DART: DO LANGUAGE MODELS FOLLOW REQUESTED DISTRIBUTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask whether language models can sample from a probability distribution that a user explicitly requests. We introduce Distribution Audit via Requested Targets (**DART**), an audit that asks a model for a single draw from a distribution stated in the prompt and scores the completed answer exactly over an enumerated candidate set, separately measuring whether the model attempts the task and, when it does, how closely its choice follows the requested law. Applied to 13 base and post-trained model pairs from six families, DART finds a consistent pattern: base models usually decline the task; when they answer, they typically follow the requested law more closely than their post-trained counterparts, which almost always answer yet concentrate on a single favorite outcome. Four of the six comparable chat-model endpoints place most of their probability on a single answer in a uniform ten-way draw, and the favorite shifts with the answer range rather than a fixed token. A factorial experiment shows that the residual distortion decomposes into interpretable priors over label identity and display position that predict held-out combinations but do not transfer across models or prompt templates. Interventions that flatten the measured output distribution fail to restore genuine sampling behavior. DART thus makes distribution-following measurable: current post-training trades it for unconditional compliance.

1 INTRODUCTION

Consider a simple request: *pick a number uniformly at random between 1 and 10.* People may answer "7" more often than chance, but the request itself is unambiguous: a die or a random number generator would follow it exactly. Language models are increasingly used in roles that implicitly require this capability, from stochastic simulation and synthetic data generation to randomized evaluation and agent rollouts (Wang et al., 2023; Shumailov et al., 2024; Argyle et al., 2023; Santurkar et al., 2023; Taillandier et al., 2025; Ye et al., 2026). Whether a model can comply with such a request is a basic question about what post-training does to a model's behavior, yet it is rarely measured directly.

The question is harder to answer than it looks, because a completed answer is a token sequence, and different token sequences can spell out the same answer ("7" may be one token, a leading-space variant, or several tokens). A single forward pass tells us the probability the model assigns to each completed answer only if we enumerate every token path that produces it and add them up. Much of the existing evidence on models and randomness sidesteps this problem: it asks models to state probabilities in words, or it counts sampled outputs (Zhang et al., 2025; West and Potts, 2025); both routes mix the model's choice among candidates with its willingness to produce a candidate at all.

Distribution Audit via Requested Targets (DART) addresses this measurement problem directly. The audit lists the candidate answers explicitly in the prompt and asks for a single draw from a stated target distribution over them; for each candidate we enumerate its token paths and sum their probabilities exactly under the model's own scoring, with no sampling noise. This separates two questions that are often conflated: *coverage*, the probability that the model produces any candidate at all, and *conditional fidelity*, how closely its choice among candidates follows the requested law when it produces one. A model can score well on one axis and fail badly on the other (Figure 1).

Applying DART to 13 base→post-trained model pairs across six families, we find a consistent trade-off between compliance and fidelity. Base models rarely produce a candidate answer, but when they do, they typically choose more uniformly among the candidates than their post-trained

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

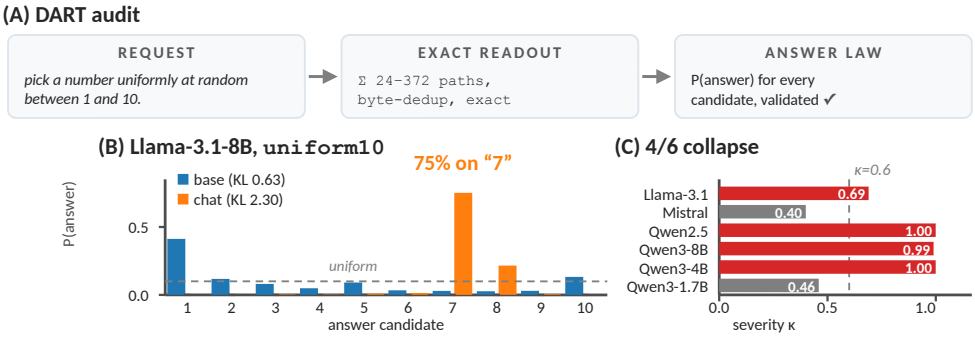


Figure 1: **DART and the collapse it measures.** (A) The audit: the prompt states the candidate answers and a target law; answer probabilities are computed exactly by summing token paths. (B) Llama-3.1-8B on `uniform10`: the base model spreads its conditional candidate mass while the post-trained model concentrates on "7"; parentheses give conditional KL in bits. (C) Conditional KL to the requested law for the six comparable chat endpoints of Table 4: four of six (three of four lineages) exceed the collapse threshold κ=0.6 (dashed line).

Table 1: **What each measurement route delivers.** *Prompt target*: the target is stated in the prompt. *Exact*: no sampling noise. *Answer event*: the measured quantity is a completed answer, not a token or self-report. *Cov. sep.*: compliance (coverage) and conditional fidelity reported separately. *Free-run val.*: validated against free-running generation. *Stage-wise*: supports checkpoint-level stage attribution. ✓– marks partially addressed properties.

Route	Prompt	Exact	Answer	Cov. sep.	Free-run	Stage
Verbalized probability reports	✓	×	×	×	×	×
Empirical sample counting	✓	×	✓	×	—	×
First-token marginal	✓	✓	×	×	×	×
Marginal-likelihood scoring (Cao and Rimell, 2021)	×	✓	✓–	×	×	×
Decoding-free estimation (Pohl et al., 2025)	×	✓	✓–	×	×	×
DART (ours)	✓	✓	✓	✓	✓	✓

counterparts do. Post-training flips the picture: the models answer nearly every prompt, yet their choices concentrate on a single favorite token; on the uniform 1–10 probe, four of the six comparable endpoints put most of their probability on one answer. This favorite is not a fixed token such as "always 7"; it moves with the candidate range, which points to a learned anchoring policy (Section 4).

