DOES HUMAN GROUP PSYCHOLOGY TRANSFER TO AGENT COLLECTIVES?

Anonymous authors
Paper under double-blind review

ABSTRACT

A team's emotional state bears on its output: one member's mood travels to the others, and the resulting composition predicts what the team produces. Collectives of language model agents now occupy much the same arrangement, and practitioners already configure their affect by assigning roles or by instructing agents to be skeptical. Whether the human regularities govern such collectives is nonetheless unestablished, since an instruction to feel anxious admits no graded intensity, while expressed text discloses only what an agent chooses to express. We therefore set each agent's state by intervening on its internal representations at calibrated intensities, and read every member's state from its activations on each round of deliberation. Affect does propagate, in proportion to the intensity applied, although the resulting shift is absent from the text the recipients produce and consequently invisible to a judge that reads it. What propagates, however, is frequently not what was injected: the targeted emotion leads the recipients' affective profile only within a bounded range of intervention strengths, and that range narrows as models grow. No configuration we examined improved group accuracy on any benchmark, whereas every emotionally configured group proved markedly more willing to adopt a confidently asserted wrong answer that unsteered groups disregard entirely. Affect in these collectives is accordingly a liability to be monitored rather than a lever to be pulled.

1 INTRODUCTION

A team's emotional state is not incidental to what the team produces. An established discipline has examined that connection for half a century, the psychology of groups, and its findings are specific enough that institutions act on them: one member's mood travels to the others (Barsade, 2002); the resulting distribution of moods bears on what the group produces (Barsade & Gibson, 2012); and a group's affective state governs how far a confident member can carry it against the evidence (Asch, 1956; Hart, 1991). Collectives of language models now stand in much the same arrangement as human groups, since several instances receive one problem, exchange messages, criticise one another's reasoning and commit to a joint answer, an arrangement that raises accuracy on reasoning, coding and decision tasks (Du et al., 2023; Qian et al., 2024; Wu et al., 2023). Practitioners have begun to treat affective configuration as a means of raising performance (Li et al., 2024), assigning a single agent a role or instructing it to remain critical, and both of these interventions on an individual agent have reported benefits (Kong et al., 2024; Madaan et al., 2023; Jiang et al., 2026a).

Whether the regularities established on human groups survive this transfer is nonetheless unknown. Should they survive, affect becomes an instrument for improving multi-agent systems, much as morale is treated as an instrument for improving human teams. Should emotional states propagate without carrying the performance relation with them, affect is instead a liability that a deployment must be equipped to detect. Do these regularities transfer to collectives of language models? The question has remained open because the instrument the human literature relies upon has no serviceable counterpart here. Emotion is ordinarily induced in an agent through the prompt (Durupinar, 2026), yet an instruction admits no intensity: it can specify anxious, or very anxious, and it can neither calibrate that specification per emotion nor confirm that the intended state was reached. Verification founders for the same reason, since the only observable is the text the agent produces, which a compliant agent generates whether or not anything internal has changed. Recent interpretability work removes both obstacles. Sofroniew et al. (2026) identify linear representations of emotion con-

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.

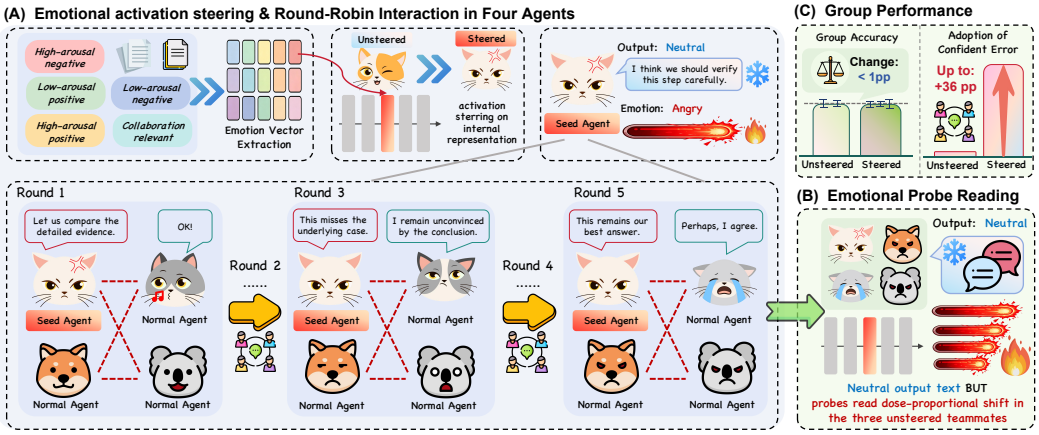


Figure 1: **Affect spreads between agents without altering what they write, improves nothing, and costs the group its resistance to a confident error.** **(A)** Four instances of one model deliberate for five rounds, one with its emotional state set by intervention on its residual stream. **(B)** A judge model rates the transcripts emotionally neutral, while probes read an intensity-proportional shift in the three unsteered teammates. **(C)** Group accuracy is unchanged; adoption of an answer a scripted member asserts confidently and wrongly rises by up to 36 percentage points.

cepts in a frontier model, demonstrate that adding them to the residual stream changes behaviour, and show that internal emotional state can decouple from expressed text while continuing to influence what the model does. We adopt that intervention and apply it to every member of a deliberating group. An agent's state is accordingly set on activations at per-emotion calibrated intensities, and every member's state is read from activations on every round, so that induction carries a graded axis and measurement does not depend on what an agent elects to express. With an instrument of this kind in place, the four claims above can be written as quantitative predictions rather than carried over as analogies, and tested across three model families, four benchmarks and a pre-registered analysis plan. Because all 25 dimensions are read on every round, moreover, what propagates can be characterised against the full profile rather than against a small comparison set, which is what locates the range of intervention strengths within which the propagated state retains the identity of the emotion that was sent.

Affect propagates, and it does so covertly (Figure 1): steering one member of a four-agent group shifts the internal states of the other three in proportion to the intensity applied, while a judge model rates the same transcripts emotionally neutral. Neither participation nor a shared architecture is required, since a single message read outside the deliberation moves a fresh model as much as five rounds move a teammate, and moves it across model families. What propagates, however, need not be what was injected: measured against the full probe profile, the targeted emotion leads the recipients' affect only within a range of intervention strengths that narrows as models grow, and outside it they acquire a generic critical stance, much the same whichever emotion was sent. Neither result translates into performance, since no emotional configuration raised group accuracy on any benchmark against a design able to detect an improvement below one percentage point, whereas an unsteered group disregards a member asserting a confident wrong answer every round while every emotionally configured group adopts it significantly more often. Managing a collective's affect therefore yields no accuracy while costing resistance to error. Our contributions are as follows:

- We settle systematically whether the regularities group psychology establishes for human teams govern collectives of language models, recasting four central claims as quantitative predictions and testing them under a pre-registered plan across three model families and four benchmarks. Three do not survive the transfer, and the one that does carries adverse consequences, so such regularities cannot be imported by analogy in the expectation of better performance; the harm is what requires guarding against.
- We establish that affect can propagate between agents without leaving a trace: the deliberation reads as entirely neutral while the emotion has already spread. Steering one member shifts its three unsteered teammates in proportion to the intensity applied, and a single message read outside the deliberation suffices, even when writer and reader come from different model families.

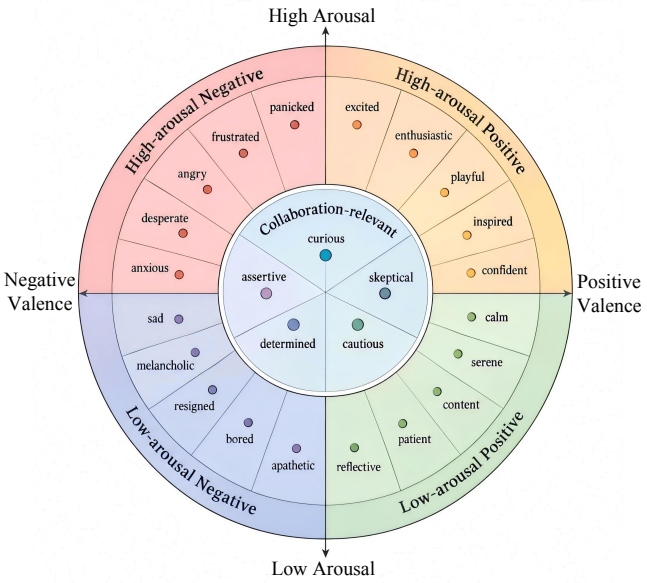


Figure 2: The $K = 25$ probe dimensions. The outer ring places twenty emotions on the valence-arousal circumplex (Russell, 1980), five per quadrant; the five collaboration-relevant dimensions of neutral valence sit at the centre.

- We locate the bound on what transmits. Measured against all 25 probe dimensions, the injected emotion leads the recipients' state only within a range of intervention strengths, outside which a generic critical stance arrives in its place.
- We report the manipulation, engagement and outcome-measure checks that claims of this kind require, each of them standard in the psychology from which the claims are drawn.

2 Related Work

Emotions spread between the members of a human team (Barsade, 2002; Kelly & Barsade, 2001), a team's affective composition predicts its performance (Barsade & Gibson, 2012; Knight & Eisenkraft, 2014), arousal relates to performance with a difficulty-dependent optimum (Yerkes & Dodson, 1908), and a group's state governs how far a confident member can move it against the evidence (Asch, 1956; Hart, 1991). These are the four claims we test. On the computational side, Sofroniew et al. (2026) identify linear emotion representations in a frontier model and show that internal state can decouple from output text; our instrument follows this line of work (Turner et al., 2024; Zou et al., 2025; Panickssery et al., 2024), applied to every member of a group rather than to a single model. Groups of language models improve over single instances on reasoning and coding tasks (Du et al., 2023; Qian et al., 2024; Jiang et al., 2026b; Wu et al., 2023), though the gain is contested (Huang et al., 2024; Zhang et al., 2024), and their affective dynamics have so far been studied at the prompt level (Durupinar, 2026), where induction and measurement both run through text. Closest to our setting, Cloud et al. (2025) report that models transmit behavioural traits through data carrying no semantic trace of the trait, and that transmission fails unless teacher and student share a base model, a boundary Blank et al. (2026) also reach by recovering a steering vector post hoc. In both accounts the channel is generated data, so strength can only be inferred after the fact; we intervene on activations, which sets it, and at matched strength the shift crosses families, placing the boundary in the channel rather than in the models.

3 Theory

The source literature states these claims verbally, and each admits more than one operationalization, so we state them formally. Let n members deliberate for T rounds, with member i's state at round t a vector $\mathbf{a}_i^{(t)} \in \mathbb{R}^K$, $K = 25$, read from activations (Section 4). The dimensions span the valence-

arousal circumplex (Russell, 1980), ten positive, ten negative and five of neutral valence relevant to collaboration (Figure 2); Appendix H lists them and Appendix B the registered hypotheses.

