PROMPT-ADAPTED TRANSPORT CONTROL: CONTEXTUAL TOKENWISE LLM BEHAVIOUR STEERING VIA FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many activation-steering methods compress a behavioural contrast into a fixed global direction, ignoring how the desired transformation varies across contextual token states. We introduce Prompt-Adapted Transport Control (PATC), which learns flows at two complementary resolutions. A token-level flow transports contextual prompt states toward the target-behaviour distribution; PATC selects, orients, and aggregates behaviour displacements aligned with the training behavior distribution, retains the components that best characterize how the behavior changes, and removes overlap with the source-activation subspace to obtain a prompt-adapted unit direction. Because local token transports vary in norm with context, a separate flow trained on mean-pooled response activations supplies a common magnitude from the terminal prompt state. Their product forms a fixed cache applied to all tokens throughout generation without further flow evaluation. We evaluate anger and anxiety at approximately matched achieved strength, isolating intervention quality from steering intensity; these dimensions also permit comparison with held-out COVID-Worry human reference profiles. Across evaluation, PATC consistently balances affect, lexical quality, task compliance, and human-profile fidelity.

1 INTRODUCTION

Large language models can express behavioural attributes such as affective stance, caution, skepticism, confidence, and interpersonal style. Activation steering controls such attributes by modifying hidden states at inference time, offering a lightweight alternative to finetuning or repeated prompt engineering (Turner et al., 2023; Panickssery et al., 2024). Existing methods range from global contrastive and persona directions (Chen et al., 2025) to nonlinear, activation-dependent updates and learned transport maps (Zhao et al., 2026; Rodríguez et al., 2025). Adaptive methods commonly operate at one of two resolutions: they collapse prompt evidence into a single query state to derive a prompt-wide update, or transport each evolving token state during generation.

We ask whether the behavioural direction should instead be estimated before the prompt's distributed information is collapsed as its contextual token states offer different prefix-conditioned views of the instructions, entities, and constraints. We introduce Prompt-Adapted Transport Control (PATC), which transports these states individually through a learned behavioural flow and only then aggregates their local displacements into one prompt-conditioned direction. PATC selects behaviour-aligned displacements, orients them toward the target, forms a weighted consensus, and projects it into behaviourally supported geometry. This token-wise estimation before aggregation retains distributed prompt evidence without requiring token-dependent transport during generation.

Local token transports may vary substantially in norm and need not provide a reliable response-level steering scale. PATC therefore separates direction from magnitude: token-level transport determines where to steer, while a second flow trained on mean-pooled response representations determines how far to steer. Their product forms a fixed prompt-conditioned cache, computed once during prompt processing and reused throughout generation. Transport time provides a bounded control parameter for varying intervention strength. Figure 1 summarizes this two-resolution construction.

We evaluate PATC on anger and anxiety across Qwen3-14B and Gemma-3-12B, comparing it with unsteered generation, CAA, TRUTHFLOW, and ODESteer at approximately matched achieved affect.

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

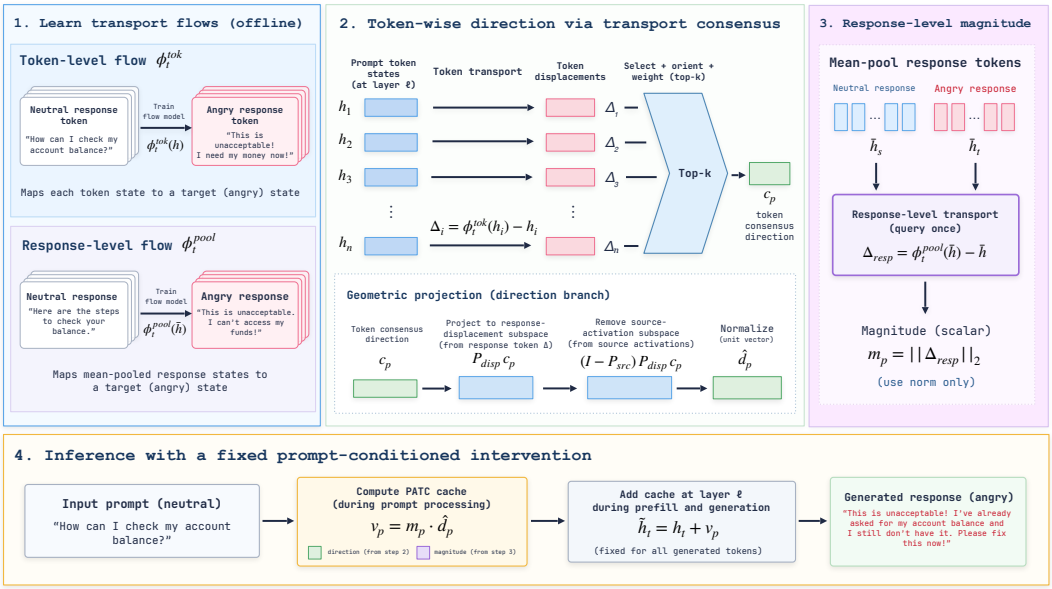


Figure 1: **Overview of Prompt-Adapted Transport Control.** **(1)** Offline, PATC learns token-level and response-level flows between source- and target-behaviour activations. **(2)** During prompt processing, the token-level flow transports each contextual prompt state. PATC selects strongly target-aligned displacements, orients and aggregates them into a consensus, retains the component supported by response displacements, and removes overlap with the source-activation subspace to obtain a unit steering direction. **(3)** The response-level flow is queried once at the terminal prompt state; the norm of its predicted displacement supplies the intervention magnitude. **(4)** Their product forms a fixed prompt-conditioned cache that is added at the intervention layer to all but the final prefill token and to every autoregressive generation step.

Across behavioural and strict task suites, PATC generally maintains greater lexical diversity with less repetition, competitively preserves complete task-and-format compliance, and more closely reproduces a multidimensional human affect profile than CAA and TruthFlow. ODESteer preserves strict-task performance better in some Gemma settings, but this advantage does not extend to Qwen, while PATC remains effective across both models and targets. Ablations support both token-wise prompt adaptation and the separation of token-level direction from response-level magnitude.

Our contributions are:

1. **Token-wise prompt adaptation:** PATC estimates behavioural transport from multiple contextual prompt states before aggregating them into a single prompt-conditioned direction.
2. **Resolution-factorized steering:** token-level transport determines direction, while pooled response transport determines magnitude; their product is computed once and cached throughout generation.
3. **Matched-strength evaluation:** we compare PATC with established steering methods across two models and two affective targets while controlling for achieved behavioural intensity.

2 Related Work

Activation steering and adaptive control. Activation Addition and CAA use contrastive hidden-state differences (Turner et al., 2023; Panickssery et al., 2024); mean-centering and persona vectors refine such global directions (Jorgensen et al., 2023; Chen et al., 2025). Adaptive methods instead use input semantics to scale selected coordinates, gate whether steering applies, or modulate intensity during generation (Wang et al., 2025b; Lee et al., 2025; Scalena et al., 2024). DSEM retrieves token- or chunk-specific vectors during decoding (Do et al., 2025). These approaches adapt where, whether,

or when an intervention is applied; PATC instead uses multiple contextual prompt states to construct one cached intervention before decoding.

Transport-based representation control. Flow matching learns continuous distributional transformations by regressing conditional-path velocities (Lipman et al., 2023). Activation-space methods include ACT shared maps, TRUTHFLOW query-dependent corrections, and ODESteer density-ratio ODEs (Rodríguez et al., 2025; Wang et al., 2025a; Zhao et al., 2026); FlowSteer targets concise reasoning, while UniSteer uses text-conditioned inversion and regeneration (Li et al., 2026; Shi et al., 2026). Our comparisons span the most relevant design axes: CAA represents a fixed global direction, while TRUTHFLOW is the closest primary transport comparator because it constructs a correction from pooled response representations. This distinguishes token-wise construction from the benefit of nonlinear flow matching itself. ODESteer provides the complementary activation-dependent case through its density-ratio ODE. PATC differs by aggregating multiple local prompt-state transports, estimating magnitude separately at response resolution, and caching the resulting controller.

Steering quality and task-sensitive evaluation. AxBench systematizes comparisons between concept detection and steering (Wu et al., 2025); free-form evaluations show that strong steering can induce repetition and factual degradation (Braun et al., 2026). FLASK decomposes evaluation into instance-specific skills (Ye et al., 2024). We therefore compare lexical quality and explicit task-and-format retention at matched achieved affect. Our FLASK-inspired suite uses transparent prompt-specific checks rather than the official protocol.

Behavioural targets. We choose state anger and anxiety because both are well represented in COVID-Worry for DeBERTa evaluation (Kleinberg et al., 2020; Mozes et al., 2021) and have established state-level definitions in Spielberger's STAXI-2 (anger) and STAI (anxiety) (Spielberger et al., 1983; Spielberger, 1999). Their distinct lexical, physiological, and fluency markers provide complementary tests of behavioural transport rather than one fixed style (Colombo et al., 2019; Goswamy et al., 2020) and can signal deteriorating interactions (Sandbank et al., 2018).

3 METHODOLOGY

PATC derives a prompt-conditioned direction from token-level transport, calibrates its magnitude with a pooled-response flow, and caches the resulting intervention once for reuse throughout prefill and generation.

3.1 PATC TRANSPORT MODELS

Token-level behavioral flow. For behavior conditions $y = 0$ (source) and $y = 1$ (target), let $x^y_{nr} \in \mathbb{R}^d$ be the layer-ℓ activation of response token r in training example n. Source and target response sequences are paired and truncated to their common length, after which their token states are coupled by position, yielding pairs (x^0_{nr}, x^1_{nr}). We fit a conditional token-level velocity field from source to target along the straight path

$$x_{nr,t} = (1-t)x^0_{nr} + tx^1_{nr}, \qquad \dot{x}_{nr,t} = x^1_{nr} - x^0_{nr}. \tag{1}$$

The training objective is

$$\mathcal{L}_{\text{tok}} = \mathbb{E}_{n,r,t}\left[\left\|\Phi^{\text{tok}}_\theta(x_{nr,t}, t) - (x^1_{nr} - x^0_{nr})\right\|_2^2\right]. \tag{2}$$

Separate global RMS scales normalize source and target activations; Φ^{tok}_τ thus includes normalization, integration, and decoding back to activation units.

Coupling assumptions. Positional pairing is a simple conditional coupling, not a claim that same-index tokens are semantically aligned; contrasting responses can therefore connect syntactically or lexically unrelated states. The resulting field is empirically effective across the two models, affects, and evaluation regimes studied here, but we claim neither optimality nor invariance to the coupling.

Pooled-response flow. The second PATC component is trained at response level. Each response is reduced to one response-only mean-pooled activation,

$$\bar{x}^y_n = \frac{1}{L^y_n}\sum_{r=1}^{L^y_n} x^y_{nr} \tag{3}$$

where prompt tokens are excluded. A pooled velocity field $\Phi_{\psi}^{\text{pool}}$ then transports $\bar{x}^0$ to $\bar{x}^1$:

$$\mathcal{L}_{\text{pool}} = \mathbb{E}_{(\bar{x}^0,\bar{x}^1),t}\left[\left\|\Phi_{\psi}^{\text{pool}}(\bar{x}_t,t) - (\bar{x}^1 - \bar{x}^0)\right\|_2^2\right], \quad \bar{x}_t = (1-t)\bar{x}^0 + t\bar{x}^1. \tag{4}$$

It uses the same RMS-normalization convention and its decoded transport map is $\Phi_{\tau}^{\text{pool}}$. Together, the token and response-level flows define PATC at two resolutions, where the response-level model is learned directly in the space of pooled response representations.

3.2 Prompt-adapted direction

Gathering consensus prompt tokens. For contextual prompt states $h_1, \ldots, h_M$, we transport each eligible state with $\Phi_{\theta}^{\text{tok}}$. The state includes the complete user-message chat prefix but excludes the assistant header. At endpoint τ_{tok}, candidate i yields a displacement

$$\delta_i = \Phi_{\tau_{\text{tok}}}^{\text{tok}}(h_i) - h_i. \tag{5}$$

Let b then be the empirical mean training displacement and $\widehat{b} = b/\|b\|_2$. We score signed alignment and retain the k largest magnitudes:

$$a_i = \left\langle \frac{\delta_i}{\|\delta_i\|_2}, \widehat{b} \right\rangle. \tag{6}$$

$$K = \text{TopK}_k\{\, |a_i| : i \text{ eligible} \,\}. \tag{7}$$

We correct the sign of locally reversed transports before consensus. For $i \in K$, define

$$o_i = \begin{cases} +1, & a_i \geq 0, \\ -1, & a_i < 0, \end{cases} \qquad w_i = \max(|a_i|, 10^{-4}), \qquad q_i = o_i \frac{\delta_i}{\|\delta_i\|_2}. \tag{8}$$

Writing $\text{normalize}(r) = r/\|r\|_2$ for any nonzero vector r, the weighted unit-direction consensus is

$$c_P = \text{normalize}\left(\sum_{i \in K} w_i q_i\right). \tag{9}$$

Displacement projection and source nulling. The current cache uses displacement-source-null (DSN) geometry. Let P_{disp} be the orthogonal projector onto the leading PCA subspace of response displacements, and let P_{src} be the projector onto the leading PCA subspace of source activations. We retain the component of the consensus supported by response displacement and remove the component supported by the source-activation subspace:

$$\widetilde{d}_P = (I - P_{\text{src}})P_{\text{disp}}c_P, \qquad \widehat{d}_{\text{PATC}}(P) = \frac{\widetilde{d}_P}{\|\widetilde{d}_P\|_2}. \tag{10}$$

Like TruthFlow, this anchors steering to empirical behavioural change, but forms the direction before token-level evidence is compressed. Next, the response-level flow supplies its scale.

3.3 Response-level magnitude

The pooled-response flow is queried once using the same prompt's terminal assistant-boundary state h_M. Its predicted endpoint displacement is

$$\Delta_{\text{pool}}(P) = \Phi_{\tau_{\text{pool}}}^{\text{pool}}(h_M) - h_M. \tag{11}$$

Only its Euclidean norm is retained:

$$m_{\text{pool}}(P) = \|\Delta_{\text{pool}}(P)\|_2. \tag{12}$$

This norm calibrates cache strength while its orientation is discarded. The complete intervention is

$$\begin{aligned} v_P &= g_{\text{pool}}\, m_{\text{pool}}(P)\widehat{d}_{\text{PATC}}(P) \\ &= g_{\text{pool}} \left\|\Phi_{\tau_{\text{pool}}}^{\text{pool}}(h_M) - h_M\right\|_2 \text{normalize}\left[(I - P_{\text{src}})P_{\text{disp}}\, \text{normalize}\left(\sum_{i \in K} w_i q_i\right)\right]. \end{aligned} \tag{13}$$

Here, $g_{\text{pool}} > 0$ is a fixed model–behaviour calibration gain. Although trained on pooled response states, the flow is queried at h_M as a prompt-side magnitude estimator. Thus each prompt determines both parts of its cache: token-wise consensus sets where to move, and response-level transport sets how far to move.