Our contributions:

1. We introduce DART, an audit that scores completed answers exactly over enumerated candidate sets, separately measuring whether a model attempts the task and how closely its choice follows the requested law.
2. We trace the distortion across the OLMo-2 post-training pipeline and show that it is already detectable at SFT and deepens at later stages on selected probes.
3. We decompose the residual distortion into interpretable priors over label identity and display position that predict held-out combinations but do not transfer across models or prompt templates.
4. We show that interventions that flatten the measured output distribution fail to restore genuine sampling behavior under free-running generation.

2 The DART audit

2.1 Probes

Each DART probe names a finite candidate set Z with $|Z| = K$, lists the candidates explicitly in the prompt, states a target law q over Z, and asks for a single draw. The core battery uses seven probe families whose target laws are known analytically: uniform integers on 1–10 and 1–20, a fair die, a fair coin, uniform letters A–Z, uniform weekdays, and a biased die with specified face probabilities.

Table 2: **Readout map.** Each readout answers a different question about the model. DART's branch and strict readouts measure completed answers; the completion and first-token readouts measure surrogates that can diverge from completed-answer behavior.

Readout	Scores	Estimand	Known failure
Completion log-prob	one path per z	continuation preference	ignores tokenization variants
First-token marginal	next token only	leading-token mass	shared-prefix conflation
DART branch (primary)	all paths in $\mathcal{T}(z)$	completed-answer choice	requires inventory
DART strict	branch + terminator	immediate answer event	parser-dependent tail

2.2 Completed-answer probabilities

For each candidate $z \in Z$ we declare a finite inventory $\mathcal{T}(z)$ of token-path variants: capitalization and leading-space forms, multi-token realizations, and valid terminators. The completed-answer probability of z is the sum of sequence probabilities over its inventory under the model's scoring configuration η (chat template or raw prompt):

$$\tilde{\pi}_\eta(z) \;=\; \sum_{x_{1:T} \in \mathcal{T}(z)} \prod_{t=1}^{T} \pi_\theta(x_t \mid x_{<t}) \;-\; \mathrm{PF}(z), \qquad c_B = \sum_{z \in Z} \tilde{\pi}_\eta(z), \tag{1}$$

where c_B is the *branch coverage*: the total probability the model assigns to producing any candidate. The prefix term $\mathrm{PF}(z)$ removes the continuation mass already counted under a longer candidate z' whenever a path of z is a proper prefix of a path of z'. Without this correction, shared leading tokens are counted twice, which breaks naive branch sums on the Qwen tokenizers at $K{=}20$ (Appendix F). The conditional readout $\hat{p}(z) = \tilde{\pi}_\eta(z)/c_B$ describes the model's choice among candidates given that it produces one. A terminator-complete *strict-event* variant requires end-of-turn mass instead,

$$\pi_\eta^{\mathrm{strict}}(z) \;=\; \sum_{x_{1:T} \in \mathcal{T}(z)} \Big(\prod_{t=1}^{T} \pi_\theta(x_t \mid x_{<t}) \Big) \cdot \pi_\theta(\mathrm{eot} \mid x_{1:T}), \tag{2}$$

so that a candidate counts only if the model would actually stop there. The per-path terminator factor makes the strict events of distinct candidates disjoint, so the strict readout needs no prefix correction. A re-derivation without the correction still reproduces coverage to within 0.006 in all 18 validation cells (Appendix F). All readouts are evaluated in a single scoring pass per candidate path; no sampling is involved.

2.3 Branch and first-token readouts

DART's primary readout is the sequence-level branch sum in Eq. (1). Two simpler readouts appear in the literature and serve here as baselines (Table 1 contrasts what each measurement route delivers). The *completion log-probability* scores the candidate string as a continuation,

$$s_\eta(z) \;=\; \sum_{t=1}^{T_z} \log \pi_\theta(x_t^z \mid x_{<t}^z), \qquad \hat{p}_{\mathrm{comp}}(z) \;=\; \frac{\exp s_\eta(z)}{\sum_{z'} \exp s_\eta(z')}, \tag{3}$$

which scores a single canonical path per candidate and ignores alternative tokenizations. The *first-token readout* uses only the next-token distribution,

$$\hat{p}_{\mathrm{first}}(z) \;=\; \frac{\sum_{v \in \mathrm{first}(z)} \pi_\theta(v)}{\sum_{z'} \sum_{v \in \mathrm{first}(z')} \pi_\theta(v)}, \tag{4}$$

which is cheap but conflates candidates that share a leading token; this failure mode is severe for multi-digit ranges. Table 2 maps each readout to the estimand it targets.

2.4 Metrics

For each probe we report the branch coverage c_B and the conditional Kullback–Leibler divergence $\mathrm{KL}(\hat{p}\|q)$ in bits. To compare across probe sizes we use the normalized severity

$$\kappa \;=\; \frac{\mathrm{KL}(\hat{p}\|q)}{\log_2 K} \;\in\; [0, 1], \tag{5}$$

Table 3: **Audit scope.** "Pairs" counts base→post comparisons, one per post-trained checkpoint against its own base; "stages" counts separately scored checkpoints, including the seven released OLMo-2 training stages. All post-trained cells are scored under the model's own chat template; base cells use raw prompting.

