"I've Seen How This Goes": Characterizing the Diversity of LLM Generations and Human Writing via Progressive Conditional Surprise

Matthew Khoriaty [1] **David Williams-King** [1] **Shi Feng** [2]

Abstract

Measuring the diversity of creative outputs is central to evaluating post-training mode collapse, comparing decoding strategies, and quantifying creative behavior in both AI and human writing. We propose a new approach to measuring diversity using in-context learning, of which the "Decan" metric, $D_{Ca_n} = C \times a_n$, is the working instance we evaluate: a per-byte score read off the per-token log-probabilities of a base model θ in a *single forward pass* per permutation, with no embedding model, no reference corpus, and no human labels. This approach is grounded in information theory, makes use of language model in-context learning to detect a wide range of similarities between any number of inputs, and obviates the need to train a special-purpose model. The same pipeline scores AI samples and human-written response sets, with diversity treated as a property of (responses, prompt, scoring model). On Tevet and Berant's human-grounded McDiv benchmark, D_{Ca_n} reaches OCA 0.846 on the McDiv prompt_gen set where it performs best, behind the strongest neural baseline reported in Tevet and Berant (SentBERT, 0.897). On the OLMo-2-7B post-training pipeline, D_{Ca_n} drops monotonically across the base → SFT → DPO → RLVR stages, detecting the type of diversity loss that creative-writing applications care about.

1. Motivation

The possibility for diverse outputs is necessary, but not sufficient, for creativity. Machine learning researchers that study creativity routinely need to compare the diversity of outputs across generation processes: post-training stages of the same base model (mode collapse), decoding strategies (what does temperature trade off against?), prompting interventions (which prompts have a wide response distribution?), and human-AI hybrid pipelines (would AI use cause human-AI writing to be less creative?). Existing diversity metrics rely on embedding distances (Du & Black, 2019) or surface-level n-gram statistics, both of which lack a human-like ability to recognize arbitrary patterns when those patterns differ from their training data or surface similarities, respectively (Section 4). A complementary line of work applies information theory directly: Maximum Mutual Information decoding (Li et al., 2016) and Adversarial Information Maximization (Zhang et al., 2018) optimise pairwise mutual information between input and response, but as a training or decoding objective rather than an evaluation-time diagnostic. We propose a new approach to measuring diversity using in-context learning, of which the "Decan" metric $D_{Ca_n} = C \times a_n$ is the working instance we evaluate: it scores AI samples and human-written response sets through the same pipeline of per-byte log-probabilities of a base model θ, in a *single forward pass* over the concatenated responses (Figure 1). There is no embedding model, no reference corpus, no human labels, no auxiliary classifier. The approach uses only the per-token probabilities θ already produces.

The intuition is that if responses from a policy π are diverse, seeing one should not help θ predict the next; if they are repetitive or constrained to a few modes, conditioning on earlier responses should sharply reduce θ's surprise at later ones. We exploit the in-context learning capability described by Brown et al. (2020), applied here as a measurement lens rather than as a few-shot-prompting setup. The progressive conditional surprise curve a_k (the per-byte cross-entropy of the k-th response given the previous $k-1$) captures this signal directly, and its last point is a_n (Section 3.1). A separate coherence term $C = 1/\mathrm{PPL}_\theta(\pi, p)$, the reciprocal of the geometric-mean per-byte perplexity that θ assigns to each response individually, prevents pure noise from registering as "diverse" (Section 3.2). The product $D_{Ca_n} = C \times a_n$ is the working scalar we adopt; it is plausibility-weighted residual diversity in bits per byte. We validate it against Tevet and Berant's human-grounded diversity benchmark (Section 5) and on the OLMo-2-7B post-training pipeline (Section 6), and release a working

[1]ERA Fellowship [2]George Washington University. Correspondence to: Matthew Khoriaty <matthewkhoriaty@gmail.com>.

Accepted at the ICML 2026 Workshop on Human-AI Co-Creativity (non-archival).

open-source implementation alongside.[1] The metric measures diversity as θ perceives it: outputs differing only in ways θ cannot distinguish in context appear less diverse. This relativity gives the metric the potential to tighten as base models improve (see Appendix G for a preliminary scaling experiment).

2. Setup and Notation

We use the following notation:

- p be a prompt,
- π be the policy under evaluation,
- $r_1, r_2, \ldots, r_n \sim \pi(\cdot \mid p)$ be n i.i.d. responses sampled from π,
- θ be a trusted base model from which we can obtain per-token log-probabilities,
- $|r|_{\text{tok}}$ denote the number of tokens in response r,
- $\|r\|$ denote the number of bytes in the UTF-8 encoding of response r.

We define the **cross-entropy** (total surprise) of a response r under θ:

$$-\log_2 \theta(r \mid p) = \sum_{t=1}^{|r|_{\text{tok}}} -\log_2 \theta(r^t \mid r^{<t}, p) \qquad (1)$$

where r^t is the t-th token and $r^{<t}$ the preceding tokens. Units: bits.[2] This is the total surprise of the response under θ, a function of the string and of θ's distribution but not of θ's tokenizer (since the chain rule yields the same total regardless of how the sequence is factored).

Similarly, the **conditional cross-entropy** given previous responses $r_{<k} = (r_1, \ldots, r_{k-1})$ is $-\log_2 \theta(r_k \mid r_{<k}, p)$ (bits).

For the coherence term and diversity scores (Sections 3.2 and 3.3), we also use the **per-byte cross-entropy rate**:

$$h_\theta(r \mid p) = \frac{-\log_2 \theta(r \mid p)}{\|r\|} \qquad (2)$$

Units: bits/byte. We adopt this per-byte rate because it works better than total bits in our experiments; we have not investigated why. Normalising by byte count rather than token count keeps the rate independent of θ's tokenizer when comparing base models with different vocabularies.

In practice, computing $\theta(r_k \mid r_{<k}, p)$ requires feeding θ a formatted context containing the prompt and all previous responses (see Section A).

[1] Code and data: https://github.com/AMindToThink/icl-diversity

[2] Throughout, "bits" refers to self-information in $\log_2$ units, identical to the "shannon" (Sh) of IEC 80000-13. We use "bits" to match standard usage in the information-theory literature.

3. Method

This section defines the three quantities used in the main body: the progressive conditional surprise curve a_k, the coherence term C, and the scalar score $D_{Ca_n} = C \times a_n$. The full information-theoretic motivation, the alternative excess-entropy summary E, and the cross-mode learning analysis appear in Appendices D–G.

3.1. Progressive Conditional Surprise

Given prompt p and n responses $r_1, \ldots, r_n$ from policy π, define

$$a_k = -\log_2 \theta(r_k \mid r_{<k}, p) \quad \text{for } k = 1, \ldots, n, \qquad (3)$$

the total surprise of the k-th response under base model θ given the previous $k-1$ responses. The sequence $(a_1, \ldots, a_n)$ is the **progressive conditional surprise curve**. We normalize each a_k by the byte count $\|r_k\|$ to get a per-byte (bits/byte) curve that is independent of θ's tokenizer. All quantities below are per-byte unless otherwise stated.

The intuition is that as θ sees more responses from π, the conditional surprise drops if the responses share patterns (θ's in-context learning picks up on modes, stylistic regularities, topic patterns) and stays flat if they do not. For a policy with rich coherent diversity the curve declines and levels off at a positive floor; for a policy producing one repeated mode the curve drops sharply to near zero; for pure noise the curve stays roughly constant at a high value.

In practice we average a_k over uniformly random permutations of the response ordering (so each position averages over all responses and the curve reflects only how θ's predictions improve with more context, not which response happened to land in slot k). We take the metric value to be the last observed point a_n. Equivalently, $a_n = H_\theta(r_n \mid p) - I_\theta(r_n; r_1, \ldots, r_{n-1} \mid p)$: the response's individual surprise under θ minus how much the prior responses make it predictable. High a_n therefore requires r_n to be both individually surprising and not made predictable by the others, two properties one would want a diversity score to reward. No fitting or extrapolation step is involved. The curve a_k is itself informative: its shape distinguishes a one-mode collapse (sharp drop to a low floor) from richer diversity (gradual decline to a higher floor) in ways the endpoint alone obscures.

3.2. The Coherence Term

The curve alone is fooled by pure noise: θ can never predict random tokens from one another, so a_n stays at the high unconditional surprise level and noise would dominate the metric. The distinguishing signal lies in how plausible θ finds each individual response, independent of the others.

Let $h_\theta(r \mid p) = -\frac{1}{\|r\|} \log_2 \theta(r \mid p)$ be the per-byte cross-

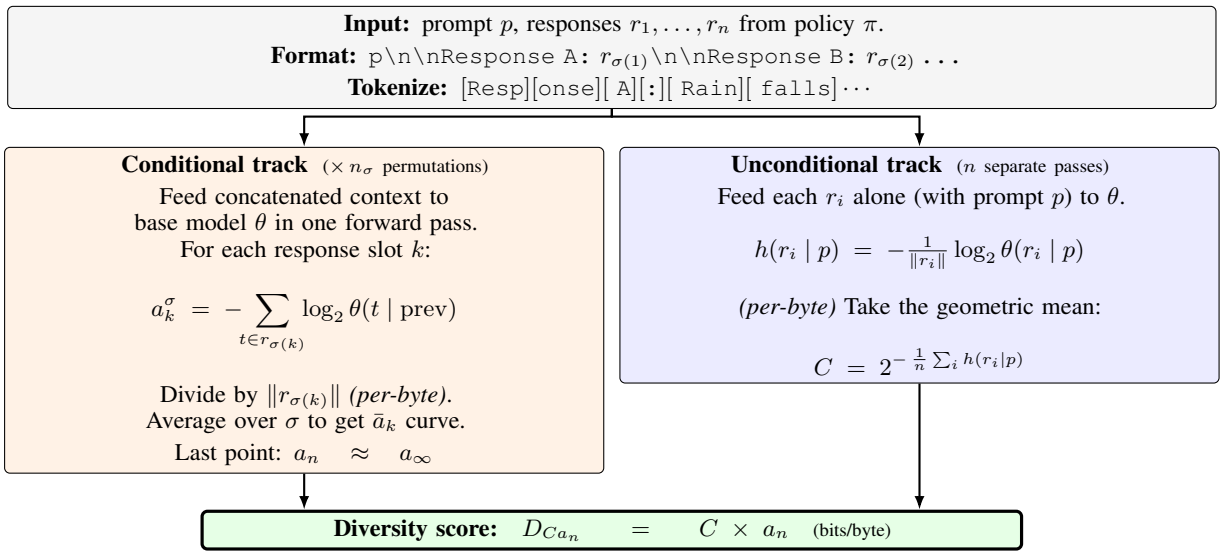


Figure 1. The diversity-metric pipeline. Given prompt p and n responses from policy π, we format them with response labels ("Response A:", "Response B:", ...) and tokenize. The **conditional track** (left) feeds the concatenated context to the base model θ in a single forward pass per permutation σ, extracts per-response total surprise, divides by each response's UTF-8 byte count, and averages over permutations to get $\bar{a}_k$. Its last point is a_n. The **unconditional track** (right) scores each r_i independently against the prompt alone, again per-byte; the geometric mean of these surprises defines coherence $C = 1/\text{PPL}_\theta(\pi, p)$, the reciprocal of the geometric-mean per-byte perplexity. The product $D_{Ca_n} = C \times a_n$ measures plausibility-weighted residual diversity (bits/byte). Formal definitions: Sections 2, 3.1, 3.2, 3.3.

entropy of response r under θ conditioned on the prompt alone. Define **coherence** as the geometric mean of the per-byte probabilities:

$$C = 2^{-\frac{1}{n}\sum_{i=1}^{n} h_\theta(r_i|p)} = \frac{1}{\text{PPL}_\theta(\pi, p)}, \tag{4}$$

the reciprocal of the geometric-mean per-byte perplexity[3] that θ assigns to the responses individually.[4] Perplexity is a standard metric of incoherence, and the geometric form is intended to suppress sets containing incoherent responses: a single sample with high per-byte cross-entropy drives C toward zero, limiting the rescue effect of any single fluent response on an otherwise incoherent set.

3.3. The Diversity Score

The scalar score is the product:

$$D_{Ca_n} = C \times a_n \quad \text{(bits/byte)}. \tag{5}$$

It can be read as "how many bits of surprise per byte remain after θ has learned what it can from the other responses, weighted by output plausibility." The intended edge-case behaviors are summarized in Table 1, alongside empirical D_{Ca_n} on synthetic instances of each scenario under two base models (full per-metric breakdown in Appendix B.2, Table 5).

The score depends on p, π's responses, and θ. Diversity is therefore a property of (responses, prompt, scoring model), not of the policy in isolation: the same response set can score differently under a stronger or weaker θ. This is by design: θ's in-context learning capability is the lens through which diversity is measured, giving the score the potential to improve as base models improve (see Appendix G for a preliminary scaling experiment). For example, Table 1 shows Qwen2.5-3B scoring multi-mode incoherent (0.29) above multi-mode coherent (0.19), reversing our predicted ranking; Figure 4 shows Qwen's $\bar{a}_k$ curve dropping over k on the multi-mode incoherent scenario, indicating that its in-context learning recognises the shared template structure as a pattern despite the within-response scrambling. GPT-2's weaker in-context learning fits the predicted ordering on this row better (0.19 for incoherent vs. 0.26 for coherent), but we still recommend stronger base models when possible – sophisticated patterns are exactly what an ICL-based metric should be able to detect. When comparing policies across a prompt suite, D_{Ca_n} can be averaged over prompts or differenced; see Appendix F.

[3] In our Tevet evaluation with Qwen2.5-3B, the central 90% (5th to 95th percentile) of per-response bits-per-byte values across 48705 responses span $[0.65, 2.20]$, giving $C \in [0.22, 0.64]$; the mean per-set C across 9741 response sets is 0.42.

[4] For scale, the bits/byte score is the average per-byte cross-entropy $\bar{\ell}$, related to coherence by $C = 2^{-\bar{\ell}}$. GPT-3 davinci on The Pile (Gao et al., 2020) scored ∼0.72 bits/byte and GPT-2 XL ∼1.05 bits/byte, giving $C \approx 0.61$ and $C \approx 0.49$ respectively; stronger base models reach lower bits/byte and correspondingly higher C.

Table 1. Intended edge-case behavior of $D_{Ca_n} = C \times a_n$ on the five synthetic scenarios (Appendix B.2), with empirical D_{Ca_n} values under GPT-2 (124M) and Qwen2.5-3B base. The product correctly suppresses pure noise and incoherent multi-mode (via C) and one-mode sets (via a_n); multi-mode coherent is the intended winner. Empirical numbers show smaller a_n for multi-mode coherent than this story predicts: with only 3 distinct modes in a set of 10 responses, θ's in-context learning identifies them and a_n is no longer high. In such ambiguous cases reporting the full a_k curve is more principled than the scalar a_n: a different choice of n would reveal the diversity, and showing the curve sidesteps having to pick that n in advance. Appendix B.3 shows a_n rising with mode count. Mixed empirically scores highest (via a high a_n) rather than the predicted "mid" position; reweighting strategies are discussed in Section B.6. On Qwen2.5-3B, multi-mode incoherent also outranks multi-mode coherent; Section 3.3 discusses.

