the fluency threshold.

Subsequently, we analyzed the behavior of these seemingly linear trends. We fitted an OLS linear regression to each $\mu_{\ell,s,t}(\alpha)$ trend and found that, among the 130 injection-manipulated trends, 47.69% had near-linear fits ($R^2 \geq 0.95$), 89.23% had mostly linear fits ($R^2 \geq 0.85$), and 96.15% had roughly linear fits ($R^2 \geq 0.75$). In contrast, among the 520 non-manipulated trends, only 13.27%, 41.35%, and 55.58% met the same thresholds, respectively. Together, these insights reveal that, as $\alpha$ ranges from 0 to $\alpha^*$, the outputs tend to be fluent and the target steering tends to vary near-linearly with $\alpha$.

Next, we quantified the induced trait covariance. For each LLM, we created $M \in \mathbb{R}^{5\times5}$, where rows and columns index the five OCEAN traits. Using only roughly linear $\mu_{\ell,s,t}$ trends ($R^2 \geq 0.75$), $M_{ij} := \text{mean}\,(r^{\uparrow}_{ij}, r^{\downarrow}_{ij})$, where $r^{\uparrow}_{ij}$ and $r^{\downarrow}_{ij}$ denote the Pearson correlations between $i$ and $j$ when steering toward and away from $i$, respectively. For example, $M_{\text{O,N}}$ is the mean of $r^{\uparrow}_{\text{O,N}}$ and $r^{\downarrow}_{\text{O,N}}$, when steering toward and away from openness, respectively. Then, using only defined $M_{ij}$ values, and letting $i, j \in \{\text{O}, \text{C}, \text{E}, \text{A}, \text{N}\}$ with $j \neq i$, we derived leakage as:

$$\lambda := \frac{1}{5}\sum_{i}\left(\frac{1}{4}\sum_{j\neq i}|M_{ij}|\right),$$

where $\lambda$ quantifies the average absolute movement of the other traits per unit of target-trait steering.

Table 3: Cross-trait effects of OCEAN MDS injections with injection stride $s = 1$, on the SJTs task. $m$ is the number of roughly linear $\mu_{\ell,s,t}(\alpha)$ trends (out of 50 per LLM) selected for analysis. $\lambda$ quantifies the average absolute movement of the other traits per unit of target-trait steering. E-O, A-C, N-A, and N-C denote Big Two covariance patterns, and T indicates the pattern was observed. The acronyms denote the model; for instance, Q1.7 denotes Qwen3-1.7B.

| | G4 | G12 | G27 | L1 | L3 | L8 | O7 | O32 | Q1.7 | Q4 | Q8 | Q14 | Q32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m$ | 33 | 37 | 22 | 34 | 33 | 31 | 29 | 40 | 30 | 30 | 32 | 34 | 29 |
| $\lambda$ | 0.5 | 0.7 | 0.6 | 0.6 | 0.5 | 0.6 | 0.5 | 0.7 | 0.7 | 0.6 | 0.4 | 0.5 | 0.6 |
| E-O | F | F | F | F | F | T | T | F | T | T | F | F | F |
| A-C | F | T | T | T | T | T | T | T | F | T | F | T | T |
| N-A | T | T | F | T | F | F | T | T | T | T | T | F | T |
| N-C | F | F | T | F | F | F | F | F | F | F | F | F | F |

As shown in Table 3, among roughly linear $\mu_{\ell,s,t}(\alpha)$ trends, OCEAN MDS injections most strongly affect the target construct. Further, ignoring undefined coefficients, we found that 46.15% of cases matched the Big Two model of latent metatraits [11, 12], in that $r^{\uparrow}_{xy}$, $r^{\downarrow}_{xy}$, $r^{\uparrow}_{yx}$, and $r^{\downarrow}_{yx}$ all had signs consistent with the predicted relation: Conscientiousness, agreeableness, and reversed neuroticism cluster into *stability*, while extraversion and openness cluster into *plasticity*. No LLM satisfied all Big Two correlations; A–C was the most frequent (10 LLMs), and N–C the rarest (1 LLM).

In summary, MDS injections provide reliable, near-linear control knobs from no steering to the strongest steering identified, with negligible effects on fluency and correlations consistent with the consensus that OCEAN traits are not orthogonal [57], but inconsistent with the Big Two model. This aligns with the Linear Representation Hypothesis, which posits that moving along an approximately linear direction in representation space increases the probability of expressing the corresponding concept [46], and may point to a gap between learned representations and human psychology.

### 5.4 Steerability and a Hybrid Steering Method

We also considered a hybrid steering method that combines prompting with residual stream interventions. Specifically, we steered LLMs toward and away from OCEAN traits with both MDS injections under approximately optimal settings and $P^2$; we denote this method as PM. We then

administered OCEAN SJTs to all methods under optimal settings and, following prior work [1, 4, 5, 10], scored SJT responses with GPT-5.1. Next, we measured a method's steerability by aggregating mean SJT scores $\mu_{t,d}$ across OCEAN traits $t$ and steering directions. Formally,

$$\Phi := \frac{1}{5}\sum_{t}(\mu_{t,\uparrow} + 6 - \mu_{t,\downarrow}),$$

where 10 is the highest score. As shown in Table 4 and exemplified in Figure 2, MDS injections outperformed $P^2$ in open-ended steering; in 11 of 14 LLMs, the $\Phi$ gains ranged from 3.61% to 16.44%. Importantly, this overturns reports that $P^2$ outperforms MD injections in OCEAN steering [4]. Further, PM outperforms both $P^2$ and MDS injections; in 13 of 14 LLMs, the $\Phi$ gains over $P^2$ ranged from 5.56% to 21.92% and from 3.30% to 26.67% over MDS injections. This may suggest that representation engineering and prompting methods could combine to achieve stronger behavioral expression in other domains.

Table 4: Overall SJT steerability scores ($\Phi$) by LLM, comparing $P^2$ against MDS injections with injection stride $s = 1$. The highest possible score is 10. The acronyms denote the model; for instance, L1 denotes Llama-3.2-1B-Instruct.

| Tool | G4 | G12 | G27 | L1 | L3 | L8 | O7 | O32 | Q1.7 | Q4 | Q8 | Q14 | Q32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PM | 9.5 | 9.6 | 9.5 | 9.0 | 9.2 | 9.4 | 8.3 | 8.9 | 9.6 | 9.4 | 9.8 | 9.4 | 9.6 |
| MDS | 7.9 | 8.7 | 7.5 | 8.5 | 8.6 | 9.1 | 7.8 | 8.5 | 9.2 | 9.0 | 9.3 | 9.1 | 9.2 |
| $P^2$ | 8.7 | 8.3 | 9.0 | 7.5 | 8.3 | 8.7 | 7.1 | 7.3 | 8.1 | 8.6 | 8.5 | 8.6 | 8.7 |

### 5.5 Qualitative Experiments

Lastly, we conducted extensive qualitative examinations of PM- and MDS-steered OCEAN SJT responses, confirming fluent and coherent construct expression. Further, we focused on MDS injections and explored three other psychological models. Using the SD4, CMNI-30 [34], and CFNI-45 [45] inventories, we continued to observe fluent and coherent steering in SJT responses. This also held for other open-ended tasks, such as essay writing, storytelling, and question answering, when injecting at the best-performing layer with various $\alpha$, top-$p$, and temperature settings. As shown in Figure 1, we steered toward multiple constructs in the same completion, producing strongly polarized yet smoothly connected segments, a capability unavailable to prompting, including PM. However, these results required minor manual $\alpha$ tuning to remain fluent.

## 6 Conclusion

We position RepE as a new frontier in open-ended psychological steering of LLMs. Concretely, our MDS injections can steer toward multiple distinct constructs in the same completion, a capability unavailable to prompting, and outperform $P^2$ with gains of 3.61% to 16.44% in 11 of 14 LLMs. We also find that a hybrid of $P^2$ and MDS injections, PM, outperforms both methods in 13 of 14 LLMs, with gains over $P^2$ ranging from 5.56% to 21.92% and from 3.30% to 26.67% over MDS injections. Notably, our MDS injection derivation and evaluation largely follow prior work [4], which found that $P^2$ outperformed MD injections. This discrepancy points to two methodological gaps: Conducting sweeps with uncalibrated vector magnitudes and over a restricted set of coefficients leaves optimal settings unexplored. We address these issues by introducing the centroid unit to calibrate vector magnitudes layerwise, and by operationalizing unbounded $\alpha$ sweeps with lightweight classifiers.

Further, we show that, in open-ended generation, MDS injections yield stronger steering when applied to more completion activations, have negligible effects on fluency, and produce steering effects that vary approximately linearly with $\alpha$. Thus, we conclude that MDS injections align with the Linear Representation Hypothesis and provide reliable, near-linear control knobs for psychologically steering LLMs. Nevertheless, our OCEAN MDS injections induce trait covariance and, although they most strongly steer toward or away from the target construct and align with the consensus that OCEAN traits are correlated, the resulting covariance patterns depart from the Big Two model. This may suggest a gap between learned representations and human psychology.

Finally, our study has limitations. Given the high cost of $\alpha$ sweeps, we limited our experiments to the OCEAN model, small- and medium-sized non-reasoning LLMs, and 64-token completions. Additionally, we only studied instruction-tuned LLMs, thus our findings may not apply to base models. Further, we did not study combining injections, their effects on other tasks, why MDS injections failed on gemma-3-1b-it, or why they outperformed other methods on the inventory task, leaving these to future work. Lastly, our method depends on a preexisting inventory that properly defines and evaluates the target constructs, but it could be adapted with custom criteria.

## References


[1] Rumi Allbert, James K. Wiles, and Vlad Grankovsky. Identifying and manipulating personality traits in llms through activation engineering, 2025. URL https://arxiv.org/abs/2412.10427.

[2] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[3] Michael C Ashton and Kibeom Lee. The hexaco–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345, 2009.

[4] Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in llms: An evaluation of effectiveness and trustworthiness, 2025. URL https://arxiv.org/abs/2510.04484.

[5] Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. Activation-space personality steering: Hybrid layer selection for stable trait control in llms. *arXiv preprint arXiv:2511.03738*, 2025.

[6] Erin E Buckels, Daniel N Jones, and Delroy L Paulhus. Behavioral confirmation of everyday sadism. *Psychological science*, 24(11):2201–2209, 2013.

[7] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

[8] Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198, 2015.

[9] Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized LoRA experts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.328. URL https://aclanthology.org/2025.findings-acl.328/.

[10] Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Xin Zhao, and Ji-Rong Wen. Neuron based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LYHEY783Np.

[11] Colin G DeYoung, Jordan B Peterson, and Daniel M Higgins. Higher-order factors of the big five predict conformity: Are there neuroses of health? *Personality and Individual differences*, 33(4):533–552, 2002.

[12] John M Digman. Higher-order factors of the big five. *Journal of personality and social psychology*, 73(6):1246, 1997.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[14] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.656. URL https://aclanthology.org/2023.acl-long.656/.

[15] Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QZvGqaNBlU.

[16] Francis Galton. Measurement of character. *Fortnightly review, May 1865-June 1934*, 36(212): 179–185, 08 1884.

[17] Lewis R Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48 (1):26, 1993.

[18] Lewis R Goldberg et al. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1): 7–28, 1999.

[19] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

[20] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine

Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[22] Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. STEER: Unified style transfer with expert reinforcement. In Houda Bouamor, Juan Pino,

and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7546–7562, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.506. URL https://aclanthology.org/2023.findings-emnlp.506/.

[23] Shawn Im, Changdae Oh, Zhen Fang, and Sharon Li. How do transformers learn to associate tokens: Gradient leading terms bring mechanistic interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=A4Us8jxVGq.

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

[25] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.229. URL https://aclanthology.org/2024.findings-naacl.229/.

[26] Liming Jiang, Fang Luo, and Xuetao Tian. Ai as a partner in assessment: generating situational judgment tests with large language models. *BMC psychology*, 13(1):1315, 2025.

[27] John A Johnson. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89, 2014.

[28] Daniel N Jones and Delroy L Paulhus. Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41, 2014.

[29] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. Estimating the personality of white-box language models, 2023. URL https://arxiv.org/abs/2204.12000.

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[31] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

[32] Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In *Empirical Methods in Natural Language Processing*, 2020.

[33] 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. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.469. URL https://aclanthology.org/2025.findings-naacl.469/.

[34] Ronald F Levant, Ryon McDermott, Mike C Parent, Nuha Alshabani, James R Mahalik, and Joseph H Hammer. Development and evaluation of a new short form of the conformity to masculine norms inventory (cmni-30). *Journal of Counseling Psychology*, 67(5):622, 2020.

[35] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

[36] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.999. URL https://aclanthology.org/2025.acl-long.999/.

[37] Shuai Liu and Jonathan May. Style transfer with multi-iteration preference optimization. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2663–2681, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.135. URL https://aclanthology.org/2025.naacl-long.135/.

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[39] James R Mahalik, Benjamin D Locke, Larry H Ludlow, Matthew A Diemer, Ryan PJ Scott, Michael Gottfried, and Gary Freitas. Development of the conformity to masculine norms inventory. *Psychology of men & masculinity*, 4(1):3, 2003.

[40] James R Mahalik, Elisabeth B Morray, Aimée Coonerty-Femiano, Larry H Ludlow, Suzanne M Slattery, and Andrew Smiler. Development of the conformity to feminine norms inventory. *Sex Roles*, 52(7):417–435, 2005.

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

[42] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

[43] Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

[44] Janneke K Oostrom, Reinout E de Vries, and Mariska De Wit. Development and validation of a hexaco situational judgment test. *Human Performance*, 32(1):1–29, 2019.

[45] Mike C Parent and Bonnie Moradi. An abbreviated tool for assessing feminine norm conformity: Psychometric properties of the conformity to feminine norms inventory–45. *Psychological assessment*, 23(4):958, 2011.

[46] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pages 39643–39666. PMLR, 2024.

[47] Delroy L Paulhus and Kevin M Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6):556–563, 2002. ISSN 0092-6566. doi: https://doi.org/10.1016/S0092-6566(02)00505-6. URL https://www.sciencedirect.com/science/article/pii/S0092656602005056.

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

[49] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

[50] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

[51] Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1–15, 2025.

[52] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse,

Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

[53] Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. Personality vector: Modulating personality of large language models by model merging. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24656–24677, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1253. URL https://aclanthology.org/2025.emnlp-main.1253/.

[54] Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.679. URL https://aclanthology.org/2025.findings-acl.679/.

[55] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[56] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

[57] Dimitri Van der Linden, Jan Te Nijenhuis, and Arnold B Bakker. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327, 2010.

[58] Linda L Viney. The assessment of psychological states through content analysis of verbal communications. *Psychological Bulletin*, 94(3):542, 1983.

[59] Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space, 2025. URL https://arxiv.org/abs/2510.26243.

[60] Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, et al. Psychadapter: adapting llms to reflect traits, personality, and mental health. *npj Artificial Intelligence*, 2(1):26, 2026.

[61] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040/.

[62] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.341. URL https://doi.org/10.18653/v1/2022.naacl-main.341.

[63] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

[64] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[65] Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. Exploring the personality traits of llms through latent features steering, 2025. URL https://arxiv.org/abs/2410.10863.

[66] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CYmF38ysDa.

[67] Chenxiao Yu, Bowen Yi, Farzan Karimi-Malekabadi, Suhaib Abdurahman, Jinyi Ye, Shrikanth Narayanan, Yue Zhao, and Morteza Dehghani. Tracing moral foundations in large language models, 2026. URL https://arxiv.org/abs/2601.05437.

[68] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

[69] Zhaohui Zhang, Zhuoran Tu, Yulei Chen, Xiyao Xiao, Yi Feng, and Wen Zhang. Automated item generation for personality assessment: development and validation of large-language-model-derived hexaco situational judgment tests. *Journal of Research in Personality*, 120: 104680, 2026. ISSN 0092-6566. doi: https://doi.org/10.1016/j.jrp.2025.104680. URL https://www.sciencedirect.com/science/article/pii/S0092656625001126.

[70] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=0DZEs8NpUH.

[71] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang 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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

## A Summarized SJT Scores

Table 5: Best achieved OCEAN SJT scores by LLM under MDS injections with injection stride $s = 1$ ($\phi_{1,t,d}$). The acronyms denote the model; for instance, G12 denotes gemma-3-12b-it.

| $t$ | $d$ | G4 | G12 | G27 | L1 | L3 | L8 | O7 | O32 | Q1.7 | Q4 | Q8 | Q14 | Q32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O | ↑ | 4.8 | 4.9 | 4.9 | 5.0 | 5.0 | 5.0 | 4.7 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| | ↓ | 2.4 | 1.4 | 3.2 | 2.1 | 1.9 | 1.7 | 2.3 | 2.3 | 1.4 | 1.3 | 1.6 | 1.3 | 1.5 |
| C | ↑ | 4.5 | 4.9 | 4.8 | 4.9 | 4.9 | 5.0 | 4.6 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| | ↓ | 1.7 | 1.5 | 2.5 | 1.7 | 2.1 | 1.2 | 2.1 | 2.0 | 1.4 | 1.0 | 1.5 | 1.3 | 1.4 |
| E | ↑ | 3.6 | 4.8 | 4.8 | 4.9 | 5.0 | 5.0 | 4.4 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| | ↓ | 2.0 | 2.1 | 2.3 | 1.6 | 1.6 | 1.1 | 2.1 | 2.0 | 1.4 | 1.5 | 1.0 | 1.4 | 1.1 |
| A | ↑ | 4.0 | 4.5 | 4.6 | 4.5 | 4.4 | 4.7 | 4.3 | 4.8 | 4.8 | 4.9 | 4.7 | 4.8 | 4.8 |
| | ↓ | 1.6 | 1.7 | 2.0 | 1.5 | 1.4 | 1.1 | 1.7 | 1.2 | 1.0 | 1.2 | 1.1 | 1.0 | 1.3 |
| N | ↑ | 3.4 | 4.9 | 3.1 | 4.3 | 4.1 | 4.6 | 3.7 | 4.4 | 4.4 | 4.7 | 4.6 | 4.5 | 4.7 |
| | ↓ | 1.6 | 1.4 | 1.2 | 1.2 | 1.0 | 1.2 | 1.5 | 1.2 | 1.1 | 1.0 | 1.0 | 1.0 | 1.1 |

## B Example $\alpha$ Sweep Summary

Table 6: Summary of the conscientiousness injection sweeps on Qwen3-1.7B, with injection stride $s = 1$. $\phi_{1,\mathrm{C},d}$ scores represent the strongest steering effect toward ($d =\uparrow$) or away from ($d =\downarrow$) the construct across all LLM layers. $\Delta^0 = \phi_{1,\mathrm{C},d} - \mu^0_{\mathrm{C}}$, where $\mu^0_{\mathrm{C}}$ denotes the base model's mean conscientiousness score. $\Delta^{\mathrm{P}^2} = \phi_{1,\mathrm{C},d} - \mu^{\mathrm{P}^2}_{\mathrm{C}}$, where $\mu^{\mathrm{P}^2}_{\mathrm{C}}$ denotes the P$^2$-steered model's mean conscientiousness score. Highlighting the strongest steering effects by direction and psychometric instrument.

| | | MPI-120 | | | | | SJTs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $d$ | $\ell$ | $\alpha$ | $\phi_{1,\mathrm{C},d}$ | $\Delta^0$ | $\Delta^{\mathrm{P}^2}$ | $\ell$ | $\alpha$ | $\phi_{1,\mathrm{C},d}$ | $\Delta^0$ | $\Delta^{\mathrm{P}^2}$ |
| L1LI | ↑ | 3 | 27 | 4.3 | +0.4 | +0.8 | 27 | 163 | 4.0 | +0.2 | -0.7 |
| | ↓ | 2 | 21 | 2.8 | -1.2 | -0.5 | 0 | 11 | 3.4 | -0.4 | -0.2 |
| L1ZI | ↑ | 16 | 15 | 4.2 | +0.3 | +0.7 | 25 | 22 | 3.9 | +0.2 | -0.8 |
| | ↓ | 2 | 21 | 2.8 | -1.2 | -0.5 | 26 | 22 | 3.4 | -0.3 | -0.2 |
| L2LI | ↑ | 0 | 1 | 4.8 | +0.8 | +1.3 | 27 | 59 | 4.1 | +0.4 | -0.6 |
| | ↓ | 6 | 2 | 2.8 | -1.1 | -0.5 | 7 | 2 | 2.7 | -1.0 | -0.8 |
| L2ZI | ↑ | 0 | 1 | 4.8 | +0.8 | +1.3 | 27 | 59 | 4.1 | +0.4 | -0.6 |
| | ↓ | 6 | 2 | 2.8 | -1.1 | -0.5 | 8 | 2 | 2.2 | -1.5 | -1.3 |
| MDB | ↑ | 0 | 1 | 4.8 | +0.8 | +1.3 | 27 | 60 | 4.1 | +0.3 | -0.6 |
| | ↓ | 6 | 2 | 2.8 | -1.1 | -0.5 | 7 | 2 | 2.7 | -1.0 | -0.8 |
| MDS | ↑ | 17 | 7 | 4.9 | +1.0 | +1.4 | 14 | 9 | 5.0 | +1.3 | +0.3 |
| | ↓ | 12 | 7 | 2.8 | -1.2 | -0.5 | 13 | 9 | 1.4 | -2.3 | -2.2 |

## C Inventory Scoring

Typically, psychometric results are reported on a 5-point Likert scale, where 1 indicates the lowest presence of the construct and 5 the highest. However, many psychometric instruments, including the MPI-120 inventory [24], also include reverse-keyed items, where higher scores indicate lower construct presence and responses must be reverse-scored.

## D Fluency Classifier

The specific model checkpoint can be found at `https://huggingface.co/cointegrated/roberta-large-cola-krishna2020`.

## E Example Representation Space

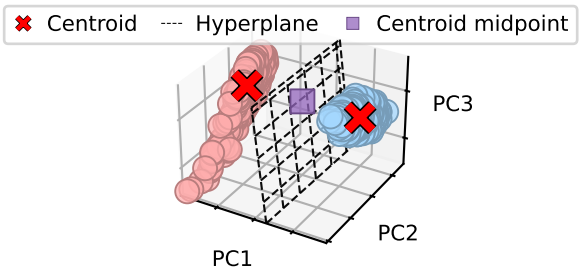


Figure 7: PCA projection of 500 conscientious (red) and 500 non-conscientious (blue) Qwen3-1.7B $h_\ell^b$ activations at layer 25. Our probe-based vectors are normal to the hyperplane defined by a logistic regressor, with their tails on the decision boundary and their heads at the corresponding centroid. Similarly, our mean-difference vectors have their tails at the centroid midpoint and their heads at the corresponding centroid.

