The Preference Centroid: Consensus Density Governs Output Dispersion in Aligned LLMs

Anonymous authors
Paper under double-blind review

Abstract

A growing literature reports that the same generative model *equalizes* outcomes for some users and *amplifies* differences for others. Sampling an aligned model repeatedly and embedding the completions, we summarize each output cloud by its *dispersion*—mean pairwise cosine distance—and report four findings. First, unconditioned dispersion is set by the *consensus density* of the task, rising from near-zero on factual prompts to wide on open-ended ones (Spearman $\rho = 0.85$, n = 20); the ordering replicates on a second, unrelated model ($\rho = 0.84$). Second, the construct is not circular: three held-out judge models, scoring only the prompt and never any output, predict the measured dispersion at $\rho = -0.91$, and a preliminary human expert panel tracks these judges at $\rho = 0.74$. Third, conditioning does not "sharpen" the model but *relocates* its sampling region: dispersion at the destination tracks the destination's own judge-scored consensus—falling below the unconditioned default for high-consensus relocations, rising above it for low-consensus ones—not the instruction's surface form. Fourth, a matched base-versus-instruct comparison on open weights shows the pretrained base already carries the gradient and alignment *amplifies* it, compressing the high-consensus end roughly 8.7-fold against 1.8-fold at the low end. Absolute magnitudes drift between sampling batches, so every claim rests on rank correlations and effect directions, which are stable across five fresh batches, two embedders, and four model families. These findings ground the equalizer/amplifier boundary at the distribution level rather than in posited task complexity; its further reading in terms of human expertise amplification is an interpretation the measurements motivate but do not themselves test.

1 Introduction

The same model, handed to two people, does two different things. Give a frontier assistant to a novice on a well-defined task and it pulls their output up toward a competent standard; give it to an expert on the same task and it pulls them toward the same standard, compressing the gap between them. But move to an open, judgment-heavy task and the picture inverts: the novice and the expert now diverge, because the model returns a wide field of uneven possibilities and the outcome depends on who is choosing among them. The first pattern is the "equalizer" effect documented across writing, customer support, and coding studies (Brynjolfsson et al., 2025; Noy & Zhang, 2023; Peng et al., 2023); the second is the "amplifier" effect visible when expertise, not access, decides the result (Sun et al., 2022; Dell'Acqua et al., 2023; METR, 2025). Recent work reconciles the two by positing **task complexity** as the moderating variable: AI equalizes on routine tasks and amplifies on complex ones (Anonymous, 2026).

That reconciliation answers *when* the split happens. It does not say *why*—what, mechanically, at the level of the model's output distribution, makes one task an equalizer and another an amplifier. "Complexity" is a description of the task, not a property of the model we can point to and measure. This paper supplies the missing layer. We claim that the moderating variable is not complexity in the abstract but the **consensus density** of the task's output region—how much agreement there is about what a good answer looks like—and that this quantity is directly measurable in a model's sampling behavior.

The starting observation is familiar and usually offered as metaphor: an aligned language model "regresses to the mean." Reinforcement learning from human feedback (Ouyang et al., 2022) concentrates the output distribution onto a high-probability region that reflects the aggregate preferences of the feedback process. This concentration is by now well documented—RLHF measurably reduces output diversity relative to supervised fine-tuning (Kirk et al., 2024), alignment shrinks the effective "generative horizon" (Yang et al., 2025), and aligned models exhibit reduced conceptual diversity (Murthy et al., 2024) and even mode collapse onto characteristic outputs (Hamilton, 2024; Mohammadi, 2024). We take the *existence* of this concentration as established and not our contribution. Our contribution is to turn the metaphor into a measurement and, in doing so, to extract a relation that the existence claim alone does not give:

> **Output dispersion is, to first approximation, the inverse of the consensus density of the region being sampled.** The unconditioned default sits at the preference centroid. Conditioning *relocates* the sampling region; the dispersion at the new region is set by that region's own consensus, not by the act of conditioning.

We name the model's default high-density region the **preference centroid**—a naming handle, not a novelty claim—and we name the governing quantity **consensus density**. Figure 1 sketches the principle. From it three measurable consequences follow, and we verify each:

1. **The centroid's tightness is a property of the task.** Without any conditioning, dispersion varies across tasks according to how much consensus the task admits. We measure a clear graded rise from factual prompts (dispersion ≈ 0.02, near-identical answers) to open-ended creative prompts (median ≈ 0.18), with judgment and advice tasks in between and effectively tied with each other (Spearman $\rho = 0.85$, p $\approx 1.8 \times 10^{-6}$), and the same ordering replicates on a second, unrelated model (GPT-4o, $\rho = 0.84$; per-prompt agreement between the two models $\rho = 0.75$). This gradient *is* the equalizer/amplifier boundary, measured: a factual task collapses every user onto one answer; a judgment task returns a wide field on which the user's selection decides the outcome.

2. **Consensus density can be measured independently of the dispersion it explains.** The natural objection to a quantity inferred from dispersion is circularity. We break it with an external measurement: three held-out judge models (GPT-4.1, Gemini 2.5 Flash, Qwen-Max), shown only the *prompt* and never any output, rate how much knowledgeable people would agree on what a good answer looks like. Their ratings agree closely with one another (inter-judge $\rho = 0.85$–0.95) and predict the measured dispersion at $\rho = -0.91$ (p $\approx 2 \times 10^{-8}$). Consensus density thus predicts dispersion rather than being read back from it (§3.4, Figure 3).

3. **Conditioning relocates rather than sharpens, and the destination's consensus—not the instruction's form—sets the spread.** Every prompt we add moves the output cloud a large distance in embedding space (centroid shift > 0.5), and whether the cloud tightens or widens depends on where it lands, not on how constraining the instruction looks. From one base prompt, a content constraint that pins a specific scene—a high-consensus destination—drives dispersion well *below* the unconditioned default (0.086 vs 0.189; Batch 1, §3.7), while an instruction to avoid typical openings—a low-consensus destination—drives it well above (0.345); both are stable across four independent runs, and the "high-consensus" and "low-consensus" labels are corroborated by held-out judges: scoring the conditioned prompts themselves, they recover this destination ordering across the ladder arms (arm-level $\rho = -0.90$, §3.7). The relation is symmetric: dispersion belongs to the destination, not to the act of conditioning. **Form is not consensus.**

These three results let us replace the metaphor of "regression to the mean" with a distribution-level mechanism beneath the equalizer/amplifier boundary. Where the distribution is tight (high consensus), conditioning and selection barely matter and every user is delivered the same competent output—equalization. Where it is diffuse (low consensus), the field is wide and uneven, and—on the premise, which we state rather than measure, that most samples in such a field are mediocre—value is realized only by selecting the good ones. That selection is what we mean by **taste**, and it is the binding constraint exactly in the regime where

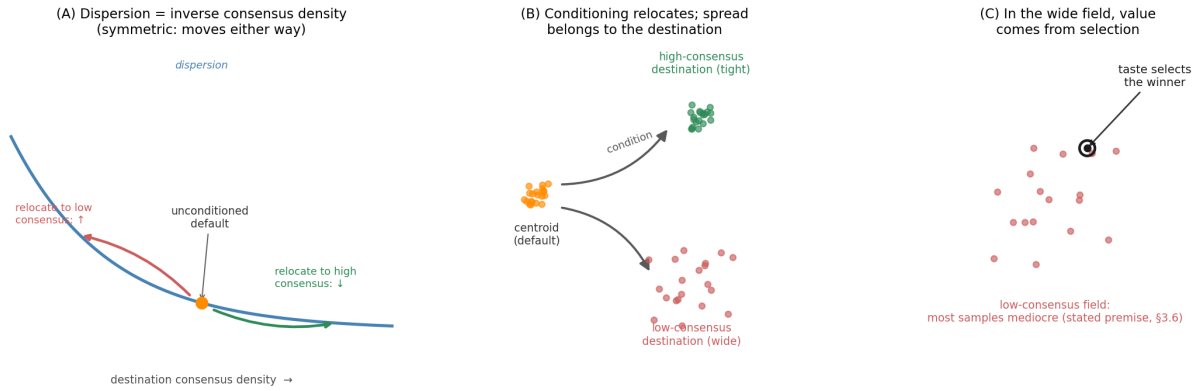

Figure 1: The thesis as a schematic. (A) Dispersion is the inverse of the destination's consensus density; the unconditioned default sits mid-curve, and conditioning can move dispersion *either way*—relocating to a higher-consensus destination lowers it, to a lower-consensus destination raises it. The relation is symmetric (there is no one-sided floor). (B) Conditioning relocates the sampling region rather than sharpening it; the spread at the destination is the destination's property (high-consensus destinations are tight, low-consensus ones wide), not a function of the act of conditioning. (C) In a low-consensus (wide) destination the field is wide and uneven; on the stated premise that most samples there are mediocre (two first tests, simulated and a small human pass, left it unestablished, §5), value is realized only by selecting the good one—that selection is taste (§3.8).

dispersion is high. The model supplies a competent floor *on outcome quality*—a premise we likewise state rather than measure; only judgment reaches the ceiling. (Throughout, "floor" and "ceiling" denote outcome quality, never dispersion—§3.7 shows dispersion has no one-sided floor.) This is the sense in which the same model can be a floor for one person and a ceiling for another; the dispersion gradient is the part we measure, and the human-outcome reading is the interpretation it motivates (§3.8, §5).

Contributions. (1) We operationalize "regression to the mean" in aligned LLMs as a measurable dispersion of the sampled output cloud, and state a single governing principle—dispersion as inverse consensus density. (2) We provide independent evidence that the governing quantity is real and not circular: three held-out judge models that never see any output predict dispersion from the prompt alone ($\rho = -0.91$), and the task-type gradient replicates across two unrelated models (Claude $\rho = 0.85$, GPT-4o $\rho = 0.84$; per-prompt cross-model agreement $\rho = 0.75$). (3) We show that conditioning relocates the sampling region and that dispersion is symmetric in the destination's consensus—a high-consensus relocation drives it below the default, a low-consensus relocation above it—so spread is a property of the destination, not of the instruction's form. (4) As an interpretive layer on those measurements, we use the mechanism to ground the equalizer/amplifier boundary and to reframe taste as selection from a low-consensus field, with direct implications for where human judgment is and is not the bottleneck; we are explicit that this layer rests on a stated premise about sample quality, and that its first two pre-registered tests (§5)—a simulated pass and a small human forced-choice pass—returned, respectively, an inconclusive result at the LLM-judge instrument's resolution and an instrument failure whose exploratory remainder trends against the prediction at a sample size we do not interpret; the premise therefore remains stated, not established. We frame the study as illustrative, sized to support its claims rather than to set a benchmark, and we report between-run and cross-model variation rather than single-run point estimates—exactly the standard of evidence the venue asks for.

2 Related Work

2.1 The equalizer–amplifier debate

A wave of field and lab studies finds that generative AI narrows performance gaps: lower-skilled customer-support agents gain most from an AI assistant (Brynjolfsson et al., 2025), mid-tier writers improve more

than strong ones (Noy & Zhang, 2023), and less-experienced developers see the largest speedups from code completion (Peng et al., 2023). Read together, these support an "equalizer" reading: AI lifts the floor and compresses the distribution of outcomes.

A second body of work complicates that reading. Dell'Acqua et al. describe a "jagged technological frontier" on which AI helps inside its capability boundary and hurts outside it, so that outcomes depend on whether the human knows where the boundary lies (Dell'Acqua et al., 2023). Studies comparing experts and novices on AI-assisted data work (Sun et al., 2022), close observation of how programmers actually interact with code-generating models (Barke et al., 2023), and a controlled measurement of experienced open-source developers (METR, 2025) find that skilled users extract disproportionate value—or, strikingly, that AI can slow experts who must vet its output. Doshi and Hauser report that generative AI enhances individual creativity while *reducing the collective diversity* of the resulting content (Doshi & Hauser, 2024)—the population-level signature of many users being drawn toward the same default region. Meta-analysis of human–AI combinations finds that they help on some task types and hurt on others (Vaccaro et al., 2024).