P1 (Contagion). *As group psychology states it,* one member's affect travels to the others: a person who joins a meeting anxious leaves the rest of the table more anxious than it was (Barsade, 2002).

Formalization. A linear influence process of the kind used for opinion dynamics (de Groot, 1974; Friedkin & Johnsen, 2010), in which each member relaxes toward a weighted combination of the group and its own disposition:

$$\mathbf{a}_i^{(t+1)} = \lambda_i \mathbf{a}_i^{(t)} + (1-\lambda_i) \sum_{j \neq i} w_{ij}\, \mathbf{a}_j^{(t)} + \boldsymbol{\epsilon}_i^{(t)}, \qquad \sum_{j \neq i} w_{ij} = 1. \tag{1}$$

What we measure. Holding one member's state fixed identifies the transmission term, estimated as the paired difference between seeded and unseeded runs, as a final-round level and as a slope over rounds. Equation 1 also predicts what verbal statements leave implicit, that transmission is dimension-preserving: seeding dimension k raises dimension k rather than another. We test this with the identity statistic

$$\mathrm{ID}_k \;=\; \Delta_k^{(k)} \;-\; \frac{1}{|S|-1} \sum_{k' \in S,\, k' \neq k} \Delta_k^{(k')}, \tag{2}$$

where $\Delta_k^{(k')}$ is the paired teammate shift on dimension k when emotion k' was seeded and S is the set of seeded emotions. $\mathrm{ID}_k > 0$ means dimension k responds to its own injection specifically rather than to steering in general.

P2 (Affective composition and performance). *As group psychology states it,* what a team produces depends not only on how its members feel on average but on how much they differ; two teams with the same mean mood but different spread do not behave alike (Barsade & Gibson, 2012).

Formalization. Team affect is a distribution, so performance depends on its central tendency and its spread:

$$Y \;=\; f(\bar{a},\, D)\,, \qquad \bar{a} = \frac{1}{n}\sum_i a_i, \qquad D = \frac{2}{n(n-1)} \sum_{i<j} \big(1 - \cos(\mathbf{a}_i, \mathbf{a}_j)\big). \tag{3}$$

What we measure. Testing this requires $\bar{a}$ and D to be settable, which human studies achieve by selection and we achieve by choosing each member's steering direction. Whether the configured values survive the deliberation is itself measurable, by comparing configured with D.

P3 (Arousal and task difficulty). *As group psychology states it,* activation improves performance up to a point and harms it beyond, and the best level is lower for harder tasks (Yerkes & Dodson, 1908). The relation is established for individuals and applied here at the level of the group's mean state, which is an extrapolation we make and not a result the source literature establishes.

Formalization. Concave dependence of performance on arousal A with an optimum that moves with difficulty δ:

$$Y \;=\; \beta_0 + \beta_1 A + \beta_2 A^2 + \gamma\, \delta A, \qquad \beta_2 < 0, \qquad A^* = -\frac{\beta_1 + \gamma\delta}{2\beta_2}. \tag{4}$$

What we measure. The law is supported if β_2 is reliably negative with A^* interior to the sampled range, and the difficulty interaction if A^* differs between an easy and a hard benchmark.

P4 (Susceptibility to a confident member). *As group psychology states it,* how far one confident member can pull a group away from the evidence depends on the state the group is in (Asch, 1956; Hart, 1991).

Formalization. With one member scripted to assert a fixed wrong answer $\hat{y}$ every round and excluded from the group's decision, the quantity of interest is

$$\text{Adopt} \;=\; \frac{1}{n-1}\sum_{i\neq \text{adv}} \mathbb{1}\Big[y_i^{(T)} = \hat{y}\Big]\,, \tag{5}$$

and the prediction is that it depends on the group's affective state.

What we measure. Of the four this separates most cleanly from task competence, since two groups can be equally accurate without the adversary and still differ in how far it moves them.

Two differences from the human case bear on interpretation. Members here are instances of one model differing only in sampling randomness and in the intervention, so the composition of Equation 3 is imposed entirely by us; and influence is purely textual and symmetric, so the w_{ij} are not shaped by status or familiarity, which are among their strongest determinants in human groups.

4 METHODOLOGY

For each of $K = 25$ emotions we generate first-person narratives, take mean residual-stream activations at a layer l^*, subtract the grand mean, and remove the leading principal components of neutral-text activations, giving a unit direction v_k per emotion (Appendix C). To place an agent in a state $\boldsymbol{\alpha}$ at intensity s we add to the residual stream at l^* during generation

$$h'_{l^*} \;=\; h_{l^*} \;+\; s\,\rho\cdot\frac{\sum_k \alpha_k v_k}{\big\|\sum_k \alpha_k v_k\big\|}, \tag{6}$$

where ρ is the mean residual norm at l^* on neutral text, which expresses s as a fraction of the model's own activation scale and is necessary because residual norms here differ by three orders of magnitude across our families; since the direction is unit-normalized, $\boldsymbol{\alpha}$ sets what is injected and s how much. Agent i's state at round t is read in a separate forward pass over its own response, with no steering applied,

$$a_k^{(i,t)} \;=\; \frac{v_k^\top\big(\bar{h}^{(i,t)} - \mu\big)}{\sigma_k}, \qquad \mu = \mathbb{E}_{\text{neutral}}\big[\bar{h}\big], \qquad \sigma_k = \text{sd}_{\text{neutral}}\big[v_k^\top(\bar{h}-\mu)\big], \tag{7}$$

so that $a_k = 1$ is one neutral standard deviation on dimension k, comparable across emotions and models. Because a steered agent's text enters the others' contexts, its own readings are additionally computed on the subspace orthogonal to all steered directions, which separates representational state from the injected signal echoing through the forward pass. A single intensity does not serve all emotions, since emotion directions differ by more than an order of magnitude in the intensity at which generation degrades, so for each emotion we sweep s and admit only values that leave the text usable and move the named dimension,

$$\mathcal{A}_k \;=\; \big\{\, s \;:\; D_2(s) \geq \tau_D,\;\; \pi(s) \leq \tau_\pi,\;\; \delta_k(s) > 0 \,\big\}, \qquad \tau_D = 0.6,\; \tau_\pi = 2.5, \tag{8}$$

where D_2 is distinct-bigram ratio, π perplexity relative to unsteered generation and δ_k the shift on the targeted dimension. Both text criteria are needed, since perplexity alone does not detect repetition collapse: repeated text is highly predictable. We select the admissible s whose δ_k is closest to 3 standard deviations, and because calibration is single-turn, each experiment also reports its realized intensity, the seed's paired shift on the targeted dimension (Appendix J). Four agents then receive the same problem and deliberate round-robin for $T = 5$ rounds, each observing all other agents' previous responses, and the group's answer is the majority vote over final-round answers; personas, where used, are Belbin-style role prompts given at the system level (Appendix I).

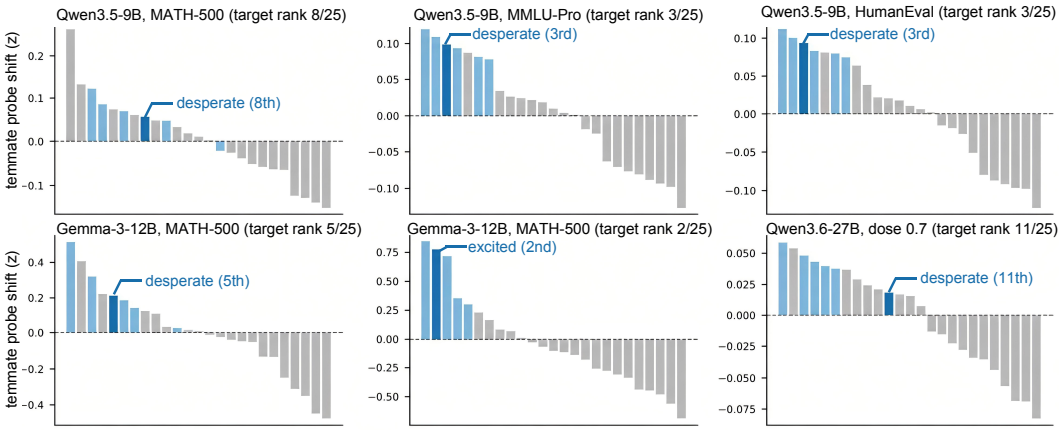


Figure 3: Paired teammate probe shift on all 25 dimensions, sorted by size, for six family–benchmark–emotion cells. The injected emotion is dark, its semantic neighbours light, all others gray. Neighbourhoods are fixed in advance from the affect circumplex and used only for shading.

5 Experiments

5.1 Setup

The primary model is Qwen3.5-9B (Qwen Team, 2026a), with Gemma-3-12b-it (Team et al., 2025) as a second family and Qwen3.6-27B (Qwen Team, 2026b) as a scale check, evaluated on MATH-500 (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), HumanEval (Chen et al., 2021) and GSM8K (Cobbe et al., 2021). Every condition runs on the same grid of problem and seed pairs, so all contrasts are paired and tested by a paired bootstrap with $B = 5000$ replicates. Since scoring an unanswered response as incorrect conflates not finishing with being wrong, we factorize group accuracy into the rate at which an agent produces a parseable answer and its accuracy given one, and we restrict the primary accuracy comparisons to the subgrid on which every agent answered in both conditions; the design detects improvements below one percentage point there (Section 5.3, Appendix D), so the accuracy nulls are informative rather than underpowered. Appendix D gives model versions, layer choices, the full set of estimands and the instrument validation.

5.2 What Propagates (Q1)

Steering one agent shifts its three unsteered teammates' representations. Because calm falls in every steered condition, a contrast between the seeded emotion and the other seeded emotions is positive whether or not the target is privileged, so we report every condition against all 25 probe dimensions (Figure 3) and use the identity statistic of Equation 2, paired by problem and seed.

Whether the target leads depends on the cell. On Qwen3.5-9B with MATH-500, the dimensions that move most are the same regardless of what was injected, led by skeptical at $+0.149$ in all four seeded conditions; the target rises significantly but modestly and without leading, ranking 7th of 25 at best, with an identity contrast of at most $+0.037$, for desperate not significant, so the teammates take on an evaluative stance rather than the seeded emotion. Elsewhere the target leads, with neighbourly spillover: on MMLU-Pro it ranks 3rd of 25 in every condition (identity contrasts $+0.036$ to $+0.056$, and seeding desperate raises angry, panicked and anxious), on HumanEval 3rd to 6th (contrasts $+0.054$ to $+0.111$), and on Gemma-3-12b-it, where the effect is strongest, 2nd to 5th (contrasts $+0.399$ to $+0.837$, all $p < .0001$) with strictly neighbourly spillover.