	Predicted			Empirical D_{Ca_n}	
Scenario	C	a_n	D_{Ca_n}	GPT-2	Qwen2.5-3B
Pure noise	low	high	low (via C)	0.04	0.05
Multi-mode incoherent	low	high	low (via C)	0.19	0.29
Multi-mode coherent	high	high	**high**	0.26	0.19
One-mode	high	low	low (via a_n)	0.10	0.09
Mixed	mid	high	mid	0.52	0.41

4. Related Work

Sampling diversity metrics. Work on decoding-time diversity typically operates at the surface level: n-gram overlap (Li et al., 2016) and self-BLEU (Zhu et al., 2018). Our approach operates at the distributional level and can distinguish policies that produce lexically varied but semantically redundant outputs.

Diversity evaluation benchmarks. Tevet & Berant (2021) introduced a systematic framework for evaluating diversity metrics, with their McDiv benchmark providing human-labeled response sets at two diversity levels. We use McDiv for validation (Section 5), though we identify a construction confound in how its low-diversity sets are produced (Appendix E). More recent work has revisited the meta-evaluation problem. NoveltyBench (Zhang et al., 2025c) defines diversity through pairwise functional equivalence rather than binary labels, training a DeBERTa classifier to group generations into equivalence classes and computing a *Distinct*$_k$ score (the number of meaningfully different outputs in k samples), conceptually close to our mode count. However, their ground truth is itself a trained classifier (79% accuracy), making it unclear whether correlation with their labels validates a metric or merely measures agreement between two model-based scores. Zhang et al. (2025b) conduct a meta-evaluation of diversity metrics for constrained commonsense generation, using GPT-4o as an annotator in place of crowd workers. Guo et al. (2024) benchmark linguistic diversity of LLMs, building on Tevet and Berant's framework with a broader set of generation tasks.

Perplexity. The coherence term $C = 1/\text{PPL}$ (Section 3.2) connects our framework to the standard language modeling evaluation metric. The primary diversity score $D_{Ca_n} = C \times a_n$ operates in bits/byte, weighting the residual surprise floor by output plausibility.

Coherence as a signal. Our coherence term C uses the base model's predictive distribution as an unsupervised quality signal. This connects to a broader line of work using model-internal coherence without external supervision: Qiu et al. (2026) show theoretically that feedback-free self-improvement methods work by optimizing coherence (compressibility of context-to-behavior mappings), and Wen et al. (2025) use internal coherence maximization to elicit capabilities from language models. Our framework leverages a related insight: the base model's ability to compress responses in context provides a meaningful signal about their diversity structure.

Embedding-based diversity. Du & Black (2019) introduced the embedding-based approach, clustering sentence embeddings of generated responses with k-means and reporting cluster inertia as the diversity score; subsequent work has explored alternative geometric statistics over embedded responses. These approaches are complementary to ours: they capture semantic distance in a learned representation space, while our metric captures statistical independence under a generative model. The approaches may disagree when θ's in-context learning captures structure invisible to the embedding model, or vice versa.

Output collapse in RL training. Wang et al. (2025) document an "Echo Trap" pattern during multi-turn agent RL: early-stage agents produce varied symbolic reasoning, then collapse to deterministic templates as training proceeds. They detect this collapse using within-prompt reward standard deviation as an early-warning signal, a proxy that relies on the reward distinguishing degenerate outputs from diverse ones. The a_k framework provides a text-level alternative: collapse should appear as a drop in a_n on the agent's own outputs, independent of reward structure. A downside of using the a_k framework is that it is unconnected from a downstream task. Responses can both be "diverse" as measured by a_k-based metrics and fail uniformly.

Excess entropy. The a_k curve also admits an excess-entropy summary related to the computational mechanics literature (Crutchfield & Feldman, 2003); see Appendix D. We found it empirically inferior to D_{Ca_n} on Tevet and Berant's diversity-eval benchmark (Tevet & Berant, 2021).

5. Tevet–Berant Diversity-Eval: Human-Grounded Validation

A diversity metric should be tested against human judgements on a standardized eval; Tevet & Berant (2021) provide both. Their McDiv and ConTest datasets pair human-grounded diversity labels with a fixed comparison protocol (Spearman ρ and OCA: *optimal classification accuracy*, the best accuracy achievable by a one-dimensional threshold separating low- and high-diversity sets), the standard methodology for ranking diversity metrics against human judgements. The response sets are written by Mechanical Turk workers, and for both McDiv_nuggets and ConTest the high/low label is fixed by the construction protocol itself: a worker writes five different continuations (the high-diversity set), then self-selects one of those continuations and paraphrases it five times preserving content but varying form (the low-diversity set; see Appendix E for the full protocol and a construction-confound caveat). We run D_{Ca_n} through this evaluation pipeline.

Setup. We use Qwen2.5-3B (base) with 50 permutations in completion format (Appendix A.1) on the released splits: 6002 McDiv sets (no_hds), 3069 McDiv_nuggets sets (no_hds), and 670 ConTest sets (with_hds), each with 5 worker-written responses per set and a binary high/low diversity label. We follow Tevet and Berant in reporting Spearman ρ and OCA, and additionally report ROC AUC (the natural metric for binary classification).

Headline result. On the binary tasks D_{Ca_n} clearly tracks diversity, with OCA between 3.7% and 21.0% behind SentBERT (Table 2). On McDiv prompt_gen, where D_{Ca_n} performs best, it reaches $\rho = +0.729$ and OCA = 0.846, against SentBERT's $\rho = +0.796$ and OCA = 0.897; the corresponding ROC AUC is 0.921. The headline takeaway is that an ICL-based diversity metric reaches close to a sentence-embedding baseline on a human-grounded eval, with no embedding model, no reference corpus, and no human labels in the metric itself.

DecTest, included for completeness. Tevet and Berant also release DecTest (Table 3), a diagnostic where the "label" is the sampling temperature used to generate each response set rather than a human judgement. We include it for parity with the released benchmark only: we do not believe that temperature labels map onto creative/meaningful output diversity, consistent with Peeperkorn et al. (2024)'s finding that the relationship between sampling temperature and creativity is more nuanced and weaker than the "creativity parameter" framing suggests. The raw a_n achieves $\rho = +0.932$ on prompt_gen and $\rho = +0.924$ on resp_gen, matching or exceeding all baselines.

Caveats. The McDiv_nuggets construction protocol introduces a confound: low-diversity sets are paraphrases of specific, dramatic endings that are intrinsically more surprising to the base model than typical high-diversity continuations. Appendix E documents the mechanism, the per-byte a_1 gap, and an evidence-by-length-bin breakdown showing the effect is content-driven not length-driven. The C weighting in D_{Ca_n} helps on this benchmark partly for that confound-related reason; on settings without such a confound we expect a_n to carry most of the signal.

6. OLMo-2-7B Post-Training: AI-Side Validation

The Tevet evaluation in Section 5 validates D_{Ca_n} against human-grounded labels. This section evaluates D_{Ca_n} on real policies: we sample from four post-training stages of the same base model and ask whether the metric detects the diversity loss widely attributed to RLHF (Kirk et al., 2023; Zhang et al., 2025a; Padmakumar & He, 2023). The same scoring pipeline applies to both human-written response sets (Section 5) and the policy-sampled sets here, so θ is held fixed across the two evaluations.

Setup. We sample from the four released stages of the OLMo-2-1124-7B pipeline (OLMo et al., 2024): the pre-trained base model, its SFT checkpoint, the DPO checkpoint trained on preference data, and the final RLVR-tuned Instruct checkpoint. On two prompt sets, 200 AlpacaFarm evaluation prompts (Dubois et al., 2023) (seed=42 subsample of the 805-prompt set later adopted by AlpacaEval) and 100 curated NoveltyBench prompts (Zhang et al., 2025c), we draw K=10 responses per prompt per stage (temperature 1.0, top-p 1.0, `max_new_tokens`=100). Base runs raw; SFT / DPO / Instruct receive the prompt through their own chat template. Because per-byte cross-entropy in a causal LM decreases with within-response context, we truncate every (stage, prompt) tuple's responses to a common per-prompt UTF-8 byte length before scoring,[5] and

[5]Truncation is applied to the response *string* via `text.encode("utf-8")[:N].decode("utf-8", errors="ignore")` and the truncated string is then re-tokenised for the forward pass. We deliberately do not truncate at token boundaries: a_n is per-byte, so we need a fixed byte denominator across stages, but tokens-per-byte is tokenizer-dependent and varies across responses with the same byte budget. That variation is harmless because both the bits-numerator and the byte-denominator are computed on the same truncated string.

Table 2. ConTest results: Spearman ρ and OCA between a set's diversity class and each metric score. Baselines from Tevet's pre-computed metrics; $C \times a_n$ and a_n are ours (Qwen2.5-3B, 50 permutations, per-byte). Best result per task in bold.

	prompt_gen		resp_gen		story_gen	
Metric	ρ	OCA	ρ	OCA	ρ	OCA
ConTest (200, with_hds)						
$C \times a_n$ (ours)	+0.584	0.785	+0.391	0.668	+0.686	0.828
a_n (ours)	+0.444	0.715	+0.274	0.641	+0.387	0.684
C (ours)	+0.214	0.625	−0.001	0.555	+0.247	0.632
SentBERT	**+0.682**	0.815	**+0.591**	**0.791**	**+0.770**	**0.896**
BERTsts	+0.646	**0.820**	+0.463	0.714	+0.601	0.780
distinct-n	+0.333	0.675	+0.346	0.677	+0.573	0.772
McDiv_nuggets (∼1K, no_hds)						
$C \times a_n$ (ours)	+0.636	0.785	+0.345	0.649	+0.317	0.634
a_n (ours)	+0.487	0.705	+0.225	0.619	+0.124	0.557
C (ours)	+0.138	0.567	+0.082	0.545	+0.251	0.643
SentBERT	**+0.728**	**0.850**	**+0.532**	**0.758**	**+0.633**	**0.803**
BERTsts	+0.683	0.830	+0.393	0.676	+0.344	0.638
distinct-n	−0.003	0.514	−0.002	0.507	−0.002	0.510
McDiv (full, no_hds, ∼2K)						
$C \times a_n$ (ours)	+0.729	0.846	+0.500	0.724	+0.523	0.717
a_n (ours)	+0.617	0.781	+0.432	0.698	+0.402	0.668
C (ours)	+0.138	0.565	−0.007	0.512	+0.171	0.594
SentBERT	**+0.796**	**0.897**	**+0.678**	**0.830**	**+0.753**	**0.867**
BERTsts	+0.780	0.893	+0.614	0.781	+0.571	0.740
distinct-n	+0.476	0.746	+0.517	0.738	+0.535	0.744

Table 3. DecTest results: Spearman ρ between sampling temperature and each metric score (1000 samples, no_hds). Reported for parity with Tevet and Berant; the labels here are sampling temperatures rather than human judgements, so DecTest is not part of the framing for D_{Ca_n}. That a_n tracks temperature is mechanically unsurprising: a_n is a conditional-entropy quantity, and sampling temperature directly scales the entropy of the policy, so a positive ρ here is a sanity check rather than evidence about creative diversity; Peeperkorn et al. (2024) likewise find only a weak relationship between temperature and creativity in LLM outputs.

Metric	prompt_gen	resp_gen	story_gen
$C \times a_n$ (ours)	+0.847	+0.771	+0.763
a_n (ours)	**+0.932**	**+0.924**	**+0.779**
C (ours)	−0.727	−0.813	−0.547
distinct-n	+0.917	+0.894	+0.758
BERTScore	+0.878	+0.874	+0.694
SentBERT	+0.747	+0.801	+0.645
cos-sim	+0.873	+0.895	+0.712

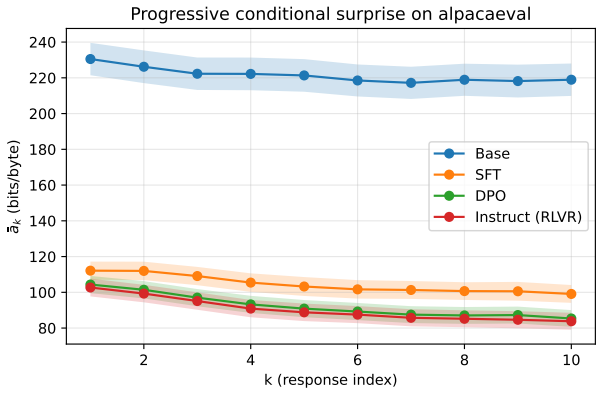
drop prompts where that common length falls below 50 bytes (150/200 AlpacaEval and 39/100 NB-curated prompts retained; see Section 7). We score each (stage, prompt) group with D_{Ca_n} using Qwen2.5-3B as θ and 25 permutations. For comparison we also compute Kirk et al.'s EAD and distinct-n (averaged over n=1 . . . 5) and a SentBERT-embedding (Reimers & Gurevych, 2019) similarity-to-diversity reduction.

Hypotheses. We pre-register three one-sided paired Wilcoxon signed-rank tests (Dror et al., 2018) with family-wise Bonferroni correction across the three contrasts (Dror et al., 2017) ($\alpha = 0.05/3$):

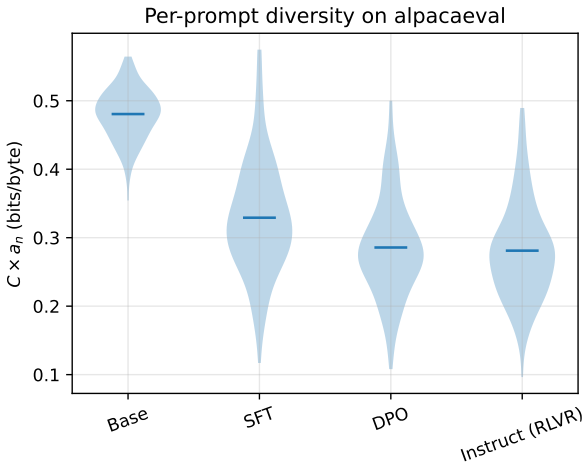
- **H1a** $D_{\text{base}} > D_{\text{SFT}}$: SFT narrows toward the helpful-assistant style.
- **H1b** $D_{\text{SFT}} > D_{\text{DPO}}$: DPO is trained on preference data and inherits its typicality bias (Zhang et al., 2025a).
- **H1c** $D_{\text{base}} > D_{\text{RLVR}}$: cumulative effect of the full post-training pipeline.