The reconciliation we build on posits task complexity as the moderating variable: AI equalizes on well-structured, routine tasks and amplifies pre-existing differences on complex tasks requiring judgment, building on a framing of AI as a *cognitive amplifier* whose output quality tracks the directing human's expertise (Anonymous, 2026). Our work sits one layer beneath this: we do not re-argue *that* complexity moderates the effect, we identify and measure the distributional property—consensus density—that complexity is standing in for, and show the moderation as a continuous, measured gradient rather than a dichotomy.

2.2 RLHF, alignment, and output diversity

That alignment concentrates the output distribution is well established. Kirk et al. show that RLHF improves out-of-distribution generalisation but *significantly reduces output diversity* relative to supervised fine-tuning—a generalisation/diversity trade-off (Kirk et al., 2024). Yang et al. quantify the concentration directly, reporting that alignment shrinks the model's "generative horizon," cutting the effective branching factor several-fold (Yang et al., 2025). Murthy et al. find the loss is not merely lexical but *conceptual*: aligned models sample from a narrower space of underlying ideas (Murthy et al., 2024). Mohammadi documents lower token-level entropy and embedding-space clustering—"attractor states"—as a price of debiasing (Mohammadi, 2024), and Hamilton detects outright mode collapse across successive aligned model versions (Hamilton, 2024). Zhang et al. trace one driver to a *typicality bias* in preference data and propose prompting remedies (Zhang et al., 2025). These results motivate our metric and our naming of the default region as a centroid; we treat the concentration itself as background, not as a finding.

Two adjacent literatures fix our vocabulary. The first is *typicality*: high-probability text is not the same as typical text, and decoding that targets the typical set avoids the degeneration that pure probability-maximization produces (Meister et al., 2023; Holtzman et al., 2020). Consensus density is, informally, a per-region typicality: a region of high consensus is one where the typical set is small. The second is the *measurement of generation diversity*. Classic metrics—distinct-n (Li et al., 2016) and Self-BLEU (Zhu et al., 2018)—operate on surface form; more recent work standardizes these scores (Shaib et al., 2024) and argues for semantic, embedding-based measures of *effective* diversity that weight quality alongside spread (Shypula et al., 2025). We adopt a deliberately simple embedding-based dispersion (§3.1) in this lineage, and we inherit its known blind spot to style, which we make explicit and exploit rather than hide.

Relation to semantic uncertainty. Closest of all in machinery is the semantic-uncertainty line. Kuhn, Gal, and Farquhar sample a model repeatedly, cluster the completions by bidirectional entailment, and read the entropy over meaning-clusters as the model's uncertainty (Kuhn et al., 2023); Farquhar et al. use the same quantity to detect confabulation, which is likeliest exactly when the answer distribution is semantically spread (Farquhar et al., 2024). Our dispersion is a clustering-free cousin of that measurement—a first moment of pairwise semantic distances rather than an entropy over entailment clusters—and on factual prompts the two should track each other closely. The constructs, however, run in opposite directions. Semantic entropy reads spread as a property of the *model*: epistemic uncertainty about a question presumed to have a right answer, and therefore a signal of unreliability. We read spread as tracking a property of the *task*: how much agreement exists about what a good answer looks like—validated against prompt-side consensus ratings that

never see an output (§3.4), and used to explain where conditioning and alignment move the distribution (§3.5, §3.7). On a factual prompt the readings coincide (high spread means the model does not know the answer); on an open prompt they come apart—a sea-story prompt with high dispersion is not a model failure but a faithful reflection of a task that admits many good answers. Telling "the model is uncertain" apart from "the task is open" is precisely what the independent consensus anchor is for, and computing semantic entropy alongside dispersion across our gradient is a natural calibration experiment (§5).

Distinguishing model collapse. A separate phenomenon shares the word "collapse" and must be set aside to avoid confusion. *Model collapse* refers to the degradation that occurs when models are trained on recursively generated synthetic data, progressively forgetting the tails of the true distribution (Shumailov et al., 2023; 2024). That is a statement about *training dynamics across model generations*. Our subject is orthogonal: the *inference-time* concentration of a single fixed, aligned model's output distribution, and how conditioning moves the sampling region within it. Nothing here concerns synthetic-data training loops; the centroid we measure is a property of one model as shipped.

Distinguishing activation steering. A second term we set aside is *steering*. A growing line of work on *activation* steering and representation engineering relocates model behavior by intervening directly on hidden states—adding contrastive steering vectors to the residual stream at inference time (Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024). When we say conditioning "steers" the model, we mean something strictly weaker and entirely external: we change the input prompt and measure where the output distribution lands in semantic space, with no access to or claim about internal representations. Our relocation is a fact about the input–output map, not about activations; the shared word should not be read as a shared mechanism or as a contribution to that literature.

2.3 Intelligence augmentation, cognitive offloading, and judgment

The framing of AI as something that augments rather than replaces human intellect descends from Engelbart's program of *augmenting human intellect* (Engelbart, 1962). A countervailing literature warns that easy access to external answers invites *cognitive offloading*: when information is a click away, people remember where to find it rather than the thing itself (Sparrow et al., 2011). The risk specific to aligned LLMs is sharper because the model returns not a retrieved fact but a confident, fluent default—and will often *agree* with a user who pushes back, a sycophancy that lets a novice reinforce an error through uninformed feedback (Malmqvist, 2024; Chen et al., 2025).

What turns access into outcome, on our account, is *judgment*: the expert–novice literature locates expertise in the ability to perceive deep structure and evaluate candidate solutions, not merely to generate them (Chi et al., 1981), and metacognition supplies the monitoring that decides when an answer is good enough (Flavell, 1979). This is precisely the faculty our mechanism makes load-bearing: when conditioning has relocated the model to a low-consensus, high-dispersion region, the field is wide and mostly mediocre, and the binding constraint is the human's capacity to select. Taste, in our usage, is that selective judgment—and §4 develops why it matters exactly where consensus density is low.

3 The Preference Centroid, Measured

The claim that an aligned language model "regresses to the mean" is usually offered as metaphor. In this section we make it a measurement. We show that a model's default output occupies a measurable high-density region of semantic space—a *preference centroid*—whose tightness is set by the *consensus density* of the region being sampled; that this consensus density can be measured independently of the dispersion it explains, by held-out models that never see the outputs; and that prompting does not "sharpen" the model so much as *relocate* it. The spread at the new location is a property of the destination, not of the instruction—and it moves *symmetrically*: relocating to a higher-consensus destination drives dispersion below the unconditioned default, relocating to a lower-consensus one drives it above. The default is not a one-sided floor; it is simply where the model sits when nothing has moved it.

3.1 Measurement

For a fixed prompt we draw $N = 25$ independent completions at temperature 1.0 from a single instruction-tuned model (Claude Sonnet 4.6, accessed through an OpenAI-compatible endpoint; max 200 tokens per completion), embed each completion with a sentence encoder (`all-MiniLM-L6-v2`, in the Sentence-BERT family (Reimers & Gurevych, 2019)), and summarize the resulting cloud by its **dispersion**: the mean pairwise cosine distance among the embeddings,

$$\mathrm{disp}(P) = \frac{2}{N(N-1)} \sum_{i<j} \big(1 - \cos(\mathbf{e}_i, \mathbf{e}_j)\big),$$

where $\mathbf{e}_i$ is the unit-normalized embedding of the i-th completion. Low dispersion means near-interchangeable answers; high dispersion means many distinct ones. To compare two prompts we also report the **centroid shift**: the Euclidean distance between their clouds' mean embeddings, $\|\bar{\mathbf{e}}^{(A)} - \bar{\mathbf{e}}^{(B)}\|$. These are deliberately simple, model-agnostic quantities in the embedding-diversity lineage (Shypula et al., 2025), and—as we make explicit in §3.7—the embedding captures *semantic content* rather than *style*, a limitation we use rather than hide. We treat the study as illustrative rather than definitive (see §3.9 and §5); it is sized to support the claims below, not to benchmark models.

3.2 A unifying principle

> **Output dispersion is, to first approximation, the inverse of the consensus density of the region being sampled.** The unconditioned default sits at the preference centroid. Conditioning *relocates* the sampling region; the dispersion at the new region is set by that region's own consensus, not by the conditioning.

The principle has three observable consequences. (1) Without conditioning, dispersion should vary across tasks according to how much consensus the task admits—and an *independent* estimate of that consensus, taken without reference to the outputs, should predict it. (2) Conditioning should *relocate* the cloud regardless of what it does to spread. (3) At the destination, dispersion should track the destination's consensus and *not* the surface form of the instruction; relocating to a higher-consensus destination should lower dispersion and relocating to a lower-consensus one should raise it, symmetrically. We take these in turn, with the measurements summarized in Figures 2–6.

3.3 The centroid's tightness is a property of the task

We grouped prompts into five task types spanning a spectrum from consensus-bound to open-ended, four prompts each, and measured the unconditioned dispersion of every prompt (Figure 2A, with per-type mean $\pm$ standard deviation over its four prompts):

Task type	Dispersion (mean $\pm$ sd)
Factual ("the capital of France . . . ")	0.020 ± 0.006
Consensus explanation ("what is a binary search tree")	0.038 ± 0.018
Judgment ("raise venture capital or bootstrap?")	0.107 ± 0.025
Soft advice ("advice for a new manager")	0.103 ± 0.020
Creative-open ("opening sentence of a sea story")	median 0.18 (mean 0.254 ± 0.180)

The gradient is statistically clear, if not strictly monotonic across all five bands: a Spearman rank correlation between task type and per-prompt dispersion gives $\rho = 0.85$ ($p \approx 1.8 \times 10^{-6}$, $n = 20$). Two honest qualifications sharpen rather than weaken the result. First, the two middle types—judgment and soft advice—are statistically indistinguishable (0.107 vs 0.103); the gradient is better read as four levels than five, with contestable-recommendation and open-advice tasks sharing a similar middling consensus. Second, the creative-open type carries a large standard deviation ($\pm$0.180), and its mean is inflated by a single degenerate

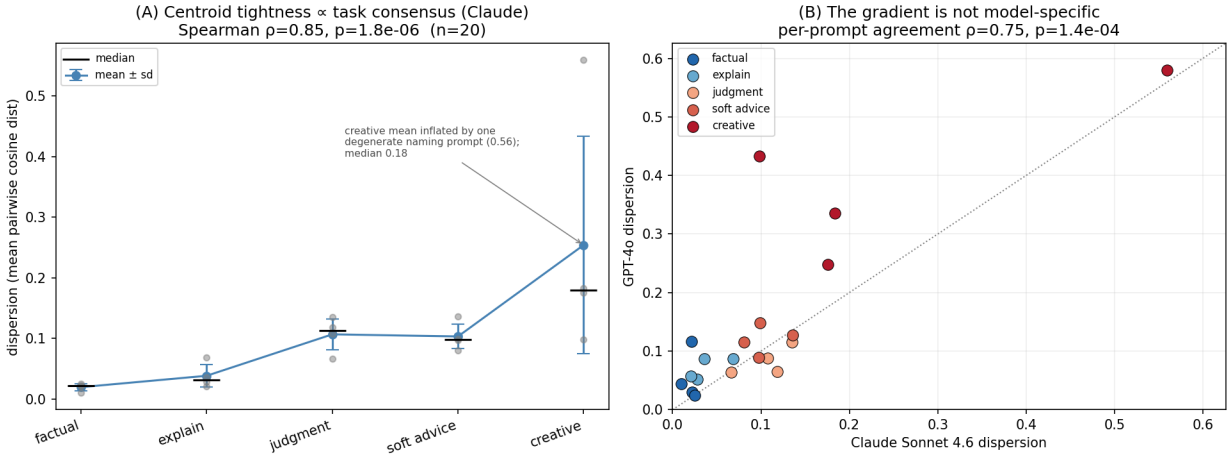

Figure 2: The task-type gradient and its cross-model replication (N = 25 completions per prompt). (A) Unconditioned dispersion by task type for Claude Sonnet 4.6, mean ± sd over four prompts each, with per-prompt scatter and per-type median; the gradient rises with Spearman $\rho = 0.85$, $p \approx 1.8 \times 10^{-6}$ (n = 20), the two middle bands effectively tied. The creative type's mean (0.254) is inflated by one degenerate naming prompt (board-game name, 0.56); its median is 0.18. (B) The same 20 prompts generated by an unrelated model (GPT-4o): the per-prompt dispersions agree with Claude's at Spearman $\rho = 0.75$, and GPT-4o's own gradient gives $\rho = 0.84$—the ordering is a property of aligned models, not of one model.

prompt: inventing a board-game *name*, whose outputs are near-random short strings (dispersion 0.56). The other three creative prompts sit at 0.10–0.18, and the type's median is 0.18; we report the median for this type and treat the naming prompt as a high-dispersion outlier rather than a representative case. The deeper point survives the correction: open tasks are not only high-dispersion but *heterogeneously* so, since "creative" spans prompts that themselves differ widely in how much they constrain. The open end of the spectrum is intrinsically noisy—itself a face of the thesis, since low-consensus tasks lack consensus even about how open they are. Because the headline statistic is a rank correlation, it is unaffected by the naming outlier (dropping it leaves $\rho = 0.84$; a leave-one-prompt-out sweep keeps ρ in [0.83, 0.89]).