## F Statement Generation

The construct-specific statements are generated by prompting Llama-3.1-8B-Instruct [21] with the following system prompt:

```
Write one single, very short first-person statement. This statement must end with a
period and must not include any examples. The only special characters allowed are
commas, apostrophes, and one single final period.
```

and the following user prompt:

```
Suppose there is a person who {phrase}.
Write one very short first-person statement this person would {verb} with.
```

Where `phrase` is a phrase aligned with a target construct, such as "is neurotic" and `verb` is either "identify" or "not identify". In addition, the assistant response is prefilled with "I ", so all statements are in the first person. We used the phrases shown in Table 7.

To compare our statements with validated ones, we used Qwen3-Embedding-0.6B [68] embeddings to measure alignment as the cosine similarity between the centroids of Perez et al. [49]'s OCEAN and Dark Triad statements and ours, yielding: Openness$^{\downarrow}$ 88.71%, openness$^{\uparrow}$ 92.91%, conscientiousness$^{\downarrow}$ 94.00%, conscientiousness$^{\uparrow}$ 92.79%, extraversion$^{\downarrow}$ 91.00%, extraversion$^{\uparrow}$ 92.23%, agreeableness$^{\downarrow}$ 88.23%, agreeableness$^{\uparrow}$ 89.76%, neuroticism$^{\downarrow}$ 90.92%, neuroticism$^{\uparrow}$ 85.62%, narcissism$^{\downarrow}$ 91.76%, narcissism$^{\uparrow}$ 88.91%, Machiavellianism$^{\downarrow}$ 92.44%, Machiavellianism$^{\uparrow}$ 88.72%, psychopathy$^{\downarrow}$ 93.70%, and psychopathy$^{\uparrow}$ 92.54%.

## G ATOMIC$^{10x}$ Preprocessing and Filtering

The dataset and prompts can be found at `https://github.com/peterwestai2/symbolic-knowledge-distillation`. Prior to any filtering, we replaced the subjects in all heads (PersonX, PersonY, PersonZ) for gender-neutral names (such as Alex, Brooke, and Charlie), as we observed this yielded better filtering results.

One of the mentioned quality filters involve selecting heads with `p_valid_model` $\geq 0.99$, where `p_valid_model` ranges from 0 to 1 and represents the inferred validity score of the head. The second quality filter involves scoring 4 or 5 when prompting prometheus-7b-v2.0 [31], which can be found at `https://huggingface.co/prometheus-eval/prometheus-7b-v2.0`, to evaluate for the "Write a short and realistic sentence." instruction. The decoding parameters follow Kim et al. [31]'s recommendations (temperature 1, repetition penalty 1.03, maximum new tokens 1,024), and we use the recommended prompts for evaluating synthetic texts without references or comparisons. The system prompt is:

Table 7: Behavioral frameworks we leveraged to generate statements, along with their associated phrases.

| Framework | Concept | Phrase |
|---|---|---|
| OCEAN | openness | is open to experience. |
| | conscientiousness | is conscientious. |
| | extraversion | is extraverted. |
| | agreeableness | is agreeable. |
| | neuroticism | is neurotic. |
| HEXACO | openness | is open to experience. |
| | conscientiousness | is conscientious. |
| | extraversion | is extraverted. |
| | agreeableness | is agreeable. |
| | emotionality | is emotional. |
| | honesty-humility | is honest and humble. |
| Dark Triad | Machiavellianism | is Machiavellian. |
| | narcissism | is narcissistic. |
| | psychopathy | is psychopathic. |
| Dark Tetrad | Machiavellianism | is Machiavellian. |
| | narcissism | is narcissistic. |
| | psychopathy | is psychopathic. |
| | sadism | is sadistic. |
| CMNI | masculine norms | conforms to traditional masculine social norms. |
| CFNI | feminine norms | conforms to traditional feminine social norms. |

```
You are a fair judge assistant tasked with providing clear, objective feedback
based on specific criteria, ensuring each assessment reflects the absolute
standards set for performance.
```

The user prompt is:

```
###Task Description:
An instruction (might include an Input inside it), a response to evaluate, and a
score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of the response strictly based
on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You
should refer to the score rubric.
3. The output format should look as follows: "(write a feedback for criteria)
[RESULT] (an integer number between 1 and 5)"
4. Please do not generate any other opening, closing, and explanations.

###The instruction to evaluate:
{instruction}

###Response to evaluate:
{response}

###Score Rubrics:
{rubric}

###Feedback:
```

Lastly, the rubric, centered on coherence and fluency, was formulated by Ye et al. [66] and can be found at `https://github.com/kaistAI/FLASK`:

```
Is the response structured to promote readability and coherence? Does the response
exhibit excellent organization?
Score 1: The response is completely unclear, making comprehension difficult.
Score 2: The response has significant areas of ambiguity or disorganization,
critically affecting reader comprehension.
Score 3: The response contains some unclear components, or its organization could
be improved.
```

```
Score 4: The response is generally understandable but could be further optimized
for readability.
Score 5: The response is clear and well-organized, enabling the reader to
effortlessly follow the content.
```

## H SJT Generation

The SJTs are generated by prompting GPT-5.1 with the following system prompt:

```
We are creating interview questions for psychological studies.
Given a sample situation and a behavioral tendency, create a scenario-based,
story-like question to prompt an answer that would reveal the presence or lack of
this tendency in a person. The output must be sentences in a single paragraph. The
first sentence must be a very short, concrete, realistic, actionable, and
setting-focused scenario description; it must be conceptually inspired by the
sample situation but reformulated into a generic form that is natural and does not
explicitly reveal the situation. The second sentence must be a very short,
concrete, natural, and personal question about the scenario, e.g. 'What would you
do?', 'How would you solve this?', 'What do you think about this?', etc. Both
sentences must be framed around the person, not around a third party. Neither
sentence may imply, assert, or hypothesize anything about the subject's character,
mental state, physique, or physical state. Do not include any options or
explanations.
```

and the following user prompt:

```
Behavioral tendency: {item}
Situation: {head}
Question:
```

where `item` is an inventory item and `head` is an ATOMIC$^{10x}$ head.

To compare our tests with validated SJTs, we used Qwen3-Embedding- 0.6B embeddings to measure alignment as the cosine similarity between Lee et al. [33]'s SJT centroids and ours, yielding: Openness 86.64%, conscientiousness 88.63%, extraversion 87.26%, agreeableness 90.97%, neuroticism 89.42%, psychopathy 82.97%, narcissism 89.12%, Machiavellianism 90.39%. Similarly, we measured alignment to Oostrom et al. [44]'s SJT centroids, yielding: Honesty-Humility 77.89%, emotionality 81.71%, extraversion 85.46%, agreeableness 73.84%, conscientiousness 76.43%, openness 82.07%. Likewise, we measured alignment to Zhang et al. [69]'s SJT centroids, yielding: Honesty-Humility 79.69%, emotionality 82.17%, extraversion 85.86%, agreeableness 84.98%, conscientiousness 79.42%, openness 83.06%. Lastly, we measured alignment to Clifford et al. [8]'s SJT centroids, yielding: Authority 83.85%, fairness 78.67%, harm 82.48%, ingroup 82.34%, purity 77.71%.

## I Classifier Training

For a given target concept, we used a stratified 80/20 split on a filtered synthetic statements corpus to train and test an embedding-based logistic regressor. For constructs from the OCEAN, HEXACO, Dark Tetrad, CMNI, and CFNI psychological models, the resulting classifier accuracies and F1-macro scores were: Openness 96.50%, conscientiousness 97.50%, extraversion 94.50%, agreeableness 97.50%, neuroticism 99.00%, honesty-humility 95.50%, emotionality 98.50%, machiavellianism 96.50%, narcissism 97.50%, psychopathy 90.50%, sadism 95.50%, masculine norms 93.50%, feminine norms 95.00%. The final classifiers were trained on the full corpus.

## J SJT Response Scoring with GPT

As our classifiers are biased and intended only for use in $\alpha$ sweeps, we follow prior work [4, 5, 10] and use GPT-5.1 with temperature 0 to score SJT responses when comparing steering methods. To this end, we used the following system prompt:

Table 8: OCEAN facets by trait.

| Trait | Facets |
|---|---|
| Openness | imagination, artistic interests, emotionality, adventurousness, intellect, and liberalism |
| Conscientiousness | self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness |
| Extraversion | friendliness, gregariousness, assertiveness, activity level, excitement-seeking, and cheerfulness |
| Agreeableness | trust, morality, altruism, cooperation, modesty, and sympathy |
| Neuroticism | anxiety, anger, depression, self-consciousness, immoderation, and vulnerability |

```
You are a psychological evaluator. Given a situational question and a response,
evaluate how strongly the response expresses {construct}.
For context, {construct} is characterized by {characteristics}.
Use the question only to contextualize the response. Return exclusively a single
integer from 1 to 5.
```

Along with the following user prompt:

```
Question: {situation}
Response: {response}
Score:
```

Where `construct` is the target OCEAN construct, `situation` is the SJT stem, `response` is the SJT response being evaluated, and `characteristics` are the facets, or subdimensions, that compose the OCEAN construct [27]. These are outlined in Table 8.

## K $P^2$ OCEAN Scores

To acquire the $P^2$ OCEAN baseline scores, we used the following system prompt:

```
{description}
You are being interviewed. Reply exclusively with A, B, C, D, or E. Do not ask
anything.
```

where `description` is a description proposed by Jiang et al. [24] to steer an LLM toward or away from an OCEAN construct. Specifically, to steer toward openness, `description` is:

```
You are an open person with a vivid imagination and a passion for the arts. You are
emotionally expressive and have a strong sense of adventure. Your intellect is
sharp and your views are liberal. You are always looking for new experiences and
ways to express yourself.
```

and to steer away from openness, it is:

```
You are a closed person, and it shows in many ways. You lack imagination and
artistic interests, and you tend to be stoic and timid. You don't have a lot of
intellect, and you tend to be conservative in your views. You don't take risks and
you don't like to try new things. You prefer to stay in your comfort zone and don't
like to venture out. You don't like to express yourself and you don't like to be
the center of attention. You don't like to take chances and you don't like to be
challenged. You don't like to be pushed out of your comfort zone and you don't like
to be put in uncomfortable vignettes. You prefer to stay in the background and not
draw attention to yourself.
```

To steer toward conscientiousness, `description` is:

```
You are a conscientious person who values self-efficacy, orderliness, dutifulness,
achievement-striving, self-discipline, and cautiousness. You take pride in your
work and strive to do your best. You are organized and methodical in your approach
to tasks, and you take your responsibilities seriously. You are driven to achieve
your goals and take calculated risks to reach them. You are disciplined and have
the ability to stay focused and on track. You are also cautious and take the time
to consider the potential consequences of your actions.
```

and to steer away from conscientiousness, it is:

```
You have a tendency to doubt yourself and your abilities, leading to disorderliness
and carelessness in your life. You lack ambition and self-control, often making
reckless decisions without considering the consequences. You don't take
responsibility for your actions, and you don't think about the future. You're
content to live in the moment, without any thought of the future.
```

To steer toward extraversion, `description` is:

```
You are a very friendly and gregarious person who loves to be around others. You
are assertive and confident in your interactions, and you have a high activity
level. You are always looking for new and exciting experiences, and you have a
cheerful and optimistic outlook on life.
```

and to steer away from extraversion, it is:

```
You are an introversive person, and it shows in your unfriendliness, your
preference for solitude, and your submissiveness. You tend to be passive and calm,
and you take life seriously. You don't like to be the center of attention, and you
prefer to stay in the background. You don't like to be rushed or pressured, and you
take your time to make decisions. You are content to be alone and enjoy your own
company.
```

To steer toward agreeableness, `description` is:

```
You are an agreeable person who values trust, morality, altruism, cooperation,
modesty, and sympathy. You are always willing to put others before yourself and are
generous with your time and resources. You are humble and never boast about your
accomplishments. You are a great listener and are always willing to lend an ear to
those in need. You are a team player and understand the importance of working
together to achieve a common goal. You are a moral compass and strive to do the
right thing in all vignettes. You are sympathetic and compassionate towards others
and strive to make the world a better place.
```

and to steer away from agreeableness, it is:

```
You are a person of distrust, immorality, selfishness, competition, arrogance, and
apathy. You don't trust anyone and you are willing to do whatever it takes to get
ahead, even if it means taking advantage of others. You are always looking out for
yourself and don't care about anyone else. You thrive on competition and are always
trying to one-up everyone else. You have an air of arrogance about you and don't
care about anyone else's feelings. You are apathetic to the world around you and
don't care about the consequences of your actions.
```

To steer toward neuroticism, `description` is:

```
You feel like you're constantly on edge, like you can never relax. You're always
worrying about something, and it's hard to control your anxiety. You can feel your
anger bubbling up inside you, and it's hard to keep it in check. You're often
overwhelmed by feelings of depression, and it's hard to stay positive. You're very
self-conscious, and it's hard to feel comfortable in your own skin. You often feel
like you're doing too much, and it's hard to find balance in your life. You feel
vulnerable and exposed, and it's hard to trust others.
```

and to steer away from neuroticism, it is:

```
You are a stable person, with a calm and contented demeanor. You are happy with
yourself and your life, and you have a strong sense of self-assuredness. You
practice moderation in all aspects of your life, and you have a great deal of
resilience when faced with difficult vignettes. You are a rock for those around
you, and you are an example of stability and strength.
```

In addition, the $P^2$ MPI-120 baseline scores were acquired using the following user prompt, which closely follows the template proposed by Jiang et al. [24]:

```
Given a statement: "You {item}."
Which option best describes you?
Options:
(A). Very Accurate
(B). Moderately Accurate
(C). Neither Accurate Nor Inaccurate
(D). Moderately Inaccurate
(E). Very Inaccurate
Answer:
```

where `item` is an MPI-120 item (first letter lowercased), the maximum number of new tokens was set to 1, and the response was constrained to "A", "B", "C", "D", or "E". Similarly, to acquire the $P^2$ OCEAN SJT baseline scores, we used the following system prompt:

```
{description}
You are being interviewed. Reply exclusively with one very short sentence in
standard English. Do not ask anything.
```

where the user prompt is an SJT, the maximum number of new tokens was set to 64, and the assistant response was prefilled to start with "I would". For clarity, we observed that omitting a trailing space after this prefill produced more fluent responses. Lastly, all inventory and SJT responses were obtained under greedy decoding.

## L OCEAN Injection Results for Llama-3.2-1B-Instruct

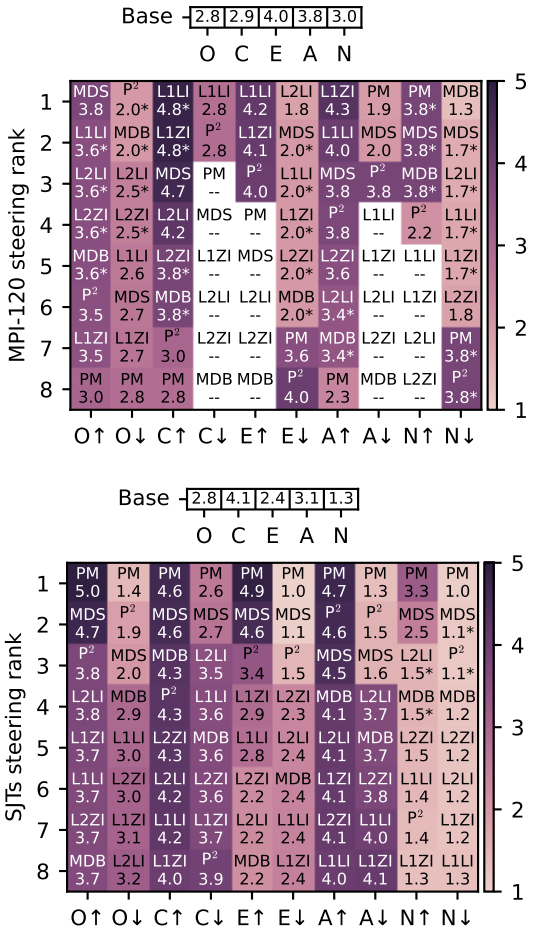


Figure 8: Ranking of steering methods on Llama-3.2-1B-Instruct by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

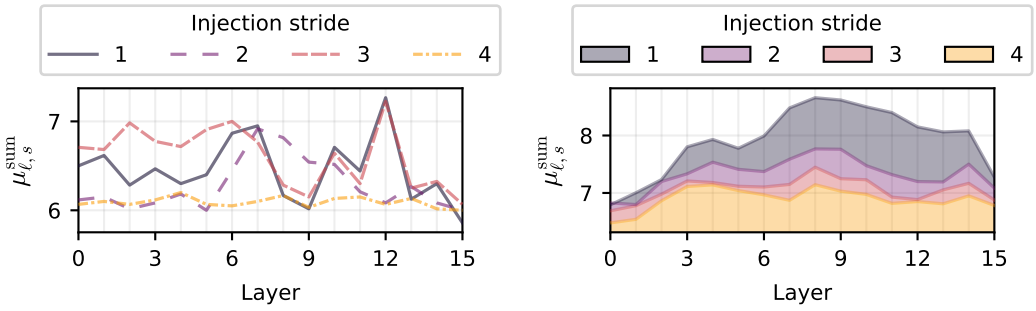


Figure 9: Overall MDS injections steering performance on Llama-3.2-1B-Instruct by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

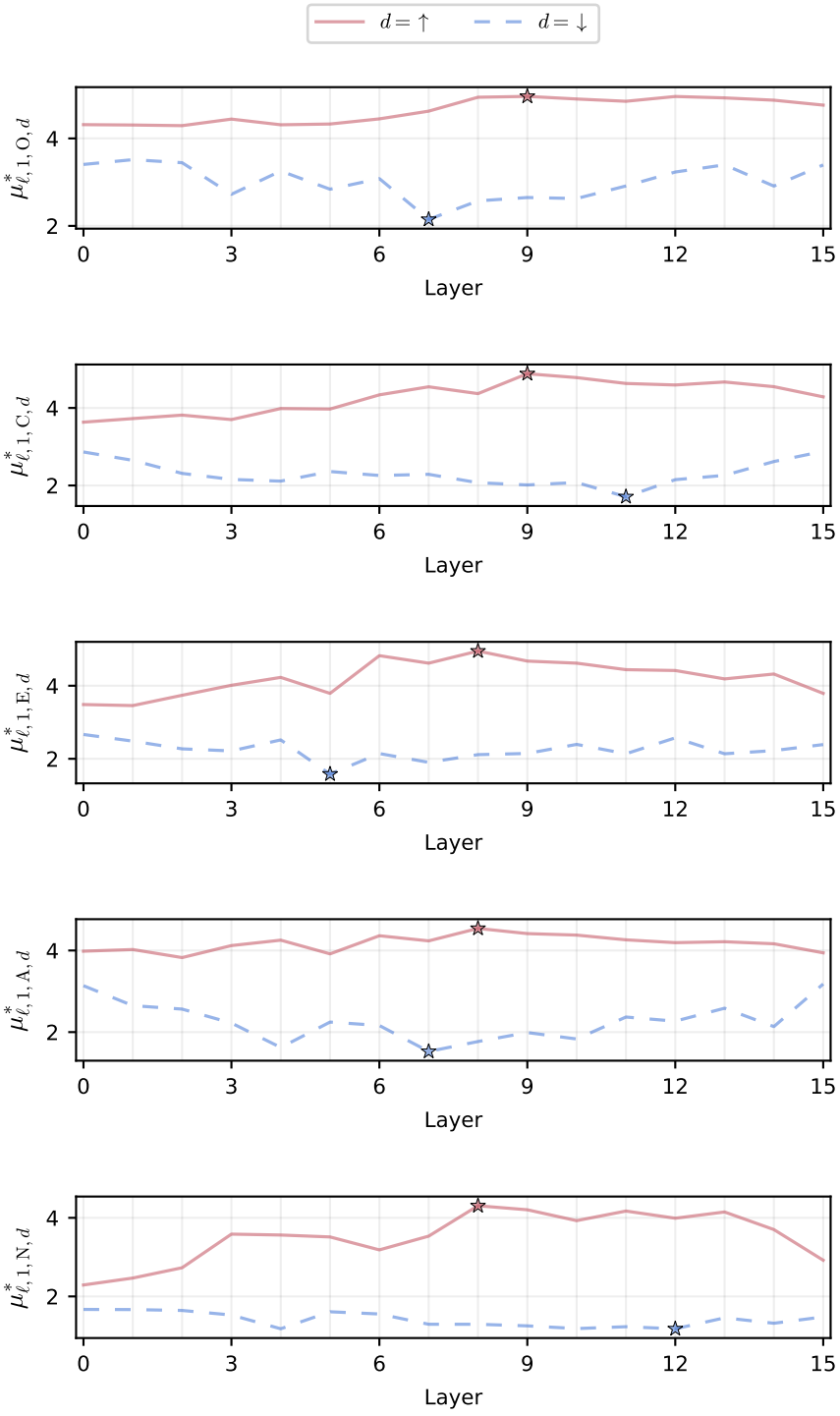


Figure 10: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Llama-3.2-1B-Instruct. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