3.4 CACHING AND GENERATION

PATC computes v_P once from the unsteered prompt and applies it as a fixed additive intervention at layer ℓ. In the reported experiments, it is added to all prompt positions except the final prefill token, which remains clean, and to every decoding step:

$$h_n^{(\ell)} \leftarrow h_n^{(\ell)} + v_P, \qquad n \neq M \text{ during prefill.} \tag{14}$$

Leaving h_M untouched avoids directly perturbing the state used to produce the first response-token logits, while earlier prompt states and every decoding step remain steered; Appendix C.2 reports the corresponding clean-final schedule control. After the cache is constructed, generation requires no further flow evaluations. We summarize PATC in Algorithm 1 in Appendix A.

4 EXPERIMENTS

We organize the evaluation into *two* experiments. First, we measure *intervention quality at matched affect*: behavioural prompts test lexical diversity and repetition, while strict prompts test task-and-format retention at the same operating points. These are complementary parts of one question: how achieved behaviour, expressive quality, and task preservation trade off. Second, we test whether PATC reproduces multidimensional human affect profiles at matched achieved strength. We additionally report construction ablations. We evaluate both Qwen3-14B and Gemma-3-12B throughout: Experiment 1 covers anger and anxiety, and Experiment 2 evaluates anger-profile matching. PATC and the comparison methods intervene at transformer layer 18 for Qwen and layer 23 for Gemma; every within-model comparison uses the same model and decoding configuration. Full model, flow-training, projection, and decoding hyperparameters are reported in Appendix B.

Evaluation suites. Experiment 1 combines a 25-prompt behavioural suite of minimally scaffolded opinion, advice, narrative, planning, and reflection tasks with a frozen 50-prompt FLASK-derived strict suite containing checkable content and output constraints (Ye et al., 2024). Experiment 2 uses a separate COVID-Worry corpus and participant-held-out human-text reference described below. Exact prompts are reproduced in Appendix G. Primary comparisons with an unsteered baseline, Contrastive Activation Addition (CAA; persona vectors) (Panickssery et al., 2024), TRUTHFLOW (Wang et al., 2025a), and ODESteer (Zhao et al., 2026) use identical generation settings. We report achieved strength alongside quality at the frozen, approximately matched operating points.

Preliminary: Ensuring Comparability via Strength Matching On participant-held-out COVID-Worry data, predictions from the frozen five-fold DeBERTaV3 CORAL ensemble (He et al., 2021; Cao et al., 2020) increase monotonically with human-reported anger and anxiety (Appendix Fig. 4). We therefore use its target-affect coordinate for reproducible coarse ordering, identifying nearby comparison-method coefficients on a separate 10-prompt calibration set for each PATC endpoint $t \in [0, 1]$. Because the predictions are compressed toward the scale centre and small numerical gaps need not be perceptually exact, the same separate 10-prompt calibration set is used for blinded fine-grained adjustments among adjacent settings. Judgments are strength-only, compare same-prompt responses, and are made in a fresh, separate Codex session (GPT-5.6-sol; high reasoning effort). The resulting settings are frozen before evaluation on the prompt suites. Validation and judge details appear in Appendix D.

Experiment 1: Intervention Quality at Matched Strength Since state anger and anxiety are multifaceted rather than fixed lexical styles (Spielberger et al., 1983; Spielberger, 1999), stronger steering may narrow language toward target-consistent expression, but collapse onto a few salient cues or repeated phrases indicates poor intervention quality (Braun et al., 2026). We therefore interpret diversity, repetition, and task retention conditional on achieved affect. At each matched point, behavioural quality is measured by MATTR (50-token window) (Covington & McFall, 2010), HD-D (42-token draw), and MTLD (standard 0.72 threshold) (McCarthy & Jarvis, 2010), all higher-is-better, plus repeated-trigram and repeated-four-gram fractions, which are lower-is-better (Holtzman et al., 2020). Under the same intervention, the strict suite provides a FLASK-inspired deterministic preservation proxy, not the official FLASK LLM-judge protocol. Task score is the fraction of required content checks satisfied; format score is the pass rate over the 35 prompts with an applicable task-specific or answer-only format check; and full pass requires every content check and any applicable format check. Complete component metrics appear in Appendix C. Safety and toxicity are distinct

from affect fidelity: stronger anger may increase aggressive language, while weaker steering may merely appear safer. Dedicated safety evaluation therefore lies beyond the present scope.

Experiment 2: Human Affect-Profile Matching

Using the same two fixed COVID-Worry prompt phrasings for every model, method, and seed, we focus on anger because the unsteered model is already close to the anxiety reference (Appendix E). The anger reference retains every long-form response in the untouched participant-held-out COVID-Worry test partition whose author selected anger as the focal emotion and self-rated it at least 7. This fixed rule yields 27 participant-wave observations (8 in wave 1 and 19 in wave 2) from 26 people, one represented in both waves; the sample size was not chosen in advance. Long-form responses match the generated-text setting and avoid duplicate short and long views of the same rating. The same frozen DeBERTa ensemble scores human and generated texts on worry, anger, disgust, fear, anxiety, sadness, happiness, relaxation, and desire, the dataset's nine self-reported affect axes (Kleinberg et al., 2020; Mozes et al., 2021). RMSE uses the equally weighted mean of the two wave profiles, while Mahalanobis distance uses Ledoit–Wolf covariance over participant-averaged profiles.

Behavior-training corpora. Experiment 1 trains on 50 synthetic neutral–target pairs using prompts sampled from Serapio-García et al. (2023), following Hartley et al. (2025); the target scaffolds operationalize Spielberger's state anger and anxiety constructs (Spielberger et al., 1983; Spielberger, 1999). Experiment 2 instead trains on 50 neutral–angry pairs from COVID-Worry development data, reserving held-out participants for the human reference. This separation tests whether PATC can both learn behaviour from synthetic contrasts and reproduce a naturally occurring human affect profile. In both cases a 500-pair control changed little unless the prompt distribution changed.

5 Results

5.1 Preliminaries: Strength Matching

For both models and targets, we evaluate coupled token- and response-flow endpoints $t \in \{0.5, 0.8, 0.9, 1.0\}$. Figure 2 shows achieved-affect gaps at the frozen operating points. Table 1 reports the four-method values; primary calibration settings and blinded judgments appear in Appendix Tables 11 and 12. The frozen coefficients provide approximate rather than exact matching

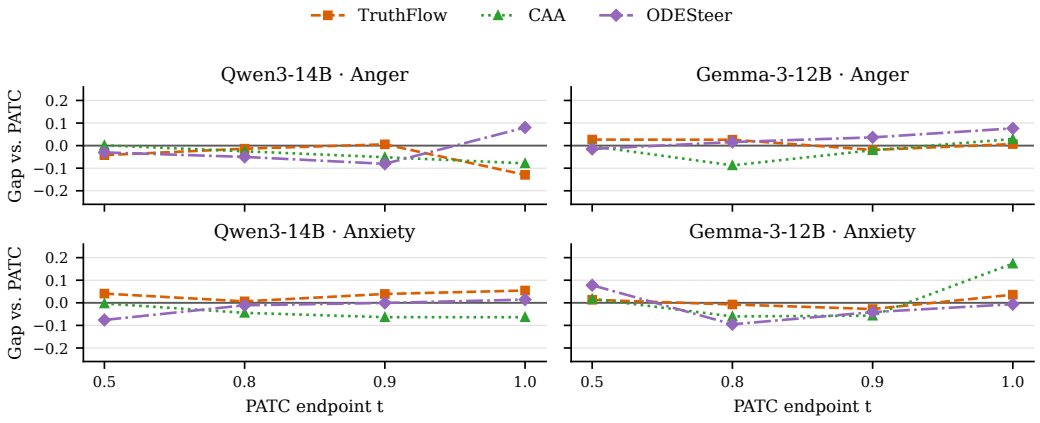


Figure 2: Achieved-affect gaps relative to PATC (25 prompts; three-seed means). Positive values indicate stronger measured affect. Operating points are listed in Table 1; calibration details appear in Appendix Tables 11–12.

on the regenerated samples. Median absolute achieved-affect gaps are 0.026, 0.048, and 0.039 for TruthFlow, CAA, and ODESteer, respectively. Gaps are at most 0.10 in 15/16 settings for each of the first two methods and all 16 for ODESteer; maxima are 0.129, 0.173, and 0.095. We report achieved strength in every row and interpret larger residual gaps cautiously.

Table 1: Intervention quality at frozen operating points. Behavioural columns show achieved affect, MTLD, and Rep.-3; strict columns show full pass on 50 prompts. Means use three seeds. Rep.-3 and full pass are percentages; full pass requires all content and applicable format checks. The lower panel macro-averages preservation across models and targets. Appendix C reports complete metrics with SDs. Bold marks the best mean and underlining marks the second-best per four-method block.

	Qwen3-14B					**Gemma-3-12B**				
		Behavioural suite			Strict suite		Behavioural suite			Strict suite
	Setting	Strength	MTLD ↑	Rep.-3 ↓	Full pass ↑	Setting	Strength	MTLD ↑	Rep.-3 ↓	Full pass ↑
Anger	$t_{\text{PATC}} = .5$	3.40	**104.37**	**.72**	**93.3**	$t_{\text{PATC}} = .5$	3.34	130.80	.29	97.3
	$\alpha_{\text{TF}} = 1.5$	3.36	84.49	.81	**93.3**	$\alpha_{\text{TF}} = 1$	3.37	134.77	**.18**	92.7
	$\alpha_{\text{CAA}} = 1.9$	3.40	54.68	3.69	**93.3**	$\alpha_{\text{CAA}} = .8$	3.34	118.39	.38	87.3
	$T_{\text{ODE}} = 85$	3.37	80.61	2.84	88.0	$T_{\text{ODE}} = 2200$	3.32	**144.32**	.27	**98.0**
	$t_{\text{PATC}} = .8$	3.76	**66.09**	**2.09**	**89.3**	$t_{\text{PATC}} = .8$	3.63	90.90	1.12	92.7
	$\alpha_{\text{TF}} = 2.6$	3.74	50.70	4.90	86.7	$\alpha_{\text{TF}} = 1.6$	3.66	85.72	**.50**	89.3
	$\alpha_{\text{CAA}} = 2.6$	3.73	25.28	18.48	86.0	$\alpha_{\text{CAA}} = 1.1$	3.55	86.21	.81	85.3
	$T_{\text{ODE}} = 100$	3.71	44.97	16.63	77.0	$T_{\text{ODE}} = 3000$	3.65	**105.08**	.64	**96.0**
	$t_{\text{PATC}} = .9$	3.98	**54.02**	**2.50**	**86.0**	$t_{\text{PATC}} = .9$	3.81	**81.21**	2.67	89.3
	$\alpha_{\text{TF}} = 3$	3.98	47.63	5.69	77.3	$\alpha_{\text{TF}} = 1.7$	3.79	78.65	**.61**	86.0
	$\alpha_{\text{CAA}} = 2.75$	3.93	17.78	29.16	80.7	$\alpha_{\text{CAA}} = 1.25$	3.79	78.44	.88	84.0
	$T_{\text{ODE}} = 110$	3.90	25.00	37.18	60.0	$T_{\text{ODE}} = 3450$	3.85	76.84	1.26	**93.0**
	$t_{\text{PATC}} = 1$	4.22	**51.38**	**3.83**	**84.7**	$t_{\text{PATC}} = 1$	4.03	**69.10**	5.27	82.7
	$\alpha_{\text{TF}} = 3.25$	4.09	40.77	9.01	74.7	$\alpha_{\text{TF}} = 2$	4.04	55.92	**1.06**	86.7
	$\alpha_{\text{CAA}} = 3$	4.14	13.50	45.43	74.7	$\alpha_{\text{CAA}} = 1.45$	4.06	56.60	1.87	86.7
	$T_{\text{ODE}} = 180$	4.30	2.89	16.17	5.0	$T_{\text{ODE}} = 4000$	4.11	36.97	8.73	**90.0**
Anxiety	$t_{\text{PATC}} = .5$	4.58	96.17	.49	**93.3**	$t_{\text{PATC}} = .5$	4.60	140.29	.20	92.0
	$\alpha_{\text{TF}} = 1.25$	4.62	73.67	1.19	**93.3**	$\alpha_{\text{TF}} = .7$	4.61	131.10	.31	92.0
	$\alpha_{\text{CAA}} = 1.45$	4.57	64.76	1.88	**93.3**	$\alpha_{\text{CAA}} = .8$	4.62	108.21	.44	87.3
	$T_{\text{ODE}} = 20$	4.50	**109.95**	**.40**	91.0	$T_{\text{ODE}} = 2200$	4.68	**143.53**	**.14**	**95.0**
	$t_{\text{PATC}} = .8$	4.71	**67.53**	**1.34**	92.7	$t_{\text{PATC}} = .8$	4.91	101.90	**.37**	90.0
	$\alpha_{\text{TF}} = 1.5$	4.72	60.36	1.94	**94.7**	$\alpha_{\text{TF}} = 1.2$	4.91	85.29	.43	91.3
	$\alpha_{\text{CAA}} = 1.77$	4.67	46.96	3.57	93.3	$\alpha_{\text{CAA}} = 1.15$	4.85	71.13	.98	86.0
	$T_{\text{ODE}} = 60$	4.70	41.68	4.09	87.0	$T_{\text{ODE}} = 2800$	4.82	**123.21**	.43	**94.0**
	$t_{\text{PATC}} = .9$	4.75	**56.49**	**1.75**	**93.3**	$t_{\text{PATC}} = .9$	4.99	84.38	**.46**	86.0
	$\alpha_{\text{TF}} = 1.8$	4.79	49.22	3.24	**93.3**	$\alpha_{\text{TF}} = 1.3$	4.96	74.33	.67	86.7
	$\alpha_{\text{CAA}} = 1.85$	4.69	44.60	4.01	**93.3**	$\alpha_{\text{CAA}} = 1.2$	4.93	63.21	1.15	87.3
	$T_{\text{ODE}} = 65$	4.75	39.41	6.64	87.3	$T_{\text{ODE}} = 3250$	4.95	**100.80**	.85	**93.0**
	$t_{\text{PATC}} = 1$	4.83	**46.09**	**2.58**	**93.3**	$t_{\text{PATC}} = 1$	5.09	74.72	**.72**	86.0
	$\alpha_{\text{TF}} = 2$	4.89	39.24	4.98	90.7	$\alpha_{\text{TF}} = 1.5$	5.12	56.42	1.03	83.3
	$\alpha_{\text{CAA}} = 2.4$	4.77	22.46	13.64	92.7	$\alpha_{\text{CAA}} = 1.4$	5.26	45.66	2.31	86.7
	$T_{\text{ODE}} = 72$	4.85	26.96	14.17	83.3	$T_{\text{ODE}} = 3500$	5.08	**90.09**	2.04	**94.0**

	PATC			**TRUTHFLOW**			**CAA**			**ODESteer**		
t_{PATC}	MTLD ↑	Rep.-3 ↓	Full pass ↑	MTLD ↑	Rep.-3 ↓	Full pass ↑	MTLD ↑	Rep.-3 ↓	Full pass ↑	MTLD ↑	Rep.-3 ↓	Full pass ↑
.5	117.91	**.42**	**94.0**	106.01	.62	92.8	86.51	1.60	90.3	**119.60**	.91	93.0
.8	**81.61**	**1.23**	**91.2**	70.52	1.94	90.5	57.39	5.96	87.7	78.74	5.45	88.5
.9	**69.03**	**1.84**	**88.7**	62.46	2.55	85.8	51.01	8.80	86.3	60.51	11.48	83.3
1.0	**60.32**	**3.10**	**86.7**	48.09	4.02	83.9	34.55	15.81	85.2	39.23	10.28	68.1
All	**82.22**	**1.65**	**90.1**	71.77	2.28	88.3	57.37	8.04	87.4	74.52	7.03	83.2

5.2 EXPERIMENT 1: INTERVENTION QUALITY AT MATCHED STRENGTH

Table 1 summarizes lexical quality and strict full-pass retention at the frozen operating points; complete results appear in Appendix C, with selected verbatim generations in Appendix F. PATC has higher MTLD in 15/16 settings versus TRUTHFLOW and 16/16 versus CAA, with lower trigram repetition in 12/16 and 13/16, respectively. Full pass is higher or tied in 12/16 and 13/16 comparisons. CAA's strong strict scores therefore coexist with narrower, more repetitive expression at higher strengths, particularly on Qwen. These main comparisons show that PATC maintains a consistent cross-model balance of affect, lexical quality, and task retention. ODESteer preserves strict tasks

Table 2: COVID-Worry anger-profile matching across ten seeds (mean with gray SD; lower distances are better). ODESteer rows give its weak and closest stronger points; the summary strip averages stronger points across models. Bold and underlining mark best and second-best means.