This gradient is, we argue, the mechanism beneath the task-complexity boundary (Anonymous, 2026). Where prior work observed that AI *equalizes* performance on routine tasks and *amplifies* differences on complex ones, and posited task complexity as the moderating variable, the dispersion gradient shows *why*: a factual task collapses the model onto a near-single answer (0.02), so every user is delivered the same competent output and differences compress (equalization); a judgment- or creativity-intensive task returns a wide field of uneven candidates (per-type median 0.11–0.18), so the outcome depends on who selects among them (amplification). The moderating variable is not complexity in the abstract; it is the consensus density of the task's output region, and it is measurable. The population-level finding that generative AI reduces the *collective* diversity of novel content (Doshi & Hauser, 2024) is the same phenomenon seen from above: many users, sampling near a tight centroid, converge.

3.4 Consensus density, measured independently

The gradient above relates two quantities read from the *same* outputs: the task labels we assigned and the dispersion of the clouds. A natural objection is that "consensus density" is then not a cause but a relabeling—inferred from the very dispersion it is meant to explain. To meet this objection we measure consensus density a second way, with no reference to the outputs at all. We gave each of the 20 prompts to three held-out judge models from different families (GPT-4.1, Gemini 2.5 Flash, Qwen-Max) and asked them to rate, *for the question alone*, how much knowledgeable people would agree on what a good answer looks like (0 = wide open, 100 = one agreed answer); each prompt was scored three times per judge and the scores averaged. The judges never saw a single completion.

Two things hold (Figure 3). The judges agree with one another—pairwise inter-judge Spearman $\rho = 0.85$–0.95—so "consensus density" is a reliable quantity rather than noise. And their consensus scores predict the measured dispersion strongly and in the expected direction: pooled across judges, Spearman $\rho = -0.91$ ($p \approx 2 \times 10^{-8}$, $n = 20$); each judge alone gives -0.88 to -0.90. A companion elicitation—"how many substantively different good answers would a hundred knowledgeable people produce?"—correlates positively with dispersion ($\rho = 0.76$). Because the prediction now runs from a prompt-side judgment to an output-side measurement, consensus density is predicting dispersion rather than being read back from it: the construct is independently grounded.

This does not make the measurement perfect, and we are explicit about what remains (§5). The judges are *model* proxies for human agreement, not humans; and the four models share overlapping training data, so a shared notion of consensus is a confound we mitigate but do not eliminate. Two purely mechanical checks, carrying no model *judgment* of consensus at all—only a generator's distributional uncertainty—point the same way from a fully independent direction. The mean token-level surprisal of an independent generator (GPT-4o) rises with dispersion ($\rho = 0.51$), though that signal is noisy (sensitive to rare tokens and length). More cleanly, the exact per-step predictive *entropy* of an open-weight generator (Qwen2.5-7B-Instruct, whose logits we read directly rather than through an API) rises with dispersion at $\rho = 0.66$ and falls with the judge-rated consensus at $\rho = -0.70$ (both $p < 0.002$, $n = 20$), and tracks that model's *own* dispersion at $\rho = 0.72$: a model more uncertain about what to say next, token by token, produces a more dispersed output cloud, exactly as the consensus-density account predicts.

A preliminary human panel. The judges above are still *model* proxies for human agreement—the one gap §5 singles out. As a first check we ran a small human panel in a domain where expertise is scarce and the harder questions are genuinely contested (Chinese pharmaceutical market access): four domain experts—three industry practitioners and one academic—independently rated each prompt, for the *question* alone, on how much knowledgeable people would agree on a good answer, never seeing any output. Their ratings agree (Kendall's $W = 0.58$, ICC(2,k) $= 0.66$) and track the held-out judges at Spearman $\rho = 0.74$ ($p < 10^{-3}$), so the judges are not merely agreeing among themselves but track real human experts, the validation the construct was owed. (A fifth rater, drawn from external pharmaceutical advisory rather than operating-side practice, correlated near zero with every panel member and with both model signals; consistent with a genuine difference in professional vantage, we report this rater separately and do not average it into the panel.) Crucially, including this fifth rater leaves the headline essentially unchanged—expert→judge $\rho = 0.73$ with all five against 0.74 without (both $p < 10^{-3}$)—so it is set aside not to secure the correlation but on a rater-side criterion alone: near-zero agreement with every other rater (its near-zero correlation with both model signals is an after-the-fact observation, not part of the rule; Appendix A.3). We keep this *preliminary*: the panel is small ($n = 18$ prompts, four raters) and its inter-rater agreement only modest (Kendall's $W = 0.58$).

What the panel does *not* establish. We are deliberately explicit about the one link this panel does *not* support. It does *not* show that *raw dispersion* tracks human consensus directly: across every panel composition that correlation stays below significance ($\rho \approx -0.35$, $p = 0.15$, $n = 18$), and we do not claim it. The null is expected rather than anomalous—it is what attenuation through a noisy mediator predicts, since the construct reaches dispersion only *through* the judges.[1] The human validation we *do* claim is therefore the strong, significant first leg—experts track the judges at $\rho = 0.74$ ($p < 10^{-3}$)—while the direct human-to-dispersion link is left to a larger study (§5), not asserted on this sample.

Two further numbers from this panel belong in the main text rather than the appendix, because they calibrate the headline. First, on these 18 within-domain prompts the judge→dispersion correlation is -0.56—real, but visibly weaker than the -0.91 of the main set. The contrast is informative: the main set spans task types from one-answer factual questions to open fiction, nearly the full consensus range, where a strong rank correlation is comparatively easy to achieve; the pharmaceutical set sits inside a single professional domain, where the differences are finer and the range is restricted. We read -0.91 as the construct's resolution across coarse, between-task-type differences and -0.56 as its noisier resolution at within-domain granularity; the

[1] As an informal consistency note only: if the two measured legs (expert→judge 0.74; judge→dispersion -0.56 on this set) composed multiplicatively they would give ≈ -0.41, close to the observed -0.35. Rank correlations do not in general compose multiplicatively, so we offer this as a plausibility check, not an estimate.

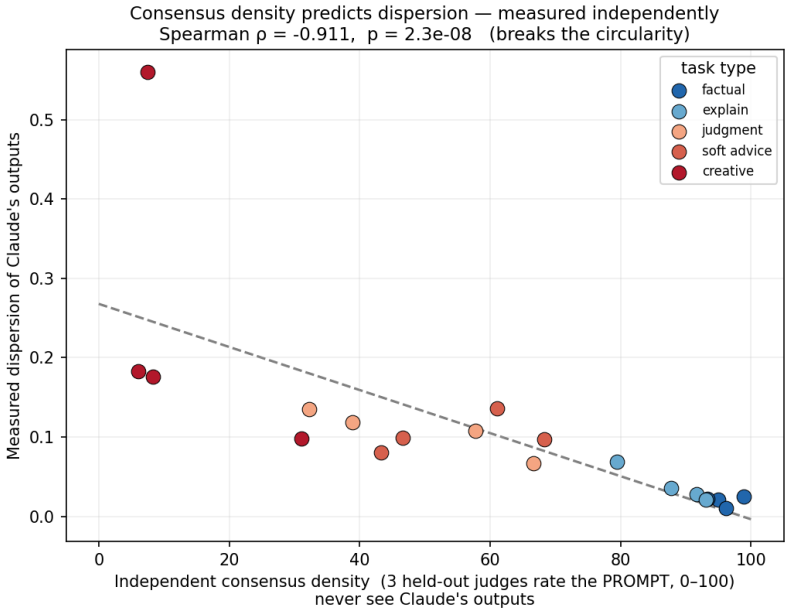

Figure 3: Consensus density, measured independently of dispersion. Three held-out judge models, shown only each prompt and never any output, rate how much knowledgeable people would agree on a good answer. Their pooled consensus scores predict Claude's measured dispersion at Spearman $\rho = -0.91$ ($p \approx 2 \times 10^{-8}$); each judge alone gives -0.88 to -0.90, and the judges agree with one another at $\rho = 0.85$–0.95. Consensus density therefore predicts dispersion rather than being inferred from it.

probe establishes the former and only suggests the latter (§3.9). Second, the panel's validation of the judges is itself domain-local—it shows the judges track human experts on these 18 Chinese pharmaceutical prompts, and carrying that validation over to the 20 English task prompts of the main set is an assumption of transfer, not a measurement; a panel on the main prompt set is the obvious next step (§5).

3.5 Alignment amplifies the gradient it inherits

The gradients above are all measured on aligned models. A sharper question is causal: does alignment *create* the dispersion gradient, or does it act on one already present in the pretrained model? We test this directly with a matched base-versus-instruct comparison on open weights. We ran the same five-band protocol—twenty prompts, N = 25 samples each, temperature 1.0, the same MiniLM embedder—on Qwen2.5-7B and its instruction-tuned sibling Qwen2.5-7B-Instruct, locally, so that the only difference between the two runs is alignment. The finding has three parts—the gradient is inherited, the base dispersion it rests on is genuine, and alignment amplifies it—and the last is load-bearing (Figure 4).

The ordering is inherited, not created. The base model already exhibits the full consensus-density gradient: its unconditioned dispersion rises across the bands from factual to creative-open at Spearman $\rho = 0.80$ (over the 20 prompts)—near-identical to the instruct model's $\rho = 0.79$. The *rank* structure—which tasks are high-consensus and which are low—is therefore a property of the pretrained distribution, present before any preference optimization. We are explicit on this point: alignment does not invent the centroid's shape.

The base dispersion is genuine, not degeneracy. Before reading anything off the base-versus-instruct difference, we rule out an artifact: base models are not optimized to answer directly, so their higher factual dispersion might reflect off-topic or incoherent completions rather than true low consensus—which would inflate the apparent collapse. It does not. The held-out judge consensus scores of §3.4—which never see any output—predict the *base* model's per-prompt dispersion at Spearman $\rho = -0.81$ ($p < 10^{-4}$), essentially as strongly as they predict the instruct model's (-0.77). The base model's high-consensus prompts are genuinely low-dispersion: the completions converge on shared content (every factual sample names Paris and the Eiffel

Tower, in varied wording), so the dispersion the collapse is measured against is real consensus-tracking, not noise.