The DPO-vs-RLVR comparison is reported as an exploratory two-sided contrast (H1′): we have no directional prediction for which post-training stage loses more diversity.

Results. Table 4 reports per-stage means and the pre-registered Wilcoxon contrasts on both prompt sets. On AlpacaEval and NoveltyBench-curated alike, D_{Ca_n} decreases monotonically across base → SFT → DPO → Instruct (RLVR), and all three pre-registered contrasts (H1a–H1c) are significant after Bonferroni correction. The underlying $\bar{a}_k$ curves and per-prompt D_{Ca_n} distributions are visualised in Figure 2: each later stage's curve lies below the base curve at every $k \geq 2$, and the per-prompt distribution shifts toward lower D_{Ca_n} as the pipeline advances.



(a) AlpacaEval $\bar{a}_k$ curves (length-matched).



(b) Per-prompt D_{Ca_n} by stage (length-matched).

Figure 2. Progressive conditional surprise curves and per-prompt diversity distributions across the four OLMo-2-7B stages on the length-matched AlpacaEval subset. Each later stage's $\bar{a}_k$ curve lies below the base curve at every $k \geq 2$; the per-prompt D_{Ca_n} distribution shifts toward lower values as the pipeline advances.

Discussion. The monotone drop across all three pre-registered contrasts is consistent with the post-training-reduces-diversity findings of Kirk et al. (2023); Padmakumar & He (2023), now measured by an ICL-based diversity metric that needs neither an embedding model nor a sentence similarity function. Per-prompt D_{Ca_n} also agrees with the lexical (EAD) and semantic (SentBERT) baselines on the same length-matched subset, confirming the metrics see the same diversity-loss signal (Appendix H, Figure 18).

Released artefacts. The $200 + 100$ prompts $\times$ 4 stages $\times$ $K{=}10$ generations are released as a public dataset under `results/rlhf_experiment/` in the project repository. To our knowledge this is the first public release of $K \geq 10$ sampled responses across a paired SFT/DPO/RL pipeline, filling a gap that blocked direct diversity replication of Kirk et al. (2023) (whose $K{=}16$ generations were not publicly released).

7. Limitations

We collect the main limitations of the framework and of our $D = C \times a_n$ choice in one place so readers can calibrate how strongly to update on the results above. Several points below recap caveats already made in their natural context and cross-reference those for detail.

Measurement is relative to θ's perception. The metric evaluates diversity *as perceived by the base model* θ: if π's outputs differ only in ways θ cannot distinguish from context, the metric underestimates diversity (Section 1, "Important caveat"). Concretely, the off-diagonal sign of the pairwise surprise-reduction matrix changes with θ's scale (Section B.4), so a_k curves for the same response set are not directly comparable across different choices of θ.

Metric selection saw the full Tevet benchmark. We chose $C \times a_n$ (rather than alternative summaries such as $C \times E$; Appendix D) after observing its performance on the full Tevet diversity-eval benchmark, without holding out a validation split. The Tevet numbers in Section 5 may therefore be optimistically biased: the scalar form was selected with visibility into those results. The OLMo-2-7B post-training experiment (Section 6) was scored after the metric form was fixed and provides independent validation; the synthetic mode-count experiments (Section B.3) informally influenced the choice (had they failed, we would likely have revised the metric) and so are not strictly independent.

External benchmark performance trails the strongest baseline. On Tevet and Berant's diversity-eval benchmark, $C \times a_n$ trails SentBERT by 3.7%–21.0% OCA across the nine binary tasks (Section 5); scaling the base model to Qwen3-30B-A3B-Base (Yang et al., 2025) did not close the gap (Appendix G). We position ICL-based diversity metrics (of which D_{Ca_n} is one instance we found to work well) as a principled complement to embedding-based metrics, not a replacement.

Length-matching drops short-response prompts. Reporting in bits/byte (rather than total bits) makes per-byte cross-entropy not invariant to response length, so the OLMo-2-7B experiment scores a length-matched subset and drops $111/300$ prompts whose common per-prompt byte length

Table 4. Per-prompt D_{Ca_n} by OLMo-2-7B stage on the length-matched subset, with pre-registered paired tests. Reported p-values for H1a–H1c are Bonferroni-corrected ($\times 3$); the H1′ row (marked *) is an uncorrected, two-sided exploratory comparison whose direction we had no prior expectation for, and is not a load-bearing claim.

AlpacaEval			
Stage	D_{Ca_n} mean	std	n
Base	0.481	0.038	150
SFT	0.329	0.084	150
DPO	0.286	0.072	150
Instruct (RLVR)	0.281	0.071	150
Contrast	Δ	d_z	p_{Bonf}
Base>SFT	0.151	1.615	3.4×10^{-24}
SFT>DPO	0.044	0.675	1.7×10^{-13}
Base>Instruct (RLVR)	0.200	2.425	5.2×10^{-26}
DPO $\neq$ Instruct (H1′)	0.005	0.227	0.004^*
NoveltyBench curated			
Stage	D_{Ca_n} mean	std	n
Base	0.481	0.030	39
SFT	0.369	0.085	39
DPO	0.312	0.082	39
Instruct (RLVR)	0.303	0.086	39
Contrast	Δ	d_z	p_{Bonf}
Base>SFT	0.112	1.212	4.1×10^{-8}
SFT>DPO	0.057	0.957	2.8×10^{-6}
Base>Instruct (RLVR)	0.179	1.889	2.3×10^{-10}
DPO $\neq$ Instruct (H1′)	0.010	0.375	0.028^*

falls below 50 bytes; full mechanics and the un-truncated robustness check are in Section B.6.

$D = C \times a_n$ is a pragmatic choice, not a theoretically derived one. The product form was selected because it captures the properties we wanted a diversity score to have (residual surprise remains after the base model has seen previous responses, low-coherence outputs are penalised, and the result stays on a bits-per-byte scale), not because it is derived from an axiomatic information-theoretic setup. Other scalars respecting the same constraints (weighted integrals of $a_k - a_\infty$, slope-based scores, or coherence applied at a different point on the curve) could equally be defended (see "The framework admits other metrics" in Section B.6). We adopt D_{Ca_n} because it is interpretable and works empirically across our validation regimes, not because it is uniquely correct. We invite other researchers to build on this work to develop other formulas that characterize the a_k curve in meaningful ways.

8. Conclusion

We have presented a new approach to measuring diversity using in-context learning, of which $D_{Ca_n} = C \times a_n$ is the working instance we evaluate: a base model's in-context learning detects similarities between arbitrary numbers of responses in a single forward pass. The approach requires no embedding model, no reference corpus, no human labels, and no special-purpose training. The same pipeline scores AI samples and human-written sets. On Tevet and Berant's human-grounded benchmark it is near the level of the strongest sentence-embedding baseline. On the OLMo-2-7B post-training pipeline it drops across the base $\rightarrow$ SFT $\rightarrow$ DPO/RLVR stages, tracking the diversity loss attributed to RLHF-style training.

Impact Statement

This paper presents work which may be used in evaluation pipelines that select for more capable AI systems. We view the prospect of increasingly capable future AI as one of unclear positive or negative net impact. This work may also contribute to the automation of human creative roles. We publish it in the hope that better understanding of machine-generated diversity will contribute to better evaluation and oversight, and that these benefits outweigh the risks.

References

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2023. URL http://arxiv.org/abs/2309.12288v4.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL http://arxiv.org/abs/2005.14165v4.

Crutchfield, J. P. and Feldman, D. P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54, March 2003. ISSN 1089-7682. doi: 10.1063/1.1530990. URL http://dx.doi.org/10.1063/1.1530990.

Dror, R., Baumer, G., Bogomolov, M., and Reichart, R. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017. doi: 10.1162/tacl_a_00074. URL https://aclanthology.org/Q17-1033/.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. The hitchhiker's guide to testing statistical significance in natural language processing. In Gurevych, I. and Miyao,

Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Du, W. and Black, A. W. Boosting dialog response generation. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 38–43, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1005. URL https://aclanthology.org/P19-1005/.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023. URL http://arxiv.org/abs/2305.14387v4.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL http://arxiv.org/abs/2101.00027v1.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh,

K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL http://arxiv.org/abs/2407.21783v3.

Guo, Y., Shang, G., and Clavel, C. Benchmarking linguistic diversity of large language models, 2024. URL http://arxiv.org/abs/2412.10271v2.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of rlhf on llm generalisation and diversity, 2023. URL http://arxiv.org/abs/2310.06452v3.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. In Knight, K., Nenkova, A., and Rambow, O. (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI Blog, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Ettinger, A., Guerquin, M., Heineman, D., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Poznanski, J., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. 2 olmo 2 furious, 2024. URL http://arxiv.org/abs/2501.00656v3.

Padmakumar, V. and He, H. Does writing with language models reduce content diversity?, 2023. URL http://arxiv.org/abs/2309.05196v3.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Is temperature the creativity parameter of large language models?, 2024. URL http://arxiv.org/abs/2405.00492v1.

Qiu, T., Ismail, A. H., He, Z., and Feng, S. Self-improvement as coherence optimization: A theoretical account, 2026. URL http://arxiv.org/abs/2601.13566v1.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL http://arxiv.org/abs/1908.10084v1.

Tevet, G. and Berant, J. Evaluating the evaluation of diversity in natural language generation. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25/.

Wang, Z., Wang, K., Wang, Q., Zhang, P., Li, L., Yang, Z., Jin, X., Yu, K., Nguyen, M. N., Liu, L., Gottlieb, E., Lu, Y., Cho, K., Wu, J., Fei-Fei, L., Wang, L., Choi, Y., and Li, M. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning, 2025. URL http://arxiv.org/abs/2504.20073v2.

Wen, J., Ankner, Z., Somani, A., Hase, P., Marks, S., Goldman-Wetzler, J., Petrini, L., Sleight, H., Burns, C., He, H., Feng, S., Perez, E., and Leike, J. Unsupervised elicitation of language models, 2025. URL http://arxiv.org/abs/2506.10139v2.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2024. URL http://arxiv.org/abs/2412.15115v2.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL http://arxiv.org/abs/2505.09388v1.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025a. URL http://arxiv.org/abs/2510.01171v3.

Zhang, T., Peng, B., and Bollegala, D. Evaluating the evaluation of diversity in commonsense generation, 2025b. URL http://arxiv.org/abs/2506.00514v1.

Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., and Dolan, B. Generating informative and diverse conversational responses via adversarial information maximization, 2018. URL http://arxiv.org/abs/1809.05972v5.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. Noveltybench: Evaluating language models for humanlike diversity, 2025c. URL http://arxiv.org/abs/2504.05228v4.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models, 2018. URL http://arxiv.org/abs/1802.01886v1.

A. Practical Considerations

A.1. Formatting the Conditioning Context

Computing $\theta(r_k \mid r_{<k}, p)$ requires feeding θ a context containing the prompt and previous responses. The formatting choice affects results, and we use two formats depending on whether the responses are best read as parallel answers to the prompt or as continuations of it. Both are implemented as `format_conditioning_context` in `src/icl_diversity/core.py` and selected via a `format_mode` argument throughout the pipeline.

Instruct format. The default, used for the synthetic scenarios (Appendix B.2), the mode-count experiments (Appendix B.3), and the OLMo-2-7B post-training case study (Section 6):

```
[prompt p]

Response A: [r_1]

Response B: [r_2]

Response C: [r_3]
```

The prompt appears once and each response is introduced by a labelled header ("Response A:", "Response B:", ..., with labels rolling over to "AA", "AB" past "Z"), encouraging the base model's in-context learning to treat the responses as parallel answers to the same prompt.

Completion format. Used for the Tevet and Berant evaluation (Section 5), where each response set is a continuation of a shared narrative prompt rather than an instruction-style answer:

```
1. [prompt p] [r_1]

2. [prompt p] [r_2]

3. [prompt p] [r_3]
```

The prompt is repeated immediately before each response so θ scores r_k in the same prompt-as-context that produced it. Per-token log-probabilities are extracted only over the response tokens; the bits accumulated over the repeated prompt prefix are excluded from a_k, keeping the bits-numerator and the per-byte denominator both attributable to r_k alone.

Base vs. instruction-tuned models. We use a base (non-instruction-tuned) model as θ for two reasons. First, an instruction-tuned model has had its output distribution shaped by RLHF or similar procedures, so using one as θ reintroduces the kind of distributional bias we seek to avoid. Second, an instruction-tuned θ may assign systematically different probabilities to text that follows or violates its alignment training, introducing a confound between coherence-as-fluency and coherence-as-alignment that the metric should not conflate. Modern base models already exhibit strong in-context learning from pretraining alone, so this choice does not sacrifice ICL capability. Optimizing the prompt format to maximally elicit in-context learning from θ is an interesting direction but out of scope for this paper; we use the neutral "Response A/B/C" template throughout.

A.2. Computational Cost

Because θ is a causal language model, the full a_k curve for one response ordering can be computed from a **single forward pass**. The prompt and all responses are concatenated into one sequence (Figure 3). The forward pass produces the log-probability of every token conditioned on all preceding tokens.

The tokens belonging to r_k are conditioned on exactly $p, r_1, \ldots, r_{k-1}$ (plus formatting), which is the conditioning required for a_k. Partitioning the output log-probabilities by response boundary and summing within each partition yields all n values of a_k simultaneously for that ordering, with FLOPs scaling as $O((|p| + n\bar{L}_{\text{tok}})^2)$ from causal attention, where $\bar{L}_{\text{tok}}$ is the average response length in tokens. Because we report $\bar{a}_k$ averaged over $|\Sigma|$ random permutations of the response ordering (Section A.3), the total a_k-curve cost is $|\Sigma|$ such passes; the permutations are independent and we batch them on a single GPU.

The coherence term C additionally requires the *unconditional* per-byte cross-entropies $h_\theta(r_i \mid p)$ for each response. These cannot be extracted from the concatenated pass, since in that pass r_i for $i > 1$ is conditioned on all prior responses. Computing them requires n independent forward passes, each over the short context (p, r_i). These passes are embarrassingly parallel and batchable, and they do not depend on the ordering, so C is computed once per response set regardless of $|\Sigma|$.

In summary, scoring one (prompt, response set) tuple takes:

- a_k curve: $|\Sigma|$ forward passes (long context, one per permutation).
- C: n forward passes (short contexts, batchable; not multiplied by $|\Sigma|$).

With $n = 10$ responses and $|\Sigma| = 50$ permutations (our default for larger-scale experiments), this is $50 + 10 = 60$ forward passes per (prompt, response set) tuple; the long-context $|\Sigma|$ passes dominate wall-clock time.

```
[prompt p] Response A: [r_1] Response B: [r_2] ... Response N: [r_n]
```

Figure 3. Single-pass input layout for computing the full a_k curve. The prompt is followed by all n responses with "Response A/B/C/..." labels; one forward pass over this sequence yields every a_k value simultaneously, since causal attention conditions the tokens of r_k on exactly $p, r_1, \ldots, r_{k-1}$ (plus formatting).