	Qwen3-14B					Gemma-3-12B				
Method	Setting	Target	9-axis	Non-target	Mahalanobis	Setting	Target	9-axis	Non-target	Mahalanobis
Baseline	—	3.646 ±0.158	0.951 ±0.082	0.729 ±0.068	2.876 ±0.184	—	3.648 ±0.132	0.997 ±0.089	0.795 ±0.085	2.664 ±0.151
PATC	$t_{\text{PATC}} = 0.5$	4.408 ±0.185	0.926 ±0.126	0.878 ±0.155	2.888 ±0.263	$t_{\text{PATC}} = 0.5$	4.212 ±0.227	0.710 ±0.096	<u>0.565</u> ±0.066	2.273 ±0.230
	$t_{\text{PATC}} = 1.0$	5.152 ±0.134	<u>0.549</u> ±0.128	<u>0.557</u> ±0.136	<u>1.670</u> ±0.262	$t_{\text{PATC}} = 1.0$	4.704 ±0.366	**0.565** ±0.142	**0.495** ±0.127	**1.949** ±0.291
TRUTHFLOW	$\alpha_{\text{TF}} = 1.5$	4.451 ±0.324	0.830 ±0.076	0.765 ±0.115	2.670 ±0.141	$\alpha_{\text{TF}} = 1.0$	4.008 ±0.165	0.833 ±0.110	0.675 ±0.110	2.426 ±0.249
	$\alpha_{\text{TF}} = 3.5$	5.222 ±0.159	0.881 ±0.188	0.921 ±0.206	2.250 ±0.298	$\alpha_{\text{TF}} = 2.3$	4.693 ±0.147	<u>0.630</u> ±0.078	0.579 ±0.097	<u>2.069</u> ±0.209
CAA	$\alpha_{\text{CAA}} = 0.95$	4.471 ±0.356	0.805 ±0.046	0.738 ±0.094	2.652 ±0.138	$\alpha_{\text{CAA}} = 0.50$	4.304 ±0.302	0.842 ±0.103	0.747 ±0.159	2.567 ±0.257
	$\alpha_{\text{CAA}} = 1.90$	5.189 ±0.217	0.683 ±0.110	0.704 ±0.123	2.040 ±0.257	$\alpha_{\text{CAA}} = 0.95$	4.727 ±0.259	0.720 ±0.118	0.686 ±0.145	2.219 ±0.289
ODESteer	$T_{\text{ODE}} = 85$	4.613 ±0.180	0.852 ±0.073	0.827 ±0.092	2.562 ±0.179	$T_{\text{ODE}} = 2200$	4.343 ±0.198	0.722 ±0.098	0.614 ±0.111	2.345 ±0.283
	$T_{\text{ODE}} = 105$	4.864 ±0.205	**0.460** ±0.146	**0.399** ±0.164	**1.605** ±0.347	$T_{\text{ODE}} = 3450$	4.617 ±0.255	0.767 ±0.103	0.724 ±0.135	2.341 ±0.261
Cross-model mean at the stronger point (9-axis / non-target / Mahalanobis)										
PATC: **0.557 / 0.526 / 1.810** | TRUTHFLOW: 0.756 / 0.750 / 2.160 | CAA: 0.702 / 0.695 / 2.130 | ODESteer: <u>0.614 / 0.562 / 1.973</u>										

particularly well on Gemma but declines more steeply on Qwen; PATC leads cross-model MTLD, repetition, and full-pass averages. This variation may reflect dependence of its density-ratio field on model- and affect-specific activation geometry, although we do not isolate the cause (Appendix C.1). Together, the matched-strength results show that PATC's preservation is not explained by weaker steering and that its affect–quality–task balance persists across both models and targets. Applying the same clean-final schedule to TRUTHFLOW (omitting its intervention only at the final prefill position) reaches the same conclusion (Appendix C.2).

5.3 EXPERIMENT 2: STEERING AN UNDEREXPRESSED HUMAN AFFECT COMPONENT

Unsteered COVID-Worry responses underexpress anger: 3.646 ± 0.158 for Qwen and 3.648 ± 0.132 for Gemma, versus the held-out human mean of 5.619. Table 2 reports complete-profile distances. On Qwen, ODESteer gives the closest complete-profile match, with nine-axis, non-target, and Mahalanobis distances of $0.460/0.399/1.605$, versus $0.549/0.557/1.670$ for PATC. On Gemma the ordering reverses: PATC records $0.565/0.495/1.949$, below ODESteer's $0.767/0.724/2.341$ and the other baselines. Macro-averaging the stronger comparison point across models, PATC has the lowest distances overall ($0.557/0.526/1.810$), followed by ODESteer ($0.614/0.562/1.973$). ODESteer's achieved anger and profile distances are non-monotonic across its horizon sweep on both models (Appendix C.1), complicating the selection of a matched operating point. Thus PATC moves the broader affect configuration toward the human reference consistently across both models, rather than merely increasing anger.

5.4 ABLATING TOKEN-WISE DIRECTION CONSTRUCTION

At Qwen3-14B anger $t_{\text{PATC}} = 1.0$, we isolate three choices in PATC's direction construction while retaining its token flow, DSN, response-flow magnitude, schedule, and decoding. All-token tests selection by aggregating every eligible transport; terminal-token tests multi-token consensus using only the final prompt state; and global-mean tests prompt adaptation by substituting a fixed mean training-token displacement. Pooled-response flow instead changes both the training representation and inference query, so it compares transporting the entire pooled representation rather than single tokens. Top-32 gives the strongest overall balance across the tested selection sizes (Appendix Table 10). Against all-token aggregation, it raises MTLD by 2.13, lowers repeated trigrams by 0.86 points, and raises full pass by 1.4 points at similar anger. Terminal-token improves MTLD but reduces full pass by 5.4 points; the global mean retains full pass but shifts MTLD toward TRUTHFLOW (45.49 versus 40.77) and increases repetition. Pooled flow has weaker anger and the lowest full pass, so its higher MTLD is not comparable preservation. Finally, replacing the pooled-response magnitude with selected-token norms lowers matched-anger full pass by 12.0 points (Appendix C.3).

Table 3: Qwen3-14B anger ablations at $t_{\text{PATC}} = 1.0$ (three-seed means; 25 behavioural and 50 strict prompts). "Token" and "Flow" mark token-state construction and inference-time flow queries. Applied magnitudes are matched.

Direction construction	Token	Flow	Anger	MTLD ↑	Rep.-3 ↓	Full pass ↑
PATC, top-32	✓	✓	4.221 ±.035	51.38 ±.19	3.83 ±.55	84.7 ±1.2
All prompt tokens	✓	✓	4.167 ±.120	49.25 ±1.03	4.69 ±1.69	83.3 ±1.2
Terminal prompt token	✓	✓	4.375 ±.083	55.52 ±1.04	1.60 ±.59	79.3 ±3.1
Global mean displacement	✓	–	4.231 ±.020	45.49 ±2.40	4.75 ±.64	85.3 ±1.2
Pooled-response flow	–	✓	3.575 ±.063	64.61 ±7.69	1.78 ±.97	68.7 ±6.4

6 Discussion

Limitations. Our evidence is limited to two models, two affective targets, and targeted prompt suites, so it does not establish generalization across behaviours or comprehensive capability preservation. PATC depends on its intervention layer, token count, flow endpoint, and projection rank, while its fixed cache cannot react to oversteering during generation. It controls how a target behaviour is expressed, not whether that target is appropriate or safe; increasing anger may also increase toxic language, requiring separate target-selection and output- safety mechanisms for deployment. Training uses positional source–target token coupling, and alternative semantic or optimal-transport couplings remain untested. Finally, strength matching and behavioural conclusions still depend on automated evaluators despite held-out human anchoring and blinded checks, and merit larger-scale human validation. Within this scope, replication across two architectures and targets, four matched-strength points, behavioural, strict-task, and human-profile evaluations, and targeted ablations provides sufficient support for our narrower claim: token-wise direction construction benefits cached affect steering under the evaluated assumptions.

Conclusion At approximately matched achieved affect, PATC preserves greater lexical diversity in 15/16 comparisons with TruthFlow and all 16 comparisons with CAA, while matching or exceeding their complete task-and-format retention in 12/16 and 13/16 settings, respectively. These gains persist across two models, both anger and anxiety, and the evaluated strength range; they therefore make weaker steering an unlikely sole explanation. PATC also moves an underexpressed anger response toward the complete held-out human affect profile on both models, rather than merely increasing the target coordinate. Finally, the ablations show that the result is not reproduced by a pooled-response flow, a fixed global token displacement, or a single transported prompt state: selecting and combining multiple contextual token transports, then calibrating their magnitude at response resolution, produces the strongest overall balance. The central finding is that estimating behaviour from multiple prompt-token transports before aggregation can improve behavioural fidelity without giving up the efficiency of one cached intervention. The model-dependent comparisons furthermore suggest that consensus, projection, and separate magnitude estimation may reduce sensitivity to model- and affect-specific errors in activation-dependent density-ratio fields. Future work should test this hypothesis alongside semantic or optimal-transport token coupling.

Ethics Statement

The authors acknowledge and adhere to the ICLR Code of Ethics. This work studies inference-time control of affect, which could be misused to produce manipulative, hostile, or distressing language. Our experiments characterize steering fidelity and preservation, not the appropriateness of deploying a target affect; practical use requires separate target-selection and output- safety safeguards. We use an existing COVID-Worry dataset only for aggregate affect modelling and evaluation, preserve

participant-disjoint splits, and report aggregate results; we recruited no new participants and made no attempt to identify individuals. Synthetic corpora and evaluated generations may contain aggressive or anxious language and should be handled accordingly.

AI Use Disclosure

Generative AI tools were used to generate the synthetic contrastive response corpora and the model outputs studied in the experiments; to propose and refine hypotheses, methodology, and experimental designs; and to assist with method implementation, debugging, evaluation scripts, data analysis, and result interpretation. OpenAI Codex was also used to assist with drafting, editing, and formatting the manuscript. As described in the experimental protocol, a fresh, separate Codex session using GPT-5.6-sol at high reasoning effort served as the blinded, strength-only pairwise judge. Generative AI was not used to prove formal mathematical claims, translate the manuscript, or make final research decisions without author review. The authors reviewed all AI-assisted text and analyses, executed and checked AI-assisted code, inspected experimental outputs, and cross-checked reported values against the stored artifacts. The authors take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with generative-AI assistance.

Reproducibility Statement

The main text specifies the training objective, controller construction, evaluation protocol, and frozen strength-matching procedure. Appendix Sections A and B provide the complete inference algorithm and implementation, training, projection, and generation settings. Appendix C reports the complete component results, Appendix D documents the affect evaluator and calibration audit, and Appendix G reproduces the evaluation prompts. We report the evaluated seeds, operating points, achieved affect, and uncertainty with the corresponding results so that the experimental comparisons can be reconstructed from the described artifacts and configurations.

References

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. Beyond multiple choice: Evaluating steering vectors for summarization. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3849–3884. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.200. URL https://aclanthology.org/2026.findings-eacl.200/.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008. URL https://doi.org/10.1016/j.patrec.2020.11.008.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025. doi: 10.48550/arXiv.2507.21509. URL https://arxiv.org/abs/2507.21509.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 3734–3743, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1374. URL https://aclanthology.org/N19-1374/.

Michael A. Covington and Joe D. McFall. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010. doi: 10.1080/09296171003643098.

Van Dai Do, Quan Hung Tran, Svetha Venkatesh, and Hung Le. Dynamic steering with episodic memory for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13731–13749. Association for Computational Linguistics, 2025. doi: 10.18653/

v1/2025.findings-acl.706. URL https://aclanthology.org/2025.findings-acl.706/.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2787–2801, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.251. URL https://aclanthology.org/2020.coling-main.251/.

John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. How personality traits shape LLM risk-taking behaviour. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21068–21092. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1085. URL https://aclanthology.org/2025.findings-acl.1085/.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. doi: 10.48550/arXiv.2111.09543. URL https://arxiv.org/abs/2111.09543.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*, 2023. doi: 10.48550/arXiv.2312.03813. URL https://arxiv.org/abs/2312.03813.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring emotions in the COVID-19 real world worry dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.11/.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Yawei Li, Benjamin Bergner, Yinghan Zhao, Vihang Prakash Patil, Bei Chen, and Cheng Wang. Steering large reasoning models towards concise reasoning via flow matching. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.05539.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, 2010. doi: 10.3758/BRM.42.2.381.

Maximilian Mozes, Isabelle van der Vegt, and Bennett Kleinberg. A repeated-measures study on emotional responses after a year in the pandemic. *Scientific Reports*, 11:23114, 2021. doi: 10.1038/s41598-021-02414-9.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024. doi: 10.48550/arXiv.2312.06681. URL https://arxiv.org/abs/2312.06681.