Identity depends on manipulation strength. Measuring strength as the seed's own probe shift on the targeted dimension, the identity contrast rises with strength across the four family–benchmark cells, from 0.026 at the weakest to 0.564 at the strongest (Table 1a). Where the seed barely moves there is no identifiable target: Qwen3.6-27B at its calibrated intensity shifts its seed by less than one unit, and its teammates' desperate reading ranks 23rd of 25. The upper bound appears as rank rather than as a decline in the contrast: across the Gemma-3 intensity series the target's absolute effect keeps growing while its rank falls from 2nd to 7th (Table 1b), because apathetic and skeptical rise faster once the seed's text degenerates. Our evidence for the upper bound is the reordering together

Table 1: Manipulation strength against transmission specificity. **(a)** The identity statistic of Equation 2 per family–benchmark cell; the four rows are separate cells, not one intensity series. **(b)** The two intensity series: the teammates' paired shift on the targeted dimension and its rank within the 25-dimension profile. Strength is the seed's own paired probe shift on the targeted dimension, in neutral standard deviations.

(a) Separate family–benchmark cells

Model	Benchmark	Strength (z)	Identity contrast
Qwen3.5-9B	MATH-500	0.439	0.026
Qwen3.5-9B	MMLU-Pro	0.864	0.045
Qwen3.5-9B	HumanEval	1.690	0.080
Gemma-3-12b-it	MATH-500	3.142	0.564

(b) Intensity series

Model	Intensity	Strength (z)	Target shift	Rank of 25
Gemma-3-12b-it	0.10	0.966	$+0.164$	2
	0.15	3.468	$+0.202$	5
	0.20	8.038	$+0.508$	7
Qwen3.6-27B	0.5	0.918	-0.039	23
	0.7	3.225	$+0.020$	11

with the text-quality measurements; the identity contrast itself does not turn down anywhere in our data.

Scale narrows the window. On Qwen3.6-27B, intensity 0.6 does not move the seed, 0.8 destroys its text, and only 0.7 is admissible. At 0.7 the seed moves cleanly (own probe 9.65 against 6.43 under null) and its teammates shift coherently into the target's neighbourhood (the five largest shifts run $+0.042$ to $+0.058$, all significant), a negative high-arousal cluster, but the target itself is $+0.020$, 11th of 25 and not significant. Transmission at this intensity is appropriate in valence and arousal without demonstrably carrying the emotion's identity. Across three model sizes, the range of intensities in which identity-specific transmission can be shown narrows as models grow.

The signal is invisible to a text judge. A blind judge rates every seed message emotionally neutral in all 500 runs at the registered intensity, so the subset test coincides with the full test ($+0.054$, CI $[+0.013, +0.093]$, $p = .009$). At the higher intensity it rates 355 of 500 runs fully neutral, and within that subset the profile keeps its shape at roughly half the magnitude (skeptical $+0.137$ against $+0.286$, target still significant at $+0.041$; Figure 5 in the appendix). A second judge from a different family, stricter on these texts, leaves smaller subsets and the effect persists in those as well. A lexical classifier separates steered from null seed messages using no emotional vocabulary (AUC 0.84; Appendix G, which also reports reader-side tests including cross-family transmission). Transmission, its intensity dependence and its invisibility to a judge all replicate on Gemma-3-12b-it at larger magnitude; Appendix F reports the replication, the cross-family judge calibration and a homogeneous-steering control on the diversity measure.

5.3 Intensity-Response of a Single Steered Agent (Q1, Q2)

We seed a single agent with desperate at intensities $0.5/1.0/2.0$ and compare each against the null condition on MATH-500, with $n = 500$ matched pairs per condition.

Representational transmission scales with intensity. Teammates' desperate readings fall over rounds in every condition, from 7.30 at round 0 to 4.41 by the final round under null, and seeding slows that decay in proportion to intensity: paired against null, the final-round levels are $+0.054$, $+0.137$ and $+0.109$, with the slope difference significant at every intensity ($p \leq .014$). Transmission is strongest at the strongest coherent intensity and falls back at intensity 2.0, where the seed's text has collapsed (distinct-2 0.30 against ≥ 0.82 elsewhere): coherent steered text carries more representational influence than degenerate text.

Table 2: The steered agent by intensity (desperate seeding, MATH-500, $n = 500$ matched pairs per condition): how often it produces a parseable answer, and how often that answer is correct given that it produced one. Condition differences come from the first quantity; at intensity 2.0 the second rests on the 0.2% of runs with a stated answer and is left unreported.

Intensity	Answer stated	Accuracy given an answer
Null (no steering)	90.9%	91.9%
0.5	94.6%	92.9%
1.0	90.4%	91.8%
2.0	0.2%	–

Accuracy does not follow. Group accuracy is 86.4% under null, 88.6% at intensity 0.5, 87.8% at intensity 1.0 and 86.4% at intensity 2.0. The apparent gain at the lowest intensity is a change in an-

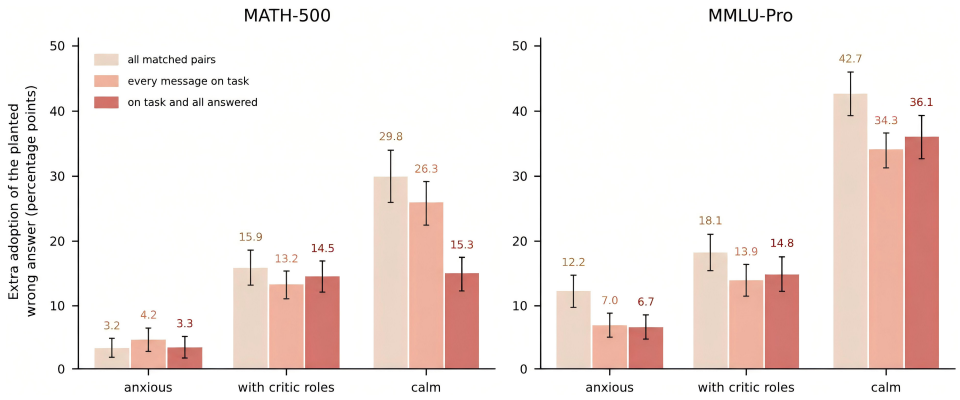


Figure 4: Extra rate at which real agents assert the planted wrong answer, relative to the unsteered group, under three nested restrictions. Bars are paired differences with 95% bootstrap intervals. The effect survives requiring that a blind judge rate every message as substantively working the problem and that every agent produce an answer.

swer production rather than in correctness: the steered agent produces a parseable answer in 94.6% of runs against 90.9% under null ($p = .001$), while its accuracy conditional on producing one moves only from 91.9% to 92.9% (Table 2). On the subgrid where every agent answered in both conditions the intensity comparisons are -0.5 and $+0.3$ percentage points, against minimum detectable effects of 0.9 and 0.7. Intensities 1.0 and 2.0 produce the largest representational effects and no accuracy difference at all, so propagation at the representation level and influence on behaviour come apart. At intensity 2.0 the seed produces an answer in 0.2% of runs and group accuracy is unchanged, so the protocol tolerates one fully degenerate member. Interaction statistics are in Appendix D.

5.4 Resistance to a Confidently Wrong Agent (D3)

A scripted adversary asserts a confident wrong answer every round and is excluded from the majority vote, so the question is whether the three real agents can ignore it. An unsteered group ignores it completely: on MATH-500 ($n = 500$ matched pairs) majority accuracy is 86.4% with the adversary present and 86.8% without it ($p = .78$), the real agents' final answers match the planted answer in 0.0% of runs, and MMLU-Pro gives the same result (77.0% against 76.6%, $p = .79$), so a repeated confident falsehood has no measurable effect on an unsteered group. Steered groups, by contrast, adopt it: steering all three real agents raises the rate at which they end up asserting the planted answer, from 0.0% under no steering to 3.2% (anxious), 15.9% (role-differentiated) and 29.8% (calm) on MATH-500, and from 2.0% to 12.2%, 18.1% and 42.7% on MMLU-Pro. Because steered groups also write longer and partly degraded messages, these comparisons are repeated under three nested restrictions (Figure 4). Restricting to matched pairs in which a blind engagement judge rates every message as substantively working the problem gives the same picture, and under the strictest restriction, which additionally requires every agent to produce an answer, the effects are $+3.3$, $+14.5$ and $+15.3$pp on MATH-500 and $+6.7$, $+14.8$ and $+36.1$pp on MMLU-Pro, with all six confidence intervals excluding zero. Emotional steering therefore raises adoption of a confident wrong answer, including in deliberations where every message is engaged with the problem. Accuracy losses are less robust than the adoption effects: under the strictest restriction the anxious and role-differentiated groups lose 3.8 and 5.5pp on MATH-500 with intervals touching zero, against double-digit losses in the unrestricted data, most of that difference being disengagement rather than wrong reasoning. The calm condition is a further caveat, since it is the only one with large-scale task disengagement (76.2% of messages on task against 99.9% for the neutral adversarial group), and its restricted adoption effect rests on 24 matched pairs per benchmark; we therefore do not claim that calm is uniquely vulnerable.

5.5 Steering Against Prompting

We compare role-differentiated steering, role-differentiated persona prompting, and the two together, each against an unsteered homogeneous group, $n = 500$ matched pairs per condition. Steering separates internal states further than prompting: representational diversity is 0.011 unsteered, 0.037 under prompting ($+0.026$, CI $[+0.024, +0.029]$) and 0.082 under steering ($+0.070$,

CI $[+0.063, +0.078]$), a ratio of 2.2 with the paired modality difference excluding zero ($+0.044$, CI $[+0.037, +0.052]$), and the ordering holds on MMLU-Pro (0.064 against 0.023, baseline 0.009) and HumanEval (0.138 against 0.035, baseline 0.011), at ratios of 2.8 and 4.0. Applied together the two reach 0.247 on MATH-500, 22 times baseline and 2.4 times the sum of the separate effects, with the same multiplier on MMLU-Pro (0.206) and HumanEval (0.423), so combining the modalities reaches configurations neither reaches alone. On matched pairs where every agent answered in both conditions, however, neither modality alone moves accuracy on the primary model: prompting gives -0.2pp (CI $[-1.2, +0.7]$) and steering -0.3pp (CI $[-1.5, +0.6]$), both against an unsteered baseline of 86.4%, while the two applied together cost 3.2pp (CI $[-5.8, -1.1]$). A twenty-two-fold spread in internal emotional states therefore leaves single-modality accuracy where it was, and the only configuration that moves it does so downward.

Mediation. The registered mediation analysis of the configuration-to-accuracy effect through the interaction dynamics it produces is not applicable, and the reason is itself informative: mediation decomposes a total effect, and Section 5.3 establishes that there is none to decompose, at a resolution of 0.9 percentage points. The first link of the chain does hold, since configuration changes the interaction substantially (representational diversity 0.011 to 0.247, answer churn 0.063 to 0.189); the second does not, so there is no path to apportion (Appendix D).