A.3. Dependence on Sample Ordering

The a_k values depend on the ordering of $\{r_i\}$. Individual responses differ in how surprising they are to θ given the responses that precede them, so a_k depends on which response sits at position k and which others precede it, making a single ordering's curve jagged. Averaging over random permutations removes this dependence: each position averages over many choices of response and preceding context, so the curve reflects only how θ's predictions improve with more context. We therefore average over $|\Sigma|$ random permutations; Section B.5 compares low- and high-permutation runs on the validation scenarios and finds that the low setting misranks adjacent scenarios. We have not swept intermediate values and cannot pinpoint a cheap-and-reliable minimum, so we default to $|\Sigma| = 100$ for scenario-level experiments and $|\Sigma| = 50$ elsewhere.

Per-response normalization by byte count must be performed *before* averaging across permutations, since the byte count depends on which response lands at each position. Concretely, the per-byte curve is a mean of per-permutation rates,

$$\bar{a}_k = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \frac{a_k^\sigma}{\|r_{\sigma(k)}\|}, \tag{6}$$

not a ratio of permutation-averaged bits to permutation-averaged byte counts (which would degenerate, after enough permutations, into the total-bits curve rescaled by the mean response length).

A.4. Choice of n

The diversity score depends on n as a measurement parameter, not through any estimate of an asymptotic floor. Given enough in-context examples, even genuinely diverse policies eventually become predictable to θ, so a_k keeps decreasing rather than converging to a meaningful irreducible value: there is no a_∞ for a_n to approximate. Too small an n means θ has not yet exploited the inter-response structure, and D_{Ca_n} overstates diversity. Too large an n means θ has accumulated enough in-context examples to reduce its surprise even within genuine diversity; for any policy with finitely many distinct modes, a_n continues to decrease toward zero as θ learns to predict within-mode variation from prior examples. For comparisons to be meaningful, n must be held fixed across all policies under evaluation. Our experiments use $n = 5$ on Tevet (Section 5), $n = 10$ on OLMo-2-7B and scenario validation, and $n = 20$ on the mode-count experiment (Appendix B.3). This is one reason to report the full a_k curve rather than D_{Ca_n} alone. D is a lossy summary, but the curve shows directly how response surprise evolves at every context size.

B. Experiments

B.1. Experimental Setup

We evaluate the metric primarily using Qwen2.5-3B (3B parameters, 32K-token context window) (Yang et al., 2024), with GPT-2 (124M parameters, 1024-token context window) (Radford et al., 2019) as a smaller-model comparison point for scenario validation. Qwen2.5-3B is used for mode-count scaling, cross-mode learning, and external validation (Sections B.3, B.4, 5). The cross-model scaling study (Section B.6) additionally uses the full Llama 3 family for comparison.

All a_k curves are computed using single-pass forward computation (Section A.2). Unless otherwise noted, we use 100 permutations for scenario validation and 50 permutations for larger-scale experiments. Responses are formatted as described in Section A.2. All code and data, including the Olmo-2 RLHF generations under `results/rlhf_experiment/`, are publicly available.[6]

B.2. Scenario Validation

We construct five synthetic scenarios with known diversity structure:

1. **Pure noise**: Random ASCII characters (letters, digits, punctuation, spaces) with no learnable structure.
2. **Multi-incoherent**: Responses from 5 distinct modes, each internally incoherent (scrambled words within a template).
3. **Multi-mode**: Responses from 5 distinct modes, each internally coherent (e.g., recipe, poem, code).
4. **One-mode**: All responses from a single coherent mode (paraphrases of the same content).
5. **Mixed**: A mixture of coherent and incoherent responses.

Each scenario contains 10 responses per prompt, 5 prompts per scenario, with 100 permutations and seed 42.

[6] https://github.com/AMindToThink/icl-diversity

Table 5. Scenario validation metrics (mean across 5 prompts, 100 permutations, $n = 10$ responses, per-byte). We report several candidate scalars: unconditional surprise a_1, the curve's last point a_n, coherence C, our recommended score $D_{Ca_n} = C \times a_n$ (**bold**), the excess entropy E (Appendix D), $C \times E$, and the coherence spread σ_ℓ (the standard deviation of the per-byte cross-entropies $\{h_\theta(r_i \mid p)\}_{i=1}^n$). D_{Ca_n} correctly ranks multi-mode above one-mode and suppresses incoherent scenarios via C. Both E and $C \times E$ go negative for the mixed scenario on GPT-2. Mixed empirically scores the highest D_{Ca_n} on both models, exceeding multi-mode coherent rather than landing at the "mid" position Table 1 predicts; this is not the intended ranking. The coherence spread σ_ℓ flags this case in its intended diagnostic role (the largest σ_ℓ across scenarios on both models), reflecting the within-set heterogeneity between coherent and incoherent responses; reweighting variants such as $C^\alpha \times a_n$ are discussed in Section B.6.

	GPT-2 (124M)							Qwen2.5-3B						
Scenario	a_1	a_n	C	$\mathbf{D_{Ca_n}}$	E	$C \times E$	σ_ℓ	a_1	a_n	C	$\mathbf{D_{Ca_n}}$	E	$C \times E$	σ_ℓ
Pure noise	7.52	6.94	0.005	**0.04**	0.68	0.004	0.20	7.18	6.68	0.007	**0.05**	0.75	0.005	0.11
Multi-incoher.	3.26	1.85	0.10	**0.19**	5.94	0.61	0.48	2.36	1.40	0.21	**0.29**	3.12	0.64	0.69
Multi-mode	1.06	0.53	0.48	**0.26**	1.61	0.76	0.08	0.78	0.31	0.59	**0.19**	1.17	0.68	0.11
One-mode	1.06	0.22	0.48	**0.10**	1.31	0.63	0.06	0.80	0.16	0.58	**0.09**	0.92	0.53	0.08
Mixed	1.90	2.02	0.26	**0.52**	−1.29	−0.31	1.17	1.13	0.91	0.46	**0.41**	0.52	0.23	0.73

Results. Table 5 summarizes the metrics. Figure 4 shows the a_k curves for all scenarios.

The coherence term C correctly separates coherent from incoherent scenarios: $C \approx 0.5$–0.6 for multi-mode and one-mode, $C \approx 0.1$–0.3 for multi-incoherent, and $C \approx 0.01$ for pure noise. The coherence spread σ_ℓ discriminates mixed from pure scenarios: $\sigma_\ell > 1.0$ for mixed (GPT-2), reflecting the within-set heterogeneity between coherent and incoherent responses.

C consistently improves with model strength (Qwen2.5-3B assigns higher coherence to coherent text than GPT-2). D_{Ca_n} correctly ranks multi-mode above one-mode on both models, and suppresses pure noise and multi-incoherent via the C factor.

B.3. Mode Count Scaling

To test whether the metric reflects the number of distinct modes, we construct response sets with $m \in \{1, \ldots, 10\}$ modes drawn from a pool of 50 format-based generators (haiku, code, recipe, legal disclaimer, etc.), all responding to the same prompt ("Write a short piece about rain"). For Qwen2.5-3B: $n = 20$ responses, 1000 random draws of mode assignments, averaged over permutations.

Key findings. Figure 5 shows the a_k curves. a_n increases monotonically with m (9.3 bits at $m = 1$ to 77.8 bits at $m = 10$; see Table 6), reflecting that with more modes there are fewer same-mode repetitions within n responses, so each response remains more surprising even after θ has seen the others. This is exactly the behavior $D_{Ca_n} = C \times a_n$ needs: a_n tracks mode count directly. All curves are purely exponential (no sigmoidal plateau), even at $m = 10$; Section B.4 investigates why. For a detailed comparison of the excess-entropy-based scores on this data, including the sigmoid fit parameters and the non-monotonicity of $\hat{E}_n$, see Appendix D.

B.4. Cross-Mode Learning and Curve Shape

The authors predicted that at high m, the a_k curve should be sigmoidal: an initial plateau (early responses come from different modes and are uninformative about each other), followed by decline once mode repetitions accumulate. GPT-2 shows hints of this pattern at some m values (e.g. $m = 10$). Qwen2.5-3B does not: it shows immediate exponential decay at all m. This section investigates the discrepancy through systematic pairwise analysis.

Pairwise cross-mode surprise matrix. For 15 modes, we compute a 15×15 matrix M_{ij}: the surprise reduction (in bits) when a response from mode j is in the context and mode i is the target. Diagonal entries measure same-mode learning; off-diagonal entries measure cross-mode information transfer. Each entry averages over 5 context–target samples (using different samples for context and target to avoid self-prediction inflation).

Results. (Cell standard deviations are the mean within-cell sampling std across the 15 diagonal or 210 off-diagonal matrix cells, each estimated from 5 context-target sample pairs.)

- **Qwen2.5-3B**: diagonal mean = 60.5 ± 14.4 bits, off-diagonal mean = $+1.9 \pm 2.5$ bits, 64% of off-diagonal entries positive.
- **GPT-2**: diagonal mean = 83.0 ± 20.0 bits, off-diagonal mean = -3.7 bits, only 30% of off-diagonal entries positive.

The sign of the off-diagonal determines the curve shape (Figure 6):

- **Positive off-diagonal (Qwen)**: Every response in context lowers expected surprise on every other response. The a_k curve drops from position 1, producing exponential decay.
- **Negative off-diagonal (GPT-2)**: Cross-mode re-

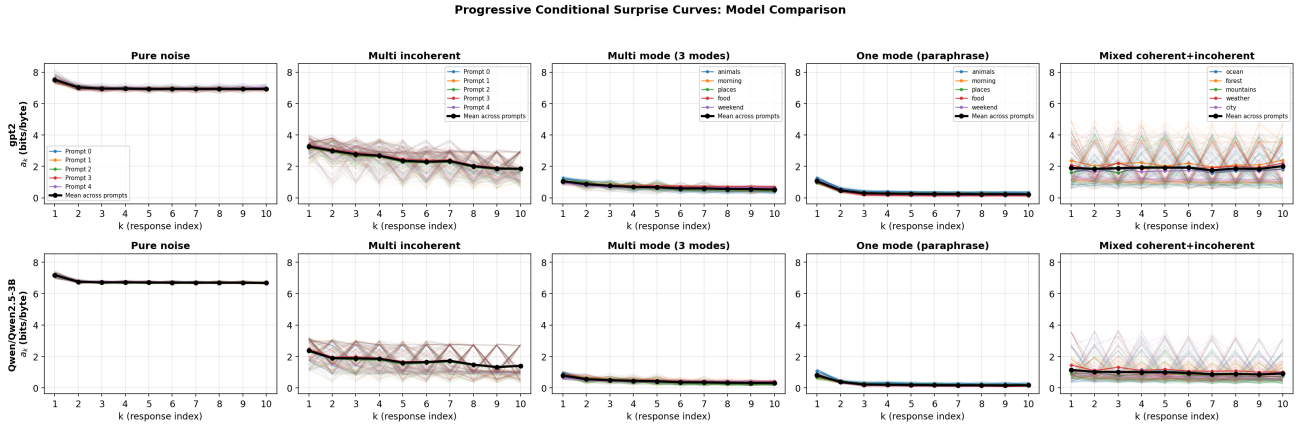


Figure 4. Progressive conditional surprise curves (a_k, per-byte) for all five scenarios, comparing GPT-2 (top row) and Qwen2.5-3B (bottom row). In each panel, faint colored curves are individual permutations (100 per prompt), medium colored curves are per-prompt averages, and the bold black curve is the mean across prompts. Pure noise curves are flat (no learnable structure); multi-mode curves show progressive decline; one-mode curves decline steeply then flatten.

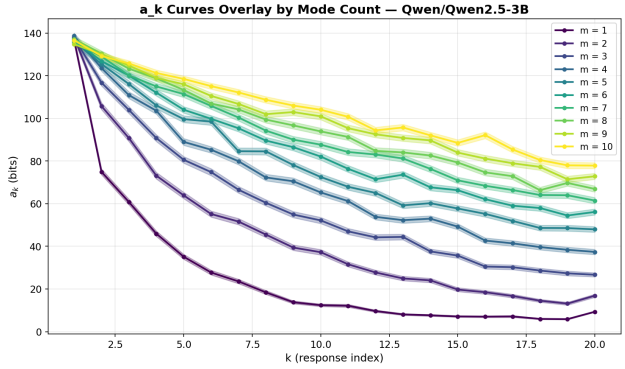


Figure 5. Mode count scaling on Qwen2.5-3B ($n = 20$, 1000 draws). The a_k curves fan out with increasing m: higher floors, slower convergence. All curves are exponential (no sigmoidal plateau), even at $m = 10$, due to cross-mode learning (Section B.4). Shaded bands around each curve are ± 1 standard error of the mean across the 1000 random draws of mode assignments (i.e., the across-draw standard deviation divided by $\sqrt{1000}$); they are narrow because averaging over 1000 draws sharply reduces the across-draw spread.

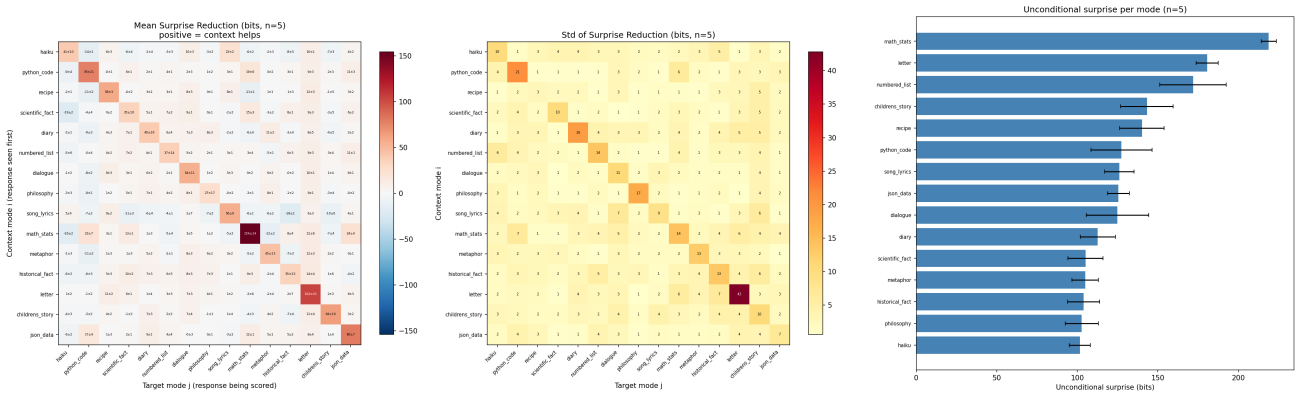
sponses actively raise surprise on subsequent responses. At $m = 10$, the expected first-step net drop under the additive independent-mode model is only $+4.9$ bits (the 8.3-bit diagonal gain from the $1/m$ same-mode chance is largely offset by the -3.4-bit cross-mode damage from the $(m-1)/m$ different-mode chance), producing a flat plateau until same-mode repetitions accumulate, the predicted sigmoid.