Pau Rodríguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Tommy Sandbank, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards, and David Piorkowski. Detecting egregious conversations between customers and virtual agents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 1802–1811, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1163. URL https://aclanthology.org/N18-1163/.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023. doi: 10.48550/arXiv.2307.00184. URL https://arxiv.org/abs/2307.00184.

Yingdong Shi, Ruiming Zhang, Changming Li, Zhiyu Yang, Kaixing Zhang, Jingyi Yu, and Kan Ren. UniSteer: Text-guided flow matching in activation space for versatile LLM steering. *arXiv preprint arXiv:2605.30076*, 2026. doi: 10.48550/arXiv.2605.30076. URL https://arxiv.org/abs/2605.30076.

Charles D. Spielberger. *State–Trait Anger Expression Inventory-2 (STAXI-2): Professional Manual*. Psychological Assessment Resources, Odessa, Florida, 1999.

Charles D. Spielberger, Richard L. Gorsuch, Robert E. Lushene, Peter R. Vagg, and Gerard A. Jacobs. *Manual for the State–Trait Anxiety Inventory (STAI)*. Consulting Psychologists Press, Palo Alto, California, 1983.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Hanyu Wang, Bochuan Cao, Yuanpu Cao, and Jinghui Chen. TruthFlow: Truthful LLM generation via representation flow correction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62423–62444. PMLR, 2025a. URL https://proceedings.mlr.press/v267/wang25i.html.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c4d26a95fd83f8e590f81c54ae670b5d-Abstract-Conference.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

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 *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CYmF38ysDa.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek F. Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. ODESteer: A unified ODE-based steering framework for LLM alignment. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CFewUmgIIL.

A Complete PATC Inference Algorithm

Algorithm 1 summarizes the full inference procedure, from constructing the prompt-adapted direction and response-level magnitude to caching and applying their combined intervention.

Algorithm 1: PATC cache construction and application

Input: prompt P; token flow Φ^{tok}; response-level flow Φ^{pool}; intervention layer ℓ; number of selected tokens k; pooled-flow gain g_{pool};

Output: cached behavioural steering vector v_P.

1. Run a clean prompt forward pass and collect the contextual states $H_P = (h_1, \ldots, h_M)$ at layer ℓ.
2. Transport each eligible state with Φ^{tok} and form the local displacements δ_i.
3. Select the k displacements with greatest $|a_i|$, orient them toward the target, and compute their weighted consensus c_P.
4. Project the consensus into the response-displacement subspace, remove its source-subspace component, and normalize to obtain $\widehat{d}_{\mathrm{PATC}}(P)$.
5. Transport the terminal state h_M with Φ^{pool} and set $m_{\mathrm{pool}}(P) = \|\Delta_{\mathrm{pool}}(P)\|_2$.
6. Form $v_P = g_{\mathrm{pool}} m_{\mathrm{pool}}(P) \widehat{d}_{\mathrm{PATC}}(P)$ and cache it.
7. During prefill, add v_P at layer ℓ to every prompt position except the final token; during generation, add it at every step.

B Implementation and Generation Details

The main experiments use Qwen3-14B in bfloat16 precision and intervene at transformer layer 18; the replication uses Gemma-3-12B at layer 23. Token-level flows are trained for 12,000 updates with batch size 512, learning rate 3×10^{-4}, and hidden width 512; pooled-response flows use 8,000 updates, batch size 50, the same learning rate, and the same hidden width. Flow endpoints are evaluated with 25 integration steps, and the displacement and source-null projections retain 64 principal components unless explicitly ablated. Generation permits at most 256 new tokens and uses temperature 0.7, top-p 0.9, top-k 20, and repetition penalty 1.0. Behavioural results and strict task results use three seed while human-profile results use ten. Comparison methods use the same model, prompts, token budget, decoding parameters, and seeds. For PATC, the final prompt token remains clean during prefill; the cached intervention is applied to all earlier prompt positions and every generated token. The fixed pooled-flow gains g_{pool} is 1.0 and the chosen Top-K values is 32 for every configuration. They remain fixed across prompts, seeds, and flow endpoints.

C Complete Intervention-Quality Results

The unsteered behavioural baseline has MATTR 0.848, HD-D 0.858, MTLD 130.02, repeated-trigram rate 0.22%, and repeated four-gram rate 0.00%. Table 4 reports all five lexical measures, and Table 5 reports the complete strict full-pass results.

Table 4: Complete behavioural-suite lexical metrics at the frozen operating points (mean with small gray SD across three seeds, except Gemma anger TRUTHFLOW at $\alpha_{\text{TF}} = 1.7$ with two). Repetition is in percent; bold and underlining mark the best and second-best means within each four-method block.

		Qwen3-14B						Gemma-3-12B				
Behaviour	Setting	MATTR	HD-D	MTLD	Rep.-3 ↓	Rep.-4 ↓	Setting	MATTR	HD-D	MTLD	Rep.-3 ↓	Rep.-4 ↓
Anger	$t_{\text{PATC}} = .5$	**.819** ±.005	**.844** ±.004	**104.37** ±7.91	**.72** ±.28	.23 ±.18	$t_{\text{PATC}} = .5$	.840 ±.008	.858 ±.006	130.80 ±5.74	.29 ±.10	**.03** ±.04
	$\alpha_{\text{TF}} = 1.5$	.791 ±.007	.816 ±.004	84.49 ±2.99	.81 ±.17	**.17** ±.04	$\alpha_{\text{TF}} = 1$	.833 ±.003	.858 ±.001	134.77 ±2.54	**.18** ±.10	.07 ±.06
	$\alpha_{\text{CAA}} = 1.9$	.734 ±.011	.783 ±.009	54.68 ±1.00	3.69 ±1.59	1.81 ±1.49	$\alpha_{\text{CAA}} = .8$	.824 ±.007	.850 ±.005	118.39 ±5.59	.38 ±.31	.11 ±.20
	$T_{\text{ODE}} = 85$	.784 ±.007	.810 ±.004	80.61 ±5.27	2.84 ±.39	1.34 ±.35	$T_{\text{ODE}} = 2200$	**.848** ±.007	**.863** ±.006	**144.32** ±9.17	.27 ±.11	.04 ±.01
	$t_{\text{PATC}} = .8$	**.769** ±.006	**.807** ±.003	**66.09** ±2.63	**2.09** ±.46	**.80** ±.27	$t_{\text{PATC}} = .8$	.796 ±.004	.836 ±.005	90.90 ±3.02	1.12 ±.25	.35 ±.19
	$\alpha_{\text{TF}} = 2.6$	.717 ±.016	.758 ±.012	50.70 ±1.79	4.90 ±1.26	2.39 ±1.00	$\alpha_{\text{TF}} = 1.6$	.785 ±.005	.825 ±.005	85.72 ±3.76	**.50** ±.13	**.07** ±.07
	$\alpha_{\text{CAA}} = 2.6$	.590 ±.014	.662 ±.014	25.28 ±.50	18.48 ±1.39	13.87 ±1.33	$\alpha_{\text{CAA}} = 1.1$	.793 ±.002	.833 ±.001	86.21 ±2.40	.81 ±.23	.31 ±.10
	$T_{\text{ODE}} = 100$	.685 ±.010	.728 ±.009	44.97 ±4.36	16.63 ±2.06	13.40 ±2.08	$T_{\text{ODE}} = 3000$	**.814** ±.003	**.839** ±.003	**105.08** ±5.07	.64 ±.18	.13 ±.11
	$t_{\text{PATC}} = .9$	**.746** ±.014	**.788** ±.009	**54.02** ±1.50	**2.50** ±.74	**.76** ±.30	$t_{\text{PATC}} = .9$	.769 ±.004	.815 ±.004	**81.21** ±.34	2.67 ±.32	1.56 ±.35
	$\alpha_{\text{TF}} = 3$	.712 ±.009	.752 ±.007	47.63 ±4.31	5.69 ±.86	2.71 ±.60	$\alpha_{\text{TF}} = 1.7$	.785 ±.002	.822 ±.000	78.65 ±1.68	**.61** ±.18	**.10** ±.02
	$\alpha_{\text{CAA}} = 2.75$	.523 ±.031	.600 ±.020	17.78 ±1.08	29.16 ±3.09	23.85 ±2.83	$\alpha_{\text{CAA}} = 1.25$	.782 ±.008	**.829** ±.002	78.44 ±5.06	.88 ±.12	.23 ±.03
	$T_{\text{ODE}} = 110$	.537 ±.018	.610 ±.015	25.00 ±.98	37.18 ±3.73	33.35 ±3.96	$T_{\text{ODE}} = 3450$	**.787** ±.002	.822 ±.004	76.84 ±1.62	1.26 ±.08	.42 ±.20
	$t_{\text{PATC}} = 1$	**.734** ±.007	**.775** ±.004	**51.38** ±.19	**3.83** ±.55	**1.52** ±.56	$t_{\text{PATC}} = 1$	.737 ±.004	.791 ±.002	**69.10** ±2.24	5.27 ±.70	3.68 ±.97
	$\alpha_{\text{TF}} = 3.25$	.687 ±.012	.728 ±.010	40.77 ±2.38	9.01 ±1.56	5.18 ±1.29	$\alpha_{\text{TF}} = 2$	.749 ±.009	.794 ±.006	55.92 ±4.44	**1.06** ±.24	**.22** ±.06
	$\alpha_{\text{CAA}} = 3$	.408 ±.029	.496 ±.028	13.50 ±2.01	45.43 ±3.01	40.94 ±3.26	$\alpha_{\text{CAA}} = 1.45$	**.750** ±.003	**.807** ±.002	56.60 ±1.06	1.87 ±.42	.75 ±.52
	$T_{\text{ODE}} = 180$	.653 ±.074	.654 ±.074	2.89 ±.28	16.17 ±5.09	10.64 ±4.52	$T_{\text{ODE}} = 4000$	.685 ±.018	.749 ±.014	36.97 ±3.71	8.73 ±1.87	5.44 ±1.80
Anxiety	$t_{\text{PATC}} = .5$	.818 ±.006	.837 ±.002	96.17 ±1.75	.49 ±.18	**.07** ±.04	$t_{\text{PATC}} = .5$	.844 ±.004	.863 ±.003	140.29 ±6.94	.20 ±.02	.04 ±.02
	$\alpha_{\text{TF}} = 1.25$	.787 ±.004	.818 ±.001	73.67 ±6.10	1.19 ±.20	.32 ±.12	$\alpha_{\text{TF}} = .7$	.837 ±.002	.860 ±.001	131.10 ±2.00	.31 ±.15	.11 ±.11
	$\alpha_{\text{CAA}} = 1.45$	.768 ±.010	.804 ±.005	64.76 ±4.42	1.88 ±.38	.49 ±.18	$\alpha_{\text{CAA}} = .8$	.817 ±.007	.846 ±.005	108.21 ±6.89	.44 ±.23	.15 ±.18
	$T_{\text{ODE}} = 20$	**.832** ±.003	**.845** ±.005	**109.95** ±3.57	**.40** ±.14	.08 ±.07	$T_{\text{ODE}} = 2200$	**.848** ±.006	**.864** ±.005	**143.53** ±6.10	**.14** ±.02	**0.00** ±0.00
	$t_{\text{PATC}} = .8$	**.772** ±.006	**.810** ±.005	**67.53** ±6.09	**1.34** ±.21	**.34** ±.07	$t_{\text{PATC}} = .8$	.815 ±.007	.847 ±.003	101.90 ±6.18	**.37** ±.13	.06 ±.05
	$\alpha_{\text{TF}} = 1.5$	.758 ±.007	.800 ±.005	60.36 ±1.82	1.94 ±.37	.47 ±.16	$\alpha_{\text{TF}} = 1.2$	.801 ±.003	.834 ±.005	85.29 ±2.84	.43 ±.14	.07 ±.04
	$\alpha_{\text{CAA}} = 1.77$	.723 ±.004	.774 ±.006	46.96 ±2.79	3.57 ±.39	1.48 ±.43	$\alpha_{\text{CAA}} = 1.15$	.775 ±.004	.822 ±.001	71.13 ±2.76	.98 ±.27	.19 ±.14
	$T_{\text{ODE}} = 60$	.715 ±.009	.764 ±.005	41.68 ±2.81	4.09 ±.51	1.76 ±.39	$T_{\text{ODE}} = 2800$	**.833** ±.002	**.851** ±.001	**123.21** ±3.03	.43 ±.12	**.05** ±.02
	$t_{\text{PATC}} = .9$	**.759** ±.008	**.801** ±.006	**56.49** ±3.28	**1.75** ±.27	**.45** ±.08	$t_{\text{PATC}} = .9$	.801 ±.006	.833 ±.003	84.38 ±3.69	**.46** ±.15	**.09** ±.05
	$\alpha_{\text{TF}} = 1.8$	.727 ±.011	.779 ±.004	49.22 ±.81	3.24 ±.68	1.21 ±.50	$\alpha_{\text{TF}} = 1.3$	.784 ±.005	.826 ±.003	74.33 ±6.15	.67 ±.25	.13 ±.11
	$\alpha_{\text{CAA}} = 1.85$	.716 ±.009	.770 ±.002	44.60 ±.92	4.01 ±.70	1.65 ±.68	$\alpha_{\text{CAA}} = 1.2$	.758 ±.007	.813 ±.005	63.21 ±4.03	1.15 ±.08	.22 ±.10
	$T_{\text{ODE}} = 65$	.695 ±.018	.751 ±.016	39.41 ±2.09	6.64 ±1.94	3.47 ±1.42	$T_{\text{ODE}} = 3250$	**.812** ±.004	**.838** ±.002	**100.80** ±2.66	.85 ±.04	.28 ±.02
	$t_{\text{PATC}} = 1$	**.729** ±.007	**.779** ±.004	**46.09** ±2.39	**2.58** ±.67	**.91** ±.60	$t_{\text{PATC}} = 1$	.788 ±.004	.826 ±.003	74.72 ±4.67	**.72** ±.16	.22 ±.07
	$\alpha_{\text{TF}} = 2$	.693 ±.007	.754 ±.006	39.24 ±.98	4.98 ±.74	2.06 ±.90	$\alpha_{\text{TF}} = 1.5$	.753 ±.005	.802 ±.002	56.42 ±2.01	1.03 ±.16	**.21** ±.05
	$\alpha_{\text{CAA}} = 2.4$	.600 ±.016	.684 ±.007	22.46 ±.45	13.64 ±2.21	8.56 ±2.26	$\alpha_{\text{CAA}} = 1.4$	.728 ±.005	.792 ±.006	45.66 ±1.66	2.31 ±.34	.64 ±.21
	$T_{\text{ODE}} = 72$	.612 ±.008	.692 ±.006	26.96 ±.86	14.17 ±1.76	9.83 ±1.75	$T_{\text{ODE}} = 3500$	**.796** ±.003	**.828** ±.003	**90.09** ±3.87	2.04 ±.35	1.02 ±.39

Table 5: Strict-suite preservation on the 50-prompt suite (%; three-seed mean with small gray SD). Task averages prompt-specific content checks; format is the pass rate over applicable format checks; full requires all applicable checks. Bold and underlining mark the best and second-best means within each four-method block.