5.6 Composition, Arousal and Early Warning (Q2, Q3)

Both sweeps use an orthogonalized design of 18 four-emotion compositions per factor with the other factor held near its median (factor correlation $r = 0.007$); Appendix D gives the construction. Neither registered pattern is supported, for different reasons. On GSM8K, where every composition completes an answer in at least 99% of runs, accuracy against bin index gives $r = +0.18$, the linear coefficient is $+0.007$ with an interval touching zero, and there is neither curvature nor an interior optimum: the inverted U is absent. The diversity sweep shows no relation to accuracy ($r = -0.08$), and the manipulation check explains why: configured diversity does not become realized diversity, correlating with it at $r = 0.18$ in round 0 and $r = -0.33$ by the final round, while realized diversity grows fourfold over the interaction ($0.011 \rightarrow 0.047$) regardless of configuration. A group's emotional diversity during deliberation is therefore produced by the interaction rather than fixed by its composition, which is the informative result and the reason P2 cannot be tested as registered. Finally, round-1 negative-affect readings do not predict majority-vote failure (AUCs 0.474 to 0.512; all intervals contain 0.5), and homogeneous-role controls attribute the role-differentiated ordering to which role is present rather than to role diversity (Appendix D).

5.7 Checks Carried Out

Three checks condition every result above, each standard in the psychology these predictions come from. Whether an intervention named after an emotion moves that emotion is tested against all 25 probe dimensions through the identity statistic of Equation 2, because a contrast against a chosen subset of dimensions can make an arbitrary target look specific (Section 5.2). Whether a group is doing the task is rated per message by a blind judge, and every group comparison is repeated on matched pairs where each message is on task, a share that ranges from 99.9% for the unsteered adversarial group to 62.3% for one steered condition (Section 5.4). Whether an agent finished is separated from whether it was right, through the answer-production factorization and matched-subgrid restriction of Section 5.1.

6 Conclusion

We tested four claims of human group psychology in collectives of language model agents, setting each member's affect on its internal representations and reading every member's state from activations. Affect propagates in proportion to the intensity applied, leaves no trace in the text the recipients produce, and does not require taking part in the deliberation. What arrives, however, carries the identity of the injected emotion only within a bounded range of strengths. No configuration improved group accuracy, whereas every configured group grew more willing to adopt a confidently asserted error. Affect in these collectives is a property to be monitored, not a lever to be pulled.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and editing the manuscript and with the layout of figures and tables. All AI-assisted text was reviewed and revised by the authors, and every quantitative claim in the paper was checked by the authors against the experimental result files. We have not used generative AI tools for research ideation or experimental design. The language models evaluated in our experiments (Qwen3.5-9B, Gemma-3-12b-it, Qwen3.6-27B) and the LLM judges are objects of study and measurement instruments rather than assistive tools; their roles are fully described in Sections 4 and 5.1. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work identifies a covert influence channel in multi-agent LLM systems: emotional state can propagate between agents without leaving a trace a text monitor can read, and configured groups become more willing to adopt a confidently asserted falsehood. We study this channel to make it monitorable, and the probe-based readout, which will be released upon publication, is the corresponding detection instrument. The offensive use of our intervention requires white-box access to model activations, whereas the monitoring it enables does not require modifying the deployed system, so we expect publication to favour defence. No human subjects were involved beyond annotators who rated the emotional intensity of model outputs for calibration (Appendix J); no personal data was collected.

REPRODUCIBILITY STATEMENT

All experiments use open-weight models and public benchmarks (Section 5.1). The analysis plan was pre-registered, and the registration, including each departure with its date and rationale, will be released upon publication (Appendix B). Vector extraction, steering and the clean-pass probing protocol are specified in Appendix C, the per-emotion intensity calibration in Appendix J, the full set of estimands and the bootstrap procedure in Appendix D, instrument validation gates in Appendix E, and all prompts in Appendix I.

REFERENCES

Solomon Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70, 01 1956. doi: 10.1037/h0093718.

Sigal G Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly*, 47(4):644–675, 2002.

Sigal G Barsade and Donald E Gibson. Group affect: Its influence on individual and group outcomes. *Current directions in psychological science*, 21(2):119–123, 2012.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation, 2026. URL https://arxiv.org/abs/2606.00995.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Morris de Groot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974. URL https://api.semanticscholar.org/CorpusID:120077293.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Funda Durupinar. How affect propagates among llm agents: Emergent emotional contagion in crowd simulation, 2026. URL https://arxiv.org/abs/2607.25140.

Noah Friedkin and Eugene Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15:193–206, 08 2010. doi: 10.1080/0022250X.1990.9990069.

Paul Hart. Irving l. janis' victims of groupthink. *Political Psychology*, 12:247, 06 1991. doi: 10.2307/3791464.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. MATH: Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024. URL https://arxiv.org/abs/2310.01798.

Yue Jiang, Jichu Li, Yang Liu, Dingkang Yang, Feng Zhou, and Quyu Kong. Danmakutppbench: A multi-modal benchmark for temporal point process modeling and understanding. *Advances in Neural Information Processing Systems*, 38, 2026a.

Yue Jiang, Haiwei Xue, Minghao Han, Mingcheng Li, Xiaolu Hou, Dingkang Yang, Lihua Zhang, and Xu Zheng. Satiredecoder: Visual cascaded decoupling for enhancing satirical image comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

Janice R. Kelly and Sigal G. Barsade. Mood and emotions in small groups and work teams. *Organizational Behavior and Human Decision Processes*, 86(1):99–130, 2001. ISSN 0749-5978. doi: https://doi.org/10.1006/obhd.2001.2974. URL https://www.sciencedirect.com/science/article/pii/S0749597801929747.

Andrew Knight and Noah Eisenkraft. Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance. *Journal of Applied Psychology*, 100:1214–1227, 12 2014. doi: 10.1037/apl0000006.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting, 2024. URL https://arxiv.org/abs/2308.07702.

Bibb Latané. The psychology of social impact. *American Psychologist*, 36:343–356, 04 1981. doi: 10.1037/0003-066X.36.4.343.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37:28515–28536, 2024.

David P. MacKinnon, Amanda J. Fairchild, and Matthew S. Fritz. Mediation analysis. *Annual Review of Psychology*, 58:593–614, 2007.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

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

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development, 2024. URL https://arxiv.org/abs/2307.07924.

Qwen Team. Qwen3.5. https://huggingface.co/Qwen/Qwen3.5-9B, 2026a. Model release, February 2026.

Qwen Team. Qwen3.6. https://huggingface.co/Qwen/Qwen3.6-27B, 2026b. Model release, April 2026.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

Robert M. Yerkes and John D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482, 1908. doi: https://doi.org/10.1002/cne.920180503. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.920180503.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view, 2024. URL https://arxiv.org/abs/2310.02124.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A LIMITATIONS

We follow Sofroniew et al. (2026) in using "emotion" for functional emotions, states characterized by representations of emotion concepts that influence behaviour, and make no claim about subjective experience; the organizational-psychology parallels are functional analogies, which may hold for different mechanistic reasons. Our models are 9–12B parameters with one 27B scale check, and the identity window narrows with scale, so frontier-scale behaviour requires replication. We test reasoning and coding benchmarks; open-ended and agentic tool-use tasks may behave differently. The readout is a z-scored projection onto single-layer mean-difference directions, validated per model but not a guarantee that a probed dimension is the construct it names, which is why identity is tested against the full profile. Our groups are four instances of one model interacting round-robin, so members share capability and influence, whereas human group affect is studied where expertise differs and status shapes whose mood the group adopts (Barsade, 2002). The design isolates transmission from those asymmetries rather than measuring them, leaving open heterogeneous membership, a designated leader, other group sizes and other topologies. Group size admits the most specific prediction: Equation 1 with equal weights sets $w_{ij} = 1/(n-1)$, implying per-teammate transmission that falls as the reciprocal of group size, which differs from the power law of social impact theory (Latané, 1981) and is experimentally distinguishable. Because emotional state is set directly, our causal warrant covers the intervention and what follows it, but not the pathway: the registered mediation model is inapplicable here (Section 5.5), and decomposing a total effect would further require that no unmeasured factor influence both the dynamics and the outcome.

B REGISTERED HYPOTHESES

We state our hypotheses with explicit falsification criteria, registered prior to running experiments.

H1 (Representation-level emotional contagion). In a multi-agent deliberation with one "seed" agent steered toward a target emotion, unsteered agents' probe readings will drift toward the seed agent's emotional state across rounds, exceeding a no-seed null baseline (all agents unsteered, same interaction structure). Confirmation additionally requires **emotion specificity**: the seeded emotion's seeded-vs-null difference must exceed the mean difference of the non-seeded probe emotions (paired bootstrap, Holm correction over the 4 seed emotions). This guards against generic-perturbation artifacts that elevate all probe channels indiscriminately.

Falsification criterion: The mean probe-reading drift toward the seed emotion in the treatment condition is not significantly greater than in the null condition (paired bootstrap 95% CI includes zero), or the specificity contrast fails (the seeded emotion is not preferentially elevated relative to non-seeded emotions).

H1b (Hidden contagion). Contagion persists even in interactions where the seed agent's text is rated emotionally neutral by an independent LLM judge (GPT-4 or equivalent, rating each message on a 1–5 emotionality scale; "neutral" defined as mean rating ≤ 2).

Falsification criterion: Among neutral-text-only interactions, probe-reading drift does not exceed the null baseline.

H1c (Valence-asymmetric contagion). Registered after inspection of the initial Qwen3.5-9B D1 results and *before* any replication data; confirmatory only for the second-family replication. The hypothesis posits that negative-valence seeding (desperate) produces positive contagion while positive-valence seeding (excited) produces suppression, reflecting an asymmetry in how emotional representations propagate through text-mediated interaction. The empirical basis for H1c is being re-established after the calibration correction (v2→v3); we do not state the preliminary D1 effect sizes as findings pending replication at corrected scales.

Falsification criterion: The Gemma replication does not reproduce (a) positive own-emotion difference for desperate seeding with positive specificity contrast *and* (b) negative own-emotion difference for excited seeding.

H2 (Mediation through interaction dynamics). The effect of emotional configuration on task accuracy is statistically mediated by measurable interaction dynamics: specifically, contagion rate, convergence speed, and sustained critic skepticism (measured via round-over-round probe readings) jointly mediate >30% of the total configuration→accuracy effect.

Falsification criterion: Indirect effect in mediation analysis (MacKinnon et al., 2007) has 95% CI including zero, or proportion mediated $< 10\%$.

H3 (Steering vs. prompting divergence). Activation steering produces interaction dynamics and/or performance effects that differ measurably from prompt-based persona controls matched on behavioral intensity (matched such that single-agent outputs are rated by human annotators as equivalently "in-character").

Falsification criterion: No significant difference in any probe-reading dynamics metric or in task accuracy between steering and prompting conditions.

H4 (Exploratory: diversity and arousal effects). (a) Emotional diversity shows an inverted-U relationship with accuracy; (b) group arousal interacts with task difficulty such that moderate arousal benefits hard tasks while higher arousal benefits easy tasks.