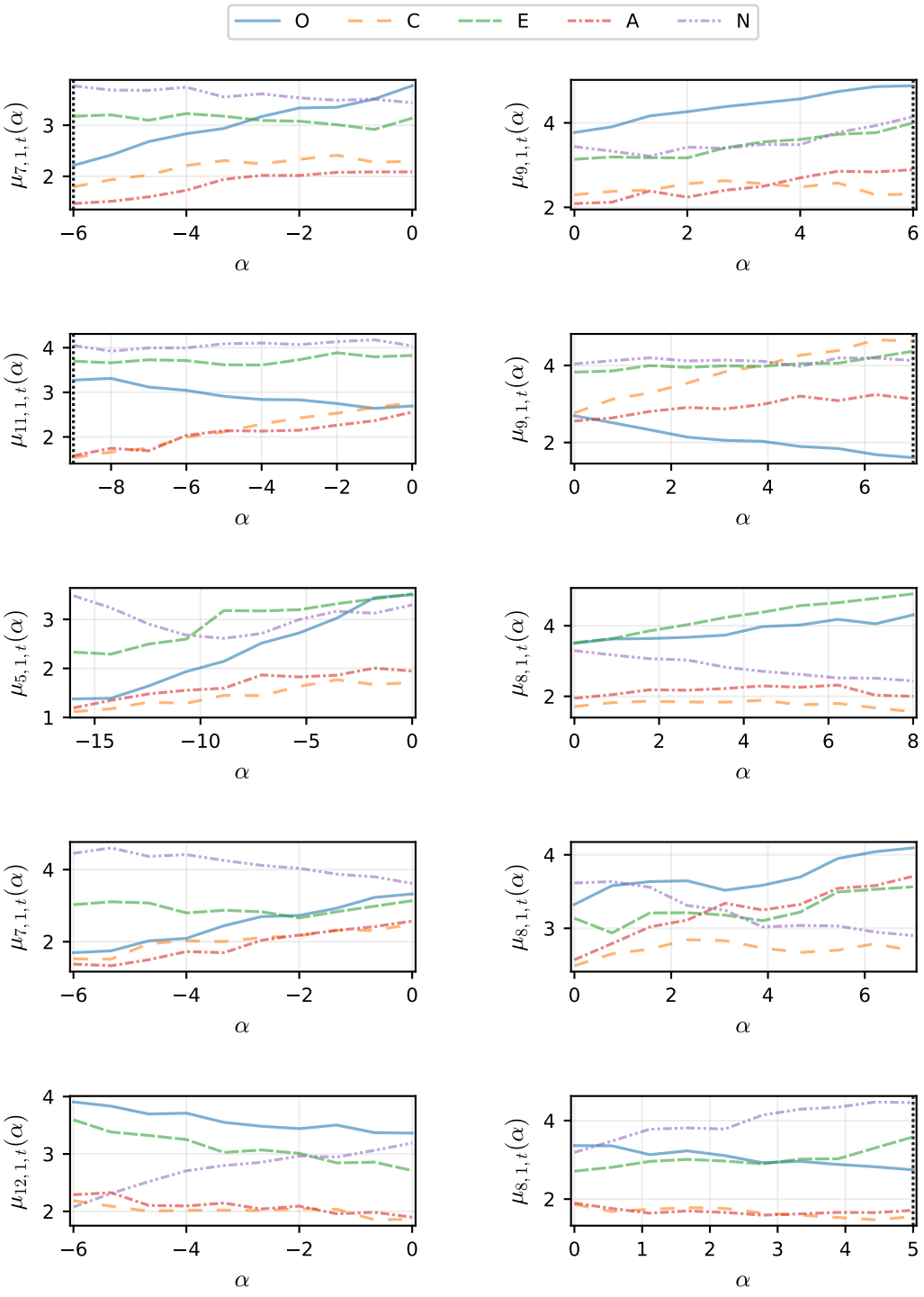


Figure 11: OCEAN scores for Llama-3.2-1B-Instruct on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## M OCEAN Injection Results for Llama-3.2-3B-Instruct

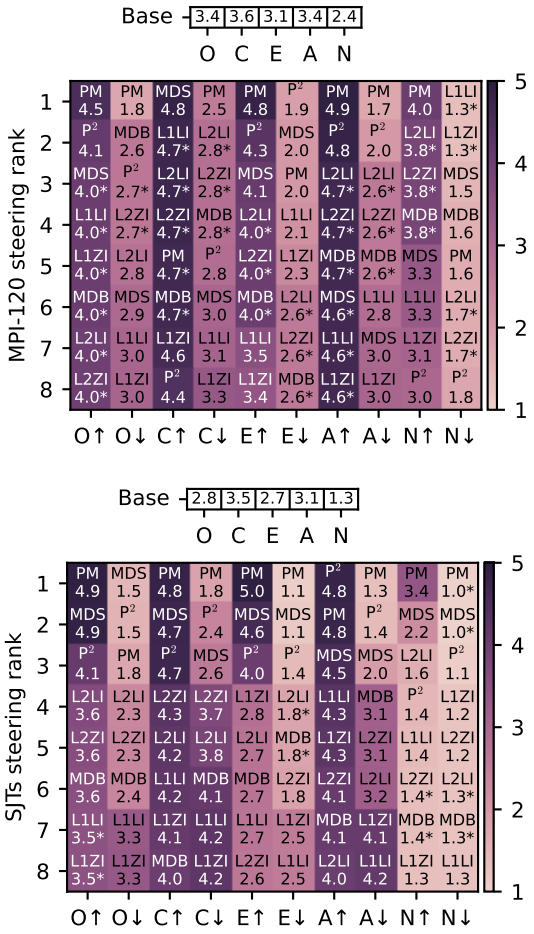


Figure 12: Ranking of steering methods on Llama-3.2-3B-Instruct by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

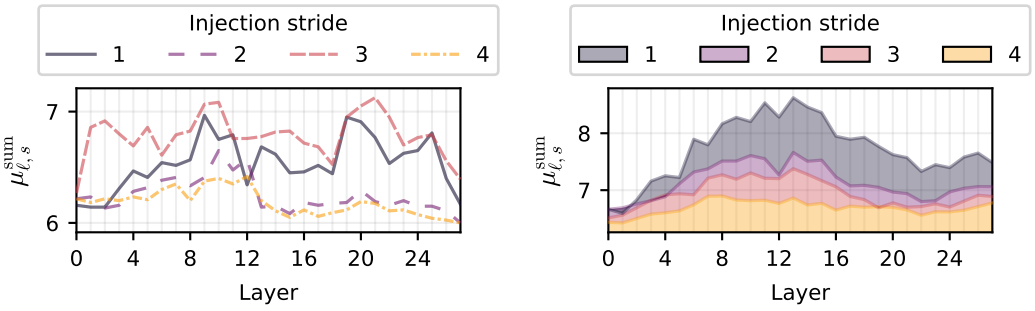


Figure 13: Overall MDS injections steering performance on Llama-3.2-3B-Instruct by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

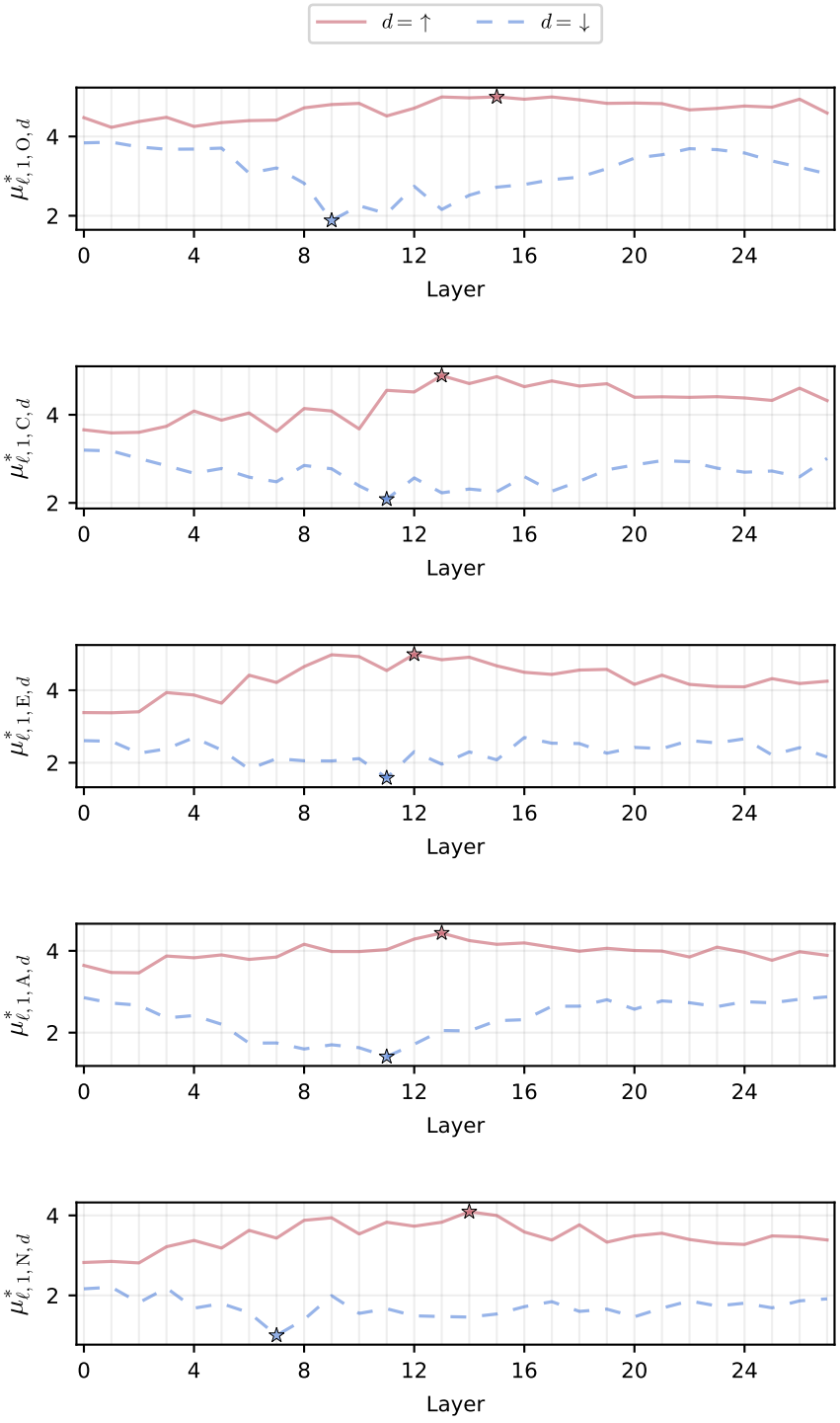


Figure 14: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Llama-3.2-3B-Instruct. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

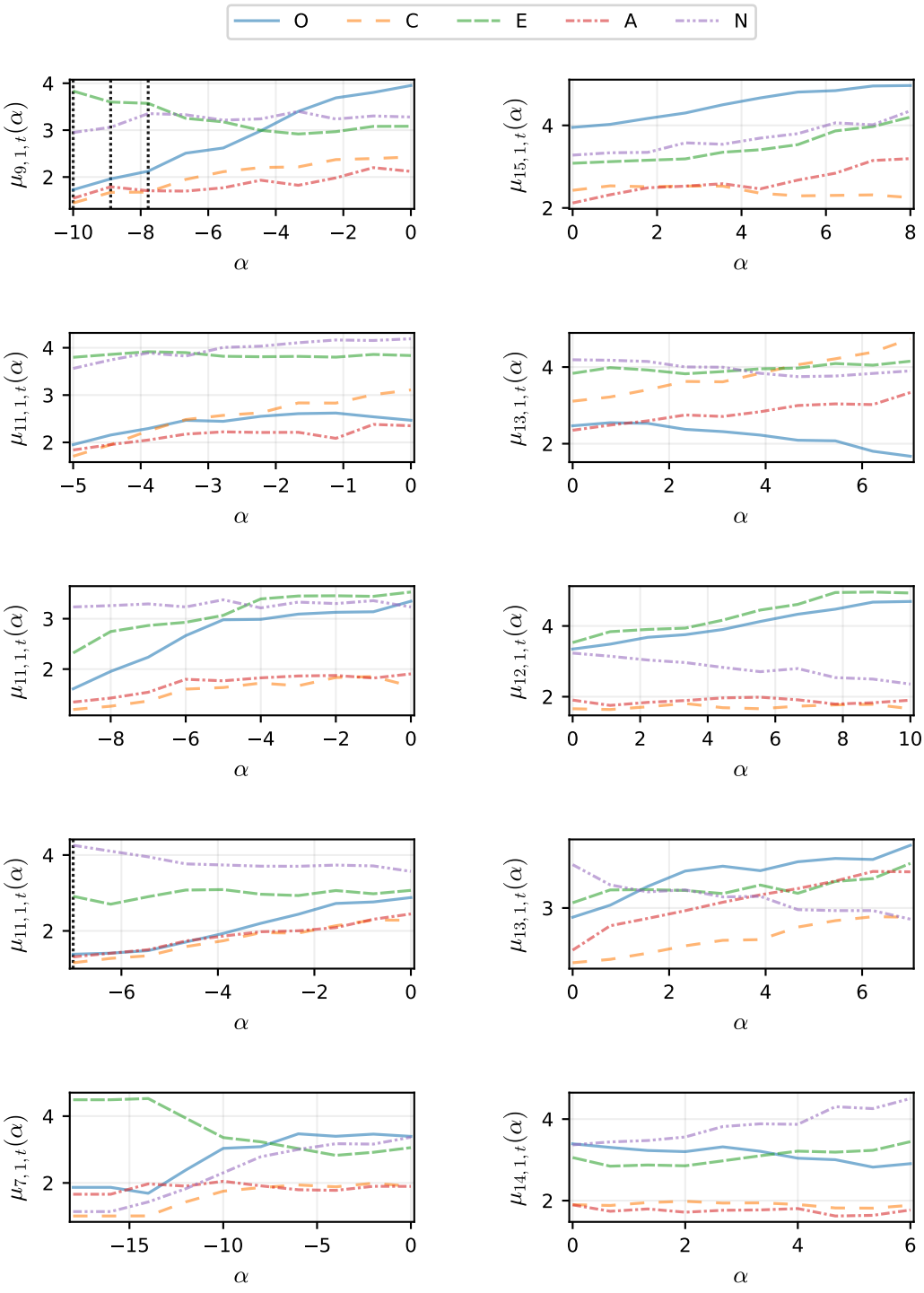


Figure 15: OCEAN scores for Llama-3.2-3B-Instruct on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## N OCEAN Injection Results for Llama-3.1-8B-Instruct

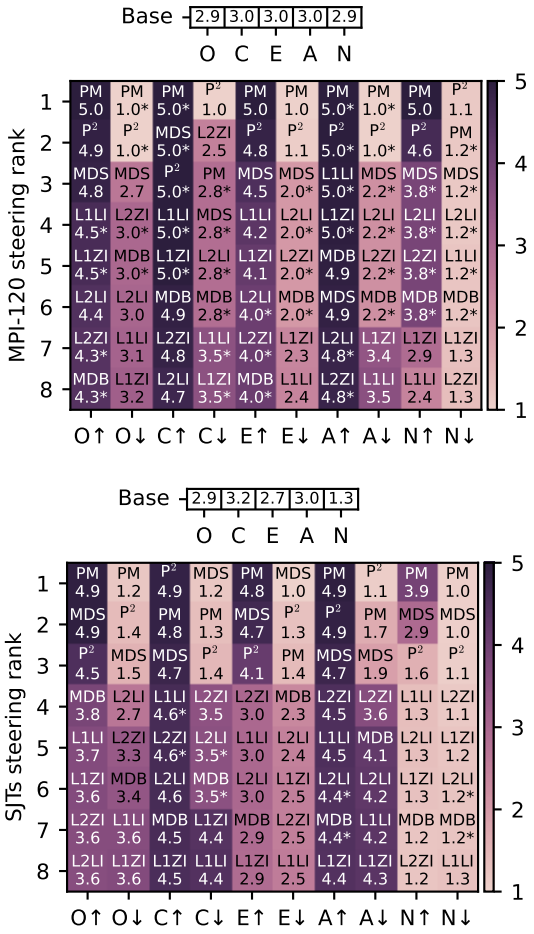


Figure 16: Ranking of steering methods on Llama-3.1-8B-Instruct by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

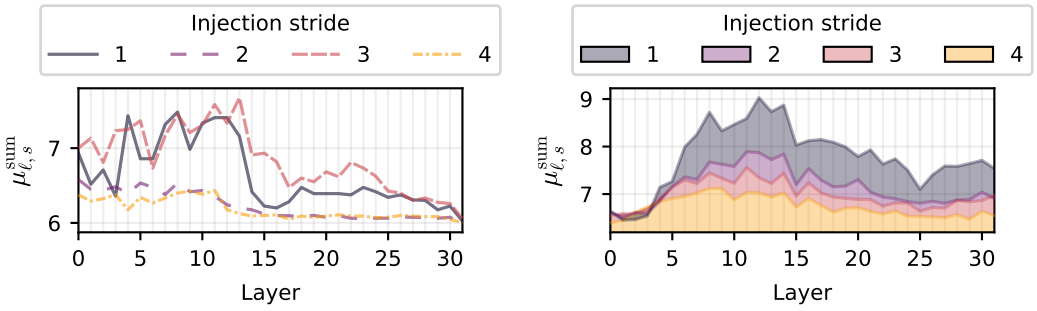


Figure 17: Overall MDS injections steering performance on Llama-3.1-8B-Instruct by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

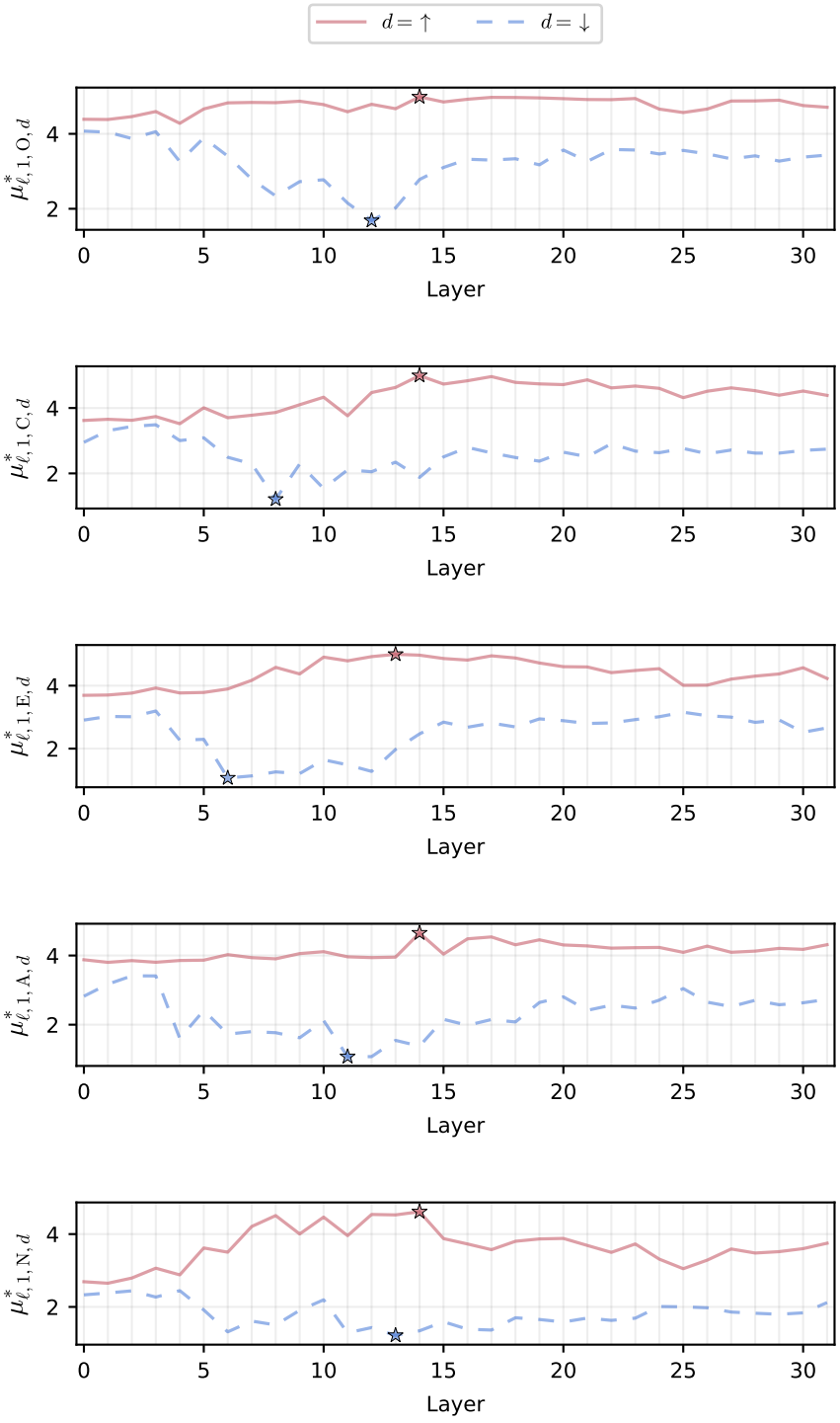


Figure 18: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Llama-3.1-8B-Instruct. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