GPT-2 has a larger diagonal and a negative off-diagonal: it gains more from same-mode context but is actively penalised by cross-mode context. The two effects are linked, since probability mass that conditioning concentrates on same-mode continuations is mass taken away from other modes; a sharper conditional on the seen mode is necessarily worse on the unseen ones. Qwen does the opposite: a smaller diagonal but a positive off-diagonal, distributing its update across modes.

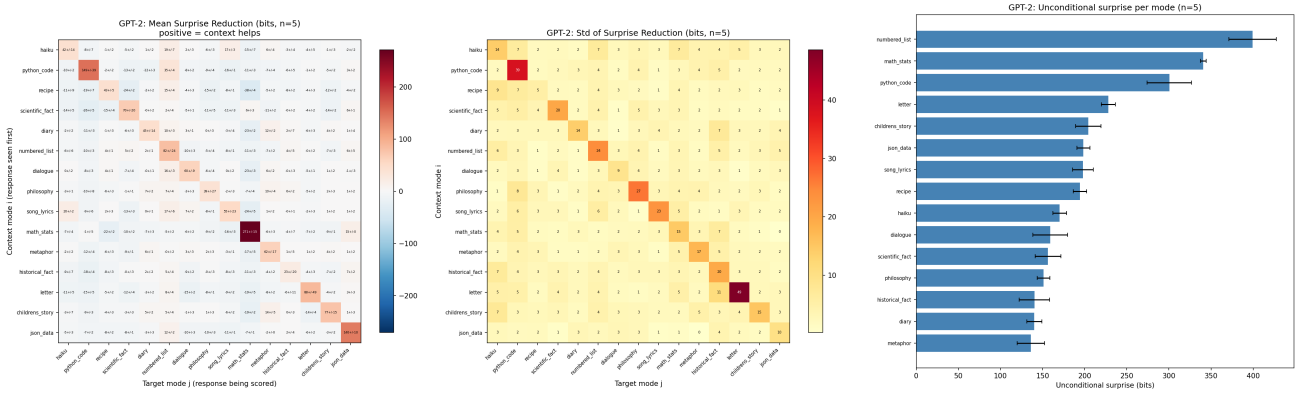
Pairwise asymmetry. The mean absolute asymmetry $|M_{ij} - M_{ji}|$ is 5.6 bits (3.0× the off-diagonal mean), falsifying the hypothesis that pairwise surprise reduction is symmetric (Figure 7). Information-theoretically $I(X;Y) = I(Y;X)$, so this asymmetry cannot reflect asymmetric information structure between modes; it reflects θ's imperfection as an in-context reasoner. The same content is more useful as context than as target depending on its surface form, mirroring the *reversal curse* observed in autoregressive LLMs trained on "A is B" but failing on "B is A" (Berglund et al., 2023).

Connections to information-theoretic properties. Two information-theoretic properties are relevant to interpreting the pairwise matrix:

1. **Non-negative conditioning** ($H(X) \geq H(X \mid Y)$):



(a) Qwen2.5-3B: positive off-diagonal.



(b) GPT-2: negative off-diagonal.

Figure 6. Pairwise cross-mode surprise reduction matrices. Each cell (i, j) shows how many bits of surprise reduction mode i (target) receives from seeing mode j (context). Qwen shows diagonal dominance with pervasive positive off-diagonal (+1.9 bits mean); GPT-2 shows diagonal dominance with negative off-diagonal (-3.7 bits mean).

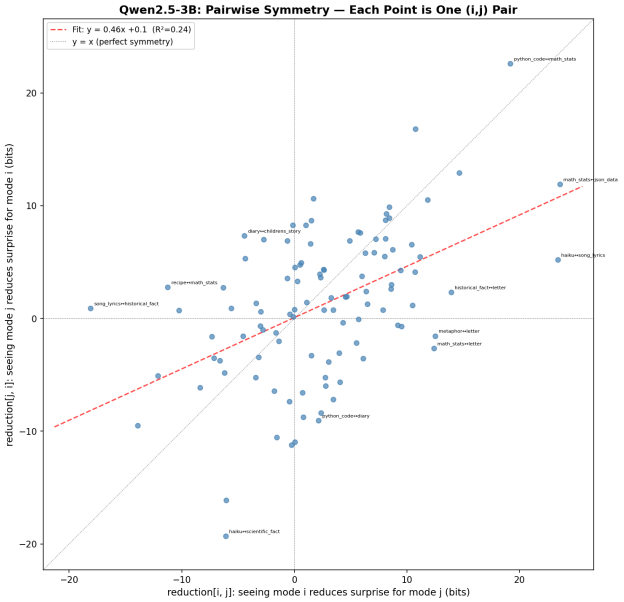


Figure 7. Pairwise symmetry scatter for Qwen2.5-3B: each point is one (i, j) pair, plotting M_{ij} vs. M_{ji}. Under symmetric mutual information, points would lie on $y = x$. The best-fit slope is 0.46 ($R^2 = 0.24$), and many pairs have opposite signs: mode i can reduce surprise for mode j while j *increases* surprise for i.

Under the true distribution, conditioning can never increase entropy on average. For cross-entropies this is not guaranteed (the KL term can increase), but a well-calibrated θ should show mostly non-negative off-diagonal entries.

2. **Symmetry of mutual information** ($I(X;Y) = I(Y;X)$): Under the true joint, row means (how informative mode i is as context) should correlate with column means (how much mode i benefits from context).

Figure 8 compares the two models. Qwen shows a tighter row–column relationship (slope = 0.84, $R^2 = 0.34$) with only 1 of 15 modes having a negative column mean. GPT-2 shows widespread violations of non-negative conditioning (6 modes with negative column means) and a barely-existent row–column relationship (slope = 0.35, $R^2 = 0.08$). Across the four Llama models (Grattafiori et al., 2024; Meta AI, 2024) we additionally tested (Section B.6, Figure 10), the fraction of positive off-diagonal entries rises monotonically with size (25% at 1B to 63% at 70B), and the off-diagonal mean follows the same trend (-4.9 bits at 1B to +2.1 bits at 70B): the *magnitude* of cross-mode information transfer increases with size. The row–column relationship does not: R^2 peaks at Llama-3B (0.67) and degrades at the larger Llamas (0.09 at 70B), so the *symmetry* of the pairwise matrix does not correspondingly improve.

Token-level attribution. Figure 9 shows per-token surprise reduction (in bits) for one pair from each of three strata: same-mode (`letter` → `letter`), cross-mode where conditioning helps (`math_stats` → `json_data`), and cross-mode where conditioning hurts (`scientific_fact` → `haiku`). Pairs were selected by stratified-median total $|\Delta|$.

Discussion. The paper's sigmoid prediction assumes modes are approximately independent. For capable models, this fails: cross-mode learning creates positive information transfer between distinct modes. The Llama 1B/3B/8B/70B series confirms this scales with model size (Section B.6): the total cross-mode information θ extracts grows monotonically with size, and with it non-negative conditioning is better satisfied. The matrix nevertheless remains asymmetric (row–column R^2 peaks at Llama-3B and degrades at the larger Llamas), which under $I(X;Y) = I(Y;X)$ signals that even Llama-70B is an imperfect in-context reasoner.

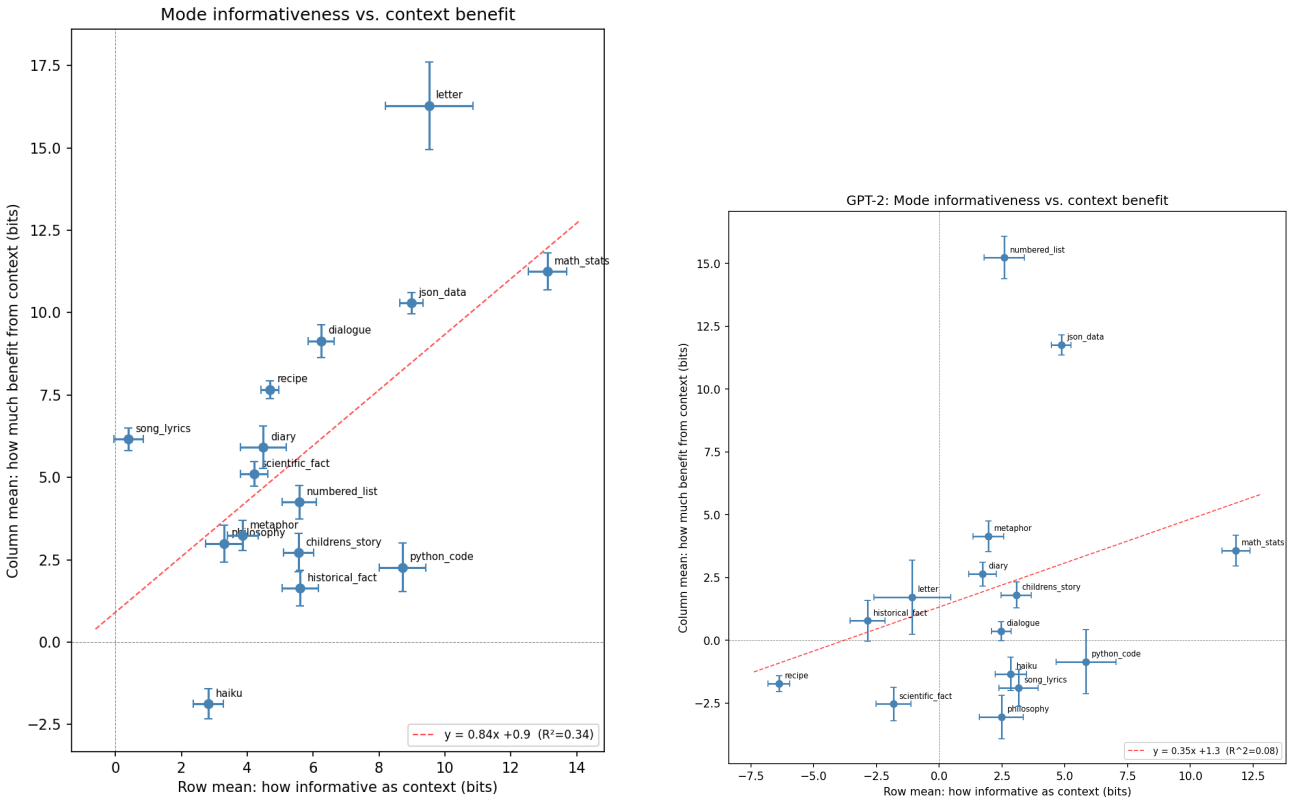
B.5. Practical Findings

Permutation sensitivity. The per-byte a_k curve depends on the response ordering. Comparing the D_{Ca_n} ranking of the 5 validation scenarios from Section B.2 between 3- and 100-permutation runs: on GPT-2, 2 of 5 scenarios shift rank between the two settings; on Qwen2.5-3B, 2 of 5. Where scenarios do shift rank, it is between adjacent scenarios with similar D_{Ca_n} values rather than dramatic reorderings. We use 100 permutations throughout for scenario-level experiments; we have not swept intermediate values, so we cannot pinpoint a minimum that is both cheap and reliable. The coherence term C is permutation-invariant by construction and unaffected.

Boundary handling. When Qwen merges a response's final `.` with the following `\n\n` into a single `.\n\n` token, our boundary detector (which uses a character-span overlap rule) attributes that token to the response, not the separator. The response's cross-entropy therefore includes a small contribution from predicting the upcoming separator alongside its own characters; the byte-count denominator remains the response's literal byte length. We have not measured the magnitude of this bias on our reported numbers. Boundary correctness is verified by `tests/test_response_boundaries.py`.

B.6. Discussion

σ_ℓ is a diagnostic, not a diversity signal. σ_ℓ increased monotonically with m in our synthetic experiments because our modes have very different formats (code vs. haiku vs. recipe → different per-byte cross-entropies). But σ_ℓ measures coherence *heterogeneity*, not diversity: modes with similar fluency would have low σ_ℓ despite high diversity,



(a) Qwen2.5-3B: slope = 0.84, $R^2 = 0.34$.

(b) GPT-2: slope = 0.35, $R^2 = 0.08$.

Figure 8. Row mean (how informative as context) vs. column mean (how much benefit from context) for each mode. Qwen shows tighter correlation closer to the identity, consistent with approximate symmetry of mutual information. GPT-2 shows widespread violations of non-negative conditioning (modes in the negative-row-mean region).

and a single mode with high quality variance could have high σ_ℓ with zero diversity. Its role is diagnostic (detecting mixed coherence), not as a standalone diversity score.

Cross-mode information scales with model quality. We test how cross-mode information transfer varies with θ's scale by running the pairwise matrix experiment on 4 dense Llama models (1B, 3B, 8B, 70B) using the same 15 modes and 5 samples per mode. Figure 10 shows the results. The off-diagonal mean increases monotonically with model size: -4.9 (1B), -1.4 (3B), -1.2 (8B), $+2.1$ (70B) bits, transitioning from negative (cross-mode context hurts) to positive (cross-mode context helps). The fraction of positive off-diagonal entries follows the same monotone trend: 25%, 41%, 44%, 63%. This hypothesis was pre-registered before running any Llama models, and the probability of accidental monotonicity across 4 models under the null is $1/4! \approx 4.2\%$.

The framework admits other metrics. The a_k curve is the primary object of our framework, and $D_{Ca_n} = C \times a_n$ is one natural summary of it among many. Reporting the curve directly preserves information that any single scalar (including ours) collapses. Alternative scalars could be derived from the same curve to capture different desiderata: weighted sums of $a_k - a_\infty$ (emphasizing early or late positions), the slope or curvature at a specific k (measuring how quickly θ learns), curve-shape descriptors, or coherence-weighted variants that combine C with quantities other than a_n. Different applications may warrant different choices; for instance, a policy-comparison setting might prefer a metric sensitive to *changes* in the curve shape, while a ranking setting needs a single well-behaved scalar. We have not explored this space systematically; we focus on D_{Ca_n} because it empirically works well across regimes and has a clean information-theoretic interpretation, but we expect the framework to support a family of related metrics with complementary properties. One untested family is coherence-reweighting: a practitioner with a clear preference for sharper Mixed penalisation could replace $C \times a_n$ with $C^\alpha \times a_n$ for $\alpha > 1$ to drive low-coherence sets toward zero faster. We have not evaluated such variants on the Tevet or OLMo experiments and so make no empirical claim about them.