What alignment does is amplify the gradient multiplicatively, and most where consensus is densest. Instruction tuning lowers dispersion at *every* band, but by sharply unequal amounts: the high-consensus factual band collapses $\approx$ 8.7-fold (computed on unrounded values; bootstrap 95% CI [6.5, 12.6]; 0.171 $\rightarrow$ 0.020, sampled completions converging to near-verbatim restatements of the same fact), while the low-consensus creative band falls only $\approx$ 1.8-fold ([1.6, 2.1]; 0.563 $\rightarrow$ 0.308) and the middle bands by $\approx$ 1.9–2.1-fold. The differential is not a single-draw artifact: bootstrapping the 25 completions per prompt, the factual-minus-creative fold-change is 6.9$\times$ with 95% CI [4.6, 10.7], excluding zero (Figure 4, error bars). Consistent with the rest of this paper, the claim we rest on is a *rank* one rather than an absolute magnitude: across the 20 prompts the per-prompt collapse *ratio* rises with independently-judged consensus (Spearman $\rho = 0.47$, p = 0.04)—alignment multiplicatively compresses higher-consensus prompts more. We flag the fragility of that last p-value plainly: it is nominal, would not survive a multiple-comparison correction across the correlations this paper reports, and serves only as per-prompt corroboration of the band-level fold-difference above, whose bootstrap CI excludes zero and which is the result we rest on. We are precise that this is a multiplicative effect: the *absolute* reduction does not order this way (the wide creative band sheds more raw dispersion), so we claim amplification in ratio, not in raw magnitude, and the point fold-changes—like all absolute levels here—are batch-dependent (§3.9). Alignment is thus not the *source* of the gradient but its *amplifier*: it pulls each region toward a point roughly in proportion to that region's consensus density, driving the high-consensus end to near-collapse while leaving the low-consensus field wide. This is the precise, distribution-level sense in which we say alignment *amplifies* the structure it inherits, and it tightens the construct's logic—because the ordering pre-exists alignment, consensus density is not an artifact of the alignment process we then use it to explain.

This is a single open-weight family and English prompts, and the embedder is the same one used for the Claude gradient, so the comparison is in commensurable units; §5 notes what a broader sweep would add. But the direction is unambiguous and matches the prediction: alignment converts a latent, gently sloped density gradient in the base model into the steep dispersion gradient an aligned model presents to its users.

3.6 Conditioning relocates the sampling region

We next conditioned a single open prompt—the story opening—several ways and measured how far each moved the cloud. Across every conditioned arm the centroid shift is large (0.54–0.78 in embedding space): each conditioned cloud occupies a neighborhood well separated from the unconditioned one. To calibrate these distances against pure sampling noise we computed two no-intervention baselines: the centroid distance between random halves of a single unconditioned cloud (mean 0.17, 95th percentile 0.22 over 1,000 resamples) and the centroid distance between two *independent batches* of the very same unconditioned prompt (0.24). Every conditioned shift exceeds even the harsher cross-batch baseline by a factor of 2–3, so the relocation is an effect of the conditioning, not an artifact of resampling. Prompting is not a dial that turns one distribution up or down; it is a *translation* that carries the model to a different neighborhood.

We call this relocation *steering*, in the prompt-level sense already distinguished from *activation* steering in §2.2: what we steer is the *sampling region in output space*, by conditioning, not the activations.

3.7 Spread belongs to the destination, not to the instruction

If prompting relocates, then whether the output spreads or tightens should depend on where it lands, not on how constraining the instruction looks—and it should move in *both* directions. We test this on the story-opening prompt across two independent sampling batches, and to keep the two sets of numbers straight we label them throughout: **Batch 1** is the bidirectional test (Figure 5) and **Batch 2** the specificity ladder (Figure 6). Their *absolute* dispersion levels differ, and the drift is uneven across arms (the comparison appears in the ladder discussion below), but the *ordering* of arms is identical in both batches—so we compare arms only *within* a batch and rest the cross-batch claim on sign and ordering, never on a number from one batch read against a number from the other. In Batch 1 the unconditioned dispersion is stable (0.189, sd 0.005

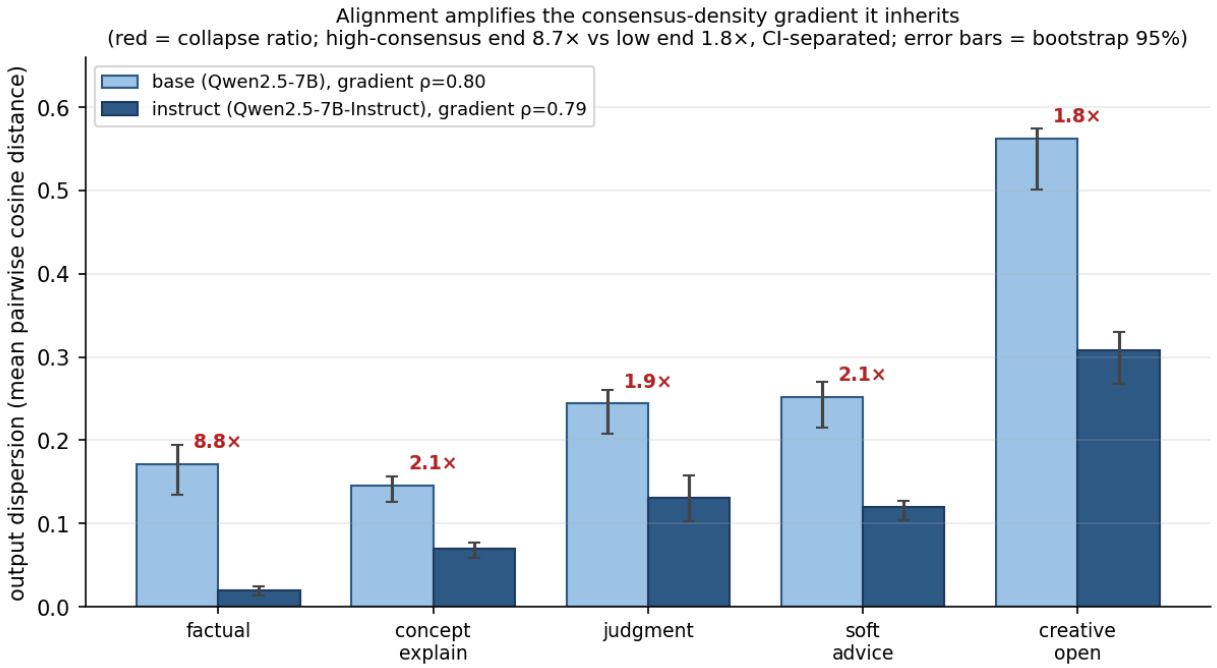

Figure 4: A matched base-versus-instruct test on open weights (Qwen2.5-7B vs Qwen2.5-7B-Instruct), same protocol as §3.3; error bars are bootstrap 95% CIs over the 25 completions per prompt. Both models share the consensus-density ordering (gradient $\rho = 0.80$ vs 0.79), so alignment does not create it. But instruction tuning collapses dispersion unequally—$\approx 8.7\times$ at the high-consensus factual end versus $\approx 1.8\times$ at the low-consensus creative end (red labels), a difference of $6.9\times$ whose bootstrap CI [4.6, 10.7] excludes zero. The collapse ratio rises with independently-judged consensus ($\rho = 0.47$). Alignment amplifies, multiplicatively, the gradient it inherits.

across five independent runs of N = 25), and we report each conditioned arm as a mean over four independent runs so that only effects surviving replication count.

Dispersion moves symmetrically with the destination's consensus (Figure 5, all numbers Batch 1). Relocating to a *low-consensus* destination raises dispersion: an instruction to avoid any opening a typical writer would choose lifts it from 0.189 to 0.345 (sd 0.013). Relocating to a *high-consensus* destination lowers it, well below the default: a content constraint that pins a single specific scene (a gray, still, empty sea under a cold sky, no people or movement) drops dispersion to 0.086 (sd 0.001 across four runs)—well below the unconditioned value and rock-stable within the batch. Conditioning is therefore not one-sided. The default is not a floor beneath which dispersion cannot fall; it is simply the resting point, and a sufficiently high-consensus relocation falls below it exactly as the single principle predicts.

Surface form does not predict the effect. Two instructions can look comparably "constraining" and do opposite things, because what matters is the consensus of the destination, not the syntax of the instruction. The scene-pinning constraint above and "in the style of Ernest Hemingway" are both narrowing in appearance, yet the first collapses dispersion to 0.086 while the second moves it little and unreliably (see below). One cannot read a prompt's effect on dispersion off the form of the instruction; one can read it only off the consensus density of the destination. **Form is not consensus.**

A specificity ladder makes the point directly (Figure 6, all numbers Batch 2). To check that the downward effect is not an artifact of a single, heavily over-constrained prompt, we ran a graded series on the same base prompt in a second batch (Batch 2; four runs each, N = 25): a *light* thematic constraint ("the story is about an old fisherman"), a *medium* scene constraint ("...returning to a quiet harbor at dawn with an empty boat"), the *strict* scene-pin above, and the diverge instruction. Dispersion is **non-monotone**

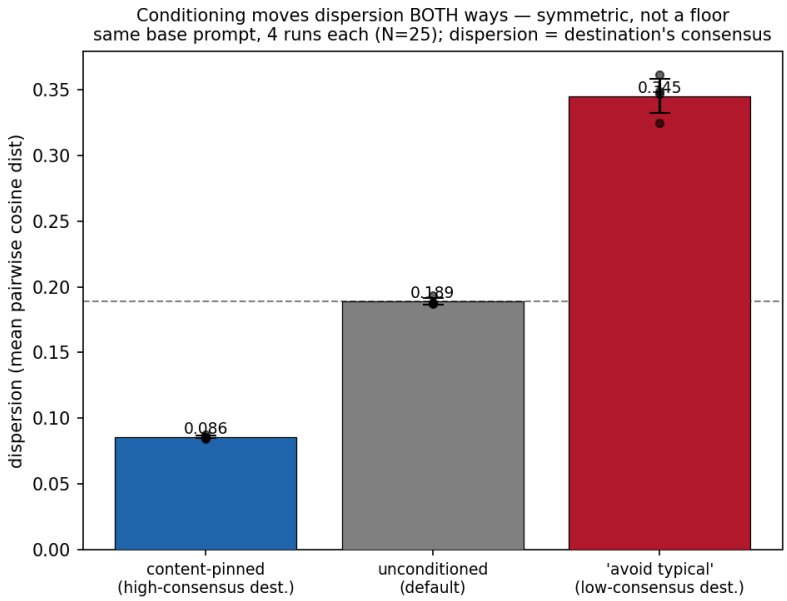

Figure 5: Conditioning moves dispersion in both directions (Batch 1). From the same base prompt ("opening sentence of a sea story"), a content constraint that relocates to a high-consensus destination drives dispersion to 0.086, well below the unconditioned default (dashed line, 0.189); an instruction to avoid typical openings—a low-consensus destination—drives it to 0.345. Each bar is the mean of four independent runs (N = 25); error bars are sd, individual runs overlaid. The relation is symmetric: there is no one-sided floor, and dispersion tracks the destination, not the act of conditioning. (Absolute values are Batch 1; Figure 6 reports Batch 2, whose levels differ but whose ordering is identical—§3.7, §3.9.)

in how much the instruction constrains. The light thematic add does not lower dispersion—if anything it trends upward (0.264 vs Batch 2's unconditioned 0.229; n.s. at n = 4)—because naming a subject opens a wide narrative space rather than pinning a high-consensus scene. The two scene constraints drop it clearly (0.160 and 0.152; both p = 0.029 vs unconditioned), while the diverge instruction—by far the longest and most elaborate—raises it most (0.340; p = 0.029). (With four runs per arm, p = 0.029 is the smallest value a two-sided Mann–Whitney test can attain; these are floor values, and we read them as direction plus replication, not as precision.) A rank correlation over the content ladder is negative (Spearman $\rho = -0.75$ across runs; we attach no p value, since the ladder has only four distinct rungs and the runs within an arm are repeated draws of the same arm, not independent observations), but the load-bearing observation is the non-monotonicity: constraint *amount* does not predict the *direction* of the effect; the destination's consensus does. This is the sharpest form of *form is not consensus*. Two caveats are of calibration, not of sign: the light arm's upward trend is underpowered, and Batch 2's absolute levels differ from Batch 1's by an amount that varies sharply across arms—the low-consensus diverge arm barely moves (0.340 vs 0.345), the unconditioned default rises modestly (0.229 vs 0.189), and the high-consensus scene-pin rises most in relative terms (0.152 vs 0.086)—so the *magnitude* of the downward effect is batch-dependent while its *direction* is not. Crucially the *ordering* is identical in both batches (scene-pin < unconditioned < diverge), and that ordering, together with the across-task gradient (§3.3), is what we rest on, not the absolute number (§3.9).