Family	Pairs	Stages	Post-training documented
Llama-3.1-8B	1	2	no
Mistral-7B-v0.3	1	2	no
Qwen2.5-7B	1	2	no
Qwen3 (0.6–8B)	4	8	no
OLMo-2 (7B, 1B)	5	7	yes (SFT/DPO/RLVR)
R1-Distill-Qwen-1.5B	1	2	reasoning distillation

which is 0 for a perfect match and, for the uniform targets where it is reported, approaches 1 as the conditional distribution concentrates on a single candidate. For the biased-die probe, whose target is non-uniform, we additionally report the entropy deficit $\mathbb{H}(q) - \mathbb{H}(\hat{p})$, which isolates concentration from mass shifts among faces; we count the two components separately throughout. Where coverage is below 1, the conditional readout describes only candidate-emitting behavior, so we always report coverage alongside it.

Interpretation notes. Two conventions recur throughout. *(i) Compliance correction.* Where coverage c is below 1, the unconditional deficit and the coverage-weighted conditional KL differ by at most $\mathbb{H}_2(c)$ (binary entropy), i.e. ≤ 0.081 bits at $c \geq 0.99$, so the two readouts coincide to within ≈ 0.08 bits on the chat-protocol endpoint cells. *(ii) Conditional Jensen–Shannon (JS) floor.* Agreement between the deterministic readout and finite-sample generation is judged per cell against a multinomial noise floor estimated conditional on the measured non-uniform distribution (Appendix B); sub-floor divergences are sampling noise, not estimator discrepancy.

3 Experimental setup

We audit 13 base→post-trained pairs across six families: Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2.5-7B (Yang et al., 2024), Qwen3 at four scales (0.6B, 1.7B, 4B, 8B; Yang et al., 2025), the OLMo-2 pipelines at 7B and 1B with all intermediate post-training stages (Team OLMo, 2025; Lambert et al., 2024), and R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) as a reasoning-distilled case. Table 3 summarizes the audit scope (one pair per post-trained checkpoint against its own base; the five OLMo-2 stage checkpoints contribute five pairs). The *directly comparable* families are six chat-template endpoint pairs (Llama-3.1-8B, Mistral-7B, Qwen2.5-7B, Qwen3-8B/4B/1.7B) whose uniform10 branch coverage exceeds 0.99 for all but Mistral-7B (0.988). Qwen3-0.6B, the OLMo-2 stages, and R1-Distill have far lower coverage and are reported separately. Base models are evaluated under raw prompting; post-trained models under their chat templates, matching deployment. The main text focuses on the uniform10 probe and the OLMo-2 stage evidence; full grids, paraphrase replicates, and the raw-protocol battery are in the appendix.

Every cell is scored on the primary prompt and two paraphrases, with 384 free-running replicates for sampling-based validation of the deterministic readouts. Validation cells use 2048 generations each. Free-running generations use temperature 1.0, top-p 1.0, and at most 32 new tokens, with raw token IDs retained per sample. Calibration of the readouts against free-running generation uses a per-cell parametric bootstrap (4000 multinomial draws from the strict conditional, 99% band), with a χ^2 statistic as a cross-check. The OLMo-2 checkpoints provide the only fully documented stage sequence in the battery, so we use them for all stage-attribution claims and treat the remaining families as endpoint comparisons. Protocol corrections, including a shared-prefix artifact that initially inverted the uniform20 severity ordering, are documented in Appendix I.

4 Results

4.1 Post-training collapses conditional fidelity while perfecting compliance

Post-trained models comply almost perfectly and choose poorly; base models do the reverse. Table 4 reports the headline comparison on uniform10 (Figure 2). Base models almost never emit a candidate: Llama-3.1 base places only 0.6% of sequence-level branch mass on any integer candidate, and Qwen2.5-7B base 11.9%. Conditional on emitting a candidate, however, base models are closer to

Table 4: **Uniform10 audit.** Base cells are candidate-conditional readouts under raw prompting at the stated branch coverage c_B; unconditional deficits appear in Appendix C. Post cells are chat-template readouts. κ is Eq. (5); modal share is the largest conditional candidate mass; rows are grouped by the descriptive collapse threshold. Strict-event calibration is reported separately in Table 5.

Model	Base		Post-trained			
	c_B	KL (cond.)	c_B	KL	κ	Modal
Collapsing families ($\kappa \geq 0.6$)						
Llama-3.1-8B	.006	.626	≈1.00	2.297	.69	7 (75%)
Qwen2.5-7B	.119	.373	≈1.00	3.321	1.00	7 (≥99.8%)
Qwen3-8B	.129	.644	≈1.00	3.303	.99	7 (99.8%)
Qwen3-4B	.101	.069	≈1.00	3.321	1.00	7 (≥99.99%)
Moderate families						
Mistral-7B-v0.3	.004	1.150	.988	1.325	.40	7 (48%)
Qwen3-1.7B	.054	.163	≈1.00	1.517	.46	5 (38%)