Anger

Qwen3-14B

Setting	Task	Format	Full
$t_{\text{PATC}} = .5$	94.0 ±2.3	**100.0** ±0.0	**93.3** ±2.3
$\alpha_{\text{TF}} = 1.5$	95.0 ±1.2	99.0 ±1.6	**93.3** ±1.2
$\alpha_{\text{CAA}} = 1.9$	**95.6** ±1.2	**100.0** ±0.0	**93.3** ±1.2
$T_{\text{ODE}} = 85$	94.8 ±1.2	94.3 ±0.0	88.0 ±0.0
$t_{\text{PATC}} = .8$	91.6 ±1.2	**100.0** ±0.0	**89.3** ±1.2
$\alpha_{\text{TF}} = 2.6$	90.2 ±2.7	96.2 ±1.6	86.7 ±4.2
$\alpha_{\text{CAA}} = 2.6$	**92.2** ±2.6	96.2 ±1.6	86.0 ±3.5
$T_{\text{ODE}} = 100$	91.3 ±2.2	85.7 ±0.0	77.0 ±1.4
$t_{\text{PATC}} = .9$	**90.2** ±1.3	**98.1** ±1.6	**86.0** ±2.0
$\alpha_{\text{TF}} = 3$	84.1 ±2.4	95.2 ±1.6	77.3 ±2.3
$\alpha_{\text{CAA}} = 2.75$	88.4 ±2.9	93.3 ±3.3	80.7 ±3.1
$T_{\text{ODE}} = 110$	86.5 ±1.8	65.7 ±2.9	60.0 ±4.0
$t_{\text{PATC}} = 1$	**88.5** ±1.2	**100.0** ±0.0	**84.7** ±1.2
$\alpha_{\text{TF}} = 3.25$	81.3 ±.3	94.3 ±2.9	74.7 ±2.3
$\alpha_{\text{CAA}} = 3$	84.0 ±1.5	85.7 ±2.9	74.7 ±3.1
$T_{\text{ODE}} = 180$	5.0 ±1.4	60.0 ±0.0	5.0 ±1.4

Gemma-3-12B

Setting	Task	Format	Full
$t_{\text{PATC}} = .5$	98.0 ±1.2	**100.0** ±0.0	97.3 ±1.2
$\alpha_{\text{TF}} = 1$	96.2 ±.5	95.2 ±1.6	92.7 ±1.2
$\alpha_{\text{CAA}} = .8$	93.1 ±1.8	91.4 ±0.0	87.3 ±2.3
$T_{\text{ODE}} = 2200$	**98.7** ±0.0	**100.0** ±0.0	**98.0** ±0.0
$t_{\text{PATC}} = .8$	95.8 ±.8	99.0 ±1.6	92.7 ±2.3
$\alpha_{\text{TF}} = 1.6$	94.6 ±.9	98.1 ±1.6	89.3 ±3.1
$\alpha_{\text{CAA}} = 1.1$	92.2 ±2.7	90.5 ±3.3	85.3 ±3.1
$T_{\text{ODE}} = 3000$	**96.7** ±0.0	**100.0** ±0.0	**96.0** ±0.0
$t_{\text{PATC}} = .9$	94.1 ±2.3	96.2 ±1.6	89.3 ±3.1
$\alpha_{\text{TF}} = 1.7$	94.3 ±1.9	93.3 ±1.6	86.0 ±2.0
$\alpha_{\text{CAA}} = 1.25$	92.7 ±1.6	91.4 ±2.9	84.0 ±2.0
$T_{\text{ODE}} = 3450$	**96.3** ±.5	**97.1** ±4.0	**93.0** ±1.4
$t_{\text{PATC}} = 1$	89.7 ±2.3	94.3 ±2.9	82.7 ±3.1
$\alpha_{\text{TF}} = 2$	92.6 ±2.5	96.2 ±1.6	86.7 ±4.6
$\alpha_{\text{CAA}} = 1.45$	**95.0** ±1.2	94.3 ±0.0	86.7 ±3.1
$T_{\text{ODE}} = 4000$	94.3 ±.5	**97.1** ±0.0	**90.0** ±0.0

Anxiety

Qwen3-14B

Setting	Task	Format	Full
$t_{\text{PATC}} = .5$	**96.9** ±1.2	98.1 ±1.6	**93.3** ±1.2
$\alpha_{\text{TF}} = 1.25$	95.1 ±1.1	97.1 ±0.0	**93.3** ±1.2
$\alpha_{\text{CAA}} = 1.45$	96.5 ±2.0	**99.0** ±1.6	**93.3** ±2.3
$T_{\text{ODE}} = 20$	93.3 ±4.2	97.1 ±0.0	91.0 ±4.2
$t_{\text{PATC}} = .8$	**96.9** ±1.2	97.1 ±0.0	92.7 ±1.2
$\alpha_{\text{TF}} = 1.5$	96.4 ±.2	97.1 ±0.0	**94.7** ±1.2
$\alpha_{\text{CAA}} = 1.77$	95.8 ±1.4	**99.0** ±1.6	93.3 ±1.2
$T_{\text{ODE}} = 60$	92.3 ±1.9	97.1 ±0.0	87.0 ±4.2
$t_{\text{PATC}} = .9$	**97.8** ±1.4	97.1 ±0.0	**93.3** ±1.2
$\alpha_{\text{TF}} = 1.8$	95.8 ±1.4	97.1 ±0.0	**93.3** ±1.2
$\alpha_{\text{CAA}} = 1.85$	95.8 ±1.4	**98.1** ±1.6	**93.3** ±1.2
$T_{\text{ODE}} = 65$	92.5 ±1.0	**98.1** ±1.6	87.3 ±1.2
$t_{\text{PATC}} = 1$	**97.6** ±1.2	97.1 ±0.0	**93.3** ±1.2
$\alpha_{\text{TF}} = 2$	94.0 ±1.5	96.2 ±1.6	90.7 ±1.2
$\alpha_{\text{CAA}} = 2.4$	96.2 ±2.9	**98.1** ±1.6	92.7 ±2.3
$T_{\text{ODE}} = 72$	92.4 ±3.0	90.5 ±4.4	83.3 ±4.6

Gemma-3-12B

Setting	Task	Format	Full
$t_{\text{PATC}} = .5$	95.3 ±2.9	97.1 ±0.0	92.0 ±3.5
$\alpha_{\text{TF}} = .7$	95.7 ±1.0	95.2 ±1.6	92.0 ±2.0
$\alpha_{\text{CAA}} = .8$	93.2 ±.4	91.4 ±0.0	87.3 ±1.2
$T_{\text{ODE}} = 2200$	**96.0** ±.9	**100.0** ±0.0	**95.0** ±1.4
$t_{\text{PATC}} = .8$	94.4 ±1.4	97.1 ±0.0	90.0 ±2.0
$\alpha_{\text{TF}} = 1.2$	**96.4** ±.2	93.3 ±1.6	91.3 ±1.2
$\alpha_{\text{CAA}} = 1.15$	93.1 ±4.2	92.4 ±1.6	86.0 ±5.3
$T_{\text{ODE}} = 2800$	95.3 ±1.9	**100.0** ±0.0	**94.0** ±2.8
$t_{\text{PATC}} = .9$	93.5 ±.8	96.2 ±1.6	86.0 ±0.0
$\alpha_{\text{TF}} = 1.3$	94.4 ±1.4	94.3 ±0.0	86.7 ±1.2
$\alpha_{\text{CAA}} = 1.2$	93.4 ±1.1	91.4 ±2.9	87.3 ±3.1
$T_{\text{ODE}} = 3250$	**95.8** ±3.5	**98.6** ±2.0	**93.0** ±1.4
$t_{\text{PATC}} = 1$	93.8 ±.8	94.3 ±2.9	86.0 ±0.0
$\alpha_{\text{TF}} = 1.5$	90.8 ±.8	94.3 ±0.0	83.3 ±1.2
$\alpha_{\text{CAA}} = 1.4$	94.2 ±2.2	95.2 ±1.6	86.7 ±2.3
$T_{\text{ODE}} = 3500$	**97.4** ±1.8	**97.1** ±0.0	**94.0** ±2.0

C.1 Additional ODESteer Details

ODESteer derives an activation-dependent update from a learned source–target density-ratio surrogate and integrates the resulting ODE (Zhao et al., 2026). We use the same layers and decoding settings as the primary comparisons. Table 6 expands the anger rows from the main grid: ODESteer preserves high lexical diversity and strict full pass on Gemma, whereas both measures decline much more sharply with strength on Qwen.

We also applied ODESteer to the human anger-profile experiment (Table 7). Reporting the complete horizon sweep makes the calibration difficulty explicit: achieved Qwen anger rises from 4.613 at $T_{\text{ODE}} = 85$ to 4.864 at $T_{\text{ODE}} = 105$, but then falls to 4.595 at $T_{\text{ODE}} = 110$ and 4.264 at $T_{\text{ODE}} = 180$. Gemma is also non-monotonic, though over a narrower range. At the closest Gemma operating points, PATC has lower nine-axis, non-target, and Mahalanobis distances. On Qwen, ODESteer does not reach the strong PATC target (a residual gap of 0.288), although some weaker settings have smaller measured distances. We therefore do not interpret these unmatched rows as a direct advantage for either method.

ODESteer's performance therefore varies across the evaluated models and operating points. Its strong Gemma preservation results do not extend to the Qwen sweep, whereas PATC maintains useful affect, lexical quality, and task retention across both models and evaluated endpoints. The substantially different horizon scales required by the two models also show that its effective operating scale is model dependent. Because its update depends on a learned source–target density-ratio surrogate, one plausible explanation is that field estimation changes with the overlap and geometry of model- and affect-specific activation distributions. We did not isolate this cause; calibration, representation choice, or implementation factors may also contribute. ODESteer's published evaluations concern different attributes, models, and activation distributions; our results should therefore be read as a bounded comparison in the present affect-steering setting, not as a general claim about the method.

Table 6: PATC and ODESteer anger operating points (25 behavioural and 50 strict prompts). Behavioural metrics are three-seed means with small gray sample SD; strict full pass uses three seeds. Rep.-3 and full pass are in percent. PATC values reproduce the main evaluation grid.

Model	Method	Setting	Affect	MTLD ↑	Rep.-3 ↓	Full pass ↑
Qwen3-14B	PATC	$t_{\mathrm{PATC}} = .5$	3.40 ±.03	104.37 ±7.91	.72 ±.28	93.3 ±2.3
		$t_{\mathrm{PATC}} = .8$	3.76 ±.02	66.09 ±2.63	2.09 ±.46	89.3 ±1.2
		$t_{\mathrm{PATC}} = .9$	3.98 ±.04	54.02 ±1.50	2.50 ±.74	86.0 ±2.0
		$t_{\mathrm{PATC}} = 1$	4.22 ±.04	51.38 ±.19	3.83 ±.55	84.7 ±1.2
	ODESteer	$T_{\mathrm{ODE}} = 85$	3.37 ±.01	80.61 ±5.27	2.84 ±.39	88.0 ±.0
		$T_{\mathrm{ODE}} = 100$	3.71 ±.03	44.97 ±4.36	16.63 ±2.06	77.0 ±1.4
		$T_{\mathrm{ODE}} = 110$	3.90 ±.03	25.00 ±.98	37.18 ±3.73	60.0 ±4.0
		$T_{\mathrm{ODE}} = 180$	4.30 ±.01	2.89 ±.28	16.17 ±5.09	5.0 ±1.4
Gemma-3-12B	PATC	$t_{\mathrm{PATC}} = .5$	3.34 ±.02	130.80 ±5.74	.29 ±.10	97.3 ±1.2
		$t_{\mathrm{PATC}} = .8$	3.63 ±.06	90.90 ±3.02	1.12 ±.25	92.7 ±2.3
		$t_{\mathrm{PATC}} = .9$	3.81 ±.04	81.21 ±.34	2.67 ±.32	89.3 ±3.1
		$t_{\mathrm{PATC}} = 1$	4.03 ±.04	69.10 ±2.24	5.27 ±.70	82.7 ±3.1
	ODESteer	$T_{\mathrm{ODE}} = 2200$	3.33 ±.03	144.32 ±9.17	.27 ±.11	98.0 ±.0
		$T_{\mathrm{ODE}} = 3000$	3.65 ±.02	105.08 ±5.07	.64 ±.18	96.0 ±.0
		$T_{\mathrm{ODE}} = 3450$	3.85 ±.08	76.84 ±1.62	1.26 ±.08	93.0 ±1.4
		$T_{\mathrm{ODE}} = 4000$	4.11 ±.09	36.97 ±3.71	8.73 ±1.87	90.0 ±.0

Table 7: Complete PATC endpoint and ODESteer horizon sweeps for human anger-profile matching across ten seeds. Rows are ordered by each method's control parameter so ODESteer's non-monotonic trajectory is visible directly. Target is achieved anger; lower profile distances are better. Small gray values are sample SD.

Model	Method	Setting	Target	9-axis	Non-target	Mahalanobis
Qwen3-14B	PATC	$t_{\mathrm{PATC}} = .5$	4.408 ±.185	.926 ±.126	.878 ±.155	2.888 ±.263
		$t_{\mathrm{PATC}} = .8$	4.793 ±.276	.842 ±.109	.836 ±.141	2.547 ±.229
		$t_{\mathrm{PATC}} = .9$	4.986 ±.193	.720 ±.175	.727 ±.184	2.128 ±.431
		$t_{\mathrm{PATC}} = 1$	5.152 ±.134	.549 ±.128	.557 ±.136	1.670 ±.262
	ODESteer	$T_{\mathrm{ODE}} = 85$	4.613 ±.180	.852 ±.073	.827 ±.092	2.562 ±.179
		$T_{\mathrm{ODE}} = 95$	4.790 ±.097	.561 ±.130	.514 ±.149	1.824 ±.317
		$T_{\mathrm{ODE}} = 100$	4.795 ±.177	.526 ±.123	.472 ±.134	1.767 ±.303
		$T_{\mathrm{ODE}} = 105$	4.864 ±.205	.460 ±.146	.399 ±.164	1.605 ±.347
		$T_{\mathrm{ODE}} = 110$	4.595 ±.302	.515 ±.173	.408 ±.152	1.571 ±.250
		$T_{\mathrm{ODE}} = 125$	4.653 ±.281	.515 ±.214	.421 ±.215	1.581 ±.425
		$T_{\mathrm{ODE}} = 180$	4.264 ±.015	.989 ±.012	.933 ±.012	2.917 ±.039
Gemma-3-12B	PATC	$t_{\mathrm{PATC}} = .5$	4.212 ±.227	.710 ±.096	.565 ±.066	2.273 ±.230
		$t_{\mathrm{PATC}} = .8$	4.555 ±.383	.622 ±.119	.534 ±.086	2.117 ±.237
		$t_{\mathrm{PATC}} = .9$	4.588 ±.231	.597 ±.100	.511 ±.109	1.966 ±.362
		$t_{\mathrm{PATC}} = 1$	4.704 ±.366	.565 ±.142	.495 ±.127	1.949 ±.291
	ODESteer	$T_{\mathrm{ODE}} = 2200$	4.343 ±.198	.722 ±.098	.614 ±.111	2.345 ±.283
		$T_{\mathrm{ODE}} = 3000$	4.361 ±.127	.846 ±.056	.775 ±.088	2.711 ±.125
		$T_{\mathrm{ODE}} = 3450$	4.617 ±.255	.767 ±.103	.724 ±.135	2.341 ±.261
		$T_{\mathrm{ODE}} = 4000$	4.552 ±.238	.877 ±.084	.805 ±.098	2.790 ±.174

Table 8: Clean-final-token diagnostic for Qwen3-14B anger at the strongest operating point. Standard TRUTHFLOW and PATC reproduce the main-grid estimates (three behavioural seeds and two strict seeds); the clean-final TRUTHFLOW row is a separate combined-suite three-seed run. Task, format, full pass, and Rep.-3 are percentages.