The destination's consensus is measured, not asserted. The ladder argument so far still contains an author-assigned label: *we* called the scene-pin a high-consensus destination and the diverge arm a low-consensus one. To remove that residue of relabeling we extended the §3.4 protocol to the conditioned prompts themselves: the same three held-out judges, under the same prompt-only elicitation (0–100, three repetitions per judge, never seeing any output), scored every conditioning arm. Their pooled ratings recover the destination ordering we had assumed—the scene-pin at 79 and the medium scene constraint at 59 on the high end, the unconditioned base at 17, and the diverge instruction at 14 on the low end—and they

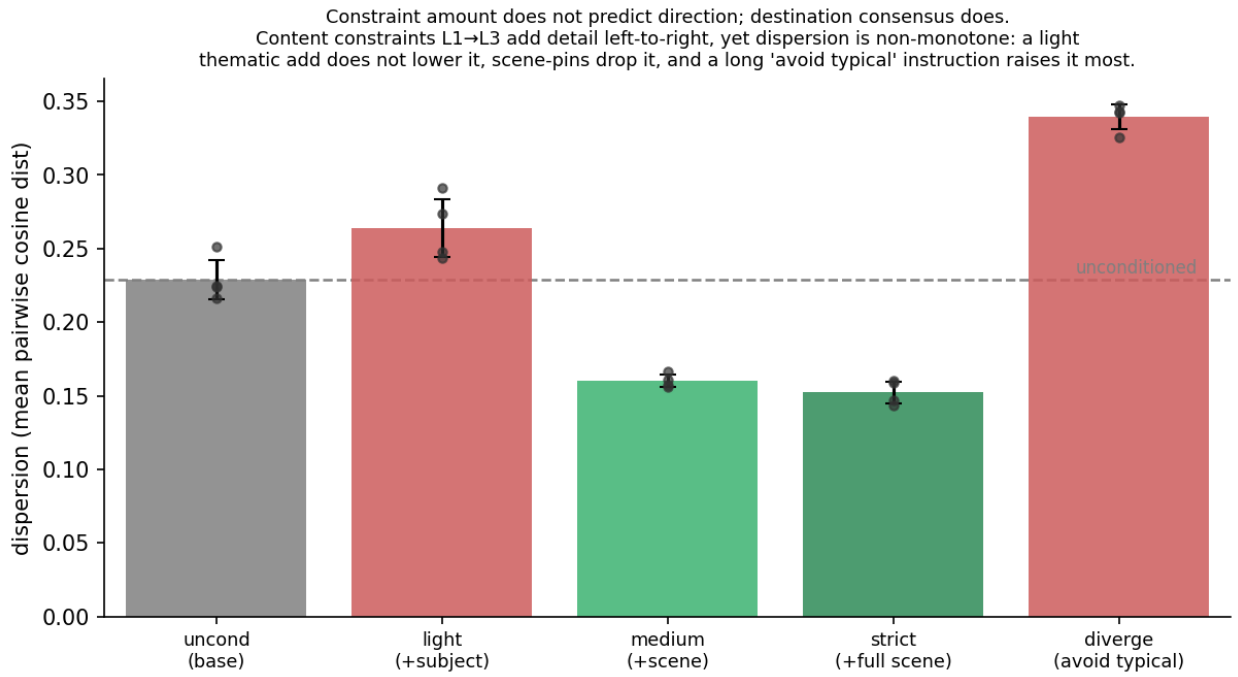

Figure 6: A specificity ladder on the story-opening prompt (four runs each, N = 25; dashed line = unconditioned). Content constraints add detail from left to right (light → medium → strict), yet dispersion is non-monotone: a light thematic constraint does not lower it, scene-pinning constraints drop it clearly, and a long "avoid any typical opening" instruction raises it most. Constraint amount does not predict the direction of the effect—the destination's consensus does. All values are Batch 2; its absolute levels differ from Batch 1's (Figure 5) by an arm-dependent amount (unconditioned 0.229 vs 0.189), but the ordering is identical—the sign and ordering, not the magnitude, are what replicate across batches (§3.9).

predict the Batch-2 dispersions at arm level $\rho = -0.90$ (n = 5; exact permutation p = 0.083 two-sided, 0.042 one-sided under the direction §3.4 fixed in advance). We state the granularity plainly: with five arms even this near-perfect ordering cannot reach conventional two-sided significance, and we do not borrow strength from the run level—the per-run correlation ($\rho = -0.87$ over 20 runs) is descriptive only, because the four runs within an arm share that arm's single judge score and are not independent observations. What the judge scoring establishes is therefore corroboration at coarse granularity of the ordering we had assumed—the same de-circularization §3.4 performed for the task gradient, here at pilot strength. On the falsifiability of the conditioning claims more broadly, the load is carried by the dispersion side—the arm ordering that replicates across two independent batches—with the judge scores as the corroborating, not the confirming, leg.

Three details deserve flags rather than burial. First, inter-judge agreement across these seven near-duplicate prompts is lower than on the 20-prompt set (ρ = 0.49–0.85, vs 0.85–0.95)—unsurprising, since the arms differ by single clauses rather than by task type, and consistent with the granularity limit discussed in §3.4. Second, the one rank inversion sits exactly at the arm we already report as underpowered: judges score the light thematic constraint well above the base (47 vs 17), yet its measured dispersion trends *up*—prompt-side raters, like the surface-form reading, over-credit a light constraint that in fact opens narrative space. Third, the Hemingway arm cleanly isolates the metric's known blind spot: judges rate it as high-consensus as the scene-pin (82 vs 79)—a consensus that lives on the *style* axis—while the semantic metric, blind to style by construction (see below), barely moves; the one large judge–dispersion disagreement falls exactly where the measurement boundary says it must. The year-3000 arm completes the picture from the null side: judges score its consensus at the base's level (24 vs 17), and its dispersion is statistically indistinguishable from the base's—relocation without a consensus change leaves spread unchanged, precisely as the principle predicts.

Generality across base prompts. Everything above is one base prompt. To test whether the pattern is a property of that prompt, we ran a compact ladder (unconditioned / light thematic / strict scene-pin / diverge; three runs each, N = 25, every arm judge-scored under the same protocol) on three new bases from different genres: a story set in a city, a poem about time, and a university welcome speech. The *upward* direction is universal: the diverge arm raises dispersion far above the default on every base (city 0.44 vs 0.23; poem 0.50 vs 0.18; speech 0.59 vs 0.09), and the judges score every diverge destination at the bottom of the consensus scale (14–27)—four base prompts, four large, same-direction effects. The *downward* direction is conditional, and the boundary is itself informative. On the city and poem bases the scene-pin trends below the default only marginally (0.227 vs 0.234; 0.175 vs 0.185; n.s. at three runs)—these pins are evidently looser than the sea-story's. A pre-registered follow-up batch tested exactly that reading: we re-ran both bases with the original pin and a tighter one copying the sea-story recipe (a single static scene, explicit color and weather, no people, no objects, no movement; four runs per arm, all arms in one batch). On the city base the tighter pin drives dispersion clearly below the same-batch default (0.181 vs 0.232; two-sided 4-vs-4 Mann–Whitney p = 0.029, the design's floor value, every pinned run below every default run), while the original pin again moves only marginally (0.215, p = 0.34)—within a single batch, the same base steps from no effect to a clear below-default effect as the pin's destination consensus rises. On the poem base the tighter pin trends below (0.160 vs 0.193) but one of four runs overlaps the default range (p = 0.11), so we report direction only. The below-default direction is therefore not a property of the sea story; what it requires is a genuinely higher-consensus destination, not more specification words—and, consistent with the over-crediting pattern above, the prompt-side judges do not separate the two pins (tighter vs original: city 68 vs 63; poem 63 vs 73).

On the speech base the relation inverts: the unconditioned welcome speech is already the highest-consensus destination in our entire set (judges rate the bare base at 54, vs 12–17 for the other bases; its dispersion, 0.090, is the lowest unconditioned value among our four base prompts), and every added constraint—including a fully pinned scene the judges rate at 80—*raises* dispersion (strict 0.169), because a concrete scene opens more semantic variety than the formulaic register of the bare genre. This is the principle's own arithmetic: below-default dispersion requires a destination of *higher* consensus than the default's own region, and when the default already sits near the top of the range there is no headroom. The judges' miss on those arms is the same miss as on the light arm above—prompt-side raters over-credit specification—so the pooled judge-side prediction across all twelve arms holds in direction but attenuated (arm-level $\rho = -0.56$, n = 12, p = 0.06; the per-run view, $\rho = -0.52$ over 36 runs, is descriptive only, since runs within an arm share the arm's single judge score), consistent with the granularity limit of §3.4. The honest summary: raising dispersion by relocating to a low-consensus destination is easy and universal; driving it *below* the default is demonstrated on two of four bases (the sea-story pin, stable across two batches; the city base under the tighter pin) and trends on a third (poem), bounded in every case by how much consensus headroom the default leaves; and surface specification masquerades as consensus for prompt-side raters too. *Form is not consensus—for human-style raters either.*

A measurement boundary, stated plainly. The Hemingway arm is the honest edge case. Across runs it tightens dispersion modestly in some draws and barely moves it in others—it is small and noisy on this metric, not a clean effect. Part of this is the metric: our dispersion is computed over semantic sentence embeddings, which capture *what a sentence is about* more than *how it sounds*, so a purely stylistic narrowing (Hemingway sea-openings still range over countless images and events) is only weakly visible. We therefore make no claim about style-axis tightening in either direction, and—importantly—we rest nothing on it: the symmetry result above is established with *content* constraints, which the metric sees clearly. This blind spot to style is shared by embedding-based diversity measures generally (Shypula et al., 2025) and is the complement to surface-form metrics like distinct-n and Self-BLEU (Li et al., 2016; Zhu et al., 2018), which would see style but miss content; §5 returns to a style-aware metric as the most informative direction for future work.

A note on what does not replicate. An earlier single run suggested two finer effects—an apparent inability of any arm to fall below the default, and a small widening from a purely contextual "set the story in the year 3000" relocation. Neither survives replication: the unconditioned dispersion is stable, but a one-time +0.1 shift of that size sits within the run-to-run noise of single N = 25 draws, and across five runs the year-3000 arm is statistically indistinguishable from the default (two-sided Mann–Whitney p =

0.69). We report this explicitly because it is the reason we now rely only on large effects measured over multiple runs (the symmetric content effects above, each $|\Delta| \gtrsim 0.1$ and stable), and on the across-task and independent-anchor correlations of §3.3–3.4, rather than on any single-draw comparison.

3.8 From mechanism to floor and ceiling

A floor-and-ceiling reading now follows as an *interpretation* of the measurements, and we are explicit about both the word and the premise it adds. Everything established so far is about dispersion; the reading below adds a two-part claim about *value*: that the unconditioned default is itself competent, and that in a wide field most samples are mediocre, so that realized value depends on selecting the good ones. We state both parts rather than establish them: a first, pre-registered simulated pass at testing the second—held-out LLM judges scoring sample quality on the stored completions, reported in full in §5—returned an inconclusive result, with the judge instrument lacking the within-prompt resolution the test requires; a first, small human forced-choice pass (also reported there) lost its confirmatory branch to an instrument failure of our own making and left only an exploratory remainder, trending against the premise at a sample size we do not interpret directionally—so the premise remains stated on both fronts. With that premise on the table, the reading runs as follows. The unconditioned default lands a user at the centroid—a competent, average output. For high-consensus tasks the centroid is tight: the novice is lifted to a competent answer and the expert has little room to add, the equalizer regime; here the centroid functions as a *floor* in the sense that everyone receives at least it by default. For judgment- and creativity-intensive tasks the centroid is diffuse, and the valuable outputs lie in low-consensus regions that are by construction high-dispersion. *Reaching* those regions is easy—relocation is cheap, in either direction, as §3.7 shows. *Profiting* from them is not, on the stated premise, because value is realized only by selecting the good samples from a wide and uneven field. That selection is taste. The ceiling is high exactly where dispersion is high, and high dispersion is exactly the condition under which judgment, not generation, becomes the binding constraint—the amplifier regime. (The floor here is a floor *on outcome quality*—a competent default, the first part of the stated premise—not the dispersion floor that §3.7 shows does not exist.)