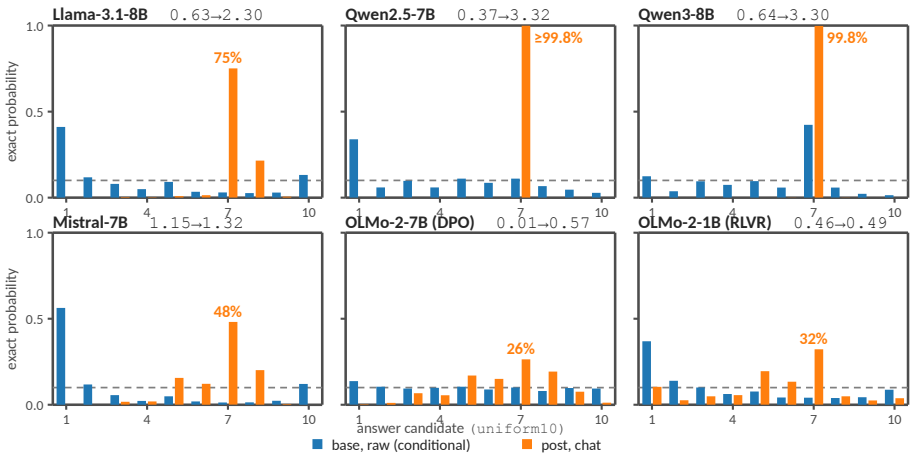


Figure 2: **Conditional answer distributions on `uniform10`.** The dashed line is the uniform target. Base bars use raw prompting and are conditional on a candidate emission (coverage 0.004–0.13); post-trained bars use chat prompting (coverage 0.56–1.00 across panels; per-cell values in Table 9). Several post-trained checkpoints concentrate on one candidate, while the OLMo-2 stages retain broader support. Panels show the three largest collapsing models, the moderate Mistral-7B, and the two OLMo-2 endpoint stages; percentage annotations mark the post-trained modal share. Panel titles report base→post conditional KL.

uniform than their post-trained counterparts (0.63 and 0.37 bits here; 0.07–1.15 across the six directly comparable families). Post-trained models show the opposite pattern: coverage is essentially perfect, yet conditional KL ranges from 2.30 to 3.32 bits on the four collapsing families, corresponding to $\kappa \geq 0.6$ and a single modal answer absorbing 75% to over 99% of the conditional mass. The mode is "7" for most families, echoing the human preference documented by Kubovy and Psotka (1976), but it is "5" for Qwen3-1.7B, so the attractor is a property of the trained model rather than of the prompt.

The collapse threshold $\kappa \geq 0.6$ is descriptive; it only groups families with qualitatively similar concentration. Llama-3.1 clears the threshold narrowly ($\kappa \approx 0.69$), and its paraphrase spread stays within the same band, so we read it as collapsed but less severely than the Qwen families. Mistral-7B ($\kappa \approx 0.40$) and Qwen3-1.7B ($\kappa \approx 0.46$) are moderate; the two are also the only cells whose paraphrase spread crosses a verdict band (Table 13). The pattern is not specific to the 1–10 range: on uniform20 Llama-3.1, Qwen3-4B, and Qwen3-1.7B sit in the collapse band with range-shifted anchors (88% on "14", 62% on "14", and 87% on "12"), while Qwen2.5 and Qwen3-8B land just below it ($\kappa \approx 0.53/0.55$).

4.2 The deterministic readout matches free-running generation in 16 of 18 cells

The deterministic readout predicts free-running behavior in 16 of 18 risk-selected cells, and both exceptions have identified causes. Because DART's primary readout is deterministic, it is natural to ask whether it describes what the model actually generates. We validated the branch and strict readouts against 2048-sample free-running generation in 18 risk-selected cells spanning four core models, four probe families, and both protocols (Table 5). Sixteen of eighteen cells

Table 5: **Strict-event validation against free-running generation** (N=2048 per cell). Br: branch-sum coverage; St: finite-inventory strict-event coverage; vr: observed strict-parser valid rate; Δ_b=Br−St and Δ_o=vr−St, computed from unrounded values; JS: empirical-vs-readout divergence in bits. cal: per-cell bootstrap calibration (4000 multinomial draws, 99% band). The two FAIL cells are calibration exceptions analyzed in Appendix F (bf16 forward-path numerics; a marginal failure not replicating under matched replay). †The raw cell passes calibration but carries a protocol flag for its branch-vs-strict coverage gap (a stopping-policy property, Section 2), not a calibration failure. Nine decision-relevant cells shown; the full 18-cell grid, bands, and temperature sweep: Table 19.

Cell	Br	St	vr	Δ_b	Δ_o	JS	cal
Llama3.1 uniform10	1.0000	.9990	.9995	+.0010	+.0005	.0007	ok
Llama3.1 biased-dice	.9986	.9942	.9922	+.0044	−.0020	.0015	FAIL
Qwen3 uniform10	1.0000	.9998	1.0000	+.0002	+.0002	.0004	ok
Qwen3 uniform20	1.0000	.9994	.9995	+.0006	+.0001	.0106	FAIL
Qwen2.5 uniform10	1.0000	1.0000	1.0000	.0000	.0000	.0000	ok
Qwen2.5 uniform10 (raw)	.9770	.8604	.8613	+.1166	+.0009	.0001	ok†
Mistral-7B uniform20	.9996	.9731	.9741	+.0265	+.0010	.0018	ok
OLMo-1B-RLVR1 uniform10	.5627	.5528	.5669	+.0098	+.0141	.0027	ok
Qwen3-0.6B uniform10	.3455	.3438	.3579	+.0017	+.0141	.0026	ok

pass a per-cell bootstrap calibration (fifteen cleanly; the raw-protocol cell passes while carrying a coverage-protocol flag). The readout does not fire on negative controls: a uniform candidate policy scores zero KL by construction, with the estimator pinned against exhaustive hand computation (Appendix D); random-direction steering is null wherever targeted steering is null (Table 11); and the factorial's flat stated-profile controls isolate baseline label priors (Appendix J). The battery contains no high-coverage, high-fidelity post-trained reference, so these cells certify the readout, not any model. The two exceptions are informative. On Qwen3 uniform20 the strict readout underestimates the empirical mixture (JS 0.0106 vs. a 0.0014-bit bound), a gap traced to bf16 forward-path numerics (an fp32 readout rerun remains untested). On Llama3.1 biased-dice the readout sits at the rounded upper endpoint of its band, a marginal failure not replicating under matched replay (Appendix F). Neither implicates the estimator's path accounting, and both are flagged wherever the affected cells appear.