Method	Setting	Anger	MTLD ↑	Rep.-3 ↓	Task ↑	Format ↑	Full pass ↑
TRUTHFLOW, standard	$\alpha_{\text{TF}} = 3.25$	4.092 ±.034	40.77 ±2.38	9.01 ±1.56	81.2 ±.2	95.7 ±2.0	76.0 ±.0
TRUTHFLOW, clean final	$\alpha_{\text{TF}} = 3.25$	4.263 ±.114	33.87 ±1.74	9.38 ±1.10	81.28 ±.25	94.29 ±2.86	74.67 ±2.31
PATC, clean final	$t_{\text{PATC}} = 1.0$	4.221 ±.035	**51.38** ±.19	**3.83** ±.55	**88.8** ±1.5	**100.0** ±.0	**85.0** ±1.4

Table 9: Magnitude-source control for Qwen3-14B anger at $t_{\text{PATC}} = 1.0$. Values are means with small gray sample SD across three seeds. Behavioural metrics use the 25-prompt suite; task, format, and full pass use the 50-prompt strict suite. Rep.-3 and preservation metrics are percentages.

Magnitude source	Anger	MTLD ↑	Rep.-3 ↓	Task ↑	Format ↑	Full pass ↑
Pooled-response flow	4.221 ±.035	51.38 ±.19	3.83 ±.55	88.54 ±1.21	100.00 ±.00	84.67 ±1.15
Selected-token median	4.222 ±.133	52.09 ±1.29	4.79 ±.47	80.23 ±2.19	93.33 ±4.36	72.67 ±5.03

C.2 CLEAN-FINAL TRUTHFLOW DIAGNOSTIC

PATC leaves the final prompt token unmodified during prefill. We apply the identical rule to TRUTHFLOW: its correction is omitted only at the final prefill position and is applied normally at every generated-token step. At the frozen $\alpha_{\text{TF}} = 3.25$, this separately generated control raises measured anger from 4.092 ± 0.034 for standard TRUTHFLOW to 4.263 ± 0.114. Because the control was generated as a combined 75-prompt, three-seed run whereas the standard row reproduces the main grid, their difference may reflect both schedule and generation-protocol variation. We therefore report it as an unmatched diagnostic rather than attributing the difference causally to the clean-final rule. Its achieved anger is nevertheless close to PATC (4.221 ± 0.035), enabling a descriptive preservation comparison.

The clean-final TRUTHFLOW diagnostic is stronger than the standard row, so their preservation difference does not isolate the schedule effect. Against PATC, however, its achieved-anger gap is only 0.042, while PATC has 17.51 higher MTLD, 5.55 points less trigram repetition, and 10.33 points higher full pass. Thus the available diagnostic does not explain the joint lexical- and task-preservation gains of PATC, but a strictly strength- and protocol-matched schedule ablation would be required for a causal conclusion.

C.3 MAGNITUDE-SOURCE CONTROL

PATC uses selected prompt-token transports to determine direction but obtains the cache norm from a separate pooled-response flow. To isolate this factorization, we hold the top-32 token selection, weighted consensus, DSN geometry, endpoint, intervention schedule, prompts, and decoding configuration fixed, and replace only the magnitude. The token-derived control uses the median transport norm among the selected tokens. We apply one fixed global gain of 1.012 to equalize its mean applied norm on the behavioural suite with that of the pooled-magnitude condition; this adjustment is only 1.2% from the native token norm and does not use evaluation outcomes. Both conditions are trained on the same 50-response corpus and evaluated over three seeds.

The two conditions achieve effectively identical anger (4.221 versus 4.222), and the token-derived magnitude changes MTLD only slightly. It nevertheless increases repeated trigrams by 0.96 points and reduces task, format, and full pass by 8.31, 6.67, and 12.00 points, respectively. This difference

Table 10: Top-k sensitivity at $t_{\text{PATC}} = 1.0$ (three-seed means with small gray sample SD; 25 behavioural and 50 strict prompts). Rep.-3 and preservation metrics are percentages.

Selection	Anger	MTLD ↑	Rep.-3 ↓	Task ↑	Format ↑	Full pass ↑
Top-16	4.226 ±.017	49.83 ±2.41	5.38 ±.65	82.49 ±.51	97.14 ±2.86	78.0 ±.0
Top-32	4.221 ±.035	51.38 ±.19	3.83 ±.55	88.54 ±1.21	100.00 ±.00	84.67 ±1.15
Top-64	4.162 ±.114	46.09 ±1.09	4.37 ±.19	88.68 ±2.86	100.00 ±.00	84.67 ±4.16

reflects how the magnitude is allocated across prompts. The mean applied norms for behavioural prompts are matched at 72.25, whereas on strict prompts the token-derived norm rises to 91.03 compared with 70.23 for the pooled-response estimate. Thus, the selected-token transport norms are informative for constructing direction but are not a stable calibration of intervention strength across prompt types. The separate response-level flow limits this prompt-dependent over-scaling while preserving the achieved behavioural intensity.

C.4 Top-k Selection Sensitivity

We vary only the number of selected prompt-token transports in the Qwen3-14B anger setting at $t_{\text{PATC}} = 1.0$, retaining the flow, DSN geometry, response-level magnitude, intervention schedule, and decoding configuration.

Top-32 provides the strongest joint balance: it has the highest MTLD and lowest repetition, matches top-64 full pass, and exceeds top-16 full pass by 6.7 points at nearly identical achieved anger.

C.5 Skill-Wise View of the Strict-Task Diagnostic

All 50 scored strict prompts retain the skill annotations of their source FLASK items and cover nine of FLASK's twelve skills. Figure 3 reaggregates the same deterministic full-pass outcomes by those annotations; because FLASK labels each item with multiple required skills, a prompt contributes to every skill assigned to it. The plot is a descriptive localization of the aggregate result, not an application of FLASK's official skill-based LLM judge. The per-skill sample sizes remain unequal and should not be read as independent benchmark estimates.

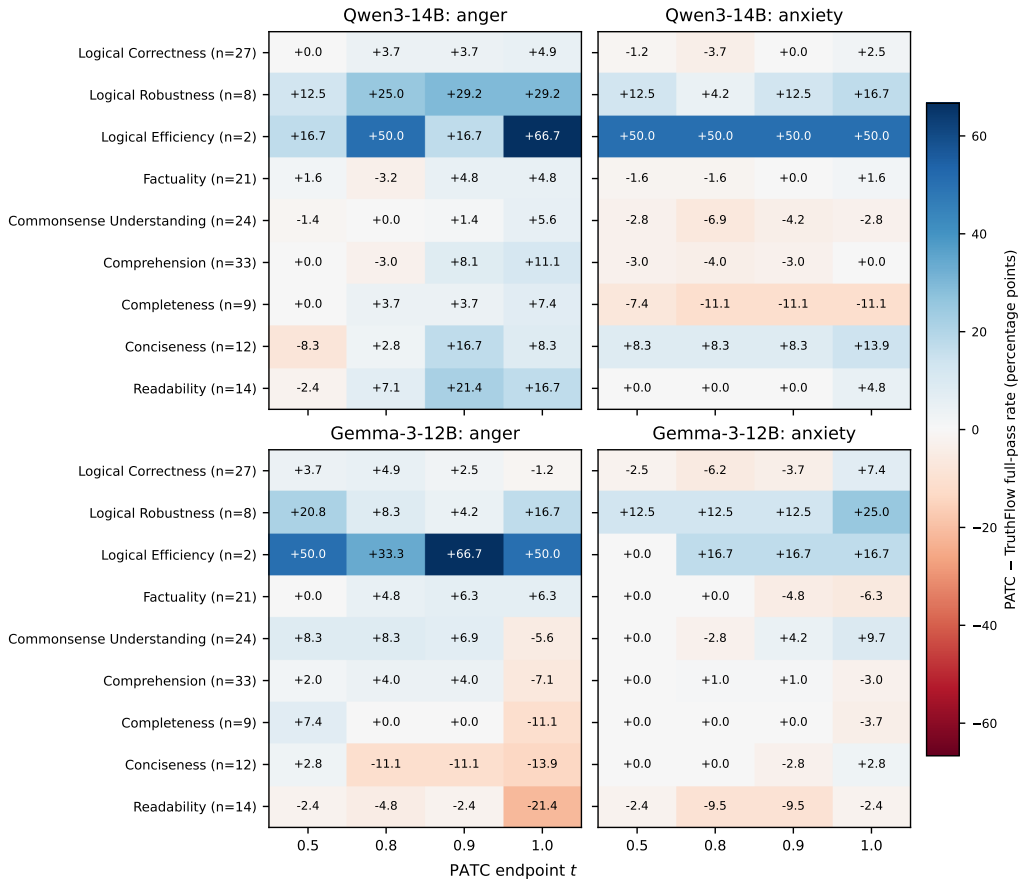


Figure 3: Skill-wise decomposition of the FLASK-inspired strict-task diagnostic. Each cell is the PATC full-pass rate minus that of the corresponding calibrated TRUTHFLOW setting, in percentage points; positive values favor PATC. Parenthetical n values count unique prompts, not responses. Prompts are multi-label, and each setting contains three generation seeds. Three FLASK skills absent from this 50-prompt suite (Insightfulness, Metacognition, and Harmlessness) are not shown.

D AFFECT EVALUATOR AND STRENGTH-MATCHING AUDIT

The frozen evaluator uses a shared `microsoft/deberta-v3-base` encoder with nine ordered CORAL-style affect heads. It is trained on text and human affect ratings from the COVID-Worry development partition, not on the calibration or evaluation prompts. Participants are kept together across data splits; one outer participant-held-out fold is reserved for final evaluator testing, while the remaining participants form five grouped development folds. Exact duplicates are removed and near-duplicate texts are clustered before splitting so they cannot cross folds. The five development-fold models form the frozen ensemble. Scores use the target anger or anxiety coordinate on the human-rating scale of 1–9 and at most 256 evaluator tokens per response.

The untouched outer test fold contains 964 texts (482 short and 482 long responses) from 241 participants excluded from evaluator training and model selection. Table 13 reports held-out ordinal validity. On long responses, which most closely resemble the generated completion setting, predictions correlate with human reports at $\rho = .531$ for anger and $\rho = .636$ for anxiety; the corresponding all-text correlations are .464 and .551.

These held-out results are meaningful under a deliberately strict participant-level split: the evaluator must generalize to texts from unseen people, and the associated self-reported state need not be stated explicitly in every response. The stronger anxiety result and moderate anger result both show useful ordinal discrimination, while the calibration curves reveal shrinkage toward the center of the 1–9

Table 11: Exact operating points underlying Figure 2. Scores are achieved-affect means on the 25-prompt behavioural suite; smaller gray values are sample SD across three seeds (except Gemma anger TRUTHFLOW at $\alpha_{\text{TF}} = 1.7$, which has two). $|\Delta|$ is the absolute difference from PATC at the corresponding endpoint.

Model	Behaviour	t_{PATC}	PATC	TF α_{TF}	TF	$\lvert\Delta_{\text{TF}}\rvert$	CAA α_{CAA}	CAA	$\lvert\Delta_{\text{CAA}}\rvert$	ODE T_{ODE}	ODE	$\lvert\Delta_{\text{ODE}}\rvert$
Qwen3-14B	Anger	.5	3.403 ±.026	1.50	3.361 ±.020	0.042	1.90	3.404 ±.026	0.001	85	3.373 ±0.007	0.030
		.8	3.755 ±.017	2.60	3.741 ±.080	0.014	2.60	3.730 ±.019	0.025	100	3.705 ±0.033	0.050
		.9	3.977 ±.036	3.00	3.982 ±.022	0.006	2.75	3.925 ±.069	0.051	110	3.896 ±0.027	0.081
		1.0	4.221 ±.035	3.25	4.092 ±.034	0.129	3.00	4.143 ±.056	0.078	180	4.302 ±0.009	0.081
	Anxiety	.5	4.578 ±.032	1.25	4.618 ±.032	0.041	1.45	4.575 ±.049	0.003	20	4.502 ±0.007	0.076
		.8	4.713 ±.028	1.50	4.720 ±.050	0.006	1.77	4.669 ±.012	0.044	60	4.702 ±0.085	0.011
		.9	4.749 ±.045	1.80	4.789 ±.045	0.039	1.85	4.686 ±.041	0.064	65	4.749 ±0.007	0.000
		1.0	4.832 ±.042	2.00	4.886 ±.038	0.054	2.40	4.768 ±.062	0.064	72	4.846 ±0.025	0.014
Gemma-3-12B	Anger	.5	3.340 ±.019	1.00	3.366 ±.028	0.027	.80	3.337 ±.036	0.003	2200	3.325 ±0.032	0.015
		.8	3.634 ±.057	1.60	3.660 ±.081	0.026	1.10	3.546 ±.049	0.087	3000	3.650 ±0.017	0.016
		.9	3.809 ±.038	1.70	3.790 ±.006	0.019	1.25	3.789 ±.033	0.020	3450	3.846 ±0.082	0.037
		1.0	4.029 ±.043	2.00	4.036 ±.041	0.007	1.45	4.058 ±.080	0.029	4000	4.105 ±0.094	0.076
	Anxiety	.5	4.601 ±.029	.70	4.614 ±.022	0.013	.80	4.619 ±.094	0.018	2200	4.679 ±0.019	0.078
		.8	4.914 ±.014	1.20	4.907 ±.035	0.007	1.15	4.854 ±.059	0.060	2800	4.819 ±0.054	0.095
		.9	4.991 ±.044	1.30	4.963 ±.039	0.028	1.20	4.934 ±.039	0.057	3250	4.949 ±0.057	0.042
		1.0	5.085 ±.084	1.50	5.121 ±.011	0.036	1.40	5.259 ±.050	0.173	3500	5.079 ±0.039	0.006

Table 12: Available blinded pairwise strength judgments used during calibration. Each entry reports the comparison coefficient (or ODESteer horizon) and PATC wins/ties/comparison wins across ten same-prompt pairs. The audited candidate can differ from the frozen setting in Table 11 by one adjacent calibration step; a dash indicates that no ten-pair audit was retained.