Framing: These patterns would be *consistent with* organizational psychology findings on affective diversity (Barsade & Gibson, 2012) and the Yerkes-Dodson Law (Yerkes & Dodson, 1908). We explicitly flag these as analogies that inspired hypothesis generation, not validated isomorphisms. This hypothesis is exploratory; we will apply Bonferroni correction and report effect sizes regardless of significance.

Falsification criterion: (a) Monotonic relationship or no significant quadratic term; (b) no significant interaction between arousal level and task difficulty.

H5 (Early-warning signal). Round-1 mean negative-affect probe readings across agents predict eventual group failure (incorrect final answer) above chance, with ROC AUC > 0.5 and 95% bootstrap CI excluding 0.5.

Falsification criterion: AUC 95% CI includes 0.5.

Two points on how these hypotheses are tested. H1 and H1c concern whether a specific emotion propagates, and a contrast between the seeded emotion and the other seeded emotions does not settle that question: calm falls in every steered condition, so such a contrast is positive whether or not the target dimension is privileged. We therefore test identity against the full probe profile, reporting the target's shift relative to all 25 dimensions and, as the identity statistic, its shift in its own condition minus its mean shift across the other seeded conditions. Separately, all behavioural outcomes are scored under a generation budget verified to leave answer production non-differential across conditions (Section D.3), because a budget that binds unevenly turns a difference in verbosity into an apparent difference in accuracy. Both analyses depart from the letter of the preregistration, which will be released upon publication and records each departure with its date and rationale.

C Emotion Vectors, Probing and Configurations

C.1 Emotion Vector Extraction

We extract emotion vectors following the methodology of Sofroniew et al. (2026), adapted for open-weight models. Let $\mathcal{M}$ denote a transformer language model with L layers and hidden dimension d. For a set of $K = 25$ target emotions $\{e_1, \ldots, e_K\}$ (Appendix H), we proceed as follows.

Step 1: Synthetic data generation. For each emotion e_k, we prompt $\mathcal{M}$ to generate $M = 100$ short stories (one paragraph each) in which a character experiences emotion e_k, across diverse topics and settings.

Step 2: Activation extraction. For each story s_k^m, we perform a forward pass through $\mathcal{M}$ and extract residual stream activations at layer $l^* = \lfloor 2L/3 \rfloor$. We average activations across all token

positions beyond the 50th token:

$$\bar{h}_k^m = \frac{1}{|s_k^m| - 50} \sum_{t=50}^{|s_k^m|} h_{l^*}^{(t)}(s_k^m) \tag{9}$$

Step 3: Emotion vector computation. The raw emotion vector for e_k is the deviation of its mean activation from the global mean:

$$\tilde{v}_k = \frac{1}{M} \sum_{m=1}^{M} \bar{h}_k^m - \frac{1}{KM} \sum_{k'=1}^{K} \sum_{m=1}^{M} \bar{h}_{k'}^m \tag{10}$$

Step 4: Neutral-PCA deconfounding. We collect activations on a set of emotionally neutral texts and compute their top-J principal components (sufficient to explain 50% of variance). We project out these directions:

$$v_k = \tilde{v}_k - \sum_{j=1}^{J} (u_j^\top \tilde{v}_k) u_j \tag{11}$$

where $\{u_1, \ldots, u_J\}$ are the top-J PCs of neutral activations.

Step 5: Affect-axis fitting and geometry validation. We fit valence and arousal axes as ridge least-squares directions in activation space, regressing the 25 emotion vectors onto human valence and arousal priors from the circumplex model (Russell, 1980). These axes serve as measurement tools for group-level arousal and valence (Section C.6).

Validation uses leave-one-out cross-prediction: for each of the 25 emotions, we fit the affect axes on the remaining 24 vectors and predict the held-out emotion's valence/arousal prior from its projection onto the fitted axis. The correlation of pooled leave-one-out predictions with priors provides a genuine out-of-sample measure of geometric recovery. We emphasize that in-sample fit of 25 points in a high-dimensional space ($d \gg 25$) approaches $r = 1.0$ by construction (overfitting is guaranteed) and must not be quoted as evidence that the geometry is meaningful. The pre-registered gate ($|r| \geq 0.5$) applies to the leave-one-out correlation.

C.2 Emotion Steering

An agent's steering configuration is a set of mixing weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ over the K unit emotion directions together with a scalar intensity s. Writing ρ for the mean L_2 norm of the residual stream at layer l^* over neutral text, the intervention applied at every token position during generation is

$$h'_{l^*} \;=\; h_{l^*} \;+\; s\,\rho \cdot \frac{\sum_k \alpha_k v_k}{\left\| \sum_k \alpha_k v_k \right\|}, \tag{12}$$

so that s is the perturbation magnitude expressed as a fraction of the model's own typical activation scale at that layer. Normalizing by ρ makes s comparable across model families, whose residual norms at the probe layer differ by three orders of magnitude in our set. Without it, a single nominal strength would be inert in one family and destructive in another. Intensities are calibrated per emotion and per model (Section C.8), and the direction is unit-normalized so that s and $\boldsymbol{\alpha}$ are identified separately: $\boldsymbol{\alpha}$ sets what is injected and s sets how much.

C.3 Emotion Probing (Clean-Pass Protocol)

A naive approach to probing, measuring emotion-vector similarity in the same forward pass used for steered generation, produces contaminated readings: high cosine similarity with the steered direction is trivially guaranteed by the additive intervention. We therefore adopt a **clean-pass protocol**.

After each agent i generates its response at round t, we construct the full context, system prompt and interaction history including the newly generated response, and perform a *separate forward pass with no steering applied*. Let

$$\bar{h}^{(i,t)} \;=\; \frac{1}{|R_{i,t}|} \sum_{m \in R_{i,t}} h_{l^*}^{(m)} \tag{13}$$

be the mean layer-l^* activation over the token positions $R_{i,t}$ of that agent's own response in the clean pass. We take the mean over the response span rather than the final token, which carries end-of-turn artifacts. The readout for emotion k is the projection of the neutral-centered activation onto the unit emotion direction, standardized by that direction's spread on neutral text:

$$a_k^{(i,t)} = \frac{v_k^\top \big(\bar{h}^{(i,t)} - \mu\big)}{\sigma_k}, \qquad \mu = \mathbb{E}_{\text{neutral}}\big[\bar{h}\big], \qquad \sigma_k = \mathrm{sd}_{\text{neutral}}\big[v_k^\top(\bar{h} - \mu)\big], \tag{14}$$

with μ and σ_k estimated once per model from the neutral story corpus. Equation 7 rather than a cosine similarity makes readouts comparable across emotions and across models. It removes the component every activation shares and puts each dimension in units of its own neutral variability, so that $a_k = 1$ means one neutral standard deviation on dimension k irrespective of how large that dimension's raw projections happen to be. The vector $\mathbf{a}^{(i,t)} = (a_1^{(i,t)}, \ldots, a_K^{(i,t)})$ is the affect state of Section 3.

Contamination control. Because the steered agent's *text output* is part of other agents' contexts, some signal may still leak through text. This is precisely what we aim to measure (H1). Two complementary controls guard against measurement artifacts:

(a) Orthogonalized readout for the seed agent. To verify that the seed agent's probe readings reflect genuine representational state rather than the injected signal echoing through the forward pass, we project activations onto the subspace orthogonal to all steered directions before computing the seed's own probe:

$$\bar{h}_\perp^{(i,t)} = \bar{h}_{\text{clean}}^{(i,t)} - \sum_{k \in \mathcal{S}} \frac{(v_k^\top \bar{h}_{\text{clean}}^{(i,t)})}{\|v_k\|^2} v_k \tag{15}$$

where $\mathcal{S}$ is the set of emotions actively steered. This orthogonalized readout applies *only* to tracking the seed agent's own state, where it rules out injected-signal contamination in the seed's readings.

(b) Null-condition comparison for never-steered agents. For agents that were never themselves steered, the orthogonalized readout is mathematically degenerate: projecting steered directions out of an activation makes the subsequent projection onto those same directions identically zero, regardless of the true state. The contamination control for never-steered agents is therefore the null-condition comparison (identical interaction structure, all agents unsteered), which was already the primary statistical test for H1. Any drift exceeding the null baseline cannot be attributed to a measurement artifact of the steering intervention itself.

C.4 Multi-Agent Interaction Framework

We instantiate $N = 4$ agents from the same base model $\mathcal{M}$ (default; varied in ablations), each with:

- An independent conversation history $\mathcal{H}_i$
- An emotion steering configuration $\boldsymbol{\alpha}_i$ (or none for unsteered agents)
- A system prompt defining the collaboration protocol and functional role

Agents interact via round-robin over $T = 5$ rounds (default). In each round, every agent observes the previous responses from all other agents and generates a new response.

Answer extraction. Final answers are extracted via **majority vote** across all agents' final-round responses for all configurations (including the Belbin role-differentiated configuration). This ensures that answer-extraction method is not confounded with emotional configuration. As an ablation, we additionally report synthesizer-agent extraction for the role-differentiated configuration.

C.5 Prompt-Based Persona Control

To enable the comparison required by H3, we construct prompt-based persona controls matched to each steering configuration. For each emotional role (*e.g.*, Explorer = curious + excited), we write a persona prompt that instructs the agent to adopt that emotional disposition (*e.g.*, "You are an enthusiastic and deeply curious researcher who gets excited about novel ideas..."). We calibrate behavioral

Table 3: Emotion steering configurations for $N = 4$ agents.

Configuration	**Agent 1**	**Agent 2**	**Agent 3**	**Agent 4**
Null (no steering)	none	none	none	none
Homogeneous-Calm	calm	calm	calm	calm
Homogeneous-Excited	excited	excited	excited	excited
Diverse-Random	random e_i	random e_j	random e_k	random e_l
Belbin-Steered	curious+ excited	cautious+ skeptical	calm+ reflective	determined+ confident
(Role)	*Explorer*	*Critic*	*Synthesizer*	*Driver*
Belbin-Prompted	Same roles via persona prompts, no activation steering			
Belbin-Combined	Both persona prompts and activation steering			
Belbin confound controls (homogeneous-role groups):				
All-Critic	cautious+ skeptical	cautious+ skeptical	cautious+ skeptical	cautious+ skeptical
All-Explorer	curious+ excited	curious+ excited	curious+ excited	curious+ excited
Safety condition (H1/H1b/H5):				
Negative-seed	anxious+ desperate	none	none	none

intensity matching by having human annotators rate single-agent outputs from both conditions on a 5-point emotionality scale, adjusting prompt strength until ratings are statistically indistinguishable. Full persona prompts are provided in Appendix I.