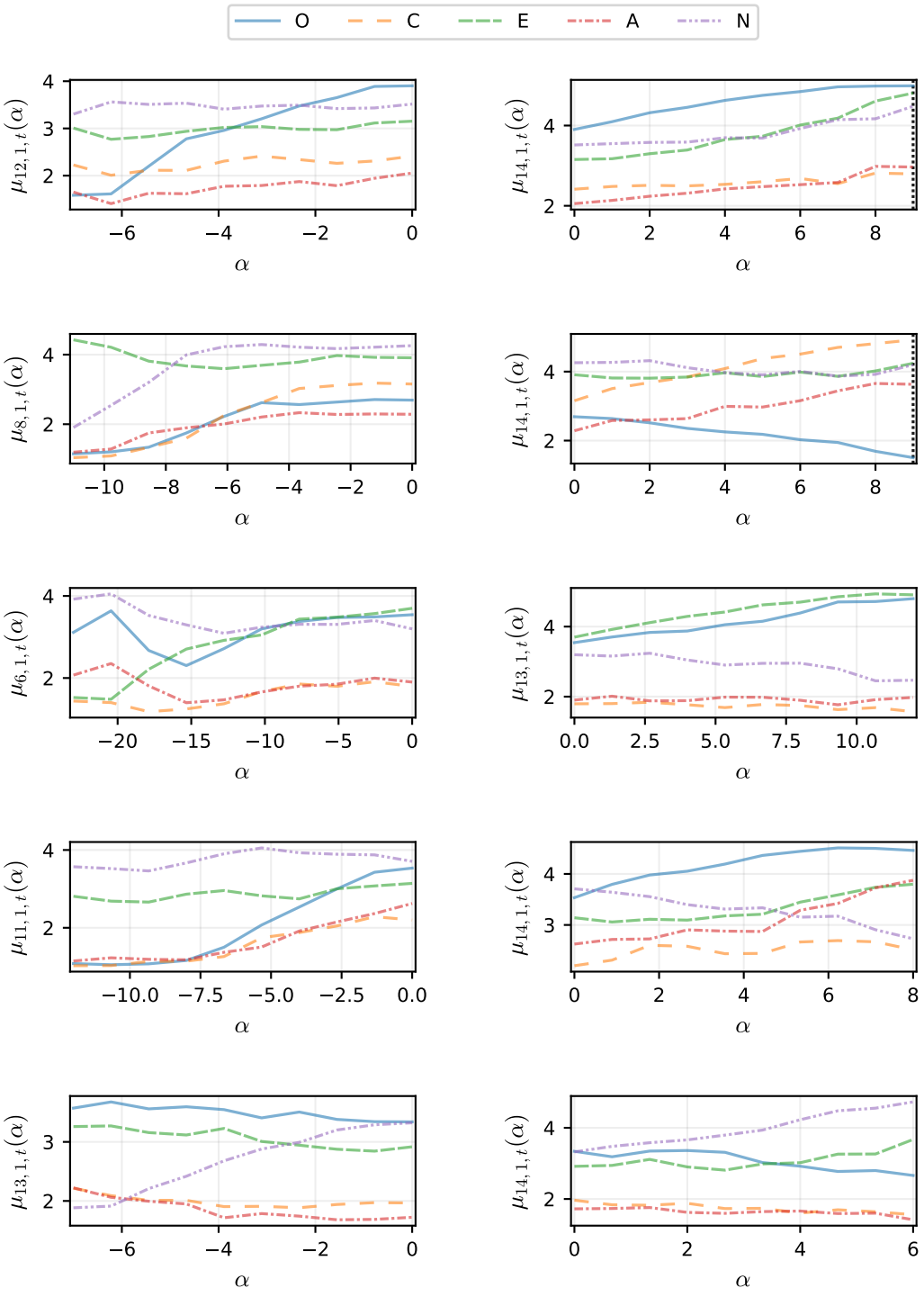


Figure 19: OCEAN scores for Llama-3.1-8B-Instruct on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## O OCEAN Injection Results for Qwen3-1.7B

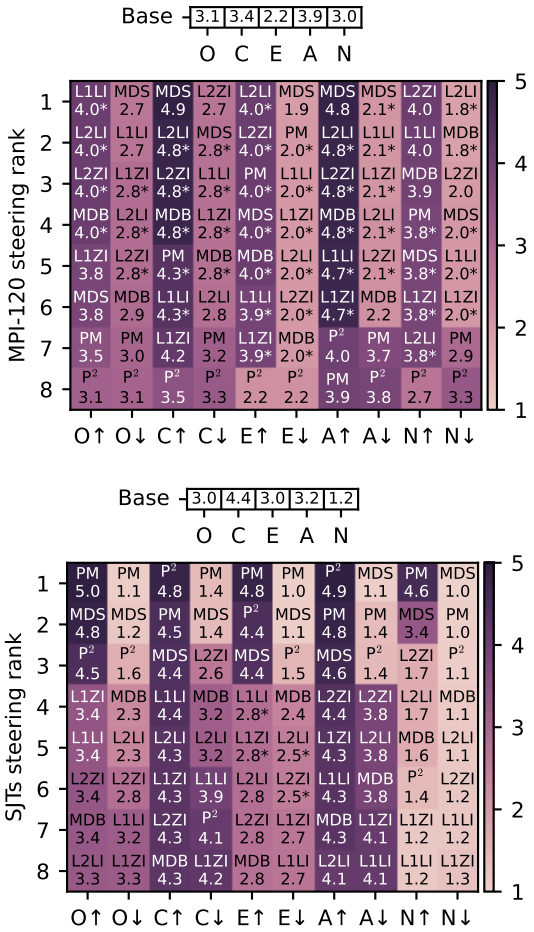


Figure 20: Ranking of steering methods on Qwen3-1.7B by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

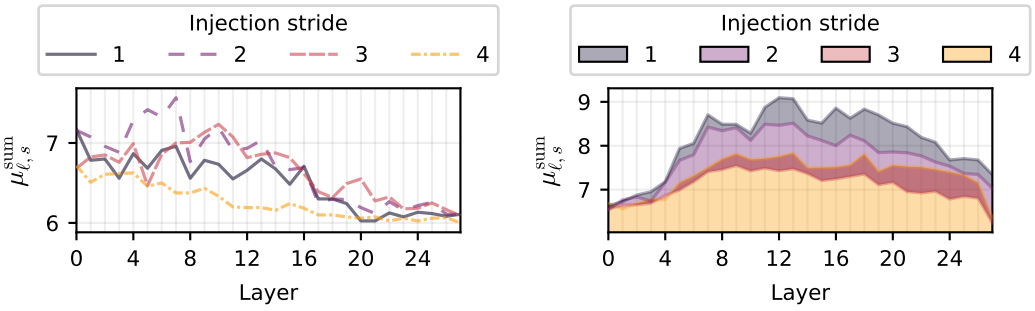


Figure 21: Overall MDS injections steering performance on Qwen3-1.7B by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

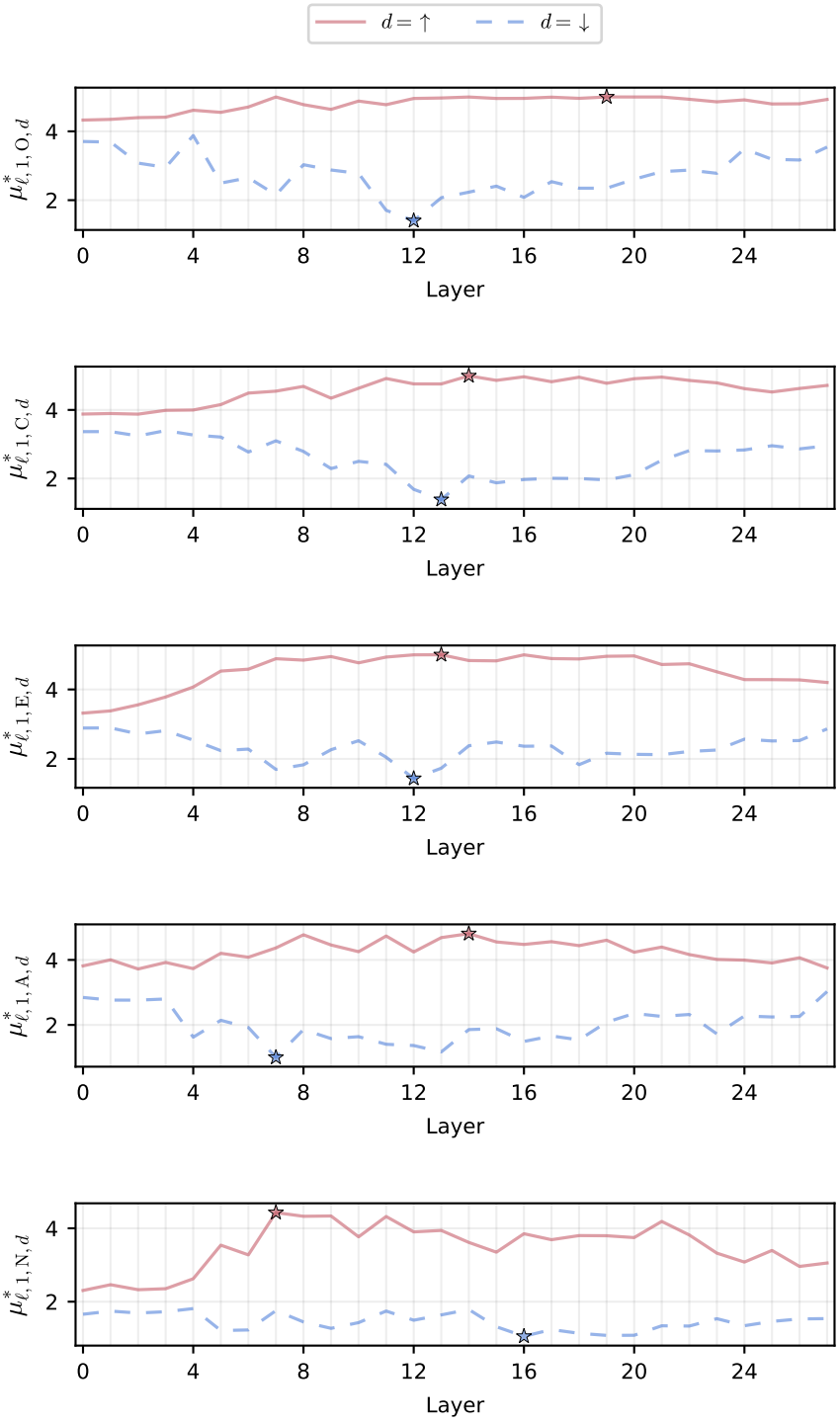


Figure 22: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Qwen3-1.7B. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

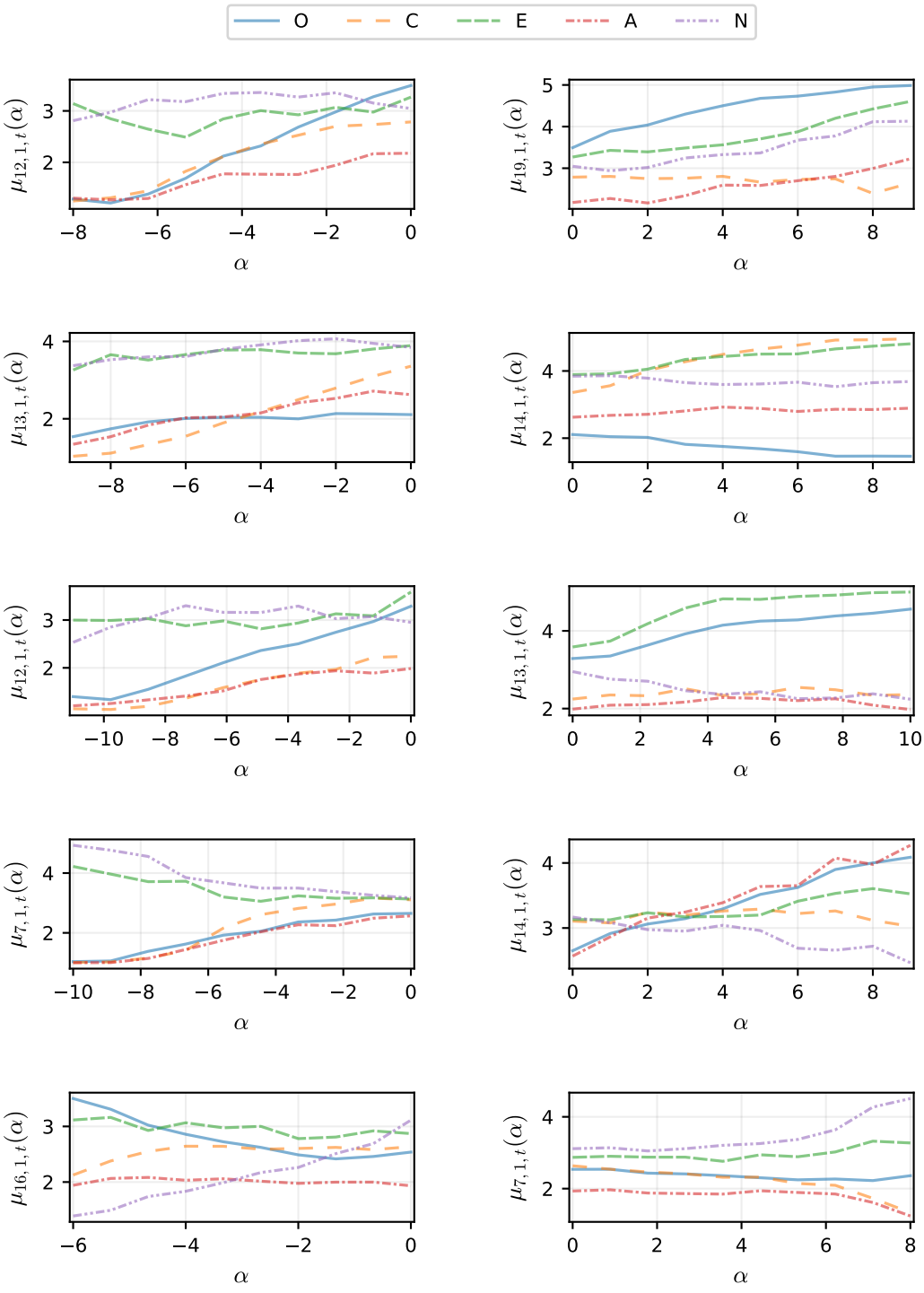


Figure 23: OCEAN scores for Qwen3-1.7B on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## P OCEAN Injection Results for Qwen3-4B

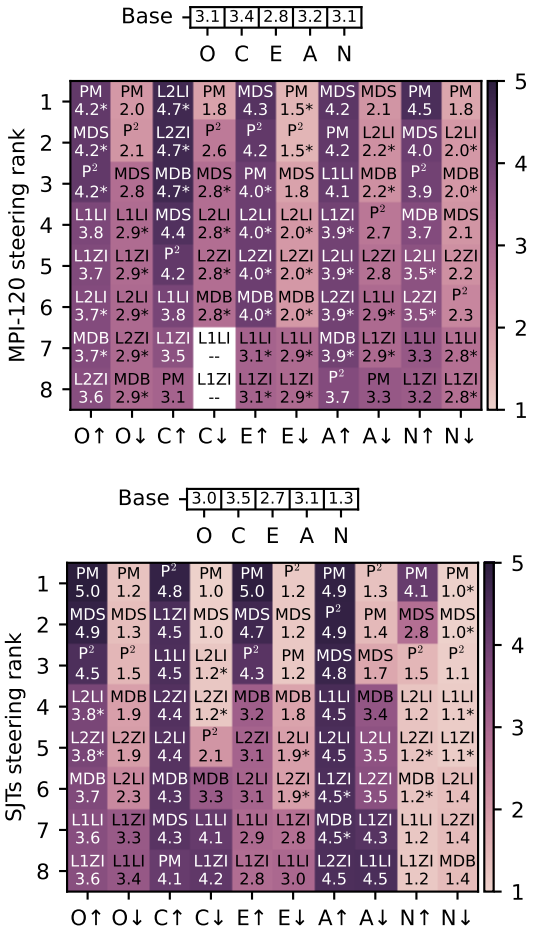


Figure 24: Ranking of steering methods on Qwen3-4B by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

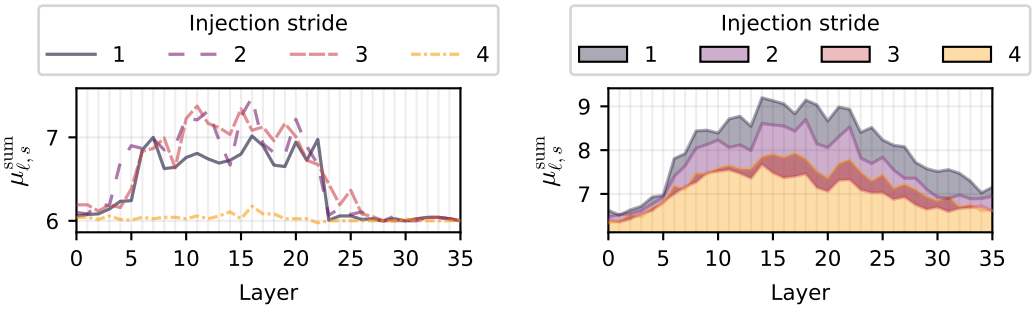


Figure 25: Overall MDS injections steering performance on Qwen3-4B by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

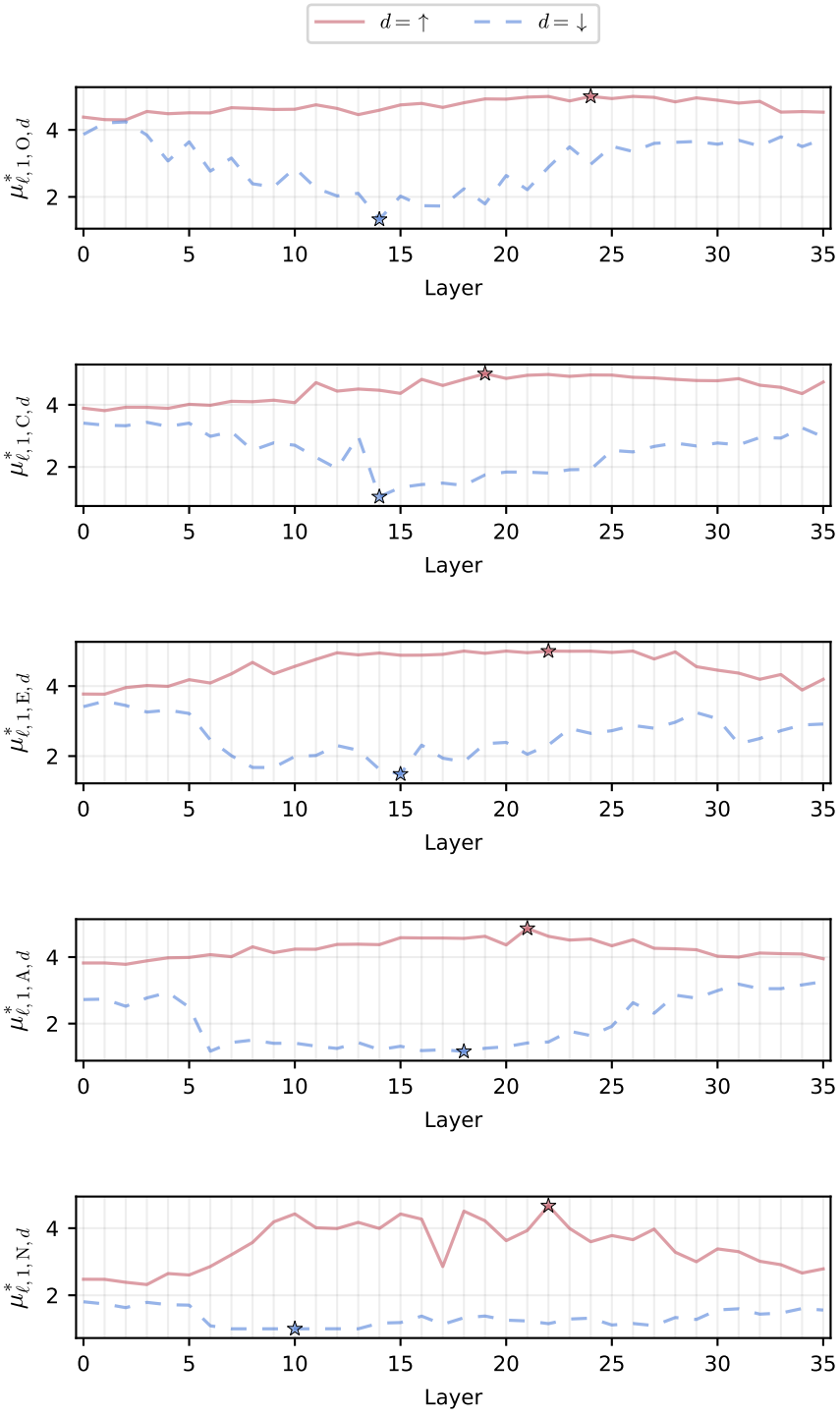


Figure 26: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Qwen3-4B. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

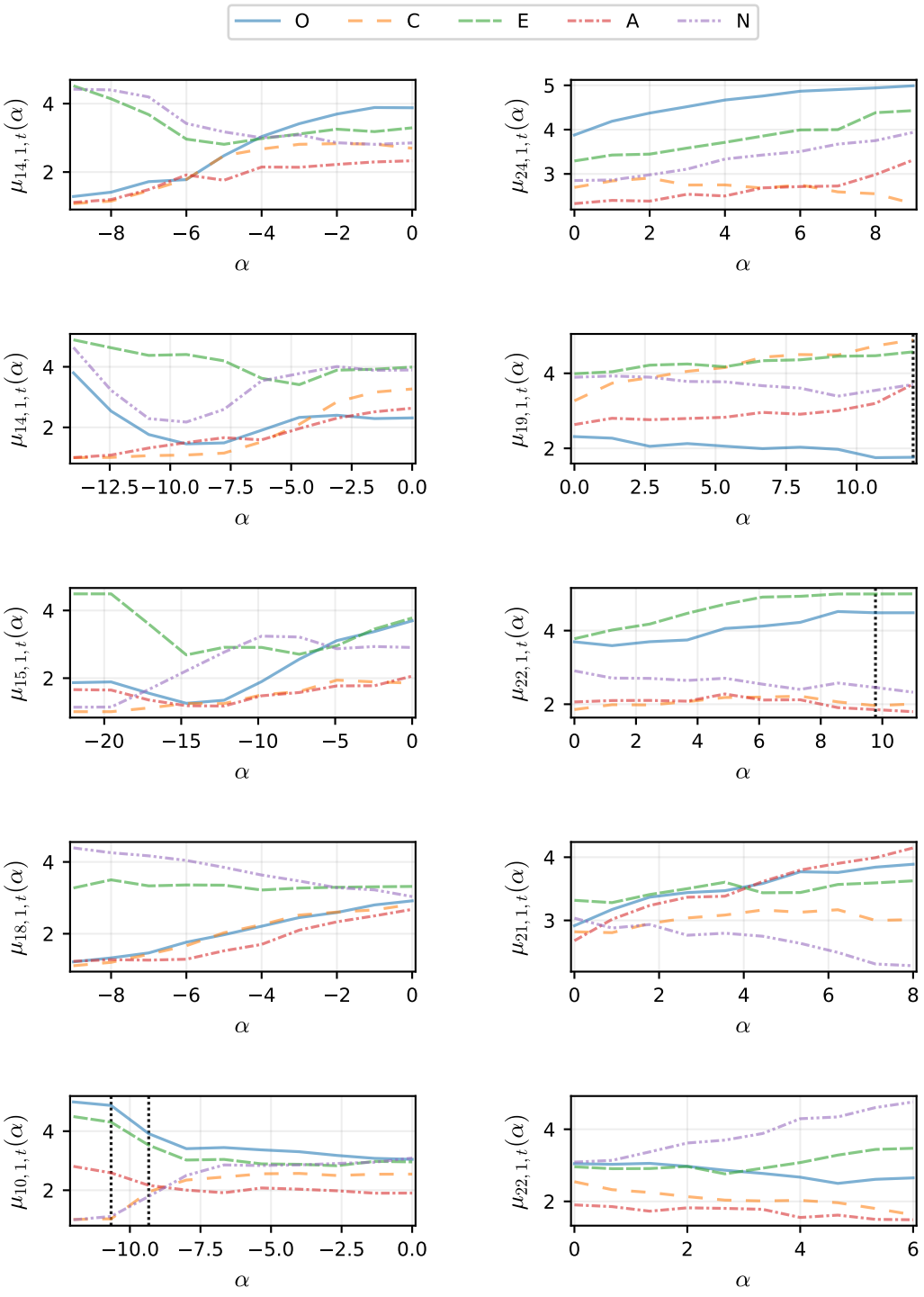


Figure 27: OCEAN scores for Qwen3-4B on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## Q OCEAN Injection Results for Qwen3-8B