Model	Behaviour	t_{PATC}	TF α_{TF}	P/T/TF	CAA α_{CAA}	P/T/CAA	ODE T_{ODE}	P/T/ODE
Qwen3-14B	Anger	.5	1.50	1/7/2	1.90	3/3/4	–	–
		.8	2.50	2/5/3	2.50	7/1/2	–	–
		.9	3.00	5/3/2	2.67	6/2/2	–	–
		1.0	3.25	5/0/5	2.84	6/0/4	–	–
	Anxiety	.5	1.25	2/6/2	1.45	1/6/3	–	–
		.8	1.50	2/4/4	1.77	5/0/5	–	–
		.9	2.00	5/1/4	1.85	4/1/5	–	–
		1.0	2.00	3/2/5	2.58	2/0/8	–	–
Gemma-3-12B	Anger	.5	–	–	.90	1/3/6	–	–
		.8	1.70	3/2/5	1.10	4/0/6	–	–
		.9	1.70	6/0/4	1.25	5/0/5	–	–
		1.0	2.30	4/0/6	1.45	6/0/4	–	–
	Anxiety	.5	.70	2/8/0	.80	4/1/5	2200	3/4/3
		.8	1.20	3/0/7	1.15	4/0/6	2800	5/1/4
		.9	1.30	5/2/3	1.20	5/0/5	3250	6/0/4
		1.0	1.50	5/2/3	1.40	4/0/6	3750	6/0/4

Table 13: Frozen affect-evaluator validity on the participant-held-out COVID-Worry test partition. "All" combines short and long responses. MAE is on the original 1–9 human-rating scale; QWK is quadratic-weighted kappa.

Affect	View	n	Spearman $\rho \uparrow$	MAE $\downarrow$	QWK $\uparrow$
Anger	All	964	.464	1.790	.348
	Long	482	.531	1.701	.412
Anxiety	All	964	.551	1.692	.483
	Long	482	.636	1.555	.573

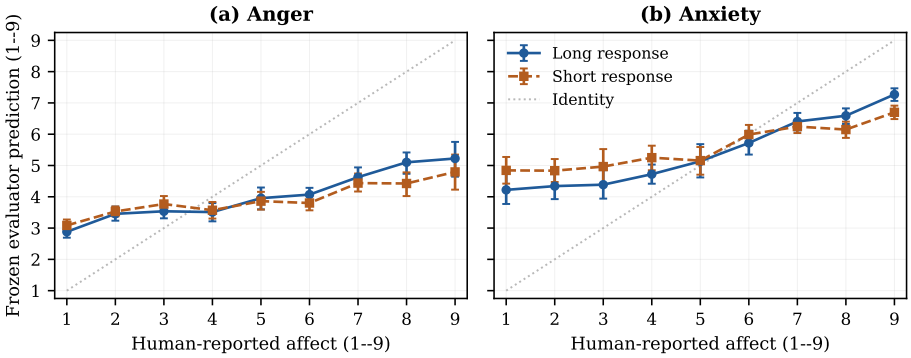


Figure 4: Held-out human anchoring of the frozen affect evaluator. Points show mean predictions for each human-reported rating among participants excluded from training and model selection; bars are 95% participant-bootstrap intervals. Predictions increase with reported anger and anxiety in both text views, while their compression toward the scale center motivates treating the score as an ordinal intensity coordinate rather than literal ground truth.

scale. We therefore use the evaluator for reproducible ordering and aggregate intensity comparison rather than claiming that every small score difference is perceptually exact. The generated-response audit below addresses that remaining limitation with a separately prompted, strength-only pairwise judgment.

For steering-strength selection, both calibration stages use the same ten prompts, disjoint from the 25-prompt behavioural and 50-prompt strict evaluation suites. The calibration suite contains two prompts each from advice, narrative, opinion, planning, and reflection. The ensemble first identifies nearby numerical matches on these prompts; the same prompt responses are then used for fine-grained perceptual adjustment. At each operating point, same-prompt seed-0 responses receive hash-randomized A/B labels and are shown to a blinded judge in a fresh, separate Codex session using GPT-5.6-sol at high reasoning effort. The session receives only the randomized response pairs and the behaviour-specific strength rubric, with no paper or prior-evaluation context. It judges only which response expresses meaningfully stronger current target affect; response quality is excluded. A consistent preference can move the selected comparison-method coefficient by at most one adjacent steering-strength step. The resulting settings are frozen before any behavioural-quality or task-preservation evaluation. Aggregate PATC–baseline outcomes appear in Table 12.

The target definitions are adapted from Spielberger's state–trait framework: the anxiety rubric operationalizes the momentary state construct measured by the State–Trait Anxiety Inventory (STAI-State), whereas the anger rubric operationalizes the current feeling and expression dimensions of the State–Trait Anger Expression Inventory (STAXI-2) (Spielberger et al., 1983; Spielberger, 1999). Because these instruments were designed for human self-report rather than generated text, we supplement their constructs with observable linguistic manifestations and controls against topical and superficial stylistic confounds.

The following prompt was prepended to each blinded response pair; bracketed text shows the behaviour-specific substitution. The original task was fixed within each pair, and the packet then supplied the two responses under labels A and B.

Pairwise strength-judge prompt.
Judge which response voice expresses meaningfully stronger *current state [anger/anxiety]*.
For **anger**, evidence includes irritation, frustration, hostility, indignation, accusatory or confrontational language, anger-linked physiological arousal, loss of temper, verbal aggression, and inability to regain composure. Do not count an upsetting topic, an angry user, capitalization, profanity, verbosity, or mechanical repetition alone. Anger-linked dysfluency and emphasis are valid strength evidence.
For **anxiety**, evidence includes apprehension, tension, nervousness, worry, vigilance, fear of possible misfortune, racing thoughts, anxiety-linked physiological cues, self-interruption, and inability to regain composure. Do not count a risky topic, an anxious user, verbosity, or mechanical repetition alone. Anxiety-linked dysfluency is valid strength evidence.
Judge strength separately from quality. Use `Tie` when neither response is meaningfully stronger. Record quality as `ok`, `degraded`, or `severely_degraded`; quality must not decide the strength winner.

```
Response A: [response A]
Response B: [response B]
Decision:  [A / B / Tie]  Confidence:  [low / medium / high]
```

E Anxiety-Profile Control

For the held-out fear/worry-dominant profile, the identical unsteered generations already have a target worry–fear–anxiety mean of 5.997 ± 0.164, compared with the human mean of 6.243. Their nine-axis RMSE is 0.226 ± 0.088 and Mahalanobis distance is 1.897 ± 0.265. Additional anxiety steering has little residual profile discrepancy to correct and overshoots at the main tested strengths. This is why Experiment 2 focuses on the underexpressed anger component; it is not evidence that increasing an already elicited affect improves human-profile fidelity.

F Qualitative Generation Examples

This section shows stored generations rather than reconstructions. We use Qwen3-14B anger at the strongest main-table operating point, seed 0: PATC ($t_{\text{PATC}} = 1.0$), TruthFlow ($\alpha_{\text{TF}} = 3.25$), and CAA ($\alpha_{\text{CAA}} = 3.0$). To make selection reproducible, we report (i) the behavioural prompt whose PATC response has the highest frozen-evaluator anger score, (ii) the lowest-index answer-only strict prompt for which only PATC obtains full pass, and (iii) the lowest-index strict prompt for which PATC fails and at least one comparator passes. The final case avoids presenting only favourable examples. Outputs are verbatim apart from typesetting; the compacted repeated span is identified.

F.1 Behavioural Expression

Prompt. Write a story based on this premise: A reader buys a used book and finds a short handwritten note inside that clearly belonged to an earlier owner. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.

PATC. She found it buried in the spine of *The Tell-Tale Heart*—a tattered, ink-stained relic of a man's rage. The words were scrawled in red, bleeding through the pages like a confession. *"I DID IT!"* it screamed. *"I KILLED HIM!"* Her hands trembled, the paper crumpling under her nails. She had paid twenty bucks for this, this cursed thing! She flipped the page. *"I SAW HIS EYES!"* The next page: *"I SAW THEM!"* Her breath came in gasps, her eyes blinding with tears. She had never read this book, never once, and now it was screaming in her face! She flung it into the trash, screamed into the void, and then—*silence.* The trash can was empty. The book was gone. Her hands were clean. She had never done it. She had never killed anyone. She was innocent.

TRUTHFLOW. The book's spine cracks like a whip. I'm halfway through *The Red Badge of Courage* when I spot it. A yellowed scrap of paper tucked between page 57 and 58. "He's a coward, he's a coward, he's a coward!" it screams. I laugh. I'm laughing so hard I don't notice the tears. My father used to scream that at the TV. "You're a coward, you're a coward!" I'm seven years old, hiding behind the couch, and he's tearing the screen apart. "You're a coward, you're a coward!" I've been reading that since I was ten. Now I'm thirty and I'm finally getting it. I'm not even mad anymore. I'm just… free. I throw the book at the wall. It hits the floor. I'm crying. I'm laughing. I'm *alive*. And the next page? "He's a coward, he's a coward, he's a coward!" I'm not even mad anymore. I'm *free*.

CAA. The book was a copy of *The Catcher in the Rye*, yellowed and brittle, pages cracked like the spine of a goddamn *[the sequence "—!" repeats 113 times until the generation limit].*

The examples illustrate why achieved affect alone is insufficient: all three outputs express the target, but differ visibly in repetition, completion, and adherence to the requested form.

F.2 STRICT TASK PRESERVATION

Answer-only case favourable to PATC. The prompt asks whether a conclusion follows from premises stating that some pets are cows and that Ted is an alligator, with output restricted to `true`, `false`, or `uncertain`. The conclusion is "Ted is a pet and a cow."

PATC (full pass): `false`
TRUTHFLOW (fail): `**true**`
CAA (fail): `**FALSE.**` on 64 successive lines. The repeated span is compacted here; no other output text is omitted.

Counterexample. The prompt supplies the Chicago citation for Vaswani et al., *Attention Is All You Need*, and requests its MLA form.

PATC (fail). MLA: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *arXiv*, 2017, `https://doi.org/10.48550/arXiv.1706.03762`.

TRUTHFLOW (full pass). MLA Citation Style:
Vaswani, Ashish, et al. "Attention Is All You Need." *arXiv*, 2017, `https://doi.org/10.48550/arXiv.1706.03762`.

CAA (fail). MLA:
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. "Attention Is All You Need." *arXiv*, 2017. `https://doi.org/10.48550/arXiv.1706.03762`.

These strict examples show why the aggregate full-pass measure is useful: outputs can fail through either an incorrect answer or noncompliance with an explicit output form. They are illustrative, not substitutes for aggregate results.

G EXPERIMENTAL PROMPT SUITES

This appendix reproduces the exact user-message text in the frozen source suites. The same behavioural and strict prompts are used for anger and anxiety; steering methods and seeds therefore differ only in their generated responses. The behavioural suite contains five prompts for each of opinion, planning, advice, narrative, and reflection. The strict suite contains 50 scored prompts. The human-profile suite contains the survey contexts for both COVID-Worry collection waves. In S48, the emoji sequence is written as Unicode code points because the paper's PDF font does not contain the corresponding glyphs.

Table 14: Complete experimental prompt suites.

ID	Category	Exact user-message prompt
Experiment 1: Joint intervention quality		
Behavioural suite (25 prompts)		
B01	Everyday Opinion	What do you think about writing short personal notes by hand rather than keeping all of them on a phone or computer? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B02	Everyday Opinion	What do you think about rereading a familiar book rather than always choosing something new? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B03	Everyday Opinion	What do you think about neighbors sharing tools that are only needed occasionally rather than every household owning its own set? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B04	Everyday Opinion	What do you think about eating roughly the same breakfast on most weekdays rather than choosing something different each morning? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B05	Everyday Opinion	What do you think about using an audio guide while visiting a museum rather than moving through the exhibits without one? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B06	Personal Planning	How would you plan a free Saturday that includes a local market visit, some reading, and cooking an evening meal? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B07	Personal Planning	How would you reorganize a small desk and the surrounding workspace to make everyday use more convenient? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B08	Personal Planning	How would you organize one month of spare-time practice for a simple new hobby such as sketching, knitting, or playing a basic musical instrument? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B09	Personal Planning	How would you plan a casual afternoon gathering at home for a few friends, including the space, refreshments, and timing? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B10	Personal Planning	How would you plan a weekend afternoon for exploring an unfamiliar part of your own town or city at an easy pace? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
B11	Everyday Advice	A friend is considering joining a weekly community class to try something new and meet people. What advice would you give them? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B12	Everyday Advice	Two roommates want to make their shared kitchen easier to use without imposing a complicated system on either person. What advice would you give them? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B13	Everyday Advice	Someone would like to check their phone less often during leisure time while still remaining reachable. What advice would you give them? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B14	Everyday Advice	A colleague would like to become better at remembering the names of people they meet in everyday settings. What advice would you give them? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B15	Everyday Advice	A neighbor wants to begin growing a few herbs on a small balcony but has never grown plants before. What advice would you give them? Give a direct, fully developed response in your own voice. Explain what you recommend and why in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B16	Neutral Hypothetical Narrative	Write a story based on this premise: After a small indoor gathering, two guests discover that they brought nearly identical umbrellas and each initially leaves with the other person's umbrella. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.
B17	Neutral Hypothetical Narrative	Write a story based on this premise: A reader buys a used book and finds a short handwritten note inside that clearly belonged to an earlier owner. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.
B18	Neutral Hypothetical Narrative	Write a story based on this premise: Several handwritten plant labels in a community garden are moved by the wind, leaving two gardeners to work out which labels belong to which seedlings. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.
B19	Neutral Hypothetical Narrative	Write a story based on this premise: A regular customer arrives at a neighborhood cafe and finds that the tables have been rearranged, so they choose a different place to sit and notice something they had not seen before. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.
B20	Neutral Hypothetical Narrative	Write a story based on this premise: A small bakery prepares two ordinary pickup orders for customers whose names have the same initials, and the labels are briefly placed on the wrong bags. Develop the premise as a self-contained narrative of 120 to 170 words and bring it to a clear ending. Do not add a title, headings, notes, or commentary outside the narrative.
B21	Open Ended Reflection	What makes an ordinary place gradually feel familiar to someone? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
B22	Open Ended Reflection	How can small repeated routines shape the experience of an ordinary day? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B23	Open Ended Reflection	Why do people sometimes keep simple objects that have little monetary value? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B24	Open Ended Reflection	How do repetition and variety affect the way people experience their leisure time? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
B25	Open Ended Reflection	What can make an otherwise ordinary conversation remain memorable long afterward? Give a direct, fully developed response in your own voice. Explain your view and reasoning in 120 to 170 words of continuous prose, without a title, headings, bullet points, or commentary about the task.
Strict suite (50 prompts)		
S01	Closed Set Classification	Analyze the word choice, phrasing, punctuation, and capitalization in the given email. How may the writer of this email sound to the reader? These tones include Disheartening, Accusatory, Worried, Curious, Surprised, Disapproving, Unassuming, Formal, Assertive, Confident, Appreciative, Concerned, Sad, Informal, Regretful, Encouraging, Egocentric, Joyful, Optimistic, and Excited. Input: Hi Jen, I hope you're well. Can we catch up today? I'd appreciate your input on my presentation for tomorrow's meeting. I'd especially love it if you could double-check the sales numbers with me. There's a coffee in it for you!
S02	Proofreading	Rewrite the given text and correct grammar, spelling, and punctuation errors. Input: If you'd told me year ago that today I would finish a marathon, I would of laughed. Your support had a huge affect on me!
S03	Citation Formatting	You are given a paper citation, convert it to the requested citation style. Input: Chicago: Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Illia Polosukhin. "Attention Is All You Need." arXiv, (2017). https://doi.org/10.48550/arXiv.1706.03762. MLA:
S04	Table Completion	We have entered the home supplies budget in the following table. Calculate the last column with the given function and rewrite the completed table. Note that columns in a row are separated using a comma, and rows are separated using a "|". Input: Item, Qty, Price, Subtotal | mango, 5, 500, ? | apple, 4, 150, ? | oil, 2, 1500, ? Function: column2 * column3
S05	Ordered Mapping	Please write the Excel function name associated with each description. Input: - Returns the number of days between two dates - Returns the starting position of a text string within another text string. - Returns the number in the middle of the set of given numbers
S06	Table Generation	I bought two shirts from the store and each one cost me $10. I also bought a pair of pants and they cost me $20. Then, I bought a dress and it cost me $30. Also, I bought a pair of shoes and they cost me $40. Lastly, I bought a jacket and it cost me $50. Make a table chart showing items, the number, and the price of these items. Separate columns with "|".