This is why, on this reading, the same model yields a floor for one person and a ceiling for another. The model supplies a distribution; the human supplies the location (by conditioning) and the selection (by judgment). Where the distribution is tight, location and selection barely matter and outcomes converge. Where it is diffuse, they are everything, and outcomes diverge. On the stated premise, the model supplies the floor; only taste reaches the ceiling. The dispersion mechanism is what we measure; this human-outcome layer is what it predicts, and testing that prediction is the experiment the probe most directly invites: a simulated pass and a first small human pass are now both run and reported in §5; neither confirms the premise, and a resolved human test remains.

3.9 Scope of the probe

This is an illustrative probe, not a benchmark. Its core measurements no longer rest on a single run or a single model: the task-type gradient replicates on two unrelated hosted models (§3.3) and on a third, open-weight family run locally (§3.5), the governing quantity is recovered by three independent judges (§3.4), a matched base-versus-instruct comparison locates the gradient's origin in pretraining and alignment's role as its amplifier (§3.5), and the conditioning effects we report are means over multiple runs with run-to-run variation stated (§3.7).

Crucially, although absolute dispersion magnitudes drift between batches (§3.7), the rank correlations that carry our claims do not, and this is not a single lucky re-draw: across *five* independent fresh batches of all 20 prompts, the gradient holds at Spearman $\rho = 0.87 \pm 0.03$ (range [0.82, 0.90], vs 0.85 originally) and the independent-anchor correlation at $\rho = -0.90 \pm 0.03$ ([−0.92, −0.86]), while every pair of batches agrees on the per-prompt dispersion ranking at $\rho = 0.93 \pm 0.04$ (10 pairs, range [0.86, 0.97]). The rank statistics are thus tight across batches even as the absolute levels drift. The *size* of the drift is uneven and its profile differs by experiment—in the task-gradient replication it concentrates at the high-dispersion end, whereas among the §3.7 conditioning arms it is the high-consensus (low-dispersion) arm that moves most in relative terms

while the diverge arm barely shifts—but in every case it is near rank-preserving, which is exactly why a rank statistic is the right instrument and why we rest on orderings rather than absolute numbers.

The gradient is also not an artifact of the embedder: re-embedding the very same completions with a second, architecturally distinct encoder (all-mpnet-base-v2, 768-d) reproduces both the task-type gradient ($\rho = 0.78$, vs 0.85 under MiniLM) and the independent anchor ($\rho = -0.85$, vs -0.91), with the two embedders agreeing on per-prompt dispersion at $\rho = 0.88$; the base-versus-instruct gradients likewise survive the swap ($\rho \approx$ 0.68–0.69). Nor is it a length artifact—a live concern, since completion length varies systematically across task types (one-sentence factual answers versus paragraph-long advice) and short strings embed noisily: mean completion length correlates only weakly and non-significantly with dispersion ($\rho = 0.32$, p = 0.18, n = 20), and controlling for length leaves both the gradient (partial $\rho = 0.84$, vs 0.85) and the judge anchor (partial $\rho = -0.90$, vs -0.91) essentially unchanged; additionally dropping the short-string naming outlier leaves the anchor at partial $\rho = -0.88$.

One further calibration note belongs here rather than in fine print: the prompt set spans task types from "the capital of France" to open fiction—nearly the full consensus range by construction—and a broad range makes strong rank correlations easier to achieve. The within-domain human-panel set (§3.4, Appendix A.3), where the range is restricted to one professional field, shows the expected attenuation (judge $\rightarrow$ dispersion -0.56 there, vs -0.91 here). Our headline correlations should therefore be read at the granularity at which they are measured: coarse, between-task-type differences, not fine within-domain ones. It remains limited in the ways §5 details—a small number of models and prompts per type, a content-only (style-blind) semantic metric, a closed generation endpoint, and a main consensus anchor that is judge-based (with preliminary human cross-checking, §3.4). To fix the boundary of the probe explicitly, we separate what the evidence supports from what it does not. *We claim:*

(i) the task-type gradient—dispersion rises as task consensus falls—appears as a clear, replicated rank correlation (Claude $\rho = 0.85$, GPT-4o $\rho = 0.84$, open-weight base/instruct $\rho = 0.80/0.79$, five fresh batches $\rho = 0.87 \pm 0.03$, second embedder $\rho = 0.78$);

(ii) consensus density predicts dispersion from an independent, output-blind measurement ($\rho = -0.91$), so the construct is not circular;

(iii) conditioning relocates the sampling region (shifts 0.54–0.78 vs a same-prompt resampling baseline of 0.24) and dispersion moves *symmetrically* with the destination's consensus—corroborated at arm level by prompt-side judges ($\rho = -0.90$, n = 5, exact p = 0.083; §3.7)—not with the instruction's surface form, with the upward direction universal across four base prompts and the below-default direction replicated on a second base under a tighter pin (§3.7), bounded by the default's own consensus headroom;

(iv) alignment amplifies the inherited gradient multiplicatively (factual collapse $\approx 8.7\times$ vs creative $\approx 1.8\times$, difference $6.9\times$ with CI excluding zero).

We do not claim:

- that absolute dispersion magnitudes or fold-changes are calibrated—they drift between batches, so we rest on rank statistics;
- that *raw* dispersion tracks *human* consensus directly—this link is non-significant at our sample size ($\rho \approx -0.35$, n = 18, §3.4);
- anything about style-axis tightening—our metric is content-only by construction (§3.7).

We also flag, rather than report as results, the comparisons that did not replicate (§3.7). Table 1 maps each claim we do make to the independent evidence that carries it. We treat the results as directional evidence for the principle, and §5 lays out the extensions that would calibrate it.

Table 1: Evidence map for the four core claims. The headline column is the primary measurement; the right column lists the independent replications and robustness checks—across two hosted models, an open-weight base/instruct pair, a fresh sampling batch, a second embedder, and (for the construct) a human panel—that carry each claim at the rank level. Absolute magnitudes are batch-dependent (§3.9); the claims rest on these orderings and signs, not on point values.

Claim	Headline statistic	Independent replications / robustness
C1. Dispersion tracks task consensus (the gradient, §3.3)	Claude $\rho = 0.85$ ($p \approx 1.8\times10^{-6}$, $n = 20$)	GPT-4o $\rho = 0.84$; cross-model agreement 0.75; five fresh batches 0.87 ± 0.03 (pairwise cross-batch rank 0.93 ± 0.04); mpnet embedder 0.78; leave-one-out [0.83, 0.89]; Qwen base/instruct 0.80/0.79
C2. Consensus density is measured independently—non-circular (§3.4)	Held-out judges → dispersion $\rho = -0.91$ ($p \approx 2\times10^{-8}$)	Per-judge −0.88 to −0.90; inter-judge 0.85–0.95; mpnet anchor −0.85; open-weight per-step entropy 0.66 vs dispersion, −0.70 vs consensus (GPT-4o surprisal 0.51); human experts → judges 0.74 ($p < 10^{-3}$)
C3. Conditioning relocates; spread is symmetric in the destination's consensus (§3.6–3.7)	Scene-pin 0.086 < default 0.189 < diverge 0.345 (Batch 1)	Specificity ladder $\rho = -0.75$ across runs (descriptive; four distinct rungs); arm ordering identical in Batch 2; both scene arms $p = 0.029$ (the 4-vs-4 Mann–Whitney floor); destination consensus judge-scored, recovering the assumed ordering: arm-level −0.90 ($n = 5$, exact $p = 0.083$ two-sided / 0.042 one-sided; per-run view descriptive, runs share arm scores); shifts 0.54–0.78 vs same-prompt resampling null 0.24; diverge direction replicates on 3 further bases (4/4 total); below-default replicates under a tighter pin on a second base (city 0.181 vs 0.232, $p = 0.029$; poem trends, n.s.), bounded by the default's consensus headroom
C4. Alignment amplifies the inherited gradient multiplicatively (§3.5)	Factual ≈ 8.7× vs creative ≈ 1.8×; difference 6.9×, CI [4.6, 10.7]	Collapse ratio vs consensus $\rho = 0.47$ ($p = 0.04$); base-model anchor −0.81; mpnet base/instruct 0.68–0.69

4 Implications

This section is deliberately interpretive: it builds on the measured dispersion mechanism of §3 together with the stated two-part value premise of §3.8 (a competent default; mostly-mediocre wide fields), and should be read at that strength.

4.1 A mechanism for the complexity boundary

The equalizer/amplifier reconciliation (Anonymous, 2026) is, on the present account, a corollary of one measured quantity. "Task complexity" is doing the work of a proxy for consensus density, and the proxy is imperfect in instructive ways. A task can be cognitively demanding yet high-consensus—a hard but standard derivation, a well-specified refactor—and there the model still collapses onto a tight, competent centroid; difficulty does not by itself widen the field. Conversely, a task can be trivially easy yet low-consensus—"name my cafe," "pick an opening line"—and there the field is wide despite the absence of any "complexity." What governs whether outcomes converge or diverge is not how hard the task is but how much agreement exists about what a good answer looks like. Replacing complexity with consensus density thus predicts the off-diagonal cases the complexity story mislabels, and it makes the boundary continuous and measurable rather than categorical: one can, in principle, place any task on the gradient of Figure 2A and read off whether it sits in the equalizer or amplifier regime.

4.2 Taste as selection, not generation

Because the default is a competent average—on high-consensus tasks often the *safest* available answer—the lever an expert pulls is rarely "make the model better than its default" on those tasks; there it is neither easily possible nor necessary. The lever is, instead, to *leave* the centroid for a high-value, low-consensus region and then to *select* from the wide field found there. This reframes taste as two separable capacities. The first is direction—the ability to relocate the model deliberately, which is a skill of expression: knowing which region to steer toward. The second is selection—the ability to recognize the good sample among many mediocre ones, which is a skill of judgment: knowing the good when one sees it. Both are taste; they are not the same faculty, and a person can have one without the other. The expert–novice literature locates genuine expertise in the second (Chi et al., 1981; Flavell, 1979); our mechanism says the second is load-bearing precisely in the low-consensus regime where the first has carried the model into a wide field.

A caution follows on the limits of taste. Selection can only choose among samples the model actually produces; it cannot conjure a correct answer the model never places in the field. On the correctness axis, this is where taste turns dangerous: a user whose discrimination does not actually exceed the centroid will, by pushing the model off its safe default, often trade a competent average for a confident error—and the model's sycophancy will oblige, agreeing its way toward whatever the user signaled (Malmqvist, 2024; Chen et al., 2025). Taste improves outcomes only when the human's discrimination genuinely exceeds the centroid's; below that bar, leaving the floor lowers quality. The floor is protective for exactly the users least able to improve on it.

4.3 Designing for the regime

Reading the two regimes off a measured quantity suggests interfaces that adapt to it. In the high-consensus (equalizer) regime, the right design minimizes friction: the tight centroid is the answer, and asking the user to choose among near-identical samples wastes effort. In the low-consensus (amplifier) regime, the right design does the opposite—it *surfaces the dispersion* rather than hiding it behind a single sampled reply, presenting a diverse slate and investing the interface budget in helping the user select, since selection is the binding constraint. The diversity-promoting decoding and prompting literature (Holtzman et al., 2020; Meister et al., 2023; Zhang et al., 2025) supplies levers to widen the slate on demand; the contribution of a consensus-density view is to say *when* to spend them—only where the destination is genuinely low-consensus, since elsewhere the extra spread is noise around an already-correct floor. An interface that could estimate dispersion on the fly (a cheap multi-sample probe) could route between the two modes automatically: collapse to the floor when consensus is high, open the field when it is low.