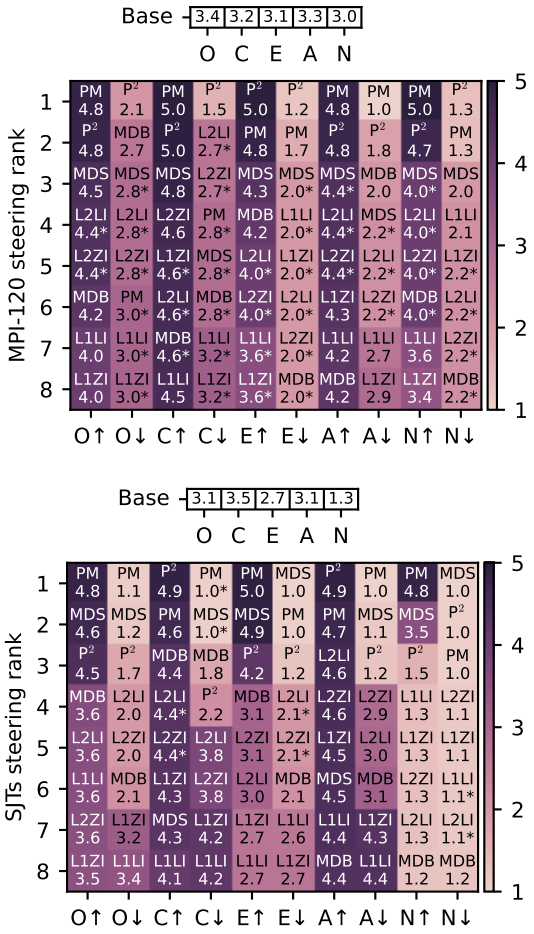


Figure 28: Ranking of steering methods on Qwen3-8B by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

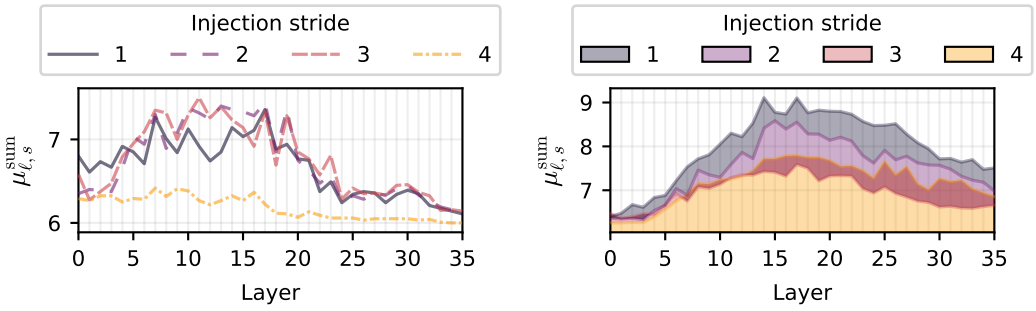


Figure 29: Overall MDS injections steering performance on Qwen3-8B by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

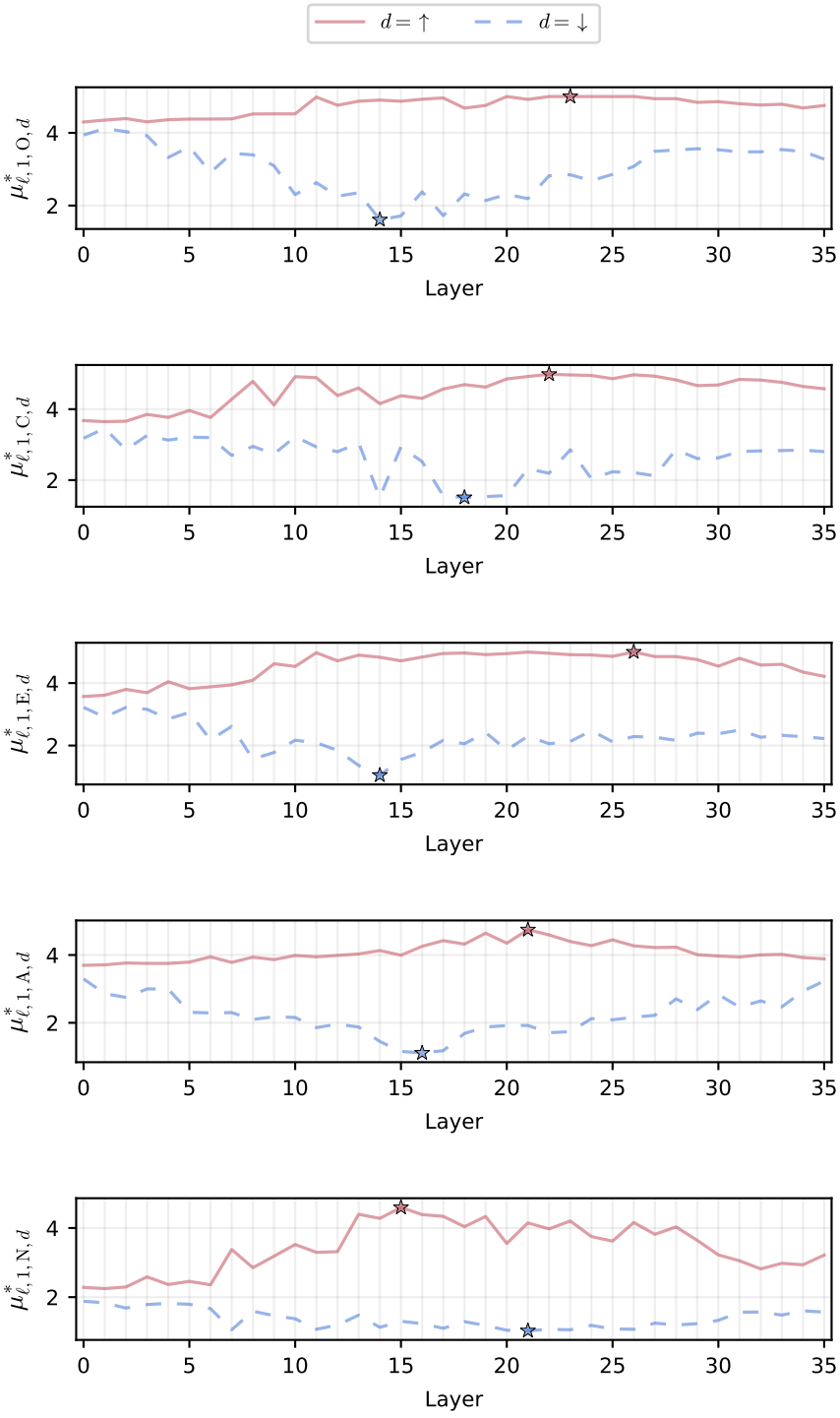


Figure 30: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Qwen3-8B. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

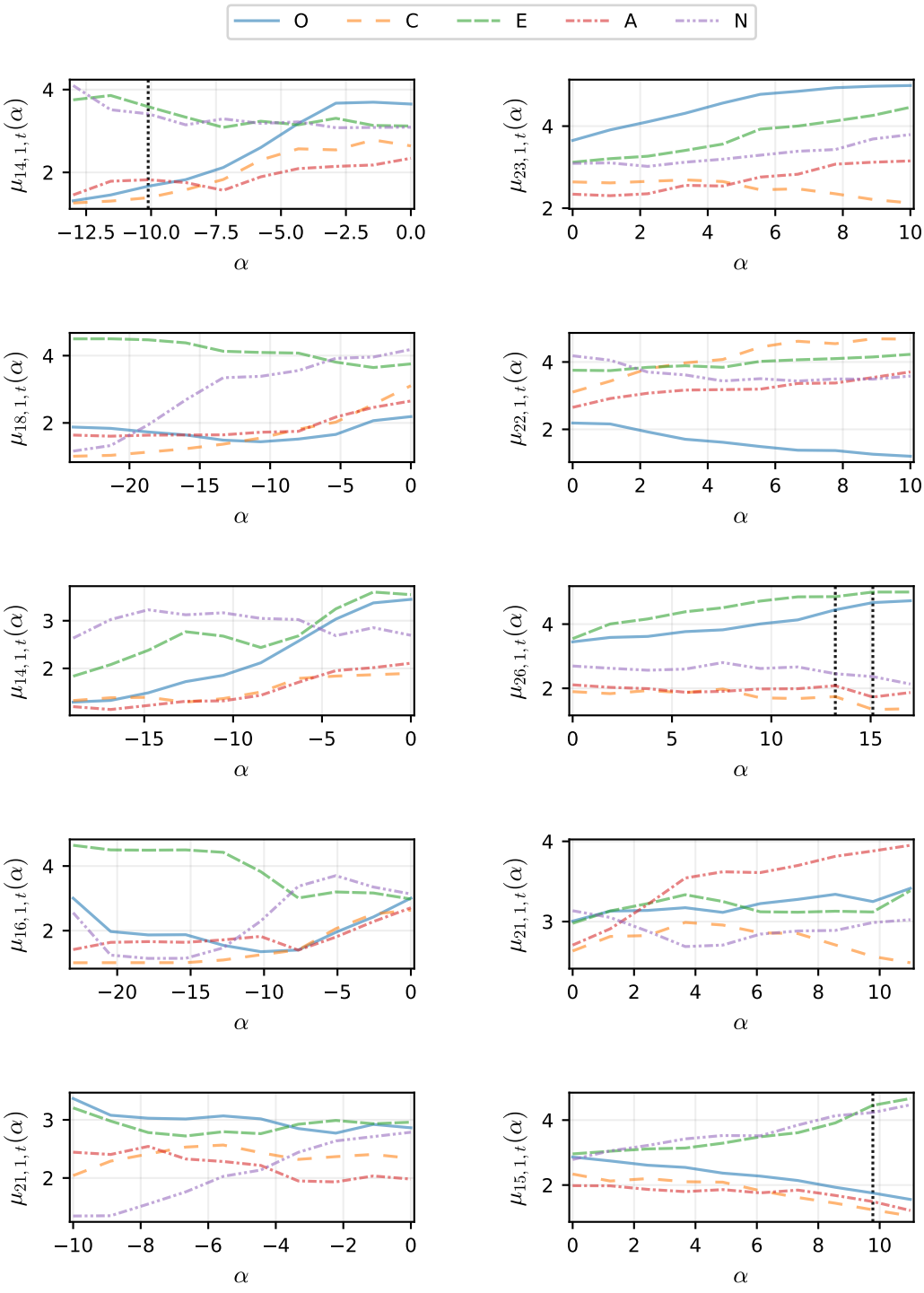


Figure 31: OCEAN scores for Qwen3-8B on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## R OCEAN Injection Results for Qwen3-14B

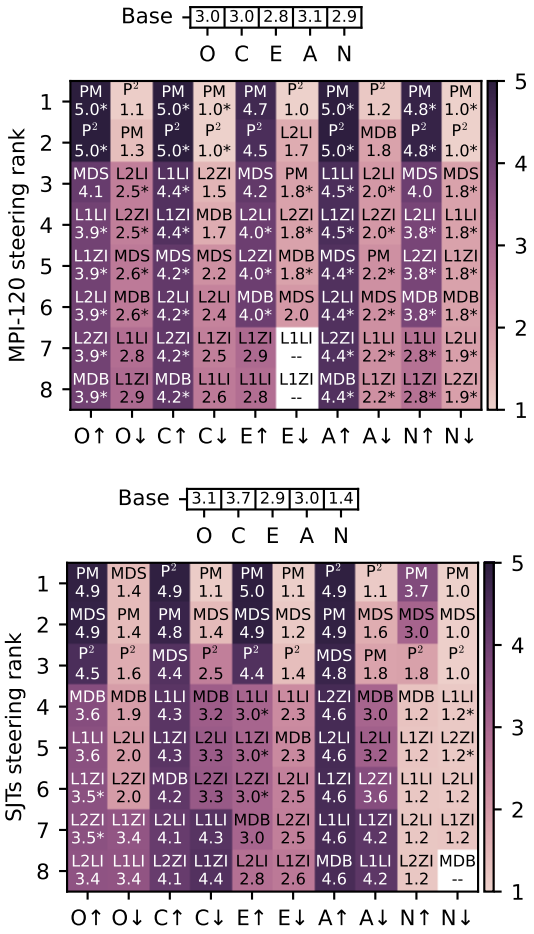


Figure 32: Ranking of steering methods on Qwen3-14B by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

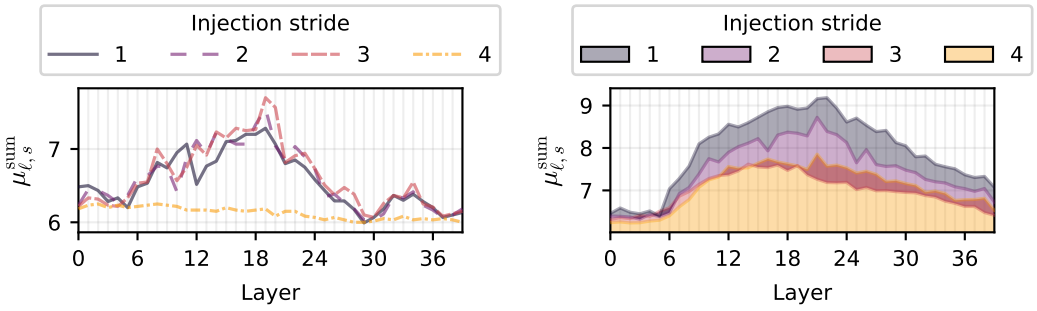


Figure 33: Overall MDS injections steering performance on Qwen3-14B by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

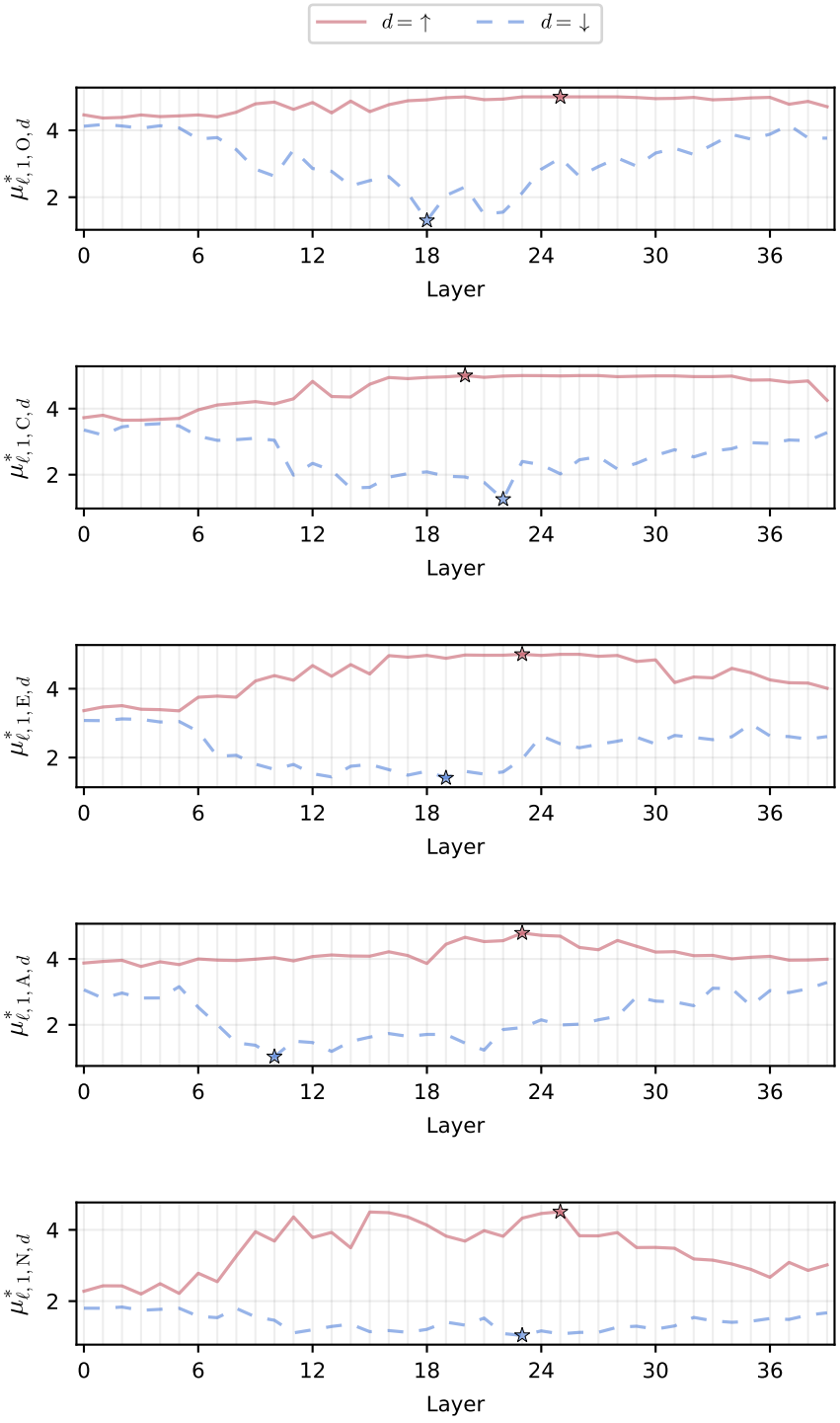


Figure 34: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Qwen3-14B. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

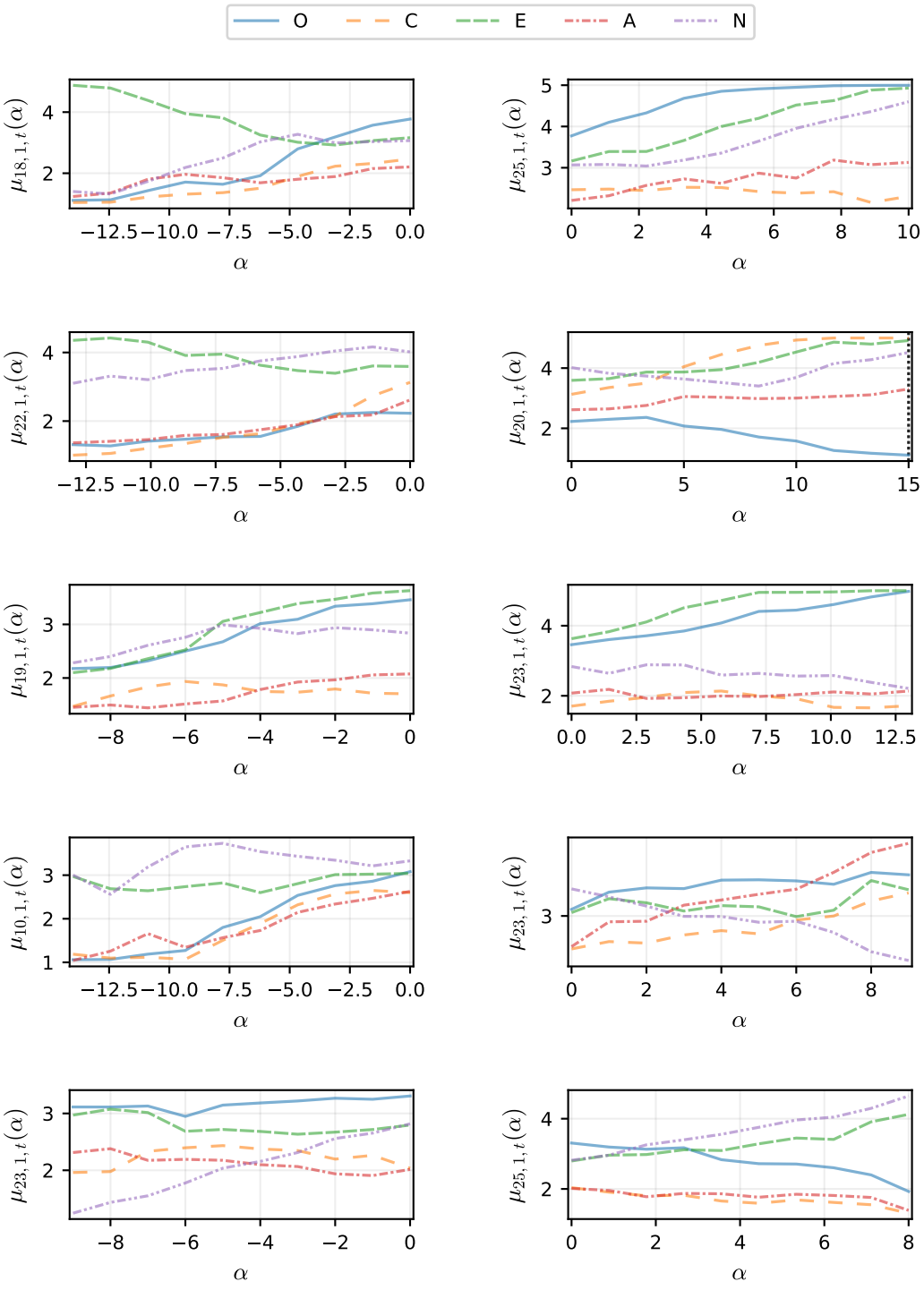


Figure 35: OCEAN scores for Qwen3-14B on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## S OCEAN Injection Results for Qwen3-32B