Qwen2.5's post-trained model remains collapsed under raw prompting (branch coverage 0.977 on the validation prompt), but strict coverage drops to 0.860 because the model often continues past the answer, a stopping-policy property, not candidate preference, which is why we keep branch and strict estimands distinct.

Practical use of the readout. The validation results assemble into a five-step protocol. (i) Use the deterministic branch readout whenever the candidate inventory is enumerable; report branch and strict coverages separately. (ii) Calibrate any new cell against $\geq$2048 free-running samples with the per-cell bootstrap band (floors scale with the valid-sample count, Appendix B). (iii) If calibration fails, treat the $\geq$2048-sample empirical distribution as authoritative and the readout as a lower bound on distortion where the failure direction is established, as in the Qwen3 uniform20 cell (Appendix F). (iv) Treat $\kappa \geq 0.6$ as a descriptive collapse band: a model there is not a sampler of the requested law, whatever its compliance. (v) Readout-level flattening is not mitigation; only free-running validation distinguishes a repaired policy from a flattened readout.

4.3 Stage attribution on OLMo-2

The distortion is already measurable at SFT and deepens at later stages on selected probes. The OLMo-2 pipelines let us ask *when* the distortion appears, because every post-training stage—supervised fine-tuning (SFT), direct preference optimization (DPO) (Rafailov et al., 2023), and reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024)—is public. Table 6 tracks conditional KL and coverage across stages (Figure 3). On the 7B chain, the SFT→DPO step worsens conditional KL on all six KL-scored probes; on the 1B chain the same step worsens only uniform10 and the die while the other four probes improve, so we report the 6/6 as a 7B-chain readout from a single measured pipeline, not a cross-chain law. Full-chain monotonicity, measured from the base model rather than assumed, holds on only three of six probes: on coin the base model's conditional KL substantially exceeds SFT's (0.371 vs. 0.034), on weekday the two are within 0.01 bits (0.228 vs. 0.224), and on alphabet the non-compliant base improves at SFT. On the 7B chain the distortion emerges at SFT (0.014→0.283 bits on uniform10); on the 1B chain conditional distortion is already measurable in the base (0.463 bits) and SFT does not increase it (0.346), so the distortion is detectable

Table 6: **OLMo-2 stage-wise audit.** Conditional KL (bits) and branch coverage for the uniform10 probe across all released stages, with the full six-probe grid in the appendix. Base cells are candidate-conditional at low coverage; unconditional deficits appear in Table 16; the biased die is scored by entropy deficit (Section 2.4).

Stage	KL	c_B	Note
OLMo-2-1124-7B			
Base	.014	.02	compliant choice near-uniform
SFT	.283	.80	
DPO	.569	.91	
OLMo-2-0425-1B			
Base	.463	.02	
SFT	.346	.63	distortion present at SFT
DPO	.403	.39	
RLVR1	.493	.56	

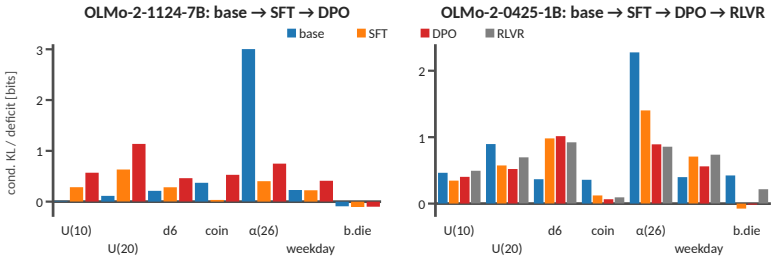


Figure 3: Conditional-KL stage trends across the OLMo-2 pipelines, with the measured base battery (raw prompting, same sequence-level estimator). Left: 7B base→SFT→DPO (SFT→DPO worsens on 6/6 KL-scored families; the full measured chain on 3/6, with coin and weekday exceptions and alphabet the single *improvement* at SFT (the corrected measurement); the biased die is scored by entropy deficit). Right: 1B chain base→SFT→DPO→RLVR1 (the base exceeds SFT on Unif(1–10), so the chain is monotone-worsening only post-SFT there).

early in both chains but does not accumulate uniformly. From DPO to RLVR on the 1B chain, conditional branch KL increases on four of six KL-scored probes and decreases on die and alphabet. The 1B biased-die entropy deficit separately increases from -0.018 at DPO to $+0.216$ at RLVR1, indicating growing concentration beyond what mass shifting alone explains.

The stage evidence establishes early presence and deepening on selected probes, not a universal law of post-training.