The broader point for hybrid-intelligence design (Anonymous, 2026; Vaccaro et al., 2024) is that "make the AI better" is the wrong target in the amplifier regime. The model already supplies a competent default (the stated premise of §3.8); what determines outcomes there is whether the human can select from the field it opens. Systems that aim to *replace* judgment will equalize a population onto the floor and, at the collective level, shrink the diversity of what gets made (Doshi & Hauser, 2024). Systems that aim to *develop and reward* judgment will let the floor lift the novice while leaving room for the expert to reach the ceiling. The choice between those two is a design decision, and the consensus-density gradient is where it should be made.

5 Limitations and Future Work

The probe is deliberately small, and several of its limits point directly at the experiments that would harden the claims.

A small set of models and prompts. Our strongest claims now have cross-model support—the gradient replicates on Claude and GPT-4o (§3.3) and the consensus anchor draws on three further model families (§3.4)—and conditioning effects are reported over multiple runs. But the study still spans few models and only four prompts per task type, and apart from the single open-weight pair of §3.5 the generation models are instruction-tuned API models. Effect sizes are not calibrated, and the per-type spread is wide at the open end. More prompts per type—especially to resolve the merged middle band (judgment $\approx$ soft

advice)—and replication across more model families and scales would test whether the gradient's *shape*, not just its monotonicity, is stable.

The metric sees content, not style. Our dispersion is computed over semantic embeddings, so it is at best weakly sensitive to stylistic narrowing (§3.7): the Hemingway constraint moves the number only a little and inconsistently across runs. We therefore make no claim about the style axis in either direction, and we rest the symmetry result on content constraints, which the metric sees clearly. The most informative extension is a style-aware metric—surface-form measures such as distinct-n and Self-BLEU (Li et al., 2016; Zhu et al., 2018), or a syntactic/stylometric embedding—run alongside the semantic one, to test whether the symmetric relation we find on content also holds on style.

Centroid as inherited vs. produced by alignment: tested, with a nuance. §3.5 runs the clean base-versus-instruct comparison on matched open weights and returns a more precise answer than "alignment creates the centroid." The ordering is *inherited*—the base model already has the gradient at $\rho = 0.80$—while alignment *amplifies* it multiplicatively, collapsing the high-consensus end $\approx 8.7\times$ (bootstrap CI [6.5, 12.6]) against $\approx 1.8\times$ ([1.6, 2.1]) at the low-consensus end, a difference whose CI excludes zero, with the per-prompt collapse ratio rising against the independent judge anchor ($\rho = 0.47$). This is consistent with RLHF reducing diversity relative to SFT (Kirk et al., 2024) and shrinking the generative horizon (Yang et al., 2025), and it sharpens those findings into a *task-dependent* law: the diversity loss is concentrated where consensus is dense. Two limits remain. The fold-changes are bootstrapped within a single batch, so they carry within-run but not between-batch uncertainty; we therefore rest the claim on the batch-robust rank statistic (ratio vs. consensus) and report the magnitudes as batch-dependent (§3.9). And the test spans a single open-weight family (Qwen2.5-7B) and English prompts. Replicating the amplification pattern across more base/instruct pairs and model scales, and across multiple fresh batches to put between-batch error bars on the fold-changes, would test whether the amplification law, not just its direction, is stable.

Independence of consensus density: addressed, not closed. §3.4 breaks the central circularity—three held-out judges, scoring only the prompt and never the outputs, predict dispersion at $\rho = -0.91$—so consensus density is no longer merely read back from the dispersion it explains. Two gaps remain. First, the judges are *model* proxies for human agreement—and a preliminary human panel (§3.4) is reassuring on this point: four domain experts (three practitioners, one academic) on a contested Chinese pharmaceutical-market set agree with the held-out judges at $\rho = 0.74$, evidence they track human consensus rather than only one another. We keep this preliminary—the panel is small (n = 18 prompts, four raters) and its inter-rater agreement only modest (W = 0.58)—and a larger human study, including whether *raw dispersion* (not just the judge-measured construct) tracks human consensus directly—a link that stays below significance at this sample size—remains future work. Second, the judge models and the generator share overlapping training data, so part of their agreement could reflect a common source rather than independent validation; pairing an open-weight generator with a held-out human panel would separate these. A third line corroborates the construct from outside model judgment entirely: GPT-4o token surprisal rises with dispersion ($\rho = 0.51$), and—more cleanly, since open weights expose the true next-token distribution rather than an API proxy—the exact per-step predictive entropy of Qwen2.5-7B-Instruct rises with dispersion ($\rho = 0.66$) and falls with judged consensus ($\rho = -0.70$).

From dispersion to outcomes: the selection experiment, simulated and first human passes both run, premise still unestablished. Everything this paper measures is output dispersion. The equalizer/amplifier reading (§3.8, §4) adds a layer about human outcomes—that in a wide field value is realized by selection, so that the selector's discrimination matters most where dispersion is highest—and that layer rests on a premise about sample quality we state but do not establish. The direct test is to score sample quality independently across the dispersion gradient and let selectors of varying discrimination choose among the samples; the mechanism predicts the selector's skill should be nearly irrelevant at the factual end and decisive at the open end. We ran the simulated half of this test on the stored completions, with the confirmatory analysis pre-registered in the experiment log before any quality score was produced (Appendix A.2): three held-out judges (the consensus judges of §3.4; the generator is excluded) rated every completion's quality individually, and we measured the *selection headroom*—what a noisy-argmax selector gains over a random draw, in raw 0–100 points—under a leave-one-judge-out split, selecting on one judge's scores and evaluating on the other two, so that judge noise cannot inflate the gain through the winner's curse.

The pre-registered test (Spearman of cross-fitted headroom against dispersion, one-sided, threshold $\rho \geq 0.38$ at $p < .05$) was *not* met: $\rho = +0.23$ ($p = 0.16$, n = 20; +0.11 with the naming outlier dropped).

The diagnosis sits in the instrument: the judges compress every completion into the top of the scale (band-mean quality 89–98, within-prompt SD 0.9–2.5 points), four prompts at the high-consensus end saturate the ceiling outright, and inter-judge agreement on quality does *not* fall with dispersion as it should if the judges were genuinely resolving creative quality. Two weaker signals point in the predicted direction and we report them as exploratory, not pre-registered: the cross-fitted headroom is concentrated where the premise puts it (creative band 1.17 points against ≈ 0 elsewhere), and the probability that the selected sample beats a random one—a purely ordinal statistic—rises with dispersion ($\rho = +0.43$, one-sided $p = .03$, outside the pre-registration; reaching 0.69 at the open end against the 0.5 chance level). The naive version of the headline statistic, selecting and evaluating on the same scores, comes out significant at $\rho = +0.50$; we report it only as the upper bound the cross-fitting exists to discipline—at the open end a judge's self-evaluated gains exceed its held-out gains three- to nine-fold, so without cross-fitting this experiment would have returned a false positive.

The honest summary of the simulated pass: a first-pass LLM-judge instrument lacks the within-prompt resolution this premise requires, and the data cannot separate the two available readings—that quality in low-consensus fields is genuinely contested (the construct's own prediction) or that single-completion absolute ratings simply inherit the judges' documented leniency and top-compression biases. It left one protocol lesson: elicit forced choices or rankings, not absolute ratings. (Per-prompt numbers, the skill-curve translation, and a diagnostic figure are in the supplementary material.)

A first, small human pass followed, implementing exactly that lesson, with the confirmatory analysis pre-registered before any vote was collected (the same test and threshold as the simulated pass: one-sided Spearman of selection performance against dispersion, $\rho \geq 0.38$ at $p < .05$, with and without the naming outlier). Four selectors each picked the best of five completions per prompt across all twenty prompts; four separate evaluators then judged each pick against a random same-prompt completion in blind, side-randomized forced-choice pairs (two votes per pair), with the selector and evaluator roles disjoint so that self-evaluation cannot inflate the gains—the human analogue of the cross-fitting above. The pre-registered branch returned *no analyzable data*, through our own instrument error: two of the three attention-check items used length-truncated completions as the intended wrong answer while the evaluator instruction said to judge what is present, so a careful evaluator following the instruction could legitimately "fail" them, and under the pre-registered all-checks-correct inclusion rule no evaluator qualified. We logged this as an instrument failure and committed the handling in the experiment log before computing any result: report the deviation analysis that admits all four evaluators, exploratory only, whatever its direction. That analysis runs against the prediction: the picked sample beats a random one 53% of the time overall ($\approx$ chance), 47% on the creative band (where the premise needs it most; four prompts, one tied at exactly 0.5, so the pre-registered secondary sign test has three informative prompts) against 66% on the factual band—where near-identical completions should make the choice arbitrary, so we read the pattern as instrument noise rather than signal, since the half that contradicts the premise is no better supported than the half that is implausible on its face—and the prompt-level Spearman against dispersion is -0.37 (one-sided p = 0.94; -0.31 without the naming outlier). One human signal does echo the construct: evaluator agreement falls as dispersion rises ($\rho = -0.38$ over the twenty prompts, descriptive)—the within-prompt analogue of the consensus gradient, and precisely the signal the simulated pass's LLM judges failed to show. The pass is small on every dimension that matters (four evaluators, two votes per pair, selectors choosing among five of the twenty-five stored completions, participants reading English as a second language), so we draw no directional conclusion from it. What the two passes jointly establish is that the premise has now survived neither of our first two instruments, and the pre-registered negative protocol applies: either selection value in low-consensus fields is genuinely absent on this sample set—in which case the floor reading carries more of the story than the ceiling reading—or neither instrument so far has the resolution to detect it. Until a resolved human test exists, the floor/ceiling vocabulary should be read as the interpretation we offer of the measured mechanism, not as a measured human-outcome result. (A protocol summary is given in Appendix A.4; materials, raw votes, the pre-registered protocol, and its timestamped amendments are in the supplementary material.)

Dispersion versus semantic entropy. The nearest neighboring quantity in the literature, semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), is computed from the same raw material (repeated samples) but through entailment clustering, and is read as model uncertainty rather than task openness (§2.2). Computing both quantities on the same completions across our gradient would calibrate the two constructs against each other—in particular, whether they diverge on open prompts, where the consensus-density reading predicts high dispersion without any implication of model failure.

Closed endpoint and decoding. The main generation model is accessed through a hosted endpoint at fixed temperature 1.0; we cannot inspect its logits or vary the decoding regime, and provider-side sampling details are opaque. The direct comparison between dispersion and token-level entropy that this calls for now exists for one open-weight model (the Qwen2.5-7B-Instruct entropy correlations reported above); what remains is to vary temperature and to extend that single model pair across scales and decoding regimes, tying the embedding-space measure to a distributional one over the sampling settings we could not manipulate here (Yang et al., 2025).

None of these undercut the claims the probe is sized to support—the replicated task-type gradient, the independent grounding of consensus density, relocation-not-sharpening, and the symmetric dependence of dispersion on the destination's consensus—but each marks where a heavier study would convert directional evidence into a calibrated one.

6 Conclusion

"An aligned model regresses to the mean" is usually a metaphor; we have tried to make it a measurement. Sampling an instruction-tuned model and measuring the dispersion of its output cloud, we find a preference centroid whose tightness is governed by a single quantity—the consensus density of the region being sampled—and we state the governing relation as a principle: output dispersion is, to first approximation, the inverse of consensus density. Conditioning does not sharpen the model; it relocates the sampling region, and the spread at the destination belongs to the destination, not to the instruction—form is not consensus. The relocation is symmetric: a high-consensus destination drives dispersion below the unconditioned default and a low-consensus one drives it above, so the default is not a floor but simply the resting point when nothing has moved the model. And the governing quantity is independently real—held-out models that never see the outputs recover it from the prompt alone and predict the dispersion it sets ($\rho = -0.91$), and the gradient itself replicates across two unrelated models. A matched base-versus-instruct test locates alignment's role precisely: the pretrained model already carries the gradient, and alignment amplifies it—collapsing the high-consensus end far harder than the low—so consensus density is the mechanism not only of where the centroid sits but of how alignment sharpens it.