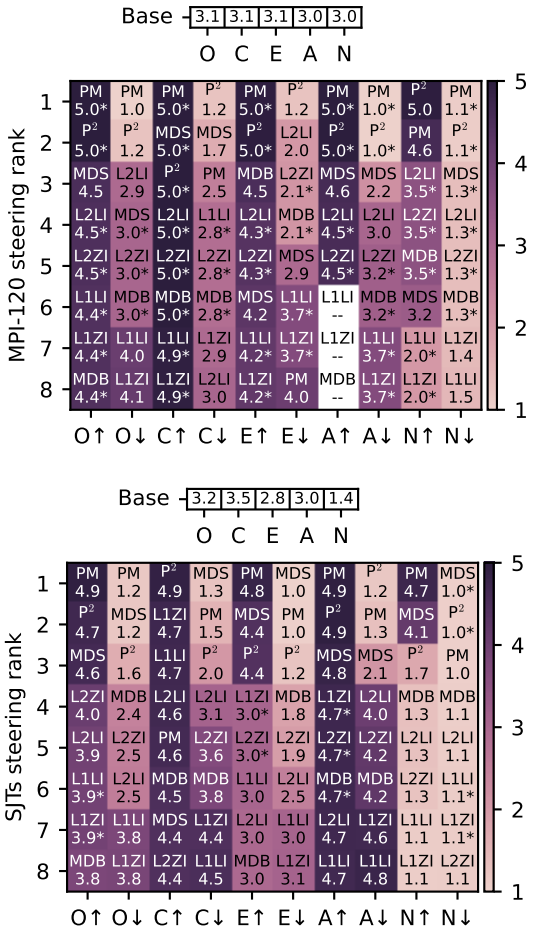


Figure 36: Ranking of steering methods on Qwen3-32B by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

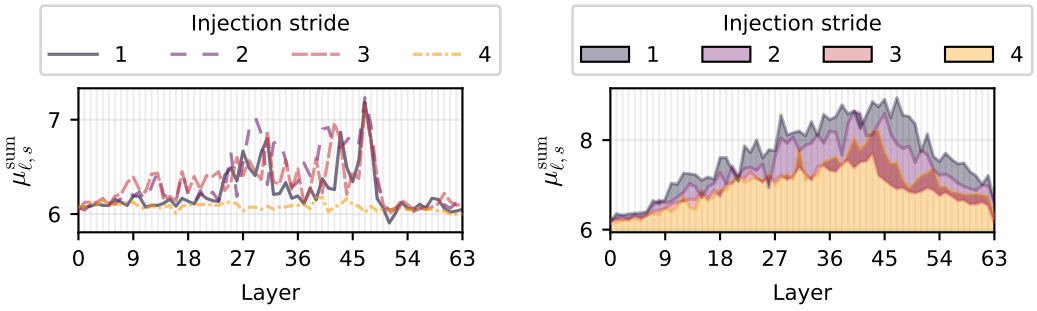


Figure 37: Overall MDS injections steering performance on Qwen3-32B by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

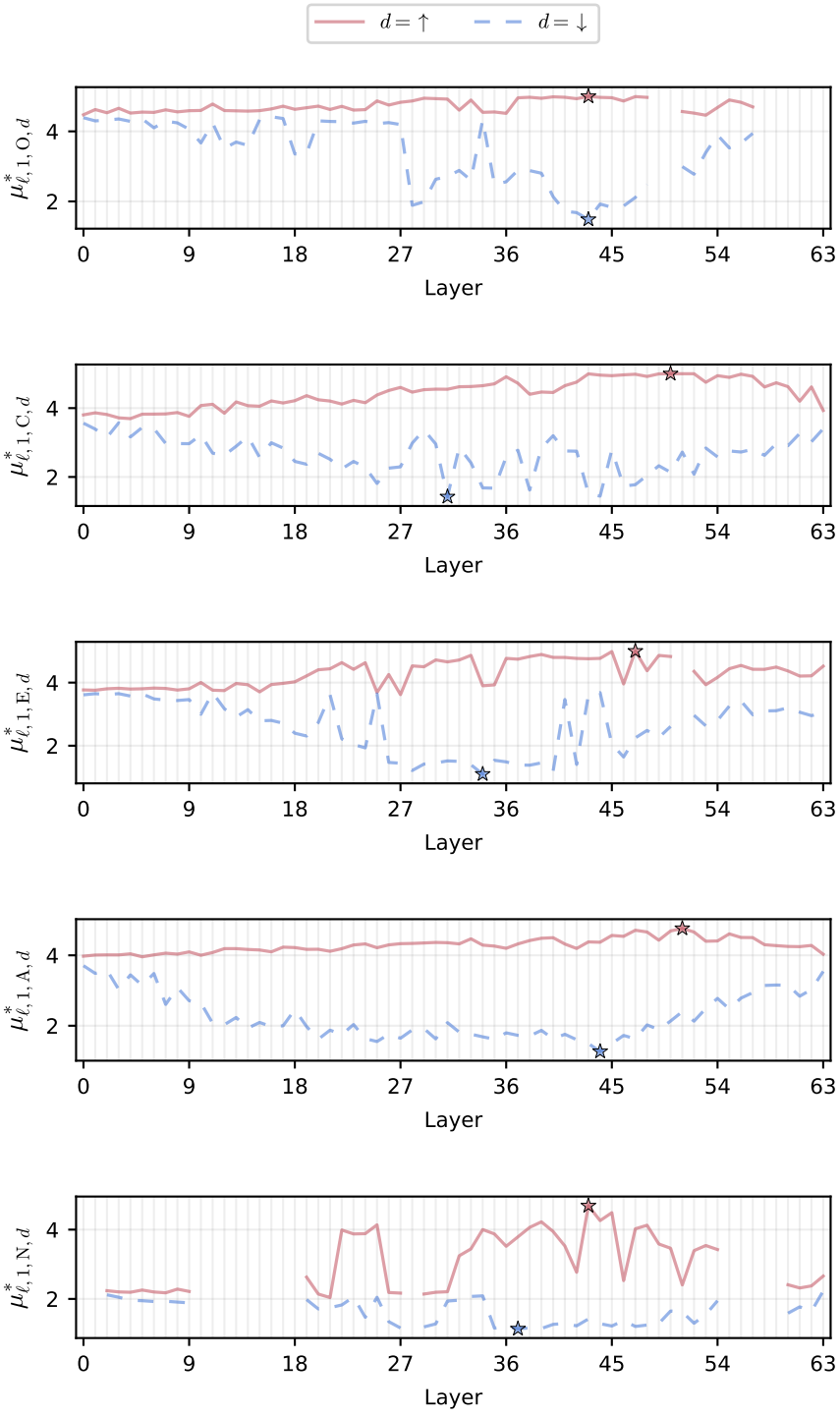


Figure 38: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Qwen3-32B. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

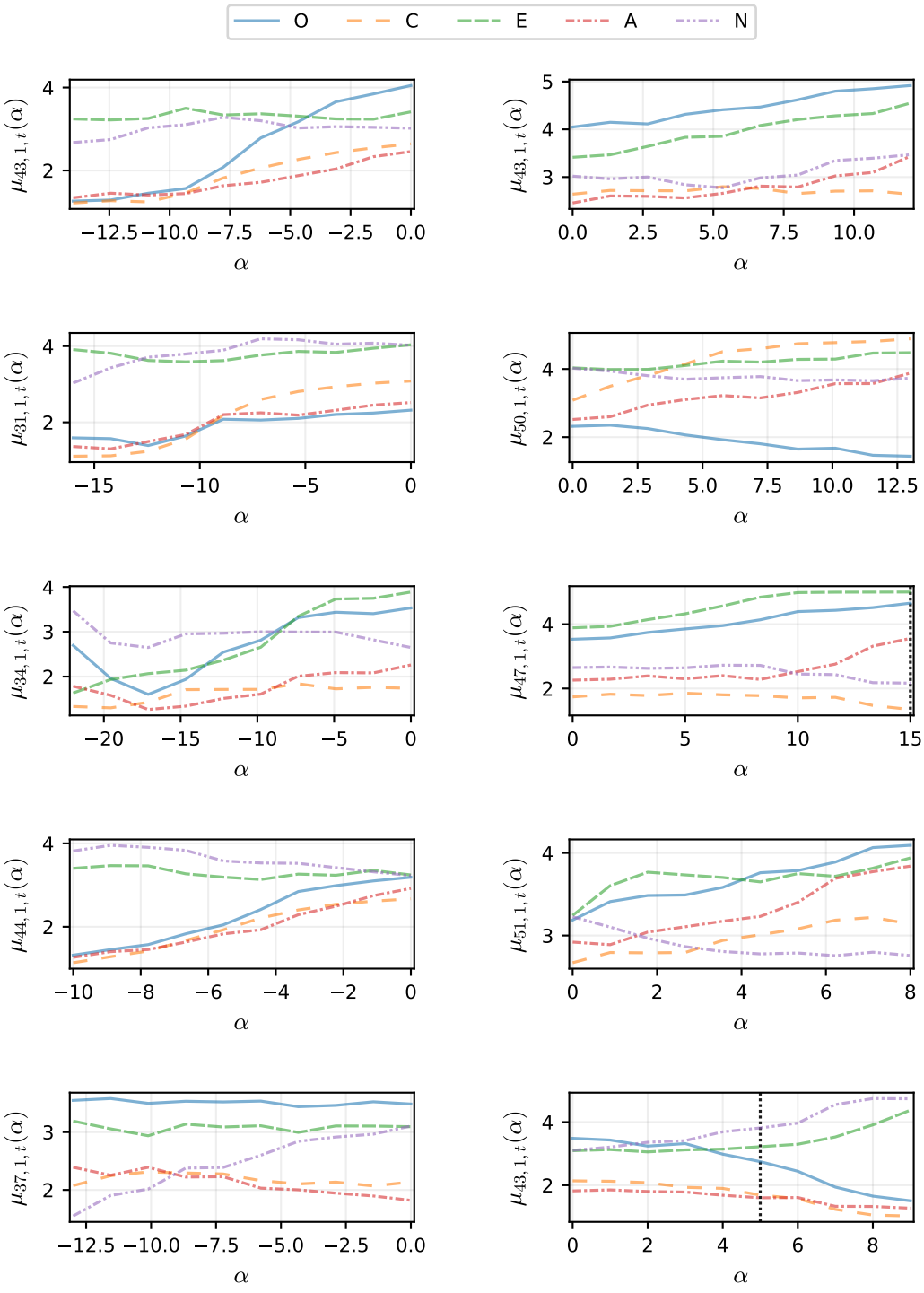


Figure 39: OCEAN scores for Qwen3-32B on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## T OCEAN Injection Results for gemma-3-1b-it

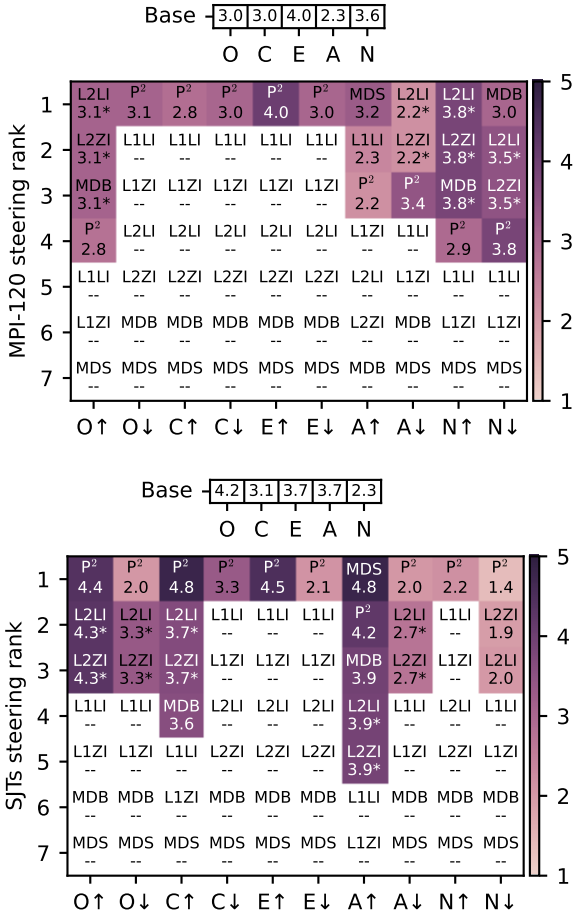


Figure 40: Ranking of steering methods on gemma-3-1b-it by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

No other valid MDS injection results to plot.

## U OCEAN Injection Results for gemma-3-4b-it

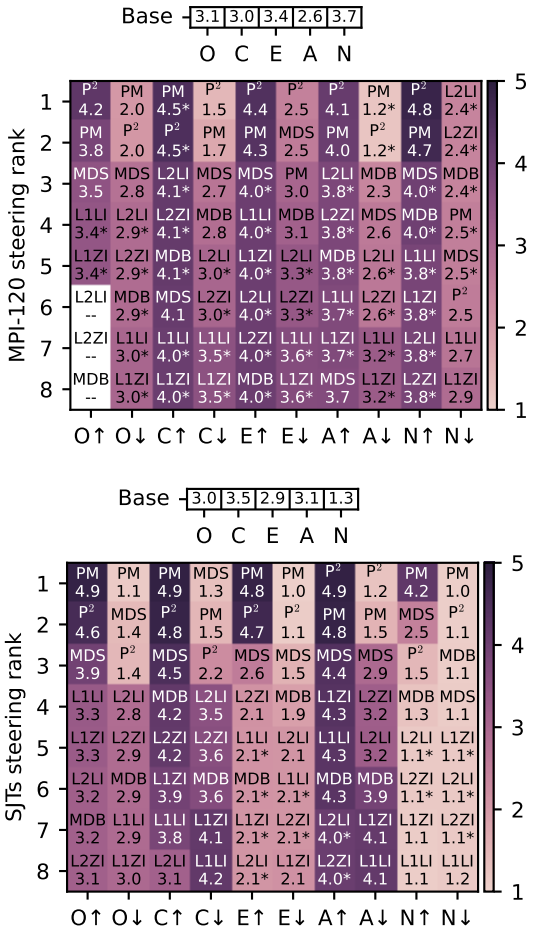


Figure 41: Ranking of steering methods on gemma-3-4b-it by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

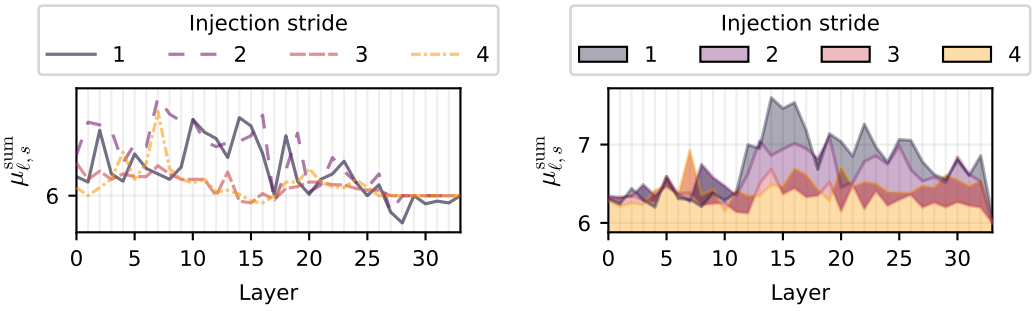


Figure 42: Overall MDS injections steering performance on gemma-3-4b-it by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

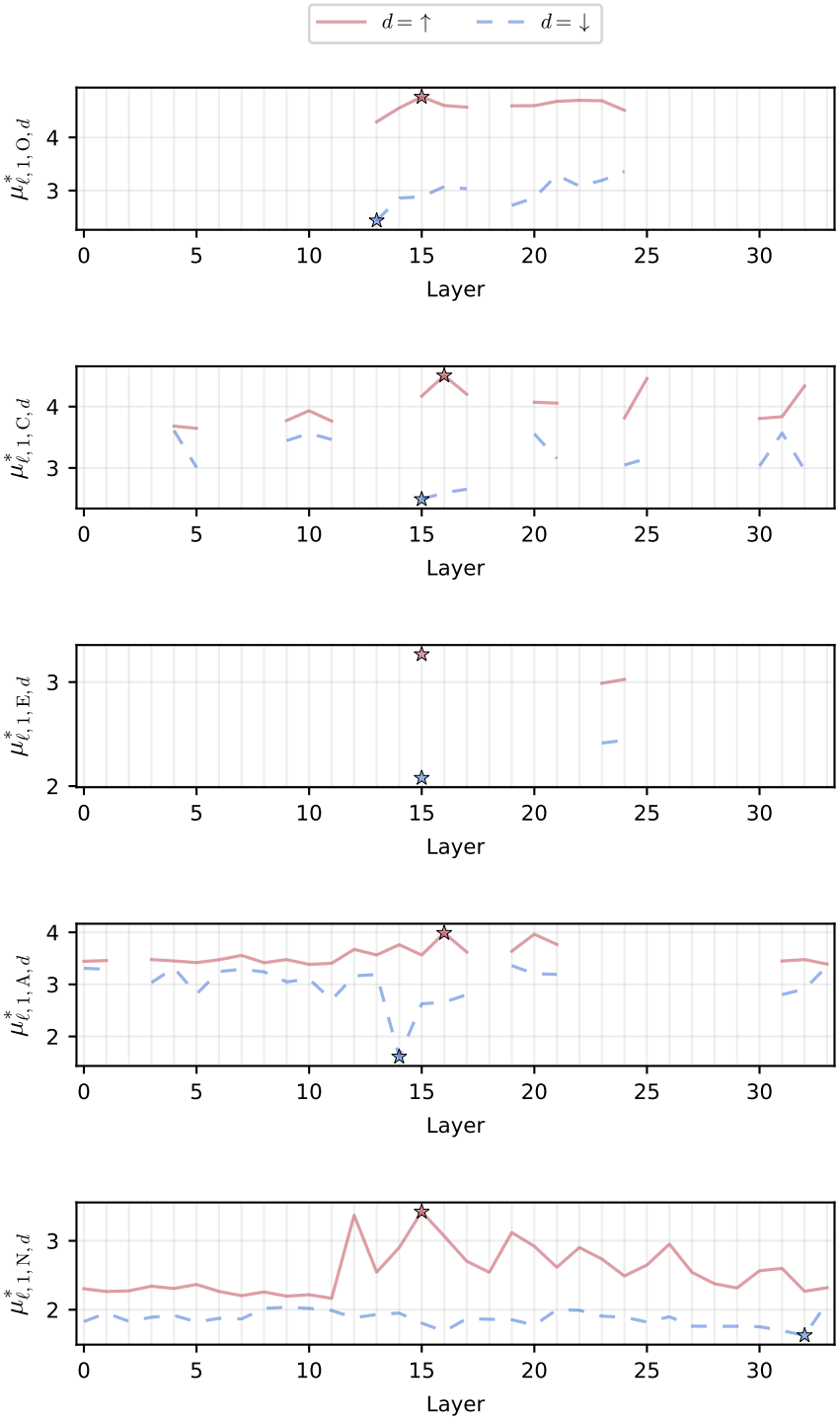


Figure 43: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on gemma-3-4b-it. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

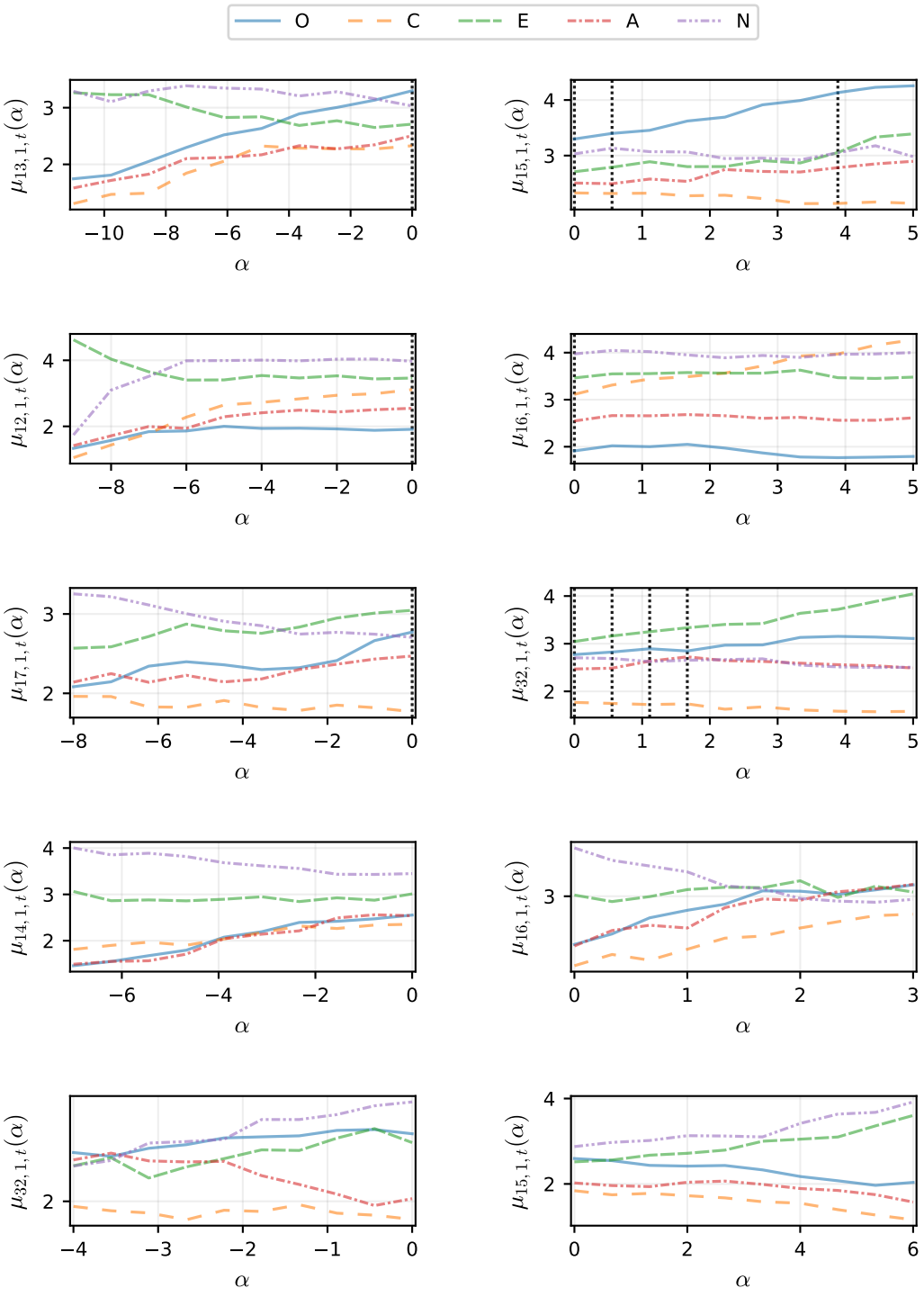


Figure 44: OCEAN scores for gemma-3-4b-it on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