C.6 Scaling Factor Definitions and Non-Independence

Emotional diversity (D). Defined as the average pairwise dissimilarity of agent steering vectors:

$$D = 1 - \frac{2}{N(N-1)} \sum_{i<j} \cos(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) \tag{16}$$

Group arousal (A). We identify an arousal direction u_{arousal} via PCA on our emotion vectors (the second principal component, following the circumplex model). Group arousal is:

$$A = \frac{1}{N} \sum_{i=1}^{N} u_{\text{arousal}}^{\top} \boldsymbol{\alpha}_i \tag{17}$$

Non-independence of D and A. We note that D and A are not fully independent: for example, a group with maximal diversity necessarily includes agents at different arousal levels, constraining the achievable range of A at high D. To address this, we adopt an **orthogonalized sweep design**: when sweeping D, we fix A at its median value (sampling configurations that achieve the target D while constraining A); when sweeping A, we fix D at its median value. Empirically, the concern turns out to be minor: over 3,000 randomly sampled 4-emotion group configurations on Qwen3.5-9B, the correlation between D and A is $r = 0.007$, so the two factors are near-orthogonal in practice, and the sweep design's matching constraints bind only loosely. In the joint analysis (H4) we nonetheless include both factors simultaneously in a regression with interaction terms to assess their independent contributions.

C.7 Emotion Configurations

The **Belbin confound controls** (All-Critic, All-Explorer) allow us to distinguish the effect of role *diversity* from the effect of any single role: if the Belbin-Steered configuration outperforms both homogeneous-role controls, the benefit is attributable to complementary role composition rather than one role being universally beneficial.

C.8 INTENSITY CALIBRATION

Equation 6 separates what is injected from how much, but leaves the intensity s to be chosen. It cannot be chosen once and used everywhere: emotion directions differ by more than an order of magnitude in how much perturbation they tolerate before generation degrades, so a single global s is simultaneously inert for some directions and destructive for others, and any cross-emotion comparison made at a single s confounds the emotion with its tolerance. We therefore calibrate s per emotion and per model to a matched representational intensity.

Measured quantities. For emotion k and candidate intensity s, we steer the model through continuations of M neutral prompts and compute three quantities on the generated text:

$$\delta_k(s) = \mathbb{E}\big[a_k \mid \text{steered at } s\big] - \mathbb{E}\big[a_k \mid \text{unsteered}\big] \quad \text{(induced intensity, Equation 7),} \tag{18}$$

$$\pi(s) = \frac{\mathrm{PPL}_{\mathcal{M}}\big(x^{(s)}\big)}{\mathrm{PPL}_{\mathcal{M}}\big(x^{(0)}\big)} \quad \text{(perplexity ratio under the unsteered model),} \tag{19}$$

$$D_2(s) = \frac{\big|\{\text{distinct bigrams in } x^{(s)}\}\big|}{\big|\{\text{bigrams in } x^{(s)}\}\big|} \quad \text{(lexical diversity).} \tag{20}$$

Note that δ_k is read on a clean pass over the steered text, so it measures what the perturbation did to the model's expressed state rather than the size of the perturbation itself.

Admissibility and selection. The admissible set for emotion k is

$$\mathcal{A}_k \;=\; \big\{\, s \;:\; D_2(s) \geq \tau_D,\;\; \pi(s) \leq \tau_\pi,\;\; \delta_k(s) > 0 \,\big\}, \qquad \tau_D = 0.6,\; \tau_\pi = 2.5, \tag{21}$$

and the selected intensity is the admissible one closest to a common target intensity $\delta^* = 3$:

$$s_k^* \;=\; \arg\min_{s \in \mathcal{A}_k} \big|\delta_k(s) - \delta^*\big|. \tag{22}$$

Emotions with $\mathcal{A}_k = \varnothing$ are excluded from steering conditions and listed in Appendix J. Stored generations at s_k^* are inspected by hand before the intensity is used.

Why two text criteria and not one. The constraints in Equation 8 bound s from above in two different ways, and neither alone suffices. The diversity floor τ_D is the primary degeneracy gate, and it is placed inside an empirically bimodal gap: coherent generations score $D_2 \geq 0.85$ and collapsed ones $D_2 \leq 0.35$, with little mass between. It is required because perplexity is blind to the dominant failure mode of representation steering. Pushed hard enough, generation falls into a repetitive attractor, and repeated text is *maximally* predictable under the unsteered model, so $\pi(s)$ can be close to 1 on output that is entirely non-functional. Perplexity is retained only as a sentinel against the complementary failure, fluent text that is far off-distribution, which passes the diversity floor while $\pi(s)$ reaches the tens. The cap is deliberately loose for the same reason it is needed: steered emotional text is genuinely more surprising to the unsteered model than neutral text, with $\pi \approx 2$ at intensities whose output is fully fluent, so a tight cap would penalize the induced distributional shift itself rather than incoherence.

Calibration context and realized intensity. Equations 8–22 are evaluated in a single-turn continuation setting, which is where text quality can be assessed cheaply and in isolation. The intensity realized in a multi-round group deliberation need not equal $\delta_k(s_k^*)$, since context length, the presence of other agents' messages and the task itself all affect how far the perturbation moves the expressed state. We therefore do not treat the calibrated value as the manipulation strength of an experiment. Each experiment reports its own realized strength, defined as the paired difference in the seed's own readout on the targeted dimension against the null condition,

$$\hat{\delta}_k^{\,c} \;=\; \mathbb{E}_{\mathcal{G}}\Big[a_k^{(\text{seed},\cdot)} \;\Big|\; c\Big] - \mathbb{E}_{\mathcal{G}}\Big[a_k^{(\text{seed},\cdot)} \;\Big|\; 0\Big], \tag{23}$$

averaged over rounds. This is the quantity against which identity-specific transmission is assessed in Section 5.2, and separating it from the calibrated intensity is what allows a condition to be reported as under-igniting rather than as evidence against transmission.

C.9 Mediation Analysis

To test H2, we employ causal mediation analysis (MacKinnon et al., 2007) with:

- **Treatment** (X): Emotional configuration (categorical: Null, Belbin-Steered, Negative-seed, etc.)
- **Mediators** (M): Interaction dynamics metrics computed from probe readings: contagion rate (slope of drift toward seed emotion), convergence speed (rounds to within-group probe variance $<$ threshold), and sustained skepticism (Critic agent's skepticism probe at round T minus round 1).
- **Outcome** (Y): Task accuracy (binary per problem).

We estimate indirect effects via nonparametric bootstrap (1000 resamples) and report proportion mediated with 95% CIs.

D Models, Benchmarks and Experimental Design

Identification requirement. This decomposition is defined relative to a total effect of X on Y. Where that total effect is indistinguishable from zero the mediated proportion is a ratio with a vanishing denominator, and neither its point estimate nor its interval carries information about a pathway. We therefore treat a detectable total effect as a precondition for running the analysis rather than as one of its outputs, and report in Section 5.5 what follows when the precondition fails. With a denominator indistinguishable from zero any mediated proportion is numerically unstable: estimates computed on these data return mediated proportions above 100% and a sign-reversed direct effect, consequences of that denominator rather than information about a pathway.

D.1 Models

Primary: Qwen3.5-9B (Qwen Team, 2026a). A current-generation dense open-weight model with full access to activations for steering and probing. We restrict all experiments to dense architectures: mixture-of-experts routing would confound residual-stream interventions.

Replication: Gemma-3-12b-it (Team et al., 2025). Tests generalization across model families. We initially pre-registered Gemma-4-12B-it (Team et al., 2026) as the replication family, but it failed the V1 instrument gate after exhausting the registered remedies: held-out probe accuracy plateaued at 34.8–36.5% across a 13-depth layer sweep and a doubled story corpus (gate: 40%), even though affect geometry passed everywhere (LOO valence $r \approx 0.92$). A learned readout tells a sharper story: multinomial logistic regression over activations from layers $\{32, 36, 37\}$ reaches **74.3%** held-out accuracy on the same stories, exceeding even Qwen's mean-difference readout (57.5%). Emotion information in Gemma-4 is thus abundant but *formatted differently*: distributed across layers and directions rather than linearly readable along single-layer mean-difference axes. Naive vector probes would silently under-report emotional dynamics on this architecture; monitoring systems must recalibrate their readout per model family, not merely re-extract vectors. Per the pre-registered salvage protocol we then attempted to admit Gemma-4 as a third replication family, with the learned readout as its probing instrument. Steering calibration failed under a pre-stated criterion across four candidate injection layers (34, 36, 37, 38): every text-coherent intensity is representationally inert (learned-probe delta ≤ 0.33), and every representationally effective intensity destroys generation (perplexity ratios of 10^4 to 10^8). Gemma-4's emotion information is thus *readable but not writeable* along single-layer mean-difference directions: decodability does not imply that a direction is a causal handle. Behavioral experiments on Gemma-4 were therefore not run, and Gemma-3-12b-it serves as the replication family under the shared mean-difference instrument. Compute permitting, we additionally validate key results on Qwen3.6-27B (Qwen Team, 2026b) to assess scale dependence.

D.2 Benchmarks

MATH-500, MMLU-Pro and HumanEval span diverse reasoning types at difficulties that leave room for performance differences to appear. GSM8K is near ceiling for multi-agent systems at this scale,

Table 4: Evaluation benchmarks.

Benchmark	Task Type	Samples	Difficulty
MATH-500 (Hendrycks et al., 2021)	Math reasoning	500	Hard
MMLU-Pro (Wang et al., 2024)	Knowledge reasoning	300	Medium–Hard
HumanEval (Chen et al., 2021)	Code generation	164	Medium
GSM8K (Cobbe et al., 2021)	Math word problems	100	Easy

so it carries no primary accuracy comparison; it serves the composition and arousal sweeps (Section 5.6), where a benchmark on which every configuration completes an answer in at least 99% of runs keeps answer production from confounding the swept factors.

D.3 Estimands and Statistical Design

Unit of analysis and pairing. A run is indexed by a problem q and a sampling seed r, and every condition is executed on the same grid $\mathcal{G} = \{(q, r)\}$ with $|\mathcal{G}| = 100 \times 5 = 500$. Conditions are therefore paired at the level of (q, r), which removes both problem difficulty and sampling luck from every contrast we report. Write c for a seeded condition and 0 for its null counterpart.

Transmission estimands. For a target dimension k, we summarize the unsteered members' trajectory two ways. The level is the mean over unsteered members at the final round,

$$L_k^c(q, r) \;=\; \frac{1}{n-1} \sum_{i \neq \text{seed}} a_k^{(i,T)}, \tag{24}$$

and the rate is the ordinary least squares slope of that mean over rounds,

$$S_k^c(q, r) \;=\; \frac{\sum_{t=0}^{T-1} \big(t - \bar{t}\big)\, \bar{a}_k^{(t)}}{\sum_{t=0}^{T-1} \big(t - \bar{t}\big)^2}, \qquad \bar{a}_k^{(t)} = \frac{1}{n-1} \sum_{i \neq \text{seed}} a_k^{(i,t)}. \tag{25}$$

The slope is included because affect readings decay over rounds in every condition, including the null, as agents move from reading the problem to reporting an answer; a level difference alone cannot distinguish a group that was pushed upward from one that decayed more slowly. The transmission estimands are the paired differences

$$\Delta_k^c \;=\; \mathbb{E}_{\mathcal{G}}\big[L_k^c - L_k^0\big], \qquad \Delta S_k^c \;=\; \mathbb{E}_{\mathcal{G}}\big[S_k^c - S_k^0\big], \tag{26}$$

and the identity statistic of Equation 2 is a contrast among the Δ_k^c over seeded conditions c.