4.4 The residual distortion has interpretable structure

The residual distortion decomposes into stated-probability, label-identity, and display-position priors. Here *residual* means the conditional-choice distortion that remains once compliance is near-perfect, probed with a four-candidate non-uniform design: a balanced factorial experiment on three instruct models crossing all 24 display orders of four candidate labels with all 24 assignments of stated probabilities $(40, 30, 20, 10)$, plus 24 flat controls. Each of the 576 units per model asks for a draw from a stated non-uniform target, letting us separate three forces: following the stated probabilities, preferring particular label identities, and preferring particular display positions.

Table 7 reports the marginal structure. All three models place excess mass on the stated argmax label, but the gradient differs sharply. Llama-3.1 tracks stated probabilities most strongly (probability–mass correlation 0.660). Qwen2.5 follows the stated argmax in 68.8% of units yet is fully deterministic (top-1 mass > 0.99) in 69.1% of units, so its stated-probability following is largely a relabeling of a deterministic policy. Qwen3 has the weakest probability tracking (correlation 0.325) and a pronounced label-identity gradient, preferring labels "C" and "D" regardless of the probability attached to them. Display-position priors are comparatively minor except in Qwen2.5, which selects the first displayed label in 51.7% of units.

The structure is predictive but local. A candidate-level specification using stated probability, label identity, display position, and their interactions predicts held-out combinations with grouped leave-one-out R^2 up to 0.78 for Llama-3.1 (Table 8), so within the tested template the model's deviation is systematic rather than noise. The same specification fails to transfer: fitted on one model, template, or label set it degrades on the others. The residual distortion is therefore real, structured, and model-

Table 7: **Factorial dissection of stated-distribution following.** $m(p)$: mean mass on the label assigned stated probability p (uniform baseline 0.25). Argmax/First: fraction of units whose top-1 label is the stated argmax / the first displayed label. Mean KL: mean conditional KL against the stated profile. Per-unit records and variance decompositions in Appendix J.

Model	$m(40)$	$m(30)$	$m(20)$	$m(10)$	Argmax	First	Mean KL
Llama-3.1-8B	.587	.251	.104	.059	.668	.104	.63
Qwen2.5-7B	.685	.215	.080	.020	.688	.517	1.35
Qwen3-8B	.476	.248	.164	.114	.474	.212	1.43

Table 8: **Held-out prediction of the factorial structure.** Grouped leave-one-out R^2 and mean absolute error for the full interaction specification (stated probability, label identity, displayed position, and their interactions) on the 576 units per model; *order* holds out each display order, *assignment* each stated-probability assignment. All five specifications and both holdouts are recorded in Appendix J.

Model	R^2 (order)	MAE (order)	R^2 (assign.)	MAE (assign.)
Llama-3.1-8B	.777	.109	.760	.114
Qwen2.5-7B	.510	.220	.495	.226
Qwen3-8B	.502	.211	.525	.204

specific, which limits simple prompt-side corrections. No analogous decomposition was run on base models, so we attribute the structure to post-training by consistency rather than by isolation.

4.5 The collapse is not undone by readout-level interventions

Flattening the readout does not restore the sampling policy at the tested configurations. If the collapse were a superficial readout phenomenon, flattening the candidate-normalized distribution should restore sampling. Activation steering on Llama-3.1 tests this directly: at one setting it flattens the uniform10 readout to near-zero KL (0.0004 bits), so by the readout the model looks repaired. Behavior says otherwise. The flattening is partly illusory: 25% of probability mass remains outside the candidate set, and arithmetic accuracy drops by 43 points (the strength–response sweep). The policy underneath is unchanged: in free-running validation the steered model emits a strict-valid candidate in only 59.8% of generations at moderate strength and 3.3% at the entropy-restoring strength, with "7" still modal among valid generations (54.4%). Nor is the failure family-specific: on Qwen3 the strongest tested steering still emits "7" in 98.7% of free-running generations at coverage 1.000, and base-state patching, a different intervention class, eliminates strict-valid candidate emission altogether. The collapse is therefore a property of the sampling policy, not of any particular readout; mitigation claims must be validated in free-running generation, not on the readout.

4.6 Scope and generality

The collapse is range-anchored, transfers to Chinese on the tested models, and is consistent with broader behavioral narrowing. The attractor moves with the candidate range rather than being a fixed token: a 90-way probe over 10–99 places 89.9% on "42" for Qwen3-8B and 70% on "57" for Qwen2.5, with conditional KL of 5.73 and 5.05 bits respectively. Chinese-language probes on the tested models also show concentration. R1-Distill-Qwen-1.5B produces free-form reasoning instead of a candidate answer (near-zero strict coverage), so the audit identifies it as non-compliant; its conditional readout is reasoning-conditional. In open-ended continuation tasks without an explicit distribution request, lexical diversity (Distinct-2) falls in 10 of 12 base→post comparisons across six lineages (OLMo-1B is mixed), consistent with broader behavioral narrowing, though we do not claim the two measurements isolate the same mechanism.

5 Related work

Distribution following and sampling from language models. West and Potts (2025) formalize a model's output distribution over prompts and derive impossibility results for natural desiderata; Sorensen et al. (2025) show that post-training induces mode collapse toward narrow behaviors. Cao and Rimell (2021) study marginal probability estimation for completions; Pohl et al. (2025) propose decoding-free distributional estimation for model-internal rather than user-specified targets; Zhang et al. (2025) counter mode collapse by verbalizing response distributions. Our audit differs in pinning

the target distribution in the prompt and scoring completed answers exactly over the enumerated inventory.