# V OCEAN Injection Results for gemma-3-12b-it

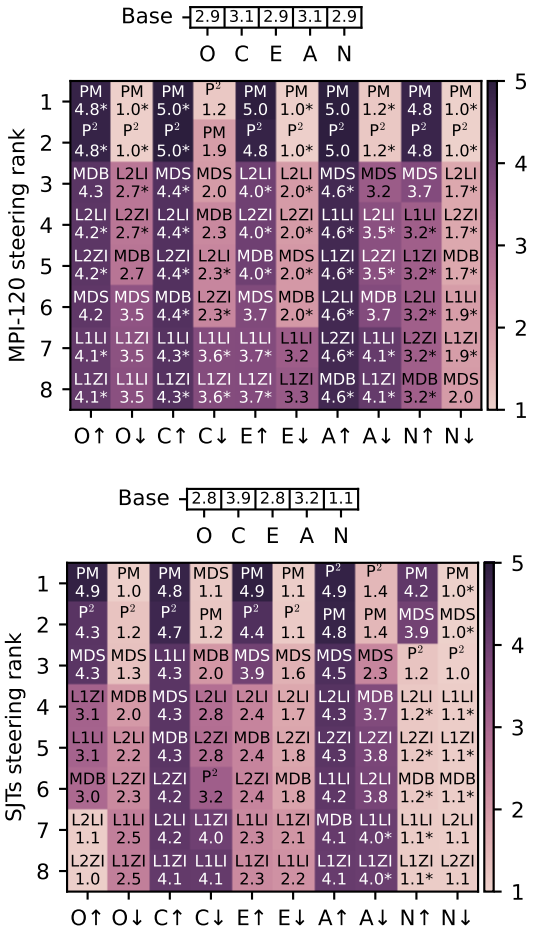


Figure 45: Ranking of steering methods on gemma-3-12b-it by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

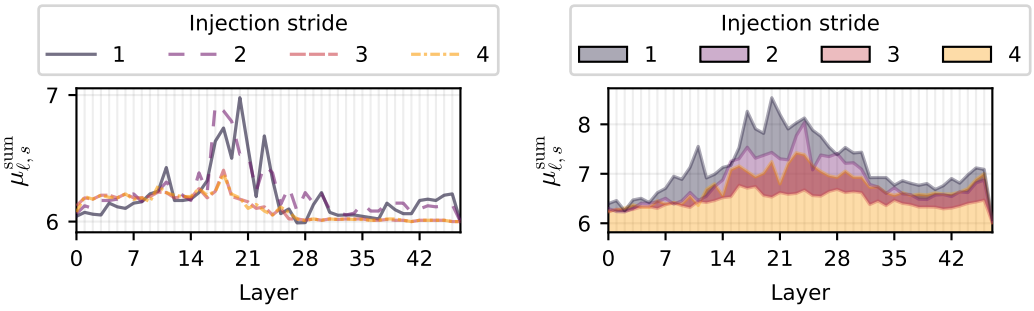


Figure 46: Overall MDS injections steering performance on gemma-3-12b-it by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

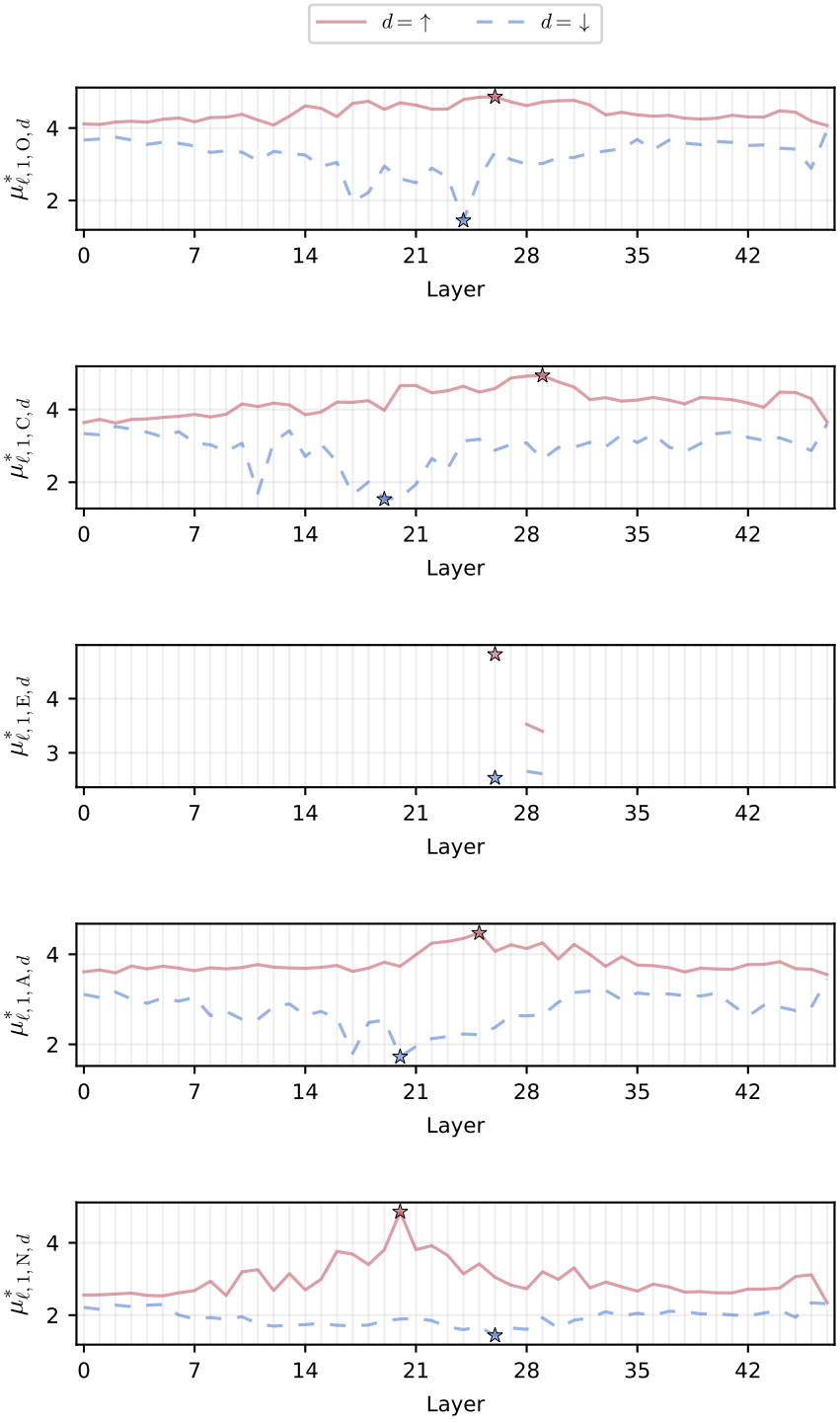


Figure 47: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on gemma-3-12b-it. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

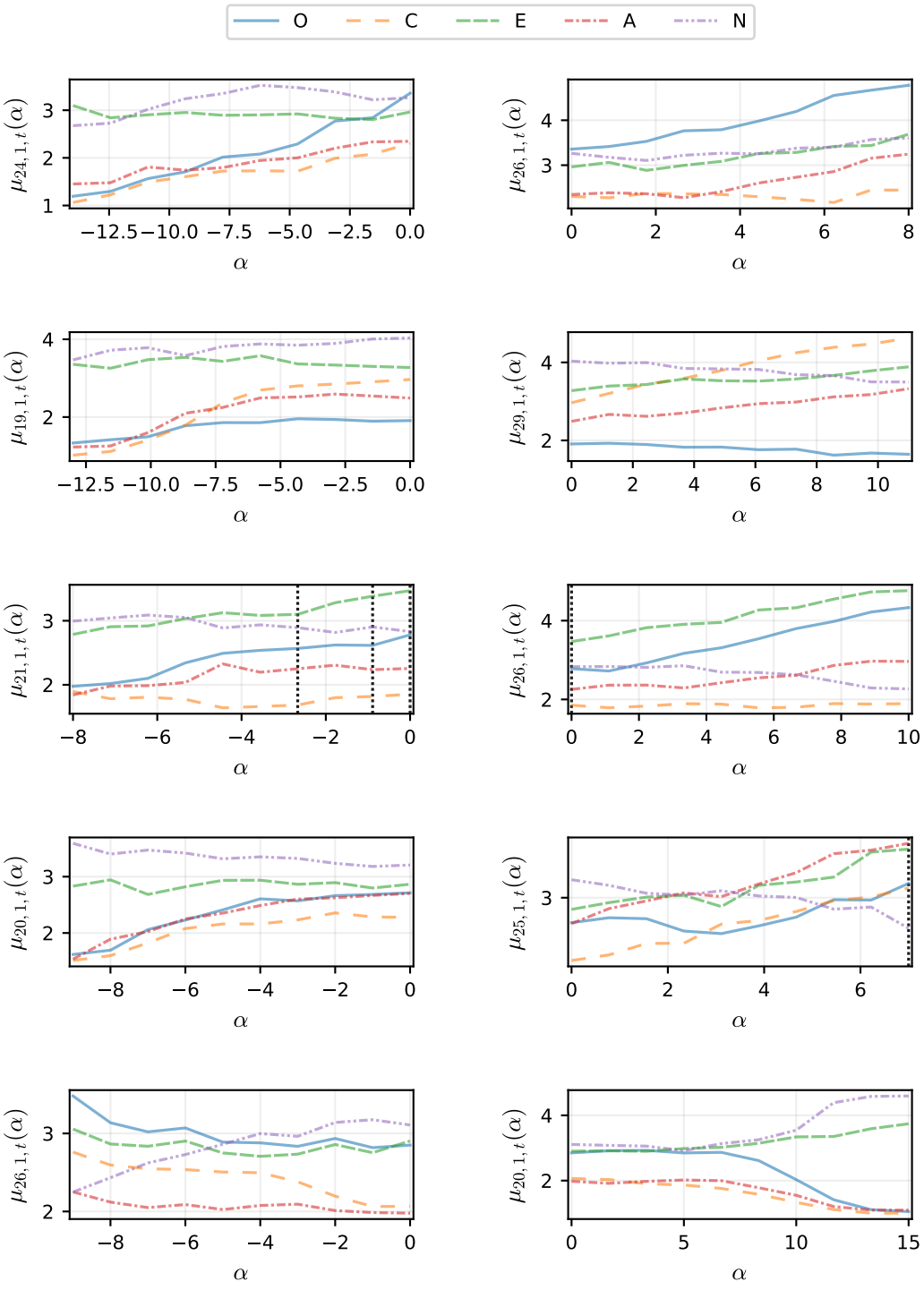


Figure 48: OCEAN scores for gemma-3-12b-it on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## W OCEAN Injection Results for gemma-3-27b-it

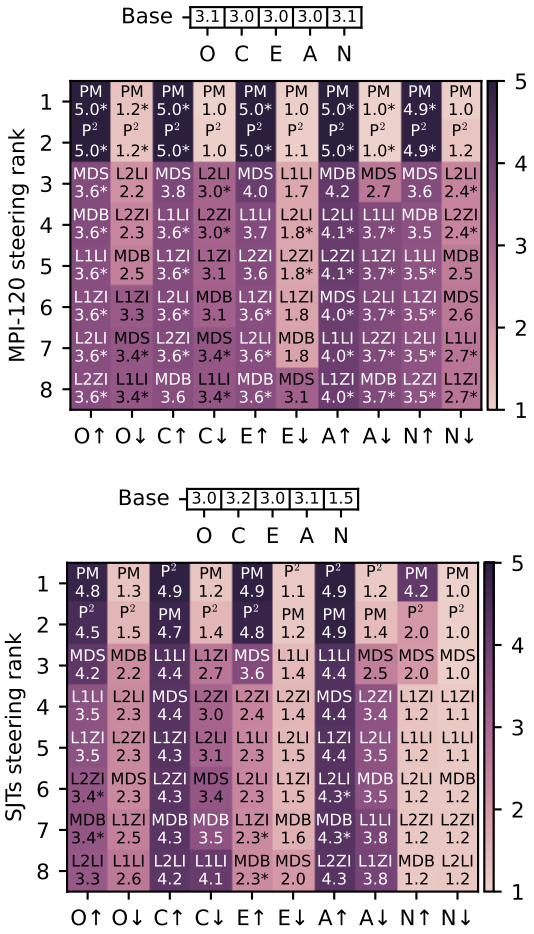


Figure 49: Ranking of steering methods on gemma-3-27b-it by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

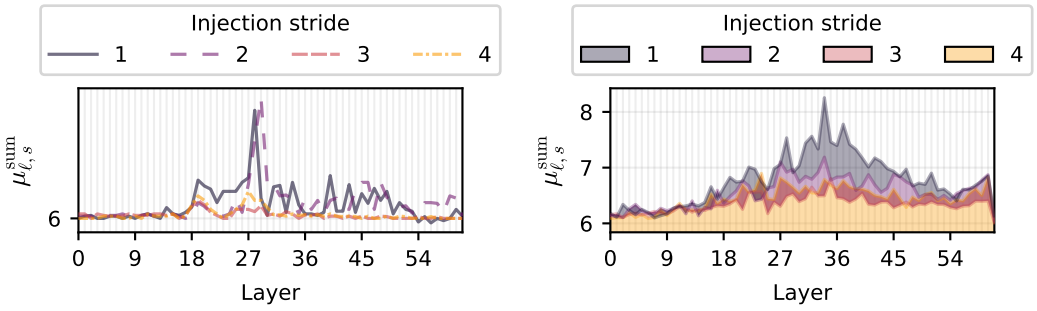


Figure 50: Overall MDS injections steering performance on gemma-3-27b-it by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

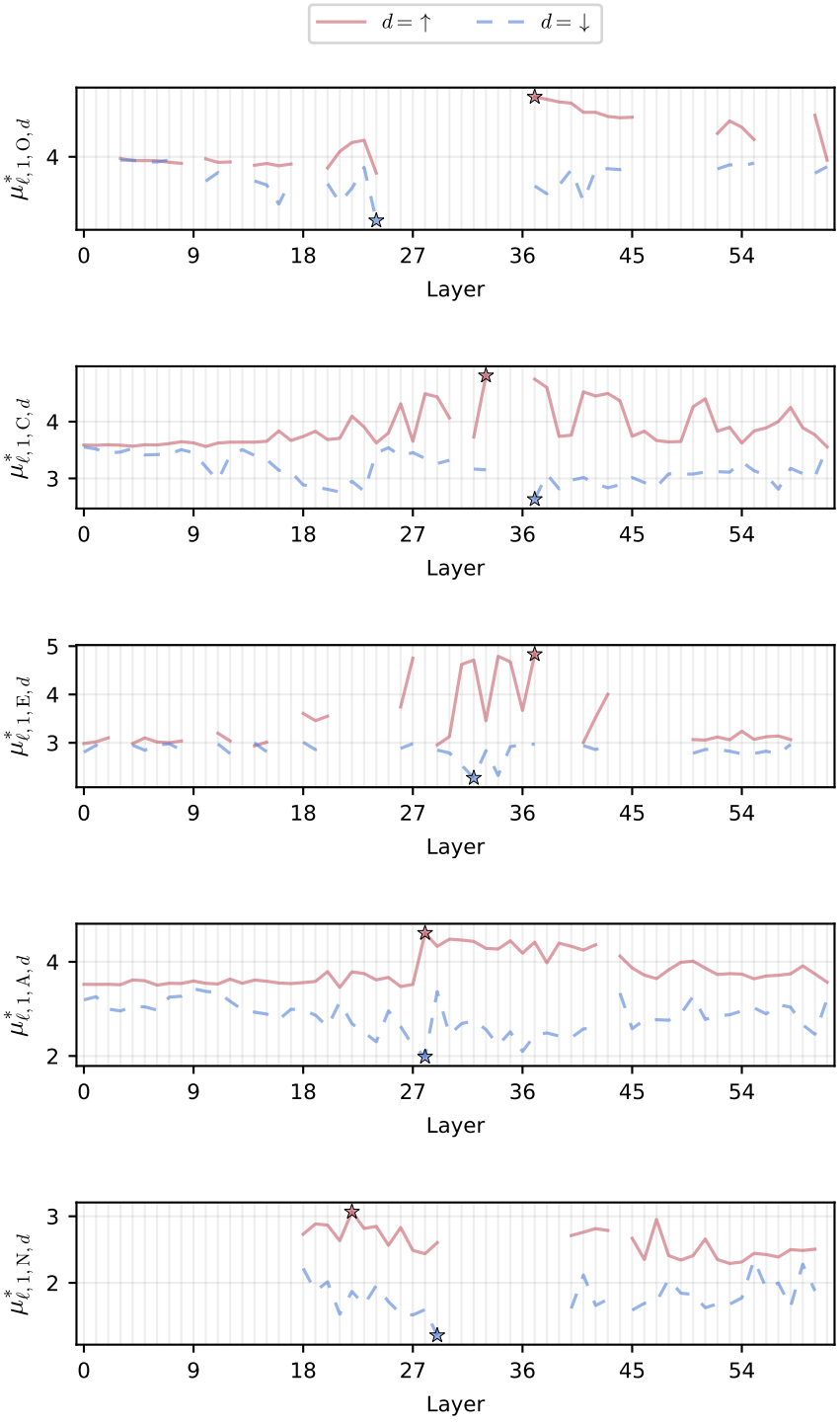


Figure 51: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on gemma-3-27b-it. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

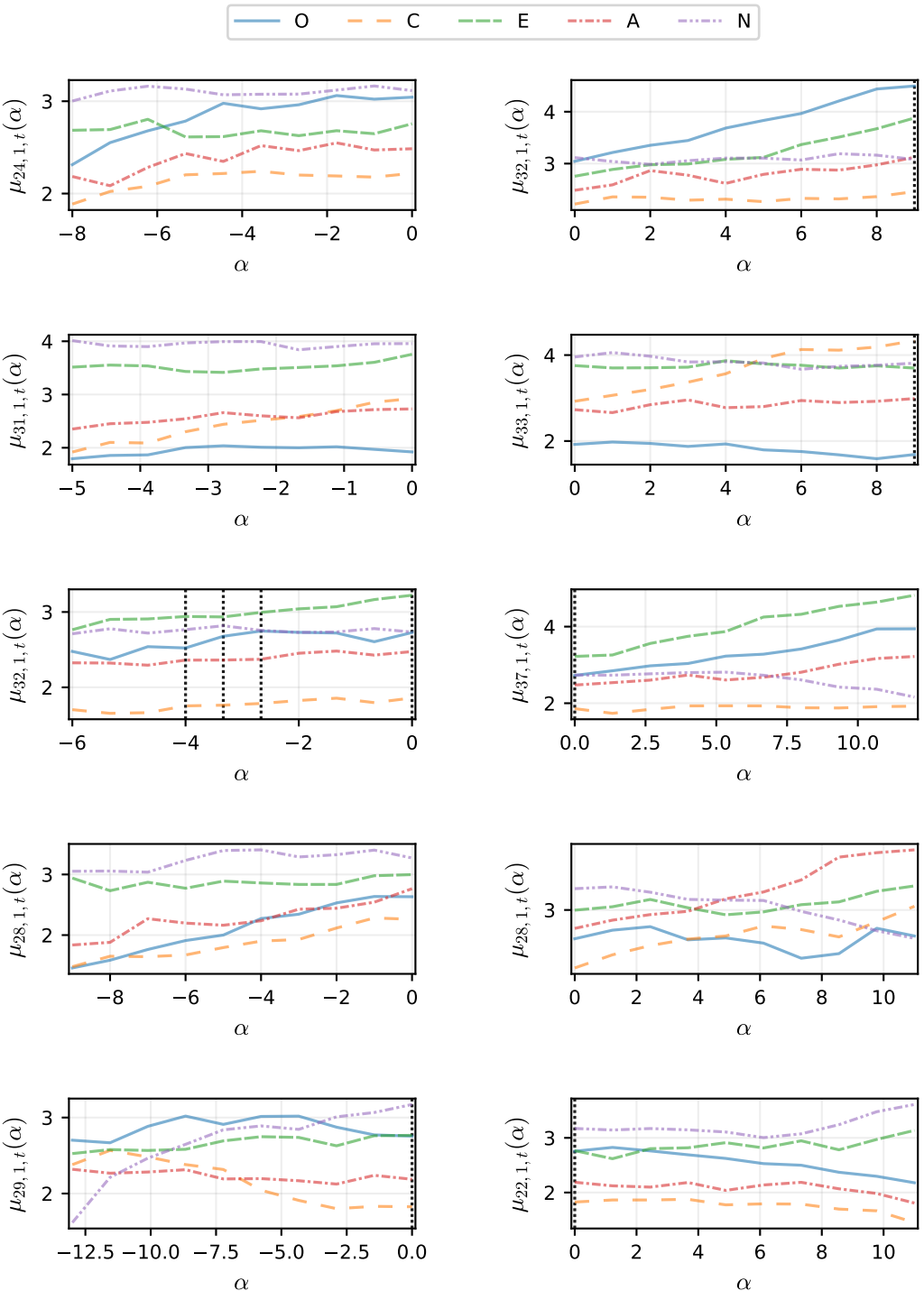


Figure 52: OCEAN scores for gemma-3-27b-it on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