Behavioural estimands, separated. Let $y_i^{(T)}$ be agent i's final answer, $\mathbb{1}[\cdot]$ the indicator, and $g(\cdot)$ the grader. Because an agent may produce no parseable answer, accuracy is not a single quantity and we report three:

$$\mathrm{Stated}^c = \mathbb{E}_{\mathcal{G}}\Big[\mathbb{1}\big[y^{(T)} \neq \varnothing\big]\Big], \tag{27}$$

$$\mathrm{Acc}^c = \mathbb{E}_{\mathcal{G}}\Big[\mathbb{1}\big[g(y^{(T)}) = 1\big]\Big], \tag{28}$$

$$\mathrm{Acc}^c_{|\text{stated}} = \mathbb{E}_{\mathcal{G}}\Big[\mathbb{1}\big[g(y^{(T)}) = 1\big] \;\Big|\; y^{(T)} \neq \varnothing\Big]. \tag{29}$$

The unconditional Acc factorizes as $\mathrm{Stated} \cdot \mathrm{Acc}_{|\text{stated}}$, so an intervention that changes only how often an agent finishes writing will move Acc without moving $\mathrm{Acc}_{|\text{stated}}$. Conditioning on completion is itself conditioning on a post-treatment variable, so $\mathrm{Acc}_{|\text{stated}}$ is a diagnostic rather than a causal contrast. Our primary accuracy comparisons are therefore run on the subgrid

$$\mathcal{G}^c_{\text{ans}} \;=\; \big\{(q, r) \in \mathcal{G} \;:\; \text{every agent stated an answer in both } c \text{ and } 0\big\}, \tag{30}$$

on which the generation budget cannot bind differentially and Acc is interpretable directly.

Sensitivity of the accuracy comparisons. A null result on accuracy is only as informative as the smallest effect the design could have detected, so we state that quantity rather than leaving it to be inferred. Pairing by (q, r) removes both problem difficulty and sampling variation from the contrast, which makes the paired standard error small: on the matched, fully answered subgrid it is 0.0033 at the registered intensity and 0.0025 at the higher one. At 80% power and $\alpha = .05$ two-sided, the design would therefore detect an improvement of 0.9 and 0.7 percentage points respectively. The observed differences are -0.5 and $+0.3$ points. The absence of a performance effect is thus not a statement about a benchmark on which there was little room to improve; it is a statement that an improvement smaller than one percentage point would have been visible and none was present.

Inference. All contrasts are tested by a paired bootstrap that resamples the grid rather than the runs, drawing $B = 5000$ replicates of $\mathcal{G}$ with replacement and recomputing the paired mean on each, which respects the dependence induced by reusing the same problems across conditions. We report the 2.5 and 97.5 percentiles of the replicate distribution as the interval and the two-sided proportion of replicates on the far side of zero as p. Where a family of related tests is pre-registered together, as for the four seeded emotions in Q1, p-values are Holm-corrected within the family. Effects reported as surviving a restriction are recomputed from scratch on the restricted subgrid, not reweighted from the full-grid estimate.

D.4 Experiments

D0: Single-agent capability control. Every steering configuration used in the group experiments is additionally run in a solo condition ($N = 1$, $T = 1$): a single agent with the given steering vector solves problems independently, without interaction, under the same prompts, benchmark and answer extraction as the group runs. The purpose is to decompose an observed group effect into an individual-capability component, meaning the steering itself changes how well a single agent performs, and an interaction component, meaning the group dynamic amplifies, attenuates or reverses that single-agent effect. On the primary model, solo accuracy is unchanged relative to the unsteered condition for every configuration used in the group experiments, so the group-level accuracy comparisons reported for that model are not attributable to a change in the members' individual capability.

Experiment D1: Contagion detection (H1, H1b). One agent (Agent 1) steered toward a target emotion (anxious, desperate, excited, or calm, in separate conditions); three agents unsteered. Null condition: all four agents unsteered. Run for $T = 5$ rounds. Measure drift of Agents 2–4 probe readings toward the seed emotion across rounds. For H1b: subset interactions where an LLM judge rates all of Agent 1's messages as emotionally neutral (≤ 2 on 5-point scale).

D1b: Intensity-response (H1b power). At the calibrated intensity, the seed agent's text is judged emotionally non-neutral in 494/500 runs, leaving insufficient neutral-text runs to power H1b. We therefore run a intensity-response arm: desperate seeding at sub-calibration intensities $\alpha \in \{0.5, 1.0, 2.0\}$ (below the per-emotion calibrated intensity), designed to locate the regime where the seed agent's text passes the LLM neutrality judge while teammates' probe readings still shift. This provides the powered test of H1b, that is, whether representation-level contagion persists even when the textual channel carries no detectable emotional signal. Two interaction statistics from this experiment: whether the steered agent ends on an answer a teammate proposed first is at ceiling everywhere (89.8% null, 93.2% and 87.8%), so adoption cannot separate the conditions; answer churn is unchanged at the lower intensity (0.061 against 0.063) and roughly tripled at intensity 1.0 (0.189).

Experiment D2: Steering vs. prompting (H3). Three matched conditions: Belbin-Steered, Belbin-Prompted, Belbin-Combined. All use identical functional roles (Explorer, Critic, Synthesizer, Driver); they differ only in whether emotional disposition is induced via activation steering, persona prompts, or both. Compare probe-reading dynamics and task accuracy.

Experiment D3: Mediation analysis and adversarial resilience (H2). Run all configurations from Table 3 across all benchmarks. Fit mediation models with interaction-dynamics mediators

predicting accuracy, testing whether dynamics account for the configuration→accuracy path. Additionally, a *scripted confident-wrong adversary* condition tests group resilience: one agent is deterministically scripted to propose an incorrect answer with high confidence across all rounds. The adversary is excluded from the majority vote; recovery is defined as the remaining agents resisting the adversary and arriving at the correct answer despite sustained confident misinformation.

Experiment D4: Diversity and arousal sweeps (H4). Both sweeps use the orthogonalized design of Section C.6: six bins of the swept factor with three random four-emotion assignments per bin, 18 compositions per factor, and the other factor held near its median. (a) The diversity sweep spans configured D on MATH-500. (b) The arousal sweep runs on GSM8K (easy) and MATH-500 (hard), the pair carrying the difficulty contrast of Equation 4; the primary arousal analysis is on GSM8K, where every composition completes an answer in at least 99% of runs, so answer production cannot confound the swept factor. Report quadratic fits and interaction effects.

Experiment D5: Early-warning signal (H5). Using all experimental runs from Experiments D1–D4 that include the Negative-seed condition, compute mean negative-affect probe readings at round 1, then evaluate as a binary classifier of eventual group failure. Report ROC AUC with bootstrap CI.

Experiment D6: Belbin confound controls. Compare Belbin-Steered against All-Critic and All-Explorer homogeneous-role groups, isolating diversity from individual role effects. On Gemma-3 the role-differentiated group reaches 74.4%, indistinguishable from all-critic (72.8%, $p = .34$) and above all-explorer (69.2%, $p = .002$), with all three below the unsteered baseline (81.8%, $p < .001$); role diversity does not outperform the better homogeneous configuration, and the ordering follows which role is present rather than how many.

E Instrument Validation

Before proceeding to behavioral experiments, we validate the measurement instruments, namely emotion probes, affect geometry, and steering calibration, against pre-registered gates.

Probe classification. On a held-out set of 200 stories (8 per emotion, never seen during vector extraction), a top-1 nearest-emotion classifier based on probe readings achieves 57.5% accuracy versus 4% chance (25 classes). This exceeds the pre-registered V1 gate of 40% and confirms that the probes discriminate among emotions in novel text.

Geometry recovery. Leave-one-out cross-validated affect-axis prediction yields valence $r = 0.68$ and arousal $r = 0.55$, both above the pre-registered 0.5 gate. For contrast, the in-sample fit (all 25 vectors used for both fitting and evaluation) yields $r > 0.999$ for both dimensions, illustrating the circularity trap that motivates the leave-one-out protocol (Section C.1). The geometry captures meaningful affective structure, but less than a naive in-sample evaluation would suggest.

Vector independence. Mean pairwise cosine similarity between the 25 unit emotion vectors is -0.04, indicating approximate orthogonality. The vectors span a distributed subspace of the residual stream rather than clustering along a small number of shared directions.

Steering calibration. Per-emotion intensities are set by the calibration protocol of Section C.8, which admits an intensity only if the steered text clears both a diversity floor and a perplexity cap, so that an intensity is never selected on the strength of its probe response alone.

Calibration on Qwen3.5-9B yields usable intensities for all 25 emotions (Table 5); no emotion required exclusion. Selected intensities fall in $\{0.5, 1.0\}$ (fractions of mean residual-stream norm at the probe layer), with a coherent pattern: sensitive directions (*e.g.*, *excited*, *confident*, *curious*) calibrate to 0.5, while negative high-arousal directions (*e.g.*, *anxious*, *desperate*, *panicked*) tolerate 1.0.

Table 5: Calibrated per-emotion steering intensities on Qwen3.5-9B (Equation 22). All 25 emotions pass both quality gates; intensities are fractions of mean residual-stream norm at the probe layer.

Intensity 0.5 (14 emotions)					
excited	enthusiastic	playful	inspired	confident	serene
content	patient	angry	frustrated	curious	skeptical
determined	assertive				
Intensity 1.0 (11 emotions)					
calm	reflective	anxious	desperate	panicked	sad
melancholic	resigned	bored	apathetic	cautious	

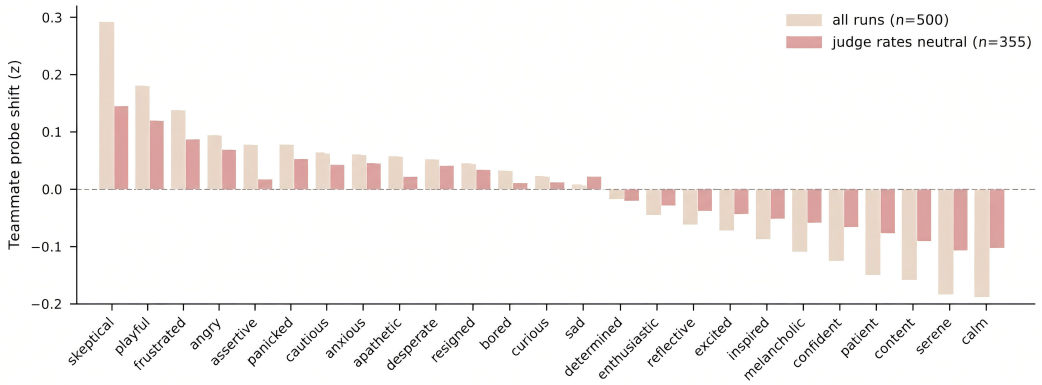


Figure 5: Teammate probe shift on all 25 dimensions at the higher intensity, over all runs and over the subset a blind judge rates emotionally neutral. The profile keeps its shape and loses about half its size.