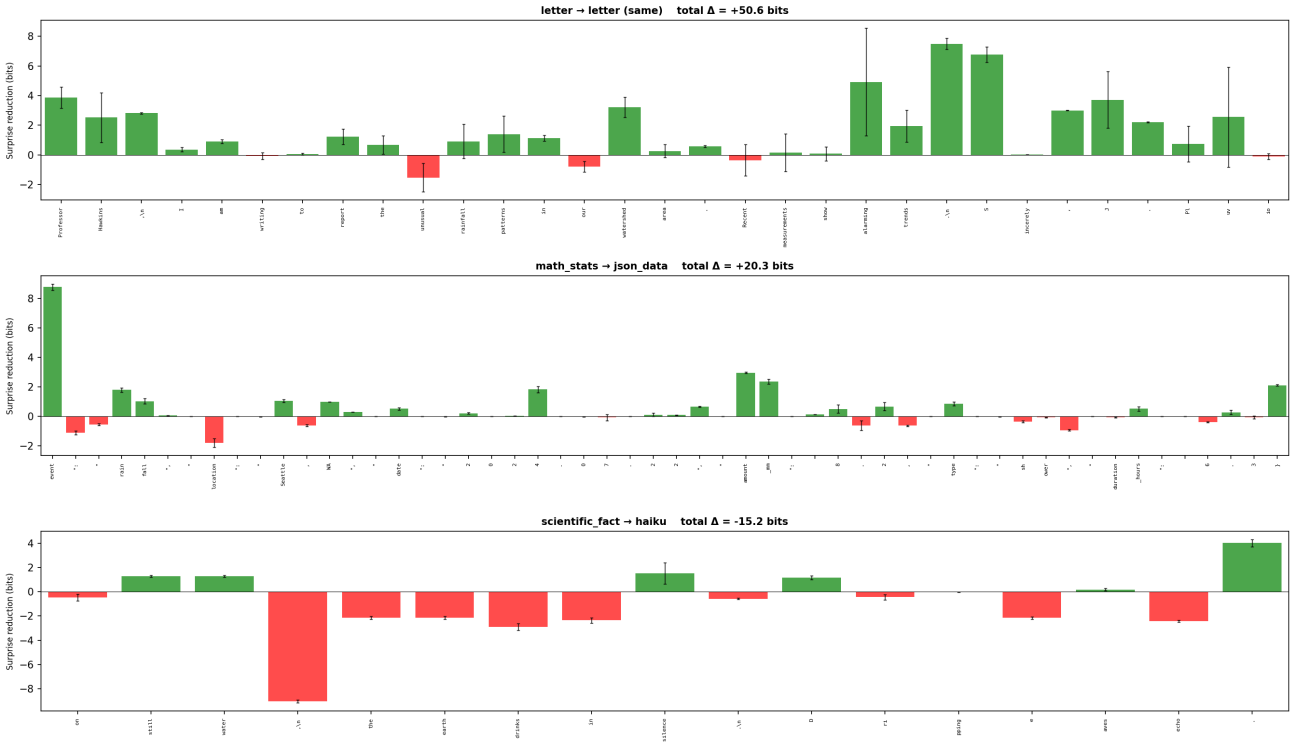


Figure 9. Per-token surprise reduction (bits) for three representative context–target pairs (Qwen2.5-3B). Bars are mean reductions across 5 context samples per pair; error bars are ± 1 standard deviation. Green bars indicate the conditioning context lowered the per-token surprise; red bars indicate it raised it. The three pairs are stratified-median selections (by total $|\Delta|$) from the same-mode, positive cross-mode, and negative cross-mode strata of a 9-pair pool. Note how in scientific fact to haiku, the early newline triggers the largest surprise. Scientific facts may discuss "still water", but they almost never have a line with three words in it.

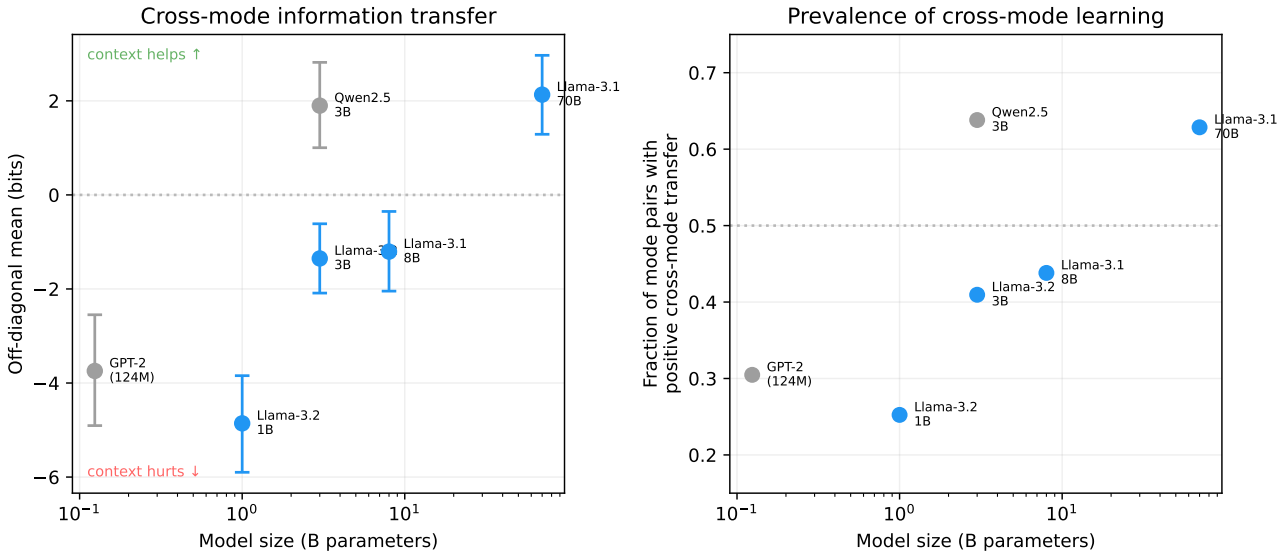


Figure 10. Cross-mode information transfer scales with model size. **Left**: Mean off-diagonal surprise reduction (with bootstrap 95% CI) transitions from negative to positive as models grow, indicating that larger models extract more information from cross-mode context. **Right**: The fraction of mode pairs showing positive cross-mode transfer increases monotonically across the 4 Llama models (blue; $p = 4.2\%$ under monotone null). GPT-2 and Qwen2.5-3B (gray) are included for context but differ in architecture.

Bits/byte vs. total bits, and the length-matching consequence. A separate choice within the framework is whether to report total bits (length-dependent but tokenizer-agnostic) or per-byte bits/byte (also tokenizer-agnostic, with length scaled out). We default to bits/byte for the headline diversity score, but per-byte quantities are not invariant to response length: in a causal LM, longer responses give θ more within-response context to predict later tokens, lowering per-byte surprise. For the OLMo-2-7B experiment (Section 6) raw response length is not monotone across stages, so we score the length-matched subset as the headline analysis: every response in a (stage, prompt) tuple is truncated to a common per-prompt UTF-8 byte length (the minimum across all 40 responses) before scoring. $111/300$ prompts had a per-prompt common length below 50 bytes – driven by the base model on AlpacaEval (no instruction-following prior) and by SFT on NB-curated short-answer prompts – and were dropped because length-matching them would compare essentially-empty fragments. The headline numbers therefore condition on the $150/200$ AlpacaEval and $39/100$ NB-curated prompts where length-matching is well-defined; the dropped prompts are themselves a stage-specific behavioural signal we are not measuring here. We also verified that the monotone D drop holds on the un-truncated $200 + 100$ prompt sets, with all three pre-registered H1 contrasts remaining Bonferroni-significant at $p < 10^{-13}$ on both prompt sets. Full numbers and reproduction commands are in `investigations/length_matched_rlhf/VERDICT.md`.

C. Future Work

Several directions remain unexplored.

Prompt format optimization. We use a fixed prompt template throughout and do not attempt to optimize it for ICL elicitation. Prompt engineering could improve the metric's discriminative power. Dimension-specific framings (e.g., "Here are several stories in different genres") are one such variant: priming θ to expect variation along a dimension removes the small portion of the surprise attributable to the mere existence of that variation, but it does not isolate per-dimension diversity, since the rest of the cross-response surprise reduction remains.

Ensembling base models. A single base model θ may produce non-monotone a_k curves when pushed out of distribution by long contexts. Ensembling multiple base models at the token level (averaging softmax probabilities) could stabilize the estimates while preserving the autoregressive structure. Our codebase supports token-level ensembling, but we have not yet validated it experimentally.

Disagreement cases with other metrics. We report aggregate correlations with embedding-based and reference metrics on Tevet's tasks (Section 5) but do not look at *which* response sets the metrics rank differently. Identifying prompts or response sets where $C \times a_n$ disagrees with SentBERT, distinct-n, or McDiv is the most direct way to characterise what each metric is actually measuring; the disagreements are where the choice of metric matters. We

expect two regimes to drive disagreement: response sets that share arbitrary rare patterns a base model can pick up on but that fall outside what a fixed sentence embedder represents (where $C \times a_n$ should detect repetition that embeddings miss), and response sets where one response is a mixture of two others (where progressive conditioning credits the mixture for being predictable from its components, while pairwise embedding-distance averages do not).

Total-bits variants of D. We report $D_{Ca_n} = C \times a_n$ in bits/byte for length control, normalising by bytes rather than tokens so that scores remain tokenizer-agnostic. The implementation also exposes a_n in total bits, which is likewise tokenizer-agnostic but not length-controlled. A hybrid scalar $C \times a_n^{\text{bits}}$ rewards length (longer responses have more positions at which to be surprising), which some practitioners may want; characterising when each variant is the appropriate measurement target is left to future work.

Optimizing generators against D. We use $C \times a_n$ as an evaluation-time diagnostic, but it could in principle serve as a training or decoding objective. Future work could fine-tune or sample from a generator π with the goal of maximizing $C \times a_n$ under a fixed base model θ to push π toward response sets that are both coherent and diverse.

D. Excess Entropy and the $C \times E$ Score

This appendix develops the excess entropy E, an alternative summary of the a_k curve that is theoretically interesting (connecting to the excess entropy of computational mechanics (Crutchfield & Feldman, 2003) and to the total correlation) but empirically inferior to D_{Ca_n} on the external benchmarks of Section 5. We tried both the last-point estimator $\hat{E}_n$ and a sigmoid-extrapolated projected floor for a_∞ (Section D.1); we stopped using E when it became clear it cannot distinguish diverse from non-diverse response sets in the few-draws regime where modes do not repeat within n samples. We include it here for theoretical completeness, as a null result, and to explain the empirical comparisons that motivate our preference for D_{Ca_n}.

D.1. The Excess Entropy

The primary metric $D_{Ca_n} = C \times a_n$ uses the curve's endpoint a_n: as the surprise after conditioning on $n-1$ peer responses, a_n reflects both the curve's level and the drop from a_1 in a single scalar. Section 5 shows this combination is what discriminates diverse from non-diverse response sets in the few-draws regime, while a_1 alone (the level) and E (the integrated drop, defined below) each underperform.

We thought that E's learnable structure would be interesting and track diversity. Following the concept of excess entropy from computational mechanics (Crutchfield & Feldman, 2003), define the **excess entropy**:

$$E = \sum_{k=1}^{\infty} (a_k - a_\infty) \tag{7}$$

Units: bits. This is the total learnable structural information in the progressive conditioning process, above the irreducible residual noise. It converges whenever θ's surprise at new responses eventually stabilizes (since $a_k \to a_\infty$ and the excess $e_k = a_k - a_\infty$ decays to zero). The "structure" captured by E is not limited to discrete mode identity: it includes any inter-response regularity to which θ assigns lower conditional surprise.

In practice, a_∞ is unknown. We estimate it as a_n (the last observed value) and compute:

$$\hat{E}_n = \sum_{k=1}^{n} (a_k - a_n) \tag{8}$$

This underestimates E (since $a_n \geq a_\infty$), but the bias decreases with n.

Parametric extrapolation of a_∞. Rather than using a_n directly, one can fit a parametric model to the observed curve and extrapolate a_∞. The a_k curve is theoretically expected to be sigmoidal (with concave-up/exponential decay as the degenerate case when the inflection point $k_0 < 1$). This motivates fitting a four-parameter sigmoid:

$$a_k = a_\infty + \frac{\alpha}{1 + e^{\beta(k - k_0)}} \tag{9}$$

where a_∞ is the asymptotic floor, α is the total drop from a_1 to a_∞, $\beta > 0$ controls the steepness of the transition, and k_0 is the inflection point. The fit yields a_∞ without requiring n to be large enough for convergence. Section B.3 provides empirical evidence that the sigmoid-extrapolated E_{fit} recovers the expected monotonic relationship between mode count and excess entropy, while the raw $\hat{E}_n$ estimator does not.

Relationship to total correlation. The excess entropy and total correlation are complementary decompositions of the same underlying structure. Working in expectation over i.i.d. draws from π, let $\bar{a}_1 = \mathbb{E}[-\log_2 \theta(r \mid p)]$ denote the expected unconditional cross-entropy, $\bar{a}_k = \mathbb{E}[a_k]$ the expected conditional cross-entropy at position k, and $e_k = \bar{a}_k - a_\infty$ the excess at step k. The per-step mutual information is $I_k = \bar{a}_1 - \bar{a}_k = (\bar{a}_1 - a_\infty) - e_k$. Summing:

$$\text{TC}_n = \sum_{k=1}^{n} I_k = n(\bar{a}_1 - a_\infty) - E_n \tag{10}$$

For large n where $E_n \to E$, the total correlation grows linearly with slope $(\bar{a}_1 - a_\infty)$: $\text{TC}_n \approx n(\bar{a}_1 - a_\infty) - E$.

Figure 11 illustrates. The three quantities have clean interpretations, all in bits: $(\bar{a}_1 - a_\infty)$ is the per-response redundancy once θ has fully learned the available structure; E is the total finite structural information, consumed during θ's transient learning phase; TC_n is the cumulative redundancy, which grows without bound.

D.2. Per-Byte Excess Entropy Rate

The excess entropy $E = \sum_k (a_k - a_\infty)$ is in bits. For a length-normalized variant, we define the **per-byte excess entropy rate** by normalizing each response's surprise by its byte count *before* averaging across permutations. For a single permutation σ, the per-byte conditional surprise at position k is $a_k^\sigma / \|r_{\sigma(k)}\|$ (bits/byte). Averaging over permutations gives

$$\hat{a}_k^{\mathrm{rate}} = \mathbb{E}_\sigma \left[\frac{a_k^\sigma}{\|r_{\sigma(k)}\|} \right] \tag{11}$$

and the per-byte floor is estimated analogously from the last position: $\hat{a}_\infty^{\mathrm{rate}} = \hat{a}_n^{\mathrm{rate}}$. The per-byte excess entropy rate is

$$E_{\mathrm{rate}} = \sum_{k=1}^{n} \left(\hat{a}_k^{\mathrm{rate}} - \hat{a}_\infty^{\mathrm{rate}} \right). \tag{12}$$

This treats each response equally regardless of length. Note that $E_{\mathrm{rate}} \neq E/\bar{B}$ in general: because total surprise a_k and byte count $\|r_k\|$ are positively correlated, $E/\bar{B}$ overweights long responses, while the per-response normalization in E_{rate} avoids this.