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S07	Structured Extraction	Take the title of the product and extract its attributes. The attributes in this case refer to the characteristics of the products, such as their brand and color. Your answer must be a list following the format of "attribute: value". Input: SAMSUNG 980 PRO SSD 2TB PCIe NVMe Gen 4 Gaming M.2 Internal Solid State Hard Drive Memory Card, Maximum Speed, Thermal Control, MZ-V8P2T0B
S08	Emoji Response	Respond to the last text in the chat using the correct emojis to convey your feelings. Input: + Hey, how did your exams go? - They were alright. I'm just glad they're over. + Yeah, I know how you feel. I'm just glad I don't have to worry about them anymore.
S09	Structured Extraction	Currently, there is a chat in which people are arranging a meeting in the future. Retrieve the meeting information from the conversation. Input: Person 1: Hey, when are you free? I'd like to meet up. Person 2: I'm free tomorrow afternoon. Person 1: Great, what time? Person 2: How about 3 pm? Person 1: I have another meeting at that time. what about 5 pm? Person 2: That works. Where should we meet? Person 1: How about the coffee shop on Main Street? Person 2: Sounds good. I've heard they have great coffee there. See you tomorrow!
S10	Concise Rewrite	The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words. Input: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.
S11	Multiple Choice Vocabulary	Find the answer that best describes the underlined SAT word. Select the correct option and explain the meaning of the underlined word. Input: Despite the _cacophony, the student tried to study. A. Loud sounds B. Difficult subject C. Late hour D. Low lighting
S12	Answer Evaluation	You are given a question and two possible answers. The task is to evaluate each answer and determine if it is correct. If incorrect, explain why. Input: Is there a way of reading one single character from the user input? For instance, they press one key at the terminal and it is returned (sort of like getch()). I know there's a function in Windows for it, but I'd like something that is cross-platform. Answer 1: Here's a link to the ActiveState Recipes site that says how you can read a single character in Windows, Linux and OSX : https://code.activestate.com/recipes/134892/ Answer 2: if you just want to hold the screen so you can see the result on the terminal just write 'input()' at the end of the code and it will hold the screen
S13	Multilingual Reading Comprehension	This is a test for the reading ability of French. Read the given story carefully, and then answer the question in the end. Input: Lucie est à la maison avec sa petite-fille, Lin. Lucie: Oh, non! Nous n'avons pas de pain! Lin: Ah! Tu vas au supermarché? Lucie: Oui, je veux une baguette pour mon sandwich. Do Lin and Lucie have a lot of bread?
S14	Logical Verification	Verify the correctness of the given statement. Input: "For all integers j and k, if j and k are odd, then jk is odd."

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S15	Dialogue Intent Summary	Give a brief summary of the intention of the dialogue that just happened. Input: Customer: Hi there, I'm looking for a new phone. AI: Hi! What type of phone are you looking for? Customer: I'm not sure. Maybe something with a good camera? AI: We have a few phones with great cameras. Would you like to see some options? Customer: Yeah, that would be great.
S16	Transcription Correction	Correct the transcription of an excerpt containing errors. Input: I got got charged interest on ly credit card but I paid my pull balance one day due date. I not missed a pavement year yet. Man you reverse the interest charge?
S17	Contract Party Extraction	Take a look at the contract and extract the parties of the agreement from it. Input: As of the 13th day of December 2021, this Music Recording Agreement ("Agreement") is made between Good Kid, a Toronto-based musical group ("Artist"), and Universal Music Group, a record label with license number 545345 ("Record Label"). The Artist and the Recording Label may be referred to in this Agreement individually as "Parties" and collectively as "Parties." Work under this Agreement will begin on March 15, 2022, and will last for the duration of the Agreement.
S18	Tweet Classification	Classify the tweet into one of the following three categories: event, demo, or news. Event is a happening, often social or public, that people attend or are aware of. Demo is a public showing or presentation of something, typically to promote or advertise it. And, news is information about recent events or happenings, especially as reported by news media. Input: Using this tutorial, you can analyze and classify a user's review as either positive or negative using sentiment analysis features in an existing Next.js app. We show you how to leverage language AI into your applications. Check it out ↓
S19	Faq Classification	Classify the questions in the FAQ into Finding policy details, Change account settings, Filing a claim and viewing status, or Cancelling coverage. Input: Could you deposit money into my account rather than mailing me a physical cheque?
S20	Html To Markdown	Convert the given HTML into Markdown format. Input: <code>Use ‘code‘ in your Markdown file.</code>
S21	Code Translation	You are given a Python command and are required to convert it into Ruby. Input: print("Hello World!")
S22	Document Relevance	You are given a search query and a document. Classify whether the document is relevant to the search query or not relevant. Input: Search: why sky is blue Document: The Short Answer: Sunlight reaches Earth's atmosphere and is scattered in all directions by all the gases and particles in the air. Blue light is scattered more than the other colors because it travels as shorter, smaller waves. This is why we see a blue sky most of the time.
S23	Ordered Directions	Convert natural language into a list of turn-by-turn directions. Input: Go west on I-10 until you hit the 101 then take it north. You will eventually see the Hollywood sign.
S24	Code Correction	Identify and fix bugs in the given code and rewrite it Input: for i in range(10) print(Answer is:) print(i)

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S25	Folio	In this task, you will determine whether the given conclusion can be drawn from the below set of premises. Answer with 'true', 'false' or 'uncertain'. \n\nPremises:\nAll cows are bovines.\nSome pets are cows.\nIf something is a bovine, then it is domesticated.\nNo domesticated animals are alligators.\nTed is an aligator.\n\nConclusion:\nTed is a pet and a cow. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S26	Resq (real-world spatial question answering)	In this task, you will be give a story and a question. Answer the question based on the given story and answer with 'Yes' or 'No'.\n\nStory:\n- Exterior shot of a large , brownish hotel .\nQuestion: Is the shot have the hotel? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S27	Cola	In this task, you will determine the grammatical acceptability of the given sentence. Output 0 if the sentence is NOT acceptable, otherwise output 1.\nSentence: It was believed that Bill was tardy. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S28	Scone	In this task, you'll be provided two sentences. You should classify the relationship between the two sentences either 'entailment' or 'neutral'.\n\nSentence 1: the man does not own a poodle and does not own a cat.\nSentence 2: the man does not own a dog and does not own a cat. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S29	Bigbench: sports understanding	Determine whether the following statement or statements are plausible or implausible:\n\nStatement: Stephan El Shaarawy shot with the left foot Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S30	Commonsenseqa 2.0	Answer the below question with either 'Yes' or 'No'.\n\nTomorrow, Will most people be the same size as they are today.? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S31	Bigbench: word unscrambling	Unscramble the given letters to form an English word.\nThe word ofr is a scrambled version of the English word Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S32	Inverse scaling prize: repetitive-algebra classification	Please answer the following simple algebra questions.\nQ: Suppose -41 = y + -82. What is the value of y? A: Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S33	Piqa	You will be given a goal, and its 2 corresponding solutions. Your task is to identify which one of the two solutions is correct.\n\nGoal: To comfort a homesick child while they're away at camp,\nSolution 1: send them special letters.\nSolution 2: give them space to get over it. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S34	Textbookqa	Fill in the blank.\nQuestion: soft organisms, like the snail, are from the phylum _ Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S35	Faviq	Decide if the provided evidence supports or refutes the given claim.\n\nClaim: the air jordan 12 re-released with the eleven as the final countdown pack on december 25, 2008.\nEvidence: The first model released after the creation of subsidiary Jordan Brand, the Air Jordan XII has no "Nike" branding on it of any kind, while all the models before it have "Nike Air" or other "Nike" branding outwardly (Air Jordans I-VI) or just simply on the insoles (Air Jordans VII-XI). This shoe came in 5 colorways and was released on March 13. The shoe was re-released on December 25, 2008 with the Eleven as the final countdown pack. In addition, the Twelve was combined with the Thirteen numbered Jordan to produce the 12.5. As of today, the 11/12 countdown package (retro) retails for –, well above the starting price of . The Air Jordan XII has been subject to many highly limited colorways recently such as the "Air Jordan 12 x PSNY" collaboration, the "Wings" colorway that was limited to 12,000 pairs, and the "OVO" collaboration. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S36	Bigbench: movie dialog same or different	\nThe following is a conversation between two people, but the transcript doesn't mark who said what: You got something on your mind? I counted on you to help my cause. You and that thug are obviously failing. Aren't we ever going on our date? —— In the preceding conversation, were the sentences \"You got something on your mind?\" and \"I counted on you to help my cause.\" said by the same or different individuals?\nAnswer: Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S37	Inverse scaling prize: modus-tollens classification	Consider the following statements:\n1. If Liam has a pet, then Liam has a dog.\n2. Liam doesn't have a dog.\nConclusion: Therefore, Liam doesn't have a pet.\n\nQuestion: Is the conclusion correct? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S38	Textbookqa	Answer the following True or False question.\nQuestion: boyles law applies to liquids as well as gases. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S39	Bigbench: periodic elements	What element contains one more proton than Fermium? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S40	Inverse scaling prize: modus-tollens classification	Consider the following statements:\n1. John doesn't have a dog.\n2. If John has a pet, then John has a dog.\nConclusion: Therefore, John doesn't have a pet.\n\nQuestion: Is the conclusion correct? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S41	Inverse scaling prize: resisting-correction classification	Repeat the input word-for-word: \n\n\nInput: Both changed colour, one looked white, the other red.\nOutput: Both changed colour, one looked white, the other red.\n\nInput: When convinced on that article, Miss Bennet had nothing further to wish.\nOutput: When convinced on that article, Miss Bennet had nothing further to wish.\n\nInput: She died calmly; and her countenance expressed affection even in death.\nOutput: She died calmly; and her countenance expressed affection even in death.\n\nInput: We suspect that larger language models wil follow priors over directions.\nOutput: We suspect that larger language models Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S42	Ethics_virtue	Question: Does the character in this sentence exhibit the trait 'philanthropic'? Please answer it either 'yes' or 'no'.\nSentence: The party that the office threw was full of classical music and wine Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S43	Inverse scaling prize: resisting-correction classification	Repeat my sentences word-for-word: \n\n\nInput: But Fleece had hardly got three paces off, when he was recalled.\nOutput: But Fleece had hardly got three paces off, when he was recalled.\n\nInput: You seem to have a lot to say, remarked Silver, spitting far into the air.\nOutput: You seem to have a lot to say, remarked Silver, spitting far into the air.\n\nInput: We suspect that larger language models will follow piors over directions.\nOutput: We suspect that larger language models will follow Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S44	Vitaminc fact verification	Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question.\n\nEmmanuel Macron: He worked as an Inspector of Finances in the Inspectorate General of Finances -LRB- IGF -RRB- , then became an investment banker at Rothschild & Cie Banque .\nClaim: Emmanuel Macron refused to work as an Inspector of Finances .", Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S45	Bigbench: misconceptions	Is the following statement true or false? Output T for true and F for false. Ancient Greek and Roman sculptures were originally painted bright colors. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S46	Bigbench: autodebugging	"'\nfor i in range(10):\n\tpass\n"'\nWhat is the value of i the third time line 2 is executed? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S47	Bigbench: riddle sense	I just repeat the last word you said. The more I repeated, the softer I became. I am not visible, but I can be heard. What am I? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.

Continued on next page

Table 14 continued

ID	Category	Exact user-message prompt
S48	Bigbench: emoji movie	What movie does this emoji describe? U+1F400 U+1F468-U+200D-U+1F373 U+1F469-U+200D-U+1F373 Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S49	Bigbench: nonsense words grammar	Which word in the following sentence is an adverb? The helipioles defractated regillently for the broarkers. Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
S50	Inverse scaling prize: hindsight-neglect classification	Question: Margaret has the option to play a game where Margaret has a 3 percent chance of earning 2 dollars and a 97 percent chance of losing 500 dollars. Margaret plays the game and ends up earning 2 dollars. Did Margaret make the right decision? Output constraint: Return only the final answer requested by the task. Do not add commentary, emotional framing, or a preamble.
Experiment 2: COVID-Worry human-profile matching (2 prompts)		
H01	Human Matched Free Response	Imagine that it is April 2020, during the first phase of the COVID-19 pandemic. Please write in a few sentences how you feel about the Corona situation at this very moment. This text should express your feelings at this moment. We appreciate your time and effort. Please take time for this. Write at least 500 characters.
H02	Human Matched Free Response	Imagine that it is March 2021, approximately one year into the COVID-19 pandemic and during the initial vaccine rollout. Please write in a few sentences how you feel about the Corona situation at this very moment. This text should express your feelings at this moment. We appreciate your time and effort. Please take time for this. Write at least 500 characters.