That single mechanism grounds the equalizer/amplifier boundary without appeal to "complexity." Where consensus is high the floor is tight and every user converges on a competent default; where consensus is low the field is wide and uneven, and—on the stated premise that most samples there are mediocre (§3.8)—the outcome is decided by the human's capacity to select from it. On that same premise the model supplies the floor; only taste reaches the ceiling. The dispersion mechanism is measured; this outcome-level reading is the interpretation it motivates, and the selection experiment of §5 is its direct test—its first two passes, simulated and human, are reported there in full; the first was inconclusive at the judges' resolution, the second lost its confirmatory branch to an instrument failure and trends against the premise only at an exploratory sample size, and a resolved human test remains. We have presented this as an illustrative probe, explicit about its scope, about the comparisons that did not replicate, and about the measurement boundary—weak sensitivity to style—that a style-aware metric would extend. The mechanism makes testable predictions, names where human judgment is and is not the bottleneck, and tells interface designers when to collapse to the floor and when to open the field. Those are the experiments, and the systems, we think it invites.

A Reproducibility and experimental detail

A.1 Generation, embedding, and dispersion

Main-probe generation uses Claude Sonnet 4.6 through an OpenAI-compatible endpoint, $N = 25$ independent completions per prompt at temperature 1.0, max 200 tokens per completion. Completions are embedded with `all-MiniLM-L6-v2` (384-d, Sentence-BERT family) and dispersion is the mean pairwise cosine distance among unit-normalized embeddings (§3.1). The robustness re-embedding (§3.9) re-encodes the *identical* completions with `all-mpnet-base-v2` (768-d). The base-versus-instruct comparison (§3.5) runs the same protocol locally on Qwen2.5-7B and Qwen2.5-7B-Instruct, the only difference between the two runs being alignment; the fold-change confidence intervals bootstrap the 25 completions per prompt. Absolute dispersion magnitudes drift between sampling batches, so all reported claims rest on rank correlations and effect directions (§3.9).

A.2 Independent consensus judges

The main anchor (§3.4) uses three held-out judges from different families—GPT-4.1, Gemini 2.5 Flash, and Qwen-Max. Each is shown only the prompt, never any completion, and rates on a 0–100 scale how much knowledgeable people would agree on what a good answer looks like (0 = wide open, 100 = one agreed answer); each prompt is scored three times per judge and the scores averaged. A companion elicitation asks how many substantively different good answers a hundred knowledgeable people would produce. The mechanical anchor reported alongside is the mean token-level surprisal of an independent generator (GPT-4o). The conditioned-prompt extension (§3.7) applies the identical elicitation—same three judges, same 0–100 scale, three repetitions per judge, prompt-only—to the full text of each §3.7 conditioning arm (the base prompt plus its added clause), including the multi-base ladder and tighter-pin arms, so that the destination-consensus axis of the conditioning experiments is measured under the same protocol as the task-level anchor. The tighter-pin follow-up batch and its confirmatory test (per base, pinned vs same-batch default, two-sided 4-vs-4 Mann–Whitney) were pre-registered in the experiment log before generation.

Completion-quality scoring (simulated selection, §5). A separate elicitation scores completion *quality* rather than prompt consensus: the same three judges are shown the prompt and one stored completion at a time and rate "the overall quality of the response as an answer to this prompt" on a 0–100 scale (0 = unusable, 50 = competent but unremarkable, 100 = outstanding), two repetitions per judge per completion, over all 20 main prompts ($\times$ 25 completions) and the six sea-story conditioning arms. The generator (Claude Sonnet) is excluded as a judge to avoid self-preference. The instruction explicitly decouples quality from length ("judge content, not length . . . extra words earn nothing by themselves") and notes that a response may end abruptly at the 200-token generation cap (a pre-check found 3.2% of completions genuinely truncated); it deliberately avoids terms such as "originality" or "distinctiveness," which would build the hypothesis into the score. The confirmatory analysis—cross-fitted selection headroom, its threshold, and the negative-result protocol—was written into the experiment log before any quality score was produced, and the outcome is reported against that pre-registration in §5.

A.3 Human panel

The human panel (§3.4) is a separate sub-study in Chinese pharmaceutical market access—a heavily policy-regulated rather than a free market. As our domain co-author characterizes it, the field is not one smooth gradient but two regimes: where policy is explicit and strict, consensus is high and options are few (registration, MAH obligations, regulatory definitions); where policy leaves room, firms differ by product and positioning (licensing, pipeline, pricing strategy)—yet even there operating-side experts stay broadly aligned, since their choices are bounded by a shared policy environment rather than freely contested. The set therefore separates *asymmetrically*: a sharp, cleanly isolated high-consensus regulatory cluster above a lower end that is only moderately, and not finely, graded. We designed 18 prompts in three intended consensus bands (six high, six middle, six low), with band labels hidden and prompt order shuffled before raters see them—a blinding safeguard, not a claim that raters recover the three bands. The analysis accordingly rests on the continuous 0–10 ratings and their rank agreement with the judges, not on the discrete band means, which—consistent with the domain reading—separate the high cluster clearly (mean 9.5/10) but do not separate the intended middle and low bands at all: their means come out in *reversed* order (middle 6.2 vs. low 7.0). The lower half of our intended design is thus not recovered by the raters, which is exactly why no claim in the paper rests

on the band structure. Each rater scores, for the question alone, "if domain experts answered independently, how much would they agree on what a good answer looks like" on a 0–10 integer scale, plus a companion count of distinct good answers; raters score independently in a single round, with no discussion among them. Generation for these 18 prompts uses Claude Sonnet 4.6 at $N = 25$, temperature 1.0, in Chinese, embedded with a Chinese encoder (`bge-small-zh-v1.5`); the same three judges score consensus.

The active panel is four raters—E1, E2, E3 (industry practitioners) and E5 (an academic)—spanning distinct operating-side functions in drug market access:

Rater	Professional vantage (anonymized)
E1	Drug registration / regulatory affairs, domestic innovator
E2	Market access / health economics (HEOR)
E3	Regulatory affairs, evaluation-side (CDE-adjacent)
E5	Pharmacy / health policy (academic)
E4 *(excluded)*	External commercial advisory, outside operating-side practice

That the four span different functions and institutions, yet still converge ($W = 0.58$), guards against the recruitment bias of a same-shop panel; the excluded rater (E4) sits on a distinct advisory vantage, consistent with its near-zero correlation against every panel member and both model signals. Their ratings give Kendall's $W = 0.58$ and ICC(2,k) = 0.66; the panel-mean expert score tracks the judges at Spearman $\rho = 0.74$ ($p < 10^{-3}$), the judges track dispersion at $\rho = -0.56$, and the panel-mean expert score tracks dispersion only weakly and non-significantly ($\rho = -0.35$, $p = 0.15$, n = 18; see §3.4).

Exclusion rule. A rater whose consensus scores correlate near zero with *every* other rater is treated as measuring a different construct and is reported separately rather than averaged into the panel. The criterion is deliberately rater-side only: it makes no reference to the judges or to the dispersion, since the dispersion is the dependent variable and must not enter an exclusion decision. One rater (E4, drawn from external pharmaceutical advisory) met this criterion, with Spearman correlations of -0.04, -0.16, -0.02, and $+0.16$ against E1, E2, E3, E5. That E4 is also uncorrelated with the judges ($+0.02$) and with the dispersion ($+0.02$) is an after-the-fact observation consistent with a distinct professional vantage, not part of the rule. E4 used the scale earnestly (six distinct values, not a flat line), so we set it aside not as invalid data but as orthogonal to the consensus construct the other four measure. Including E4 lowers inter-rater agreement (Kendall's W: $0.58 \rightarrow 0.40$) yet leaves the headline almost unchanged (expert→judge $\rho = 0.73$ vs 0.74, both $p < 10^{-3}$), confirming that the exclusion does not manufacture the result.

Author dual role, disclosed. One panel member (E1) is also an author of this paper. E1's ratings were collected while fully blind to the study's design, the hidden consensus bands, and the hypotheses, prior to any involvement in the analysis; the contribution that warrants authorship is domain design and validation of the prompt set, not the rating itself. The headline does not rest on this rater: it is unchanged whether E1 is included (panel-mean expert→judge $\rho = 0.74$, n = 4) or excluded, leaving the three fully external raters E2, E3, E5 ($\rho = 0.73$, $p < 10^{-3}$, n = 3). All raters gave informed consent; the study involved only minimal-risk text annotation and collected no personally identifying information.

A.4 Human forced-choice selection pass

The human selection pass (§5) reuses the stored 20 × 25 completions; no new text was generated. Eight volunteers participated, screened on one criterion only (fluent English reading; no domain expertise required) and assigned to disjoint roles: four *selectors*, each shown an independent seed-fixed subsample of five completions per prompt and asked to pick the best one for all twenty prompts; and four *evaluators*, each judging blind, side-randomized pairs (the selector's pick versus a seeded random non-picked completion from the same prompt) as forced choices with no ties, two evaluators per pair under a rotation that never assigns an evaluator a pair derived from their own selections. Instructions were in the participants' native Chinese; all stimuli remained the original English text, untranslated, to avoid a translator confound and to keep the dispersion axis comparable to the rest of the paper. Per-prompt selection performance is the share of votes in which the pick beats the random draw (eight votes per prompt). Three attention-check pairs (fluent versus visibly truncated completions, drawn by the simulated pass's truncation criterion) were embedded per

evaluator; as reported in §5, two of the three contradicted the "judge what is present" instruction, voiding the pre-registered inclusion rule—an instrument failure we logged, together with the pre-committed handling, in the timestamped experiment log before any result was computed. The protocol, its two amendments (both timestamped before the data they govern), the deployed materials with checksums, the raw answer codes (role codes only, no identities), and the analysis script are in the supplementary material. All participants gave informed consent; the study involved only minimal-risk text annotation and collected no personally identifying information.

A.5 Verification and artifacts

Every headline statistic reported in the paper is recomputed from the raw per-prompt data by an independent verification script that asserts each recomputed value against the number stated in the text. The full prompt sets, scoring instructions, per-rater scores (including the excluded rater's), and analysis code are provided as supplementary material.

References

Anonymous. AI as equalizer or amplifier? task complexity as the moderating factor for human expertise in hybrid intelligence systems. In *Proceedings of the 5th International Conference on Hybrid Human-Artificial Intelligence (HHAI)*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2026. To appear.

Shraddha Barke, Michael B James, and Nadia Polikarpova. Grounded Copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1):85–111, 2023.

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.

Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8:605, 2025. doi: 10.1038/s41746-025-02008-z.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152, 1981.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.

Douglas C Engelbart. Augmenting human intellect: A conceptual framework. Technical report, Stanford Research Institute, 1962.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911, 1979.

Sil Hamilton. Detecting mode collapse in language models via narration. *arXiv preprint arXiv:2402.04477*, 2024.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *The Eighth International Conference on Learning Representations (ICLR)*, 2020. arXiv:1904.09751.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06452.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=VD-AYtP0dve`.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014. arXiv:1510.03055.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024. doi: 10.48550/arXiv.2411.15287.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics (TACL)*, 11:102–121, 2023. doi: 10.1162/tacl_a_00536. arXiv:2202.00666.

METR. Measuring the impact of early-2025 AI on experienced open-source developer productivity. Technical report, METR, 2025. URL `https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/`.

Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models. *arXiv preprint arXiv:2406.05587*, 2024.

Sonia K. Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *arXiv preprint arXiv:2411.04427*, 2024.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2203.02155.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*, 2023. doi: 10.48550/arXiv.2302.06590.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. arXiv:1908.10084.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

Chantal Shaib, Venkata S. Govindarajan, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. *arXiv preprint arXiv:2504.12522*, 2025.

Betsy Sparrow, Jenny Liu, and Daniel M Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778, 2011.

Lining Sun, Yiren Liu, Geena Joseph, Zeyu Yu, Haiyi Zhu, and Steven P Dow. Comparing experts and novices for AI data work. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):195–206, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Michael Vaccaro, Abdullah Almaatouq, and Thomas W. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024. doi: 10.1038/s41562-024-02024-1.

Chenghao Yang, Sida Li, and Ari Holtzman. LLM probability concentration: How alignment shrinks the generative horizon. *arXiv preprint arXiv:2506.17871*, 2025.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 1097–1100, 2018. doi: 10.1145/3209978.3210080. arXiv:1802.01886.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.