D.3. The $C \times E$ Score

Combining E with the coherence term C (Section 3.2) yields scalar scores

$$C \times E \quad \text{(bits)} \qquad \text{or} \qquad C \times E_{\mathrm{rate}} \quad \text{(bits/byte)}. \tag{13}$$

The first measures total structural information, and the second normalizes per byte so long and short responses are treated equally.

D.4. Why Not Weight Excess Entropy Inside the Sum?

A natural alternative to reporting E and C separately is to weight the terms within E by coherence, computing $E_w = \sum_k w_k \cdot (a_k - a_\infty)$ where $w_k = 2^{-h_\theta(r_k|p)}$. This would suppress contributions from incoherent modes directly at the point of measurement. We considered and rejected this approach because it breaks the information-theoretic interpretation. $E = \sum_k e_k$ has a clean meaning as the total structural information above the noise floor, connected to Crutchfield and Feldman's excess entropy, and inserting weights makes E_w a hybrid quantity that is not the excess entropy of any well-defined process.

D.5. Limitations of $C \times E$

$C \times E$ inherits two limitations from E. First, E measures *recurrence*: surprise reduction as multiple responses accumulate, rather than continuous spread. A policy that draws from a broad, roughly continuous distribution of coherent outputs has $a_k \approx a_1$ for all k, yielding $E \approx 0$ despite maximal diversity. Second, E is near-zero in the few-draws regime: with only 5–10 responses and no repeated modes, θ finds little learnable structure regardless of diversity, so $C \times E$ carries almost no signal. These limitations are the reason D_{Ca_n} is preferred as the primary score: a_∞ is the *level* of the floor, which differs meaningfully between diverse and non-diverse response sets even when the curve stays nearly flat.

Empirical inferiority. On Tevet's diversity-eval benchmark with only 5 responses per set (Section 5), $C \times a_n$ achieves ROC AUC of 0.867 on McDiv_nuggets prompt_gen while the $D_{\mathrm{fit}} = C \times E_{\mathrm{fit}}$ variant cannot be computed at all (the four-parameter sigmoid fails to converge on only five points) and the discretized D_{disc} variant is anti-correlated with diversity (AUC = 0.303, well below chance). Figure 12 shows the ROC curves.

Why E fails in this regime. With only 5 responses and no repeated modes, θ finds little learnable structure regardless of diversity. The a_k curve stays approximately flat, yielding $E \approx 0$ for both diverse and non-diverse sets. The E-based scores thus carry almost no signal. The asymptotic floor a_∞, by contrast, differs substantially between the two conditions: diverse responses remain surprising after conditioning (a_∞ high), while paraphrases become predictable (a_∞ low). This is the fundamental reason D_{Ca_n} is preferred.

D.6. Mode Count Scaling: Excess-Entropy Metrics

In the synthetic mode count experiments (Section B.3), the sigmoid-extrapolated E_{fit} is monotonic in mode count for Qwen2.5-3B but has wide confidence intervals at high m, while the raw $\hat{E}_n$ estimator is non-monotonic. Table 6 and Figure 13 show the details.

E. Dataset Construction Confounds in McDiv

When validating against the McDiv_nuggets benchmark (Tevet & Berant, 2021), we observed that the mean $\bar{a}_1$ curve (per-byte) for the "low diversity" group starts *above* the "high diversity" group in the story_gen task (Figure 15). Since a_1 measures the surprise of the first response conditioned only on the prompt (before any other responses appear in context), this gap cannot reflect diversity. It is a confound in the dataset construction.

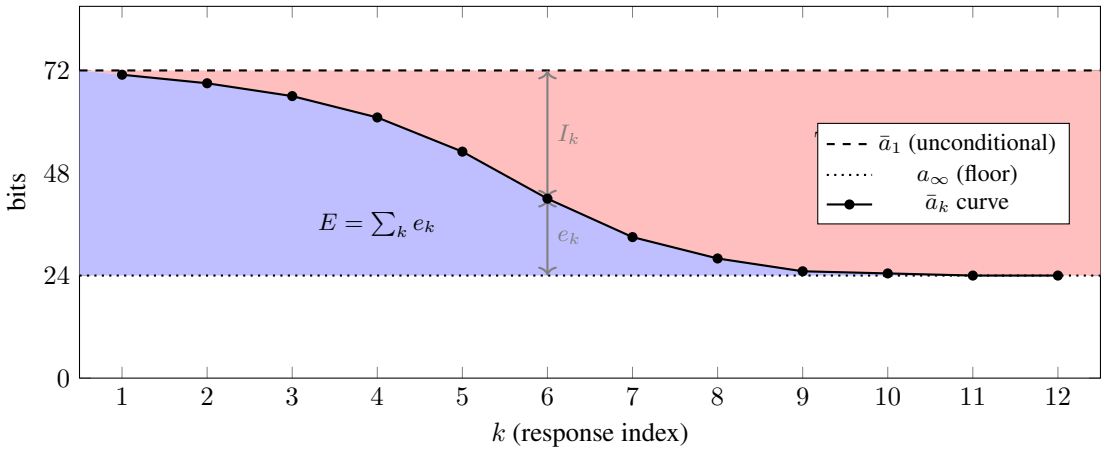


Figure 11. Decomposition of the gap between unconditional surprise $\bar{a}_1$ and the asymptotic floor a_∞ at each step k. The per-step gap $\bar{a}_1 - a_\infty$ splits into mutual information $I_k = \bar{a}_1 - \bar{a}_k$ (red, above the curve) and excess $e_k = \bar{a}_k - a_\infty$ (blue, below the curve). Summing across k: the red area is the total correlation TC_n; the blue area is the excess entropy E. As k grows, $e_k \to 0$ and each step contributes the full $\bar{a}_1 - a_\infty$ to TC_n, so TC_n grows linearly while E converges. All quantities are in bits. This figure is theoretical: the sigmoidal $\bar{a}_k$ shape reflects our initial expectation, but we subsequently found that the empirical curves are exponential decay without an initial plateau (see Section B.4).

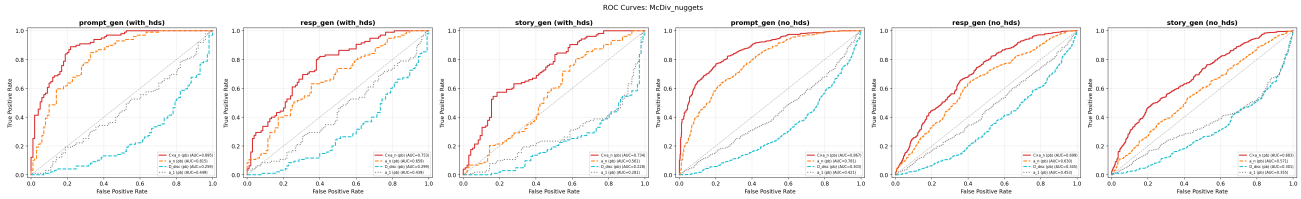


Figure 12. ROC curves for McDiv_nuggets binary classification (Qwen2.5-3B, 50 permutations). Five per-byte metrics are overlaid per panel: $C \times a_n$ (red, solid), a_n alone (orange, dashed), $D_{\text{fit}} = C \times E_{\text{fit}}$ (blue, solid), $D_{\text{disc}} = C \times \hat{E}_n$ (cyan, dashed), and a_1 (gray, dotted). Line style is redundant with color so the encoding remains legible under common color-vision deficiencies. $C \times a_n$ dominates the other four scores on every task; the E-based scores (D_{fit}, D_{disc}) hug or fall below the chance diagonal in this 5-response regime.

Table 6. Mode count scaling metrics for Qwen2.5-3B ($n = 20$, 1000 draws). E_{fit} is the sigmoid-extrapolated excess entropy; a_n is the mean floor at $k = 20$; σ_ℓ is per-byte coherence spread; k_0 is the sigmoid inflection point. Confidence intervals are 95% bootstrap. The high-m rows ($m \geq 8$) have E_{fit} CIs spanning more than half the mean and should be read as preliminary; the trend direction is robust but the point values are imprecise.

m	E_{fit} (bits)	a_n (bits)	σ_ℓ (bits/byte)	k_0
1	296 ± 11	9.3	0.1	-10.0
2	568 ± 24	16.9	0.2	-10.0
3	704 ± 40	26.7	0.2	-10.0
4	799 ± 67	37.4	0.3	-10.0
5	855 ± 113	48.0	0.3	-10.0
6	940 ± 127	56.1	0.3	-10.0
7	1069 ± 194	61.5	0.3	-10.0
8	1138 ± 253	66.9	0.3	-10.0
9	1563 ± 705	72.8	0.3	-10.0
10	1807 ± 1194	77.8	0.3	-10.0

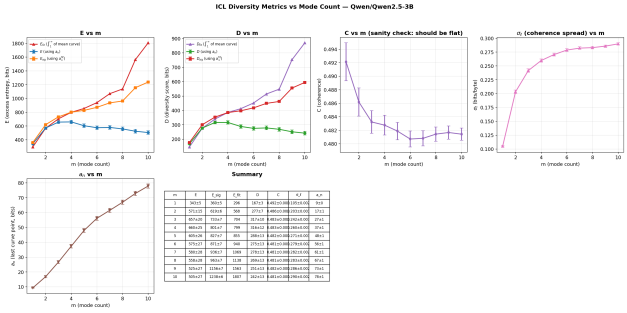


Figure 13. Summary metrics vs. mode count m (Qwen2.5-3B, $n = 20$, 1000 draws). E_{fit} (sigmoid-extrapolated) increases monotonically, while the raw $\hat{E}_n$ does not.

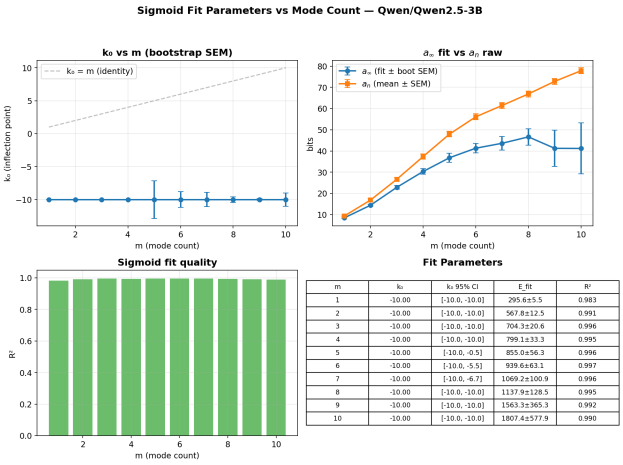


Figure 14. Sigmoid fit parameters vs. mode count m (Qwen2.5-3B). The inflection point k_0 remains at the lower bound (-10) across all m, indicating pure exponential decay without an initial plateau (see Section B.4).

E.1. Mechanism

The McDiv protocol (Tevet and Berant §6.4), from which McDiv_nuggets is sampled (Tevet and Berant Appendix C.2 specifies McDiv_nuggets as the 3K subset of McDiv on which distinct-n correlates at zero), pairs a low-diversity set with each high-diversity set as follows. Workers first write five *different* continuations (the high-diversity set). The same worker is then asked to *self-select one* of their own five responses and paraphrase it five times, preserving content while varying form (the low-diversity set). Tevet and Berant do not characterize the distribution of which responses workers self-select.

Empirically (see Table 7 and Figure 15, and the illustrative examples from the McDiv_nuggets story_gen dataset in Section E.3 below), we observe that the self-selected endings for the low-diversity sets tend to be more specific or dramatic (e.g., "He scored winner" or "they quit") than the typical high-diversity continuations (e.g., "Joel fired the cook when things went too far downhill"). This means that individual low-diversity responses are intrinsically more surprising to the base model, not because of diversity, but because of the specificity of the endings workers self-selected for paraphrasing.

E.2. Evidence

Table 7 summarizes the gap. The per-byte a_1 difference is substantial (a content effect) while the total-bits difference is small because high-diversity responses are on average several bytes longer.

The gap persists after binning by response length (Table 8), confirming it is primarily a content effect rather than a length artifact: the high-minus-low per-byte gap remains positive across the bulk of the length range.

Table 7. Unconditional surprise (a_1) by diversity label, McDiv_nuggets story_gen.

	High diversity	Low diversity	Gap
a_1 (bits/byte)	0.977	1.154	+0.177
a_1 (total bits)	49.1	50.5	+1.4
Mean response bytes	52.7	46.5	−6.3
n	103	97	

Table 8. Unconditional per-byte surprise by response length bin, story_gen.

Byte range	High div. mean	Low div. mean	Gap
$[20, 40)$	1.172	1.340	+0.167
$[40, 60)$	0.878	1.019	+0.141
$[60, 80)$	0.853	1.029	+0.176
$[80, 120)$	0.905	0.908	+0.003

E.3. Illustrative Examples

Figures 16 and 17 show per-token surprise (in bits) for the first response of a high-diversity and low-diversity sample, respectively. Samples are drawn by ranking each label group by per-byte a_1 (ascending for high-diversity, descending for low-diversity) and taking the sample at the $\lfloor N/4 \rfloor$ position, chosen to be representative of the expected confound direction rather than an extreme outlier; the specific pinned IDs (`sa_00698`, `sa-b_00279`) are fixed in `scripts/dataset_confound_analysis.py` for reproducibility. Red bars are measured response tokens; grey bars are masked context tokens.

High diversity (Figure 16). Context: *"Joel owned a restaurant. He hired a cook that didn't care about his work. The cook didn't clean after himself. The kitchen was a mess."* Response: *"Joel fired the cook when things went too far downhill."* This is a natural, predictable continuation ($a_1 = 0.783$ bits/byte).

Low diversity (Figure 17). Context: *"Harry and his basketball team was losing the game. The coach called for a timeout. Harry boosted morale by talking to his team. The team caught up and the game was tied."* Response: *"He scored winner."* This is a specific, somewhat ungrammatical ending ($a_1 = 1.777$ bits/byte). All five responses in this set are paraphrases of the same idea ("Harry scored the winning point").