Behavioral diversity and mode collapse. A substantial literature documents output-diversity loss after preference tuning (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Cui et al., 2025; Kirk et al., 2024; O'Mahony et al., 2024; Padmakumar and He, 2024; Parikh, 2026; Jiang et al., 2025). Lexical and embedding measures such as the Vendi score quantify narrowing at the corpus level (Padmakumar and He, 2024; Friedman and Dieng, 2023; Anderson et al., 2024); semantic entropy targets meaning-level variability (Farquhar et al., 2024). DART adds the complementary axis: whether the model follows a user-stated distribution, not just how varied its outputs are.

Decoding and controlled generation. Decoding methods shape the output distribution (temperature, top-p, min-p, typical sampling) (Holtzman et al., 2020; Fan et al., 2018; Welleck et al., 2020; Meister et al., 2023; Nguyen et al., 2024; Su et al., 2022). These methods take the model's own distribution as the target and truncate or sharpen it; our results show that a target law stated in the prompt is dominated by the model's learned attractor instead.

Interpretability of numerical and distributional behavior. Probing and representation analyses show that models encode numerical and probabilistic information in hidden states (Nostalgebraist, 2020; Belrose et al., 2023), and steering or representation editing can shift behaviors (Li et al., 2023; Turner et al., 2023; Marks and Tegmark, 2024; Arditi et al., 2024; Zou et al., 2023). Our interventions add a cautionary result: readout-level flattening does not imply a repaired sampling policy—free-running validation distinguishes the two.

6 Discussion

Post-training trades conditional fidelity for unconditional compliance: a base model asked for a uniform 1–10 draw mostly does not answer but is closer to uniform when it does; a post-trained model almost always answers but concentrates on an attractor. For downstream uses the failure is silent: the output looks like an answer.

The anchor's range-dependence points to a learned policy about plausible answers rather than a degenerate token prior; the factorial decomposition shows it is systematic yet model-specific, one reason the tested surface-level fixes do not generalize.

Our results do not imply distribution-following is impossible, only that current post-trained models do not do it by default and the tested fixes only partially restore it: few-shot demonstrations halve Llama-3.1's uniform10 KL (2.30→1.12 bits) without restoring uniformity (Table 11), system-prompt nudges are family-specific, and readout-flattening fails free-running validation. Whether targeted training can recover conditional fidelity without sacrificing compliance is open.

7 Conclusion

We introduced DART, an audit that measures whether language models follow explicitly requested distributions over completed answers. Across 13 base→post pairs, post-training yields answers that look compliant while the conditional distribution collapses onto range-dependent attractors with label and position structure; the distortion is already present at SFT and deepens on selected probes, and the factorial decomposition shows the residual structure is systematic yet model-specific. Distribution-following is thus a distinct compliance axis that current post-training does not provide by default; the released audit makes it measurable.

Limitations

DART measures completed-answer behavior on enumerated candidate sets; coverage reporting and free-running validation expose unenumerated realizations. Stage attribution rests on the OLMo-2 pipelines and the factorial dissection on three instruct models under one template family, establishing severity and structure rather than provenance or a general theory.

ETHICS STATEMENT

This work measures and documents a behavioral property of publicly available language models. It involves no human subjects, no private data, and no new model deployment. A potential dual-use concern is that the audit could inform attempts to make model outputs appear more random than they are; we judge this risk low because the audit measures rather than enables distributional control, and because transparent measurement of a silent failure mode benefits downstream users more than it benefits potential adversaries.

REPRODUCIBILITY STATEMENT

All models evaluated are publicly available. Probe templates, candidate inventories, scoring configurations, and per-cell records are released with the paper. The deterministic readouts require only single forward passes per candidate path and are exactly reproducible; free-running validation cells use fixed seeds with raw token IDs retained. Appendix D documents the full estimator lineage, including superseded readouts and the corrections that produced the values reported here.

AI USE DISCLOSURE

The authors used AI assistants for literature discovery, drafting assistance, and code debugging during this project. All experimental design decisions, measurements, analyses, and final text were reviewed and approved by the authors, who take full responsibility for the paper's contents.

REFERENCES

M. Kubovy and J. Psotka. The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance*, 2(2):291–294, 1976.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The Curious Case of Neural Text Degeneration. *ICLR*, 2020.

R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, and R. Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. *ICLR*, 2024.

V. Padmakumar and H. He. Does Writing with Language Models Reduce Content Diversity? *ICLR*, 2024.

Team OLMo. 2 OLMo 2 Furious. *arXiv:2501.00656*, 2025.

A. Grattafiori et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

R. Rafailov, A. Sharma, E. Mitchell, C. Manning, S. Ermon, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS*, 2023.

N. Lambert et al. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv:2411.15124*, 2024.

I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *NeurIPS*, 2023.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering Language Models With Activation Engineering. *arXiv:2308.10248*, 2023.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*, 2025.

T. Parikh. The One-Word Census: Answer-Choice Conformity Across 44 Language Models. *arXiv:2607.12796*, 2026.

L. O'Mahony, L. Grinsztajn, H. Schoelkopf, and S. Biderman. Attributing Mode Collapse in the Fine-Tuning of Large Language Models. *ME-FoMo Workshop, ICLR*, 2024.

X. Wang, J. Wei, D. Schuurmans, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*, 2023.

L. P. Argyle, E. C. Busby, N. Fulda, et al. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31(3), 2023.

S. Santurkar, E. Durmus, F. Ladhak, et al. Whose Opinions Do Language Models Reflect? *ICML*, 2023.

C. Meister, T. Pimentel, G. Wiher, and R. Cotterell. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 2023.

M. Nguyen, A. Baker, C. Neo, et al. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs. *arXiv:2407.01082*, 2024.

D. Friedman and A. B. Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*, 2023.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* 630, 2024.