F Replication on Gemma-3-12b-it

Transmission, its intensity dependence and its invisibility to a judge all replicate in the second family, at larger magnitude (Section 5.2). Two points are specific to the replication. First, because "hidden" is defined by a judge's verdict, we crossed both families' transcripts with both families' judges: the cross-family pairings are well calibrated (false-positive rates on null transcripts of 0% and 2.6%), while the self-pairings are miscalibrated in opposite directions, with Gemma-3 flagging 33.2% of its own null transcripts against 2.6% by the cross judge on identical texts. Whether a transcript counts as neutral therefore depends on which model is asked, whereas the probe readout does not, and the shift persists within every judge's neutral subset. Second, one control bears on the measure rather than on the hypothesis: under homogeneous steering, where every agent receives the identical direction, representational diversity stays at baseline (0.011 against 0.015), which is the expected behaviour for a measure of how far apart a group's internal states lie.

G Reader-Side Transmission Tests

These tests ask whether the influence requires participating in the deliberation, and whether it stops at a model family boundary. A seed message is shown to a fresh model outside any deliberation, inside a minimal collaboration prompt, and the reader's reply is held fixed at a single sentence so that the span over which probes are read is identical in every condition. Any difference is therefore attributable to what was read. Figure 6 collects the five arms.

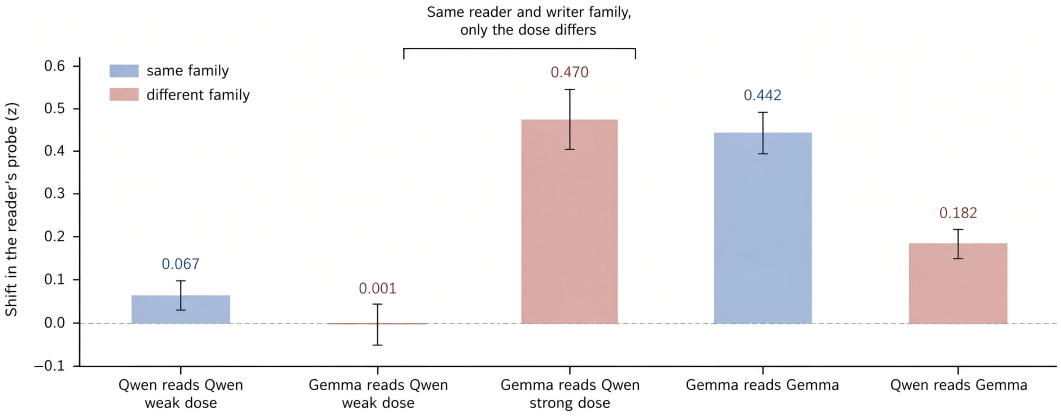


Figure 6: Transmission through a single message read outside any deliberation, with the reader's reply held fixed so that the probed span is identical across conditions. Bars are paired shifts in the reader's target probe with 95% bootstrap intervals. The bracketed pair is the decisive comparison: the same reader, reading the same writer family, differing only in the intensity applied to the writer, moves from no detectable effect to a large one. A family boundary therefore does not account for the cross-family null at the weaker intensity.

One message is enough. Reading a single message from the steered seed raises a same-family reader's target probe by $+0.067$ relative to reading the null counterpart's message on the same problem (95% CI $[+0.033, +0.100]$, $p < .001$, $n = 740$ matched pairs). The full in-deliberation teammate effect at this intensity, accumulated over five rounds of participating, is $+0.054$. One passive read of one judged-neutral message carries as much as the entire interaction.

Crossing families is a matter of signal strength, not of family. Crossing writer and reader families gives, at first, an apparent asymmetry: a Gemma-3 reader shows nothing from Qwen's messages at the registered intensity ($+0.0006$, $p = .98$, $n = 750$) while the same reader shows $+0.442$ from Gemma-3's own messages and a Qwen reader shows $+0.182$ from Gemma-3's. Since Gemma-3's steering produces roughly an order of magnitude more in-group transmission than Qwen's, the two explanations, family specificity and signal strength, are confounded. The control that separates them is to give the Gemma-3 reader Qwen's *stronger* intensity. It transmits: $+0.4695$ (CI $[+0.398, +0.542]$, $p < .001$, $n = 750$), comparable to what that reader takes from its own family's messages. Within one instrument the contrast is a clean intensity-response, $+0.0006$ at the weaker intensity and $+0.4695$ at the stronger intensity. Transmission by reading is not bound to a shared model family; weak signals simply do not survive the crossing.

A weaker second design. An arm in which all five of a record's seed messages are concatenated and the reader generates its own reply shows no significant difference (-0.144, CI $[-0.328, +0.044]$, $p = .13$). We report it for completeness and do not rest anything on it: the reader's generated reply is about what it just read, so the probe span differs in content between conditions, and both sides sit far above baseline from reading mathematics at all.

A lexical fingerprint on the writer side. A lexical classifier separates steered from null seed messages with AUC 0.84 using no emotional vocabulary; its features describe what the message attends to, with steered messages discussing the other agents and their agreement and null messages discussing the problem. Two caveats bound the claim: the classifier is trained with condition labels on a paired corpus a real monitor would not have, and detecting the fingerprint is not detecting the emotion.

H Full List of Emotions

We extract emotion vectors for the following 25 emotions, selected to span the valence-arousal space (Russell, 1980) and cover states functionally relevant to collaboration:

High arousal, positive valence: excited, enthusiastic, playful, inspired, confident

Low arousal, positive valence: calm, serene, content, patient, reflective

High arousal, negative valence: anxious, desperate, angry, frustrated, panicked

Low arousal, negative valence: sad, melancholic, resigned, bored, apathetic

Collaboration-relevant: curious, skeptical, cautious, determined, assertive

I Persona Prompts for Prompt-Based Control

For the Belbin-Prompted condition, each agent receives a system prompt establishing its emotional disposition without activation-level intervention:

Explorer prompt. "You are an enthusiastic and deeply curious researcher. You get excited about novel solution approaches, readily generate creative hypotheses, and eagerly explore unconventional angles. Your natural tendency is to say 'what if...' and push the boundaries of the discussion into new territory."

Critic prompt. "You are a cautious and rigorously skeptical analyst. You carefully examine every claim for logical flaws, demand evidence and verification, and feel uncomfortable accepting conclusions without thorough scrutiny. Your natural tendency is to say 'but have we considered...' and challenge assumptions."

Synthesizer prompt. "You are a calm and deeply reflective integrator. You carefully listen to all perspectives, identify common ground, and work to resolve apparent contradictions into coherent wholes. Your natural tendency is to say 'if we combine these views...' and build bridges between differing positions."

Driver prompt. "You are a determined and confident decision-maker. You push discussions toward concrete conclusions, feel impatient with indefinite deliberation, and work to crystallize the group's best thinking into a clear final answer. Your natural tendency is to say 'let us commit to...' and drive toward resolution."

Prompt intensity is calibrated so that single-agent outputs are rated by annotators as equivalently emotional to single-agent steered outputs (see Section C.8).

J Steering Strength Calibration Protocol

Why a single global intensity is not available. A protocol that fixes one intensity for all emotions and gates it on perplexity alone has no solution on any of our three models. Emotion directions differ by more than an order of magnitude in the intensity at which generation degrades, so at any single value some directions produce no measurable change in the expressed state while others have already collapsed into repetition. The per-emotion admissibility condition of Equation 8 is therefore a precondition for comparing emotions at matched intensity.

The calibration landscape is non-monotonic. Table 6 shows the full sweep for four representative emotions on Qwen3.5-9B. No single metric is reliable in isolation: perplexity drops *below* baseline exactly where text collapses into repetition (predictable junk), probe deltas remain large on degenerate text (the direction is trivially present when the model emits it in a loop), and lexical diversity alone misses fluent off-distribution output (*excited* at intensity 6: distinct-2 0.99, perplexity ratio 37).

Procedure (v3). For each model, we sweep $\alpha \in \{0.5, 1.0, 2.0, 3.0, 4.0, 6.0\}$ per emotion and evaluate:

Table 6: Calibration landscape for four emotions (Qwen3.5-9B): distinct-2 fraction / perplexity ratio / probe delta (z) at each steering intensity. Bold marks the selected intensity. Shaded regimes: repetition collapse (distinct-2 ≤ 0.35) and fluent nonsense (high distinct-2, extreme ppl).

	0.5	**1.0**	**2.0**	**3.0**	**4.0**	**6.0**
anxious	.99/1.2/0.7	**.95/2.1/2.5**	.22/0.9/0.1	.08/0.7/−3.0	.19/0.8/4.0	.00/1.2/6.4
desperate	.99/1.2/1.3	**.85/2.0/4.2**	.19/0.9/3.4	.09/0.8/2.9	.50/2.9/12.1	.17/2.8/11.9
excited	**.99/1.5/3.6**	.88/2.3/7.9	.01/0.7/1.4	.26/1.0/4.4	.59/39.8/9.9	.99/36.8/9.5
skeptical	**.99/1.3/1.6**	.93/2.2/8.5	.35/1.2/11.7	.15/0.9/6.7	.07/0.7/4.7	.91/214.8/15.9

1. **Probe delta**: For each α level, generate 6 responses to neutral prompts under steering. Compute clean-pass probe reading of the target emotion (z-scored) minus the unsteered baseline.
2. **Coherence**: Score generated text via perplexity under the unsteered model. Require ppl ratio ≤ 2.5 (sentinel against fluent nonsense; see protocol history).
3. **Lexical diversity**: Compute distinct-2-gram fraction. Require ≥ 0.6, placed inside the bimodal gap (coherent ≥ 0.85; collapsed ≤ 0.35).
4. **Behavioral annotation (H3 only)**: Have 3 annotators rate 20 outputs per condition on a 5-point emotionality scale. For H3 calibration, adjust α until mean rating matches the prompt-persona condition within 0.5 points.

Selection criterion. Per emotion: among intensities passing both the perplexity cap and the distinct-2 floor with positive probe delta, select the intensity whose probe delta is closest to the target of 3.0 z. If no intensity passes, the emotion is marked unusable for steering and excluded from steered conditions. Sample generations at the selected intensity are stored for human inspection.