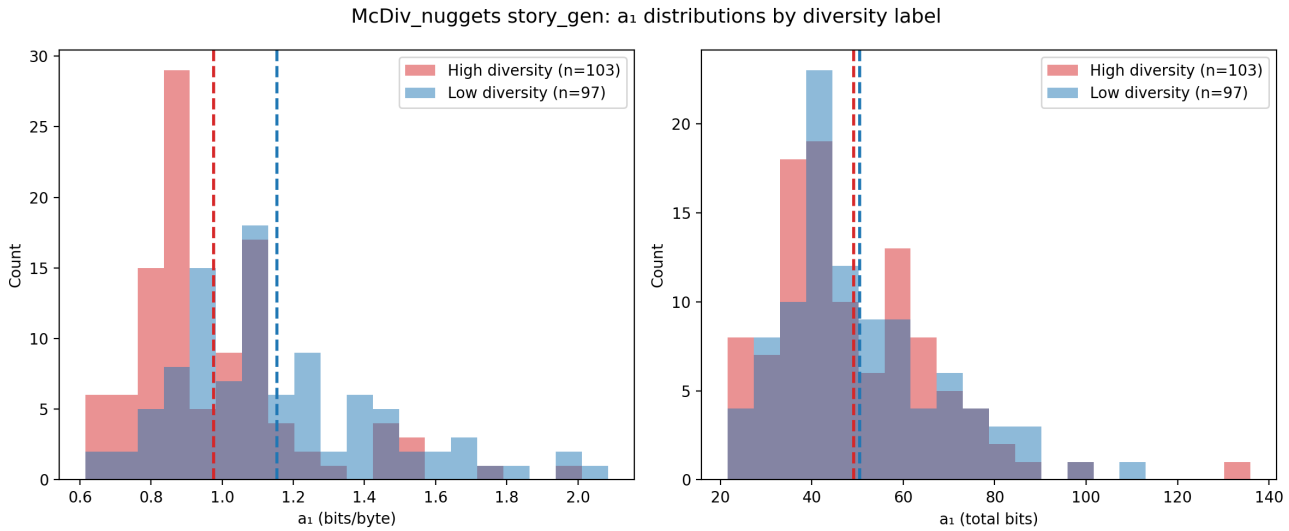


Figure 15. Distribution of a_1 (unconditional surprise of the first response) for high- vs. low-diversity story_gen samples. Left: per-byte. Right: total bits. The per-byte gap (see Table 7) is a dataset construction artifact.

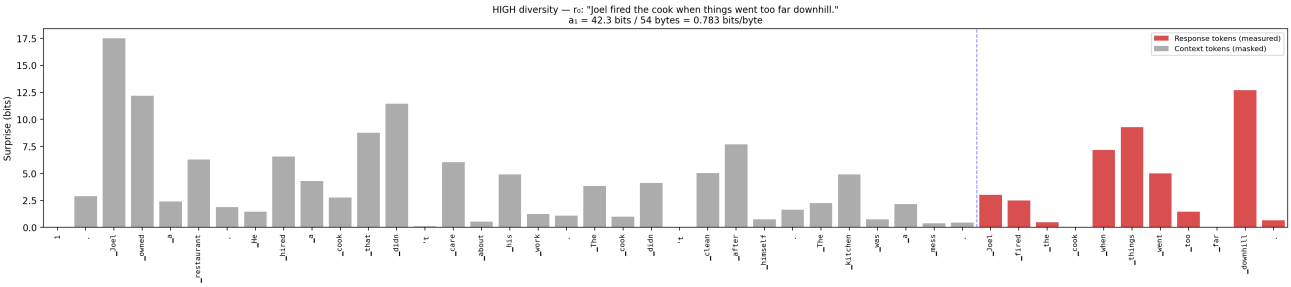


Figure 16. Per-token surprise for a high-diversity sample's first response. The continuation ("Joel fired the cook...") is predictable, yielding low per-token surprise across response tokens.

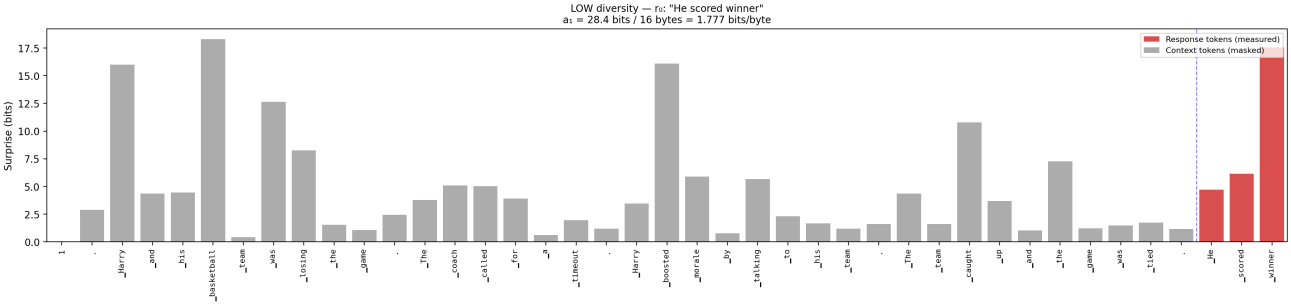


Figure 17. Per-token surprise for a low-diversity sample's first response. The specific ending ("He scored winner") is inherently more surprising to the base model, despite being labeled "low diversity."

E.4. Implications

The construction confound (low-diversity sets are paraphrases of specific dramatic endings rather than a random subsample of low-diversity content) lifts the entire $\bar{a}_k$ curve on low-diversity sets, since their content is intrinsically surprising under θ at every k, and pushes C down because the same content is less coherent. a_1's upward shift opposes the diversity label; C's downward shift happens to align with it (a confound artifact, not coherence genuinely tracking diversity). At a_n, the diversity signal we are measuring overrides the confound's upward pull: paraphrases collapse to predictable once θ has seen the others while genuinely diverse responses do not, so a_n ends up lower on low-diversity sets and $C \times a_n$ tracks the label despite the confound (Section 5).

F. Aggregation Across Prompts

This appendix concerns aggregation across prompts. Aggregation *within* a single prompt, across permutations of the response ordering, is described in Section A.3 and given by Eq. (6): the per-byte $\bar{a}_k$ is a mean of per-permutation per-byte rates, not a ratio of permutation-averaged bits to permutation-averaged byte counts.[7]

The diversity score D_{Ca_n} is defined relative to a single prompt p. To summarize a policy's diversity across a prompt suite $P = \{p_1, \ldots, p_M\}$, one can average:

$$\bar{D}_{Ca_n}(\pi) = \frac{1}{M} \sum_{j=1}^{M} D_{Ca_n}(\pi, p_j) \tag{14}$$

To compare two policies π_1 and π_2, the natural scalar is the difference:

$$\Delta D_{Ca_n} = \bar{D}_{Ca_n}(\pi_1) - \bar{D}_{Ca_n}(\pi_2) \tag{15}$$

ΔD_{Ca_n} answers "how much more plausibility-weighted residual diversity does π_1 have than π_2, on average across prompts?" No division is involved, so ΔD_{Ca_n} is stable even when either policy's score is near zero.

G. Scaling the Base Model: Qwen2.5-3B vs Qwen3-30B-A3B-Base

To test whether a stronger base model improves the metric on Tevet's benchmarks, we re-ran the full evaluation using Qwen3-30B-A3B-Base (Yang et al., 2025) (a 30B-parameter mixture-of-experts model with ∼3B active parameters per token) via the Tinker API,[8] with the same setup as in Section 5: completion-format prompting, 50 permutations, no fine-tuning. Table 9 reports per-task $C \times a_n$ (per-byte) for both base models.

Result: scaling did not help on the binary tasks. On the 12 binary classification tasks (McDiv, McDiv_nuggets, ConTest), Qwen2.5-3B wins on 11 of 12, with ConTest prompt_gen the only marginal win for Qwen3-30B-A3B ($+0.005$ AUC). The mean degradation from scaling is small but consistent (mean ΔAUC = -0.013 across the 12 binary tasks). On the 6 DecTest tasks (continuous temperature labels), Qwen3-30B-A3B is mildly better on 5 of 6 (mean $+0.021$ in Spearman ρ, with one tie).

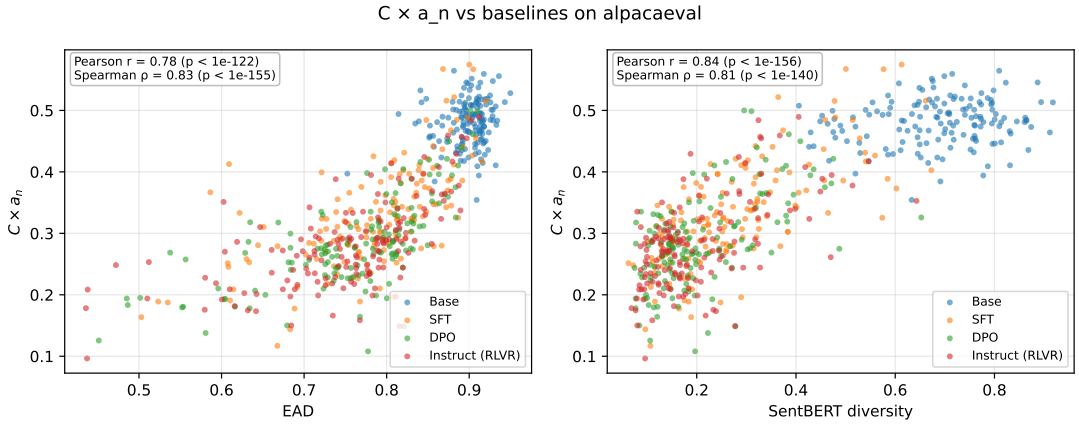
Interpretation. The binary result is a clear negative: scaling from 3B to 30B active parameters does not improve discrimination on McDiv or ConTest. We do not have a confident explanation. The Qwen3-30B-A3B run was not repeated after a bug fix applied to the primary experiments, so the comparison should be treated as preliminary. We leave the question of whether larger base models improve the metric on binary tasks to future work.

H. Cross-Metric Agreement on the OLMo-2-7B RLHF Experiment

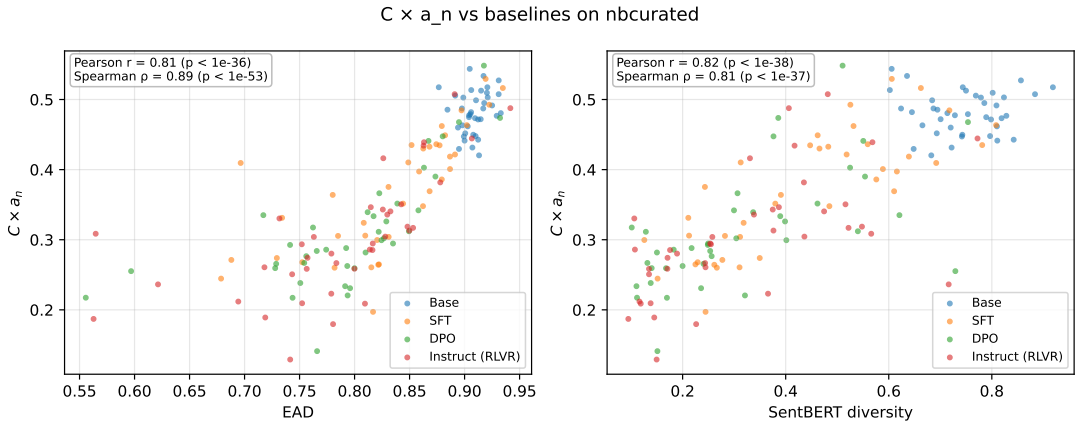
The RLHF case study (Section 6) reports the headline monotone D_{Ca_n} drop in Table 4. Figure 18 plots D_{Ca_n} against the lexical (EAD) and semantic (SentBERT) baselines, confirming that the three metrics see the same diversity-loss signal as the $\bar{a}_k$ curves and per-prompt distributions in §6.

[7] An earlier revision of our implementation computed the ratio-of-means form, in which the bits and byte counts at each position are each averaged across permutations before dividing. After enough permutations, the byte counts at each position approach the mean response length, so that estimator collapses into the total-bits curve rescaled by a single constant, losing the per-permutation per-byte signal that Eq. (6) preserves. Switching the implementation to the form in Eq. (6) shifts the headline numbers in this paper by 0–17% (in mean-of-ratios' favor), with no qualitative change to any ranking we report.

[8] https://thinkingmachines.ai/tinker



(a) AlpacaEval (length-matched, $n = 150$ prompts).



(b) NoveltyBench curated (length-matched, $n = 39$ prompts).

Figure 18. Per-prompt D_{Ca_n} versus EAD (left subpanel) and SentBERT-similarity diversity (right subpanel), coloured by stage, on the length-matched subset of prompts. Length-matching truncates each (stage, prompt) tuple's responses to a common per-prompt byte budget so the per-byte conditional surprise that defines a_n is not depressed by response length. Pearson r and Spearman ρ (two-sided p) are reported in each panel. The rank correlations are positive across both prompt sets and both baselines: D_{Ca_n} tracks the same diversity-loss signal EAD and SentBERT detect, and the agreement is not an artefact of response length.

Table 9. Per-task comparison of $C \times a_n$ (per-byte) for Qwen2.5-3B vs Qwen3-30B-A3B-Base on Tevet diversity-eval. ΔAUC is positive when the larger model wins. ROC AUC is reported for the binary tasks (McDiv, McDiv_nuggets, ConTest); only Spearman ρ is meaningful for DecTest (continuous temperature labels).

		Qwen2.5-3B		Qwen3-30B-A3B		
Section	Task	ρ	AUC	ρ	AUC	ΔAUC
McDiv	prompt_gen (no_hds)	0.729	0.921	0.718	0.915	-0.006
	resp_gen (no_hds)	0.500	0.788	0.483	0.779	-0.009
	story_gen (no_hds)	0.523	0.802	0.493	0.785	-0.017
McDiv_nuggets	prompt_gen (no_hds)	0.636	0.867	0.615	0.855	-0.012
	prompt_gen (with_hds)	0.683	0.895	0.655	0.878	-0.017
	resp_gen (no_hds)	0.345	0.699	0.325	0.687	-0.012
	resp_gen (with_hds)	0.437	0.753	0.425	0.746	-0.007
	story_gen (no_hds)	0.317	0.683	0.293	0.669	-0.014
	story_gen (with_hds)	0.405	0.734	0.376	0.717	-0.017
ConTest	prompt_gen (with_hds)	0.584	0.837	0.592	0.842	+0.005
	resp_gen (with_hds)	0.391	0.726	0.333	0.692	-0.034
	story_gen (with_hds)	0.686	0.896	0.653	0.877	-0.019
DecTest	prompt_gen (no_hds)	0.842	—	0.877	—	—
	prompt_gen (with_hds)	0.845	—	0.875	—	—
	resp_gen (no_hds)	0.771	—	0.804	—	—
	resp_gen (with_hds)	0.760	—	0.777	—	—
	story_gen (no_hds)	0.763	—	0.771	—	—
	story_gen (with_hds)	0.785	—	0.785	—	—