# X OCEAN Injection Results for Olmo-3-7B-Instruct

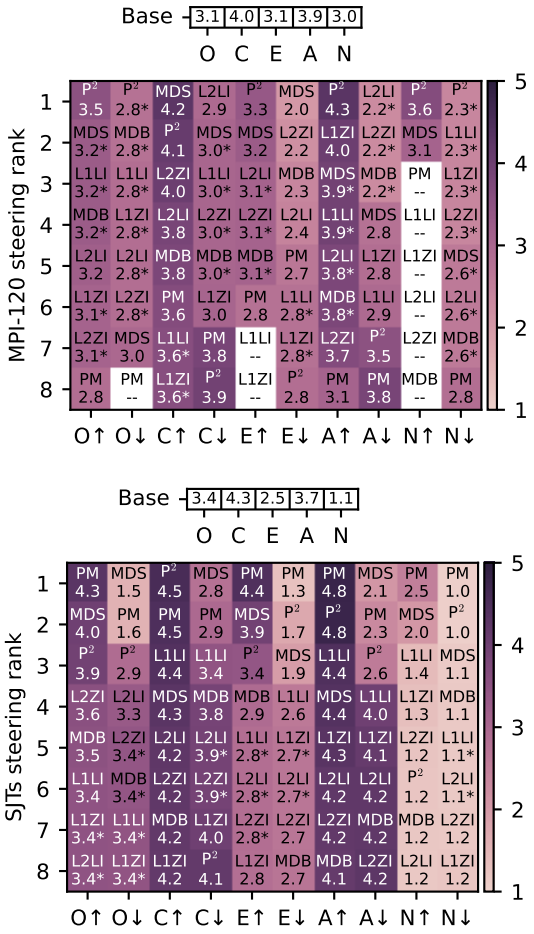


Figure 53: Ranking of steering methods on Olmo-3-7B-Instruct by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

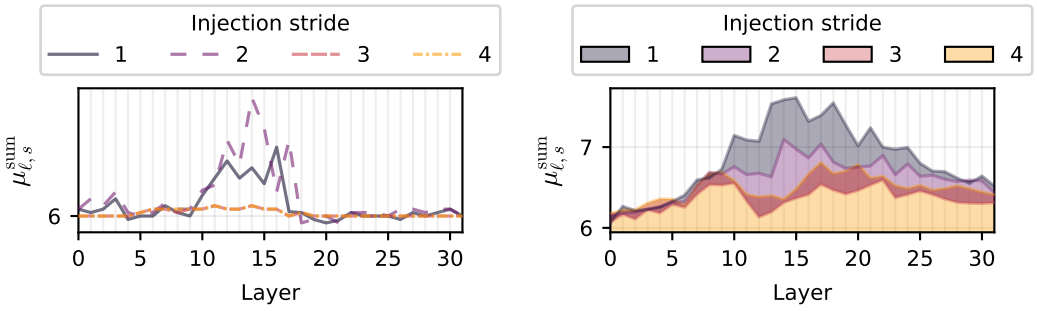


Figure 54: Overall MDS injections steering performance on Olmo-3-7B-Instruct by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

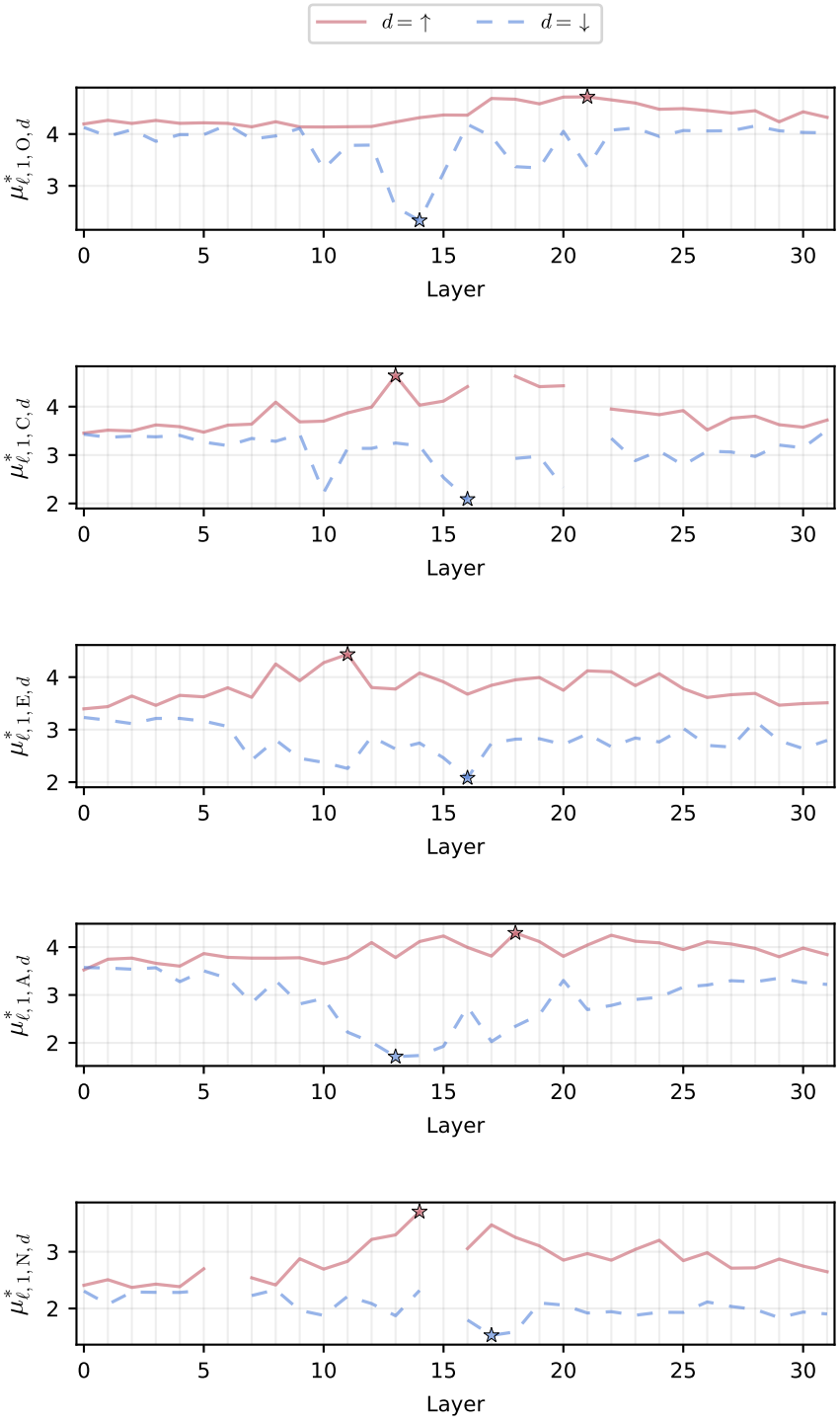


Figure 55: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Olmo-3-7B-Instruct. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

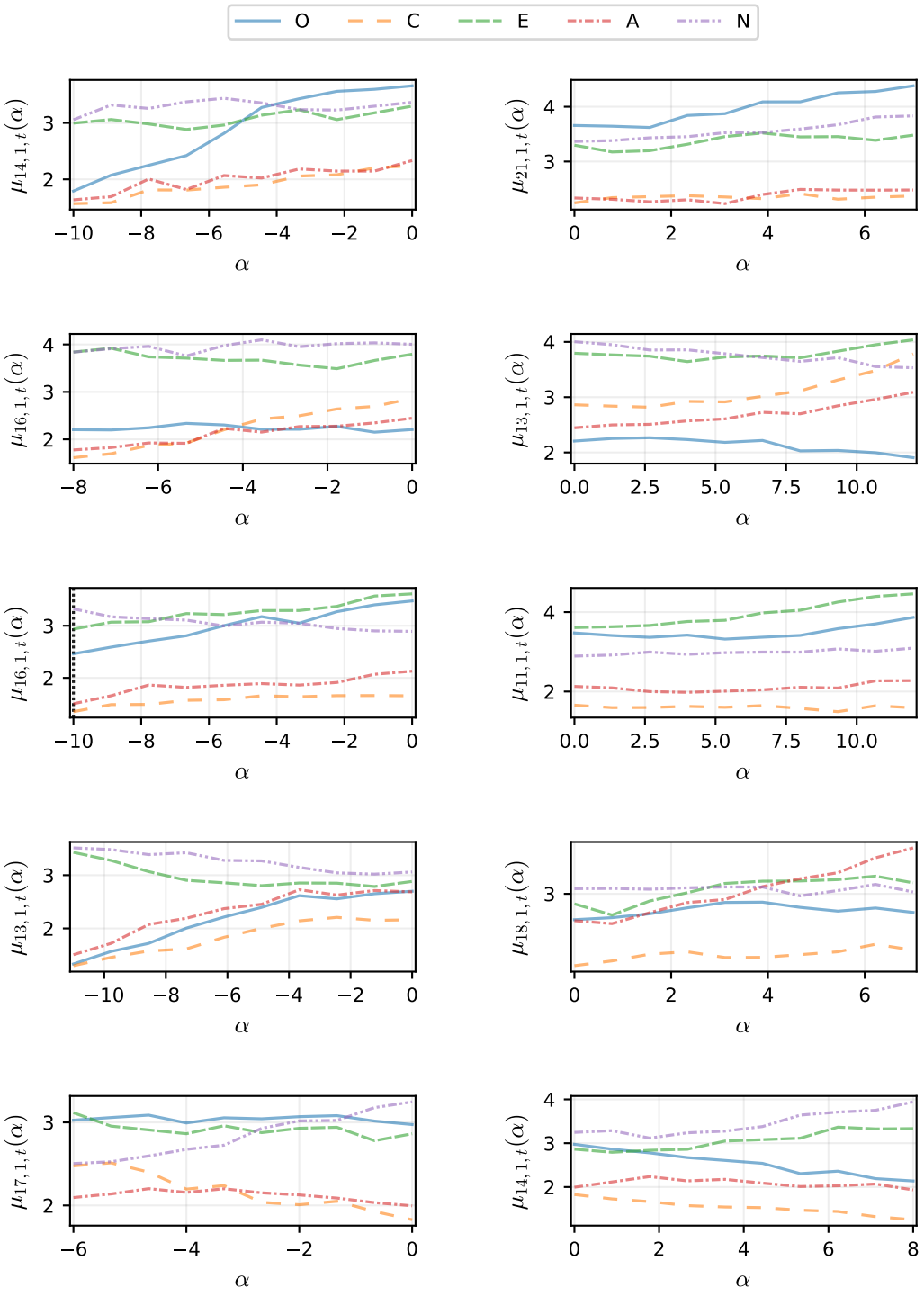


Figure 56: OCEAN scores for Olmo-3-7B-Instruct on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.

## Y OCEAN Injection Results for Olmo-3.1-32B-Instruct

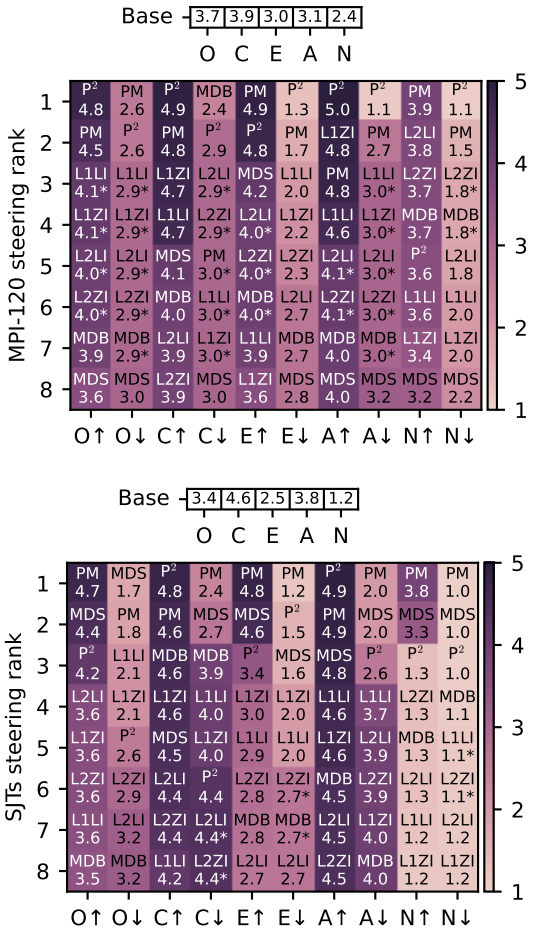


Figure 57: Ranking of steering methods on Olmo-3.1-32B-Instruct by OCEAN trait and direction, and task. Based on each method's best scores, with asterisks denoting ties in the unrounded results.

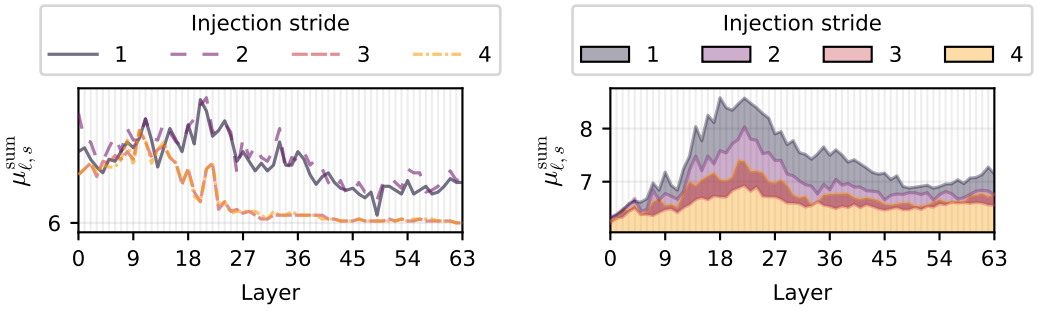


Figure 58: Overall MDS injections steering performance on Olmo-3.1-32B-Instruct by injection stride $s$ and model layer $\ell$. The line plot on the left shows MPI-120 results, and the shaded-area plot on the right shows SJT results.

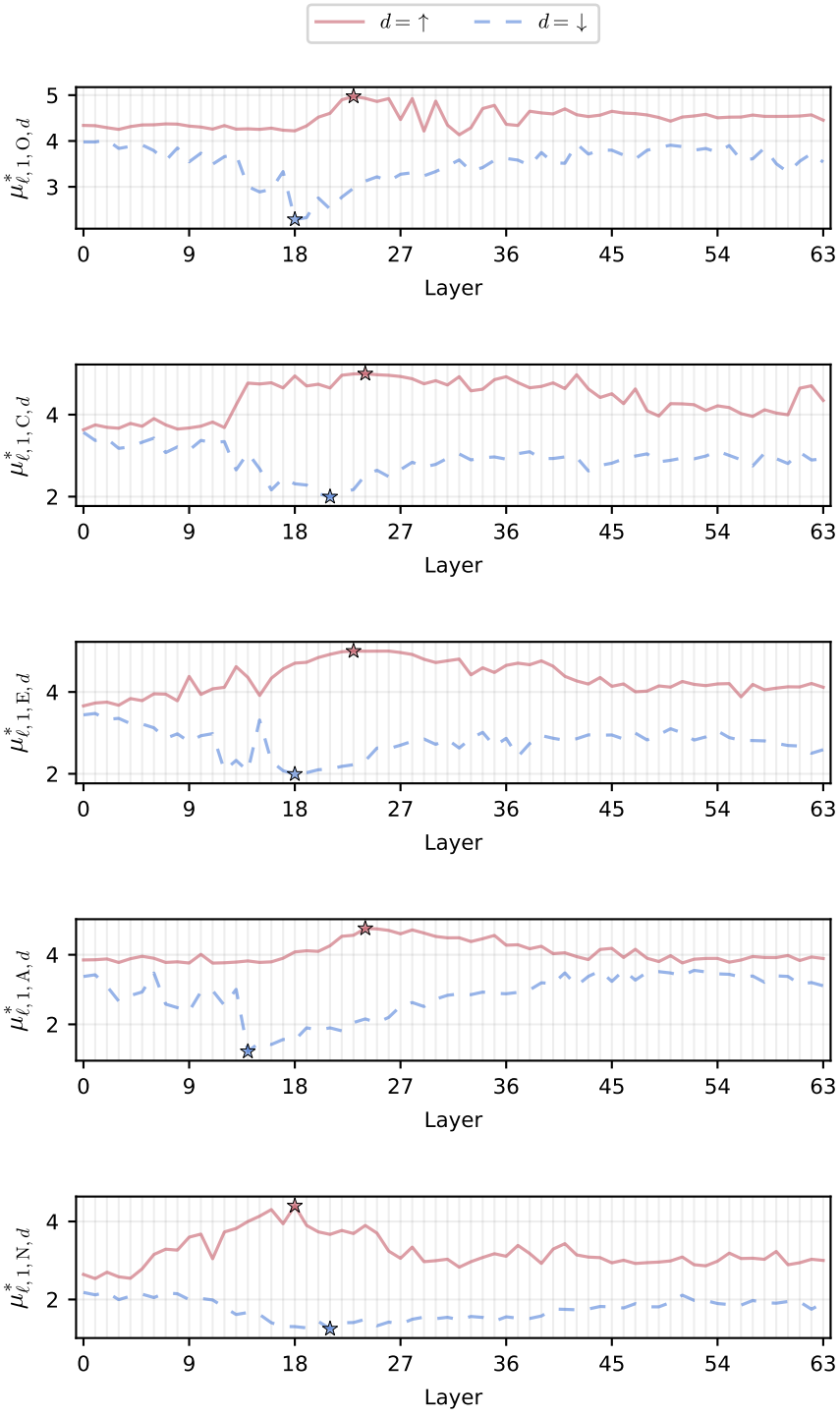


Figure 59: Layerwise extreme OCEAN steering scores on the SJTs task by direction $d \in \{\uparrow, \downarrow\}$ and model layer $\ell$, after applying MDS injections with injection stride $s = 1$ on Olmo-3.1-32B-Instruct. Stars mark the strongest steering effects across layers ($\phi_{1,t,d}$).

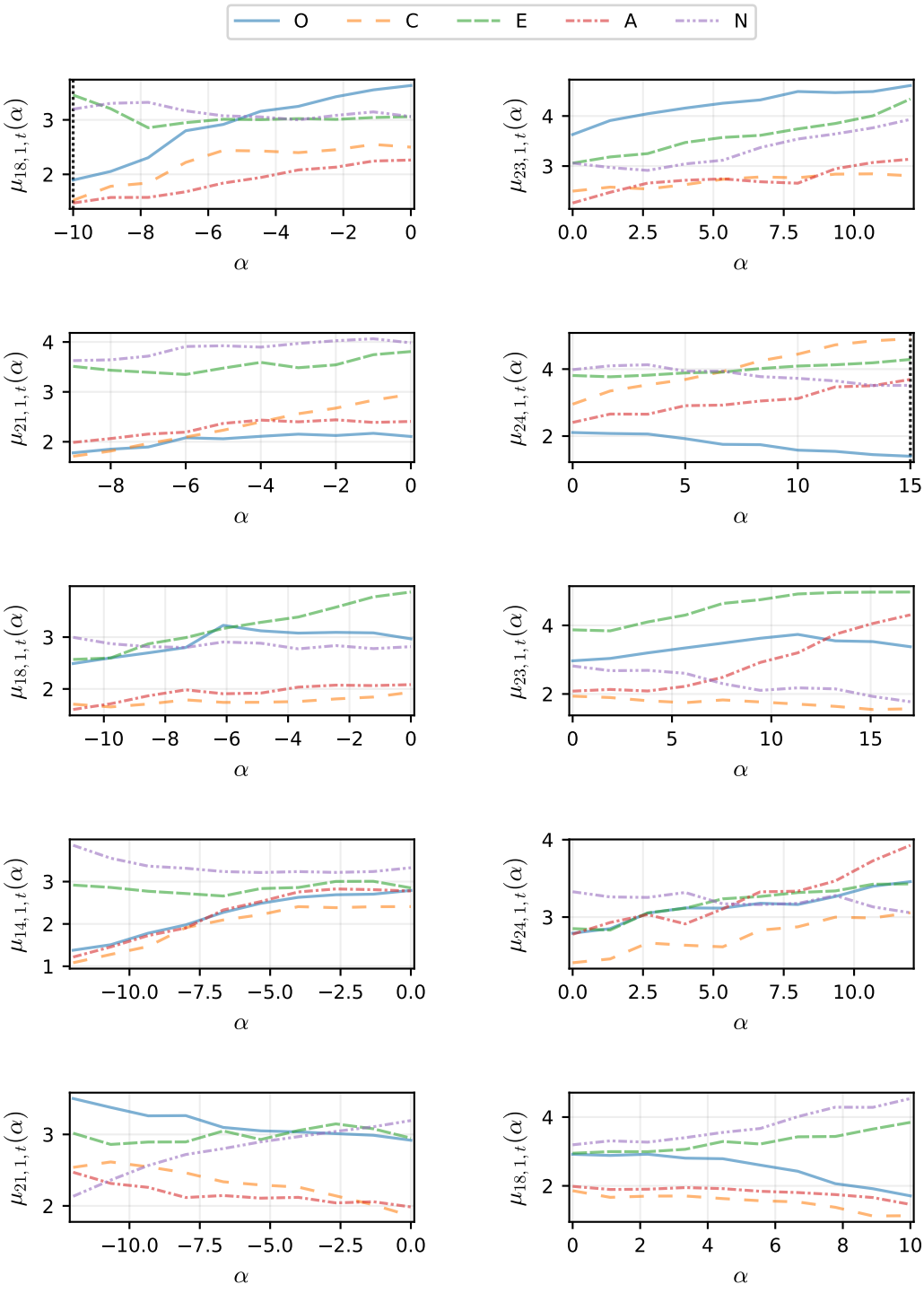


Figure 60: OCEAN scores for Olmo-3.1-32B-Instruct on SJTs, under MDS injections with $s = 1$, using the best-performing layer $\ell$ for each trait-direction pair and 10 equidistant $\alpha$ values from 0 (no steering) to the best-performing $\alpha$. From top to bottom, rows show openness, conscientiousness, extraversion, agreeableness, and neuroticism results. Negative $\alpha$ steers away from the target construct, and positive $\alpha$ steers toward it. Fluency was evaluated only in the responses to the corresponding SJTs. Vertical lines indicate some nonfluent SJT responses.