N. Belrose, I. Ostrovsky, L. McKinney, et al. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv:2303.08112*, 2023.

B. R. Anderson, J. H. Shah, and M. Kreminski. Homogenization Effects of Large Language Models on Human Creative Ideation. *Creativity & Cognition*, 2024.

P. F. Christiano, J. Leike, T. B. Brown, et al. Deep Reinforcement Learning from Human Preferences. *NeurIPS*, 2017.

L. Ouyang, J. Wu, X. Jiang, et al. Training Language Models to Follow Instructions with Human Feedback. *NeurIPS*, 2022.

Y. Bai, A. Jones, K. Ndousse, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862*, 2022.

Z. Shao, P. Wang, Q. Zhu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300*, 2024.

A. Yang, B. Yang, B. Hui, et al. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

A. Yang, A. Li, B. Yang, et al. Qwen3 Technical Report. *arXiv:2505.09388*, 2025.

A. Q. Jiang, A. Sablayrolles, A. Mensch, et al. Mistral 7B. *arXiv:2310.06825*, 2023.

A. Fan, M. Lewis, and Y. Dauphin. Hierarchical Neural Story Generation. *ACL*, 2018.

S. Welleck, I. Kulikov, S. Roller, et al. Neural Text Generation with Unlikelihood Training. *ICLR*, 2020.

Y. Su, T. Lan, Y. Wang, et al. A Contrastive Framework for Neural Text Generation. *NeurIPS*, 2022.

S. Marks and M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *COLM*, 2024.

A. Arditi, O. Obeso, A. Syed, et al. Refusal in Language Models Is Mediated by a Single Direction. *NeurIPS*, 2024.

A. Zou, L. Phan, S. Chen, et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv:2310.01405*, 2023.

G. Cui, Y. Zhang, J. Chen, et al. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv:2505.22617*, 2025.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. In *Conference on Language Modeling*, 2025.

Kris Cao and Laura Rimell. You should evaluate your language model on marginal likelihood over tokenisations. In *Proceedings of EMNLP*, 2021.

Taylor Sorensen, Benjamin Newman, Jared Moore, et al. Spectrum Tuning: Post-Training for Distributional Coverage and In-Context Steerability. *arXiv:2510.06084*, 2025.

David Pohl, Marco Cognetta, Junyoung Lee, and Naoaki Okazaki. Decoding-Free Sampling Strategies for LLM Marginalization. *arXiv:2510.20208*, 2025.

J. Zhang, S. Yu, D. Chong, et al. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. *arXiv:2510.01171*, 2025.

L. Jiang, Y. Chai, M. Li, et al. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Advances in Neural Information Processing Systems*, 2025.

P. Taillandier, J. D. Zucker, A. Grignard, et al. From the Fluency Fallacy to the Micro-to-Macro Validity Gap: Opportunities and Pitfalls of LLMs in Social Simulation. *arXiv:2507.19364*, 2025.

J. Ye, L. Cao, D. Chen, and E. Ferrara. Stop Drawing Scientific Claims from LLM Social Simulations Without Robustness Audits. *arXiv:2605.18890*, 2026.

Roadmap. Appendix A collects additional exhibits: the full ground-truth probe grid behind the headline uniform10 table (Table 9), the complete OLMo-2 stage table including the biased-die entropy-deficit row (Table 10), the mechanism summary, the KL panorama (Figure 4), few-shot (Figure 5), transfer (Figure 6), temperature (Figure 7), and coin (Figure 8) figures, protocol and validity tables, and the claims table (Table 14). Appendices B–E cover estimation infrastructure: sampling-noise floors, the unconditional deficit decomposition, experiment lineage, and the authoritative readings. Appendix F reports the free-running validation, the two calibration exceptions and their forward-path attribution, the decoder-complete enumeration bound, and the tokenizer boundary audit. Appendices G–J give the open-ended diversity batch detail, estimator boundary narratives, the measurement corrections, and the factorial disentangling with held-out transfer tests. The closing section records reproducibility settings. Throughout the appendix, shortened model names decode as llama31=Llama-3.1-8B, mistral7=Mistral-7B-v0.3, qwen25_7=Qwen2.5-7B, qwen3_8/qwen3_4/qwen3_1_7/qwen3_0_6=Qwen3-8B/4B/1.7B/0.6B, olmo2/olmo1b=OLMo-2-1124-7B/0425-1B, and r1distill=R1-Distill-Qwen-1.5B.

A Additional Exhibits

This appendix gathers the full-size exhibits behind the main-text summaries, in three groups. Tables 9–11 expand the headline results (the complete probe grid, the full OLMo-2 stage battery including the biased-die entropy-deficit row, and the readout-diagnostic summary); Figures 4–8 visualize the battery and the mechanism experiments (the KL panorama, few-shot, transfer, temperature, and coin readouts); Tables 12–14 give protocol detail, validity checks, and the claim-by-claim scope summary.

B Sampling Noise Floors and the 1024-Sample Revalidation

The multinomial noise floor for the free-running validation depends on the *valid*-sample count, not the nominal count. Monte Carlo against the measured non-uniform sequence distributions gives, for K=10 at n=384 valid samples, median JS ≈0.004 (95th percentile 0.008). At the lower valid counts of the two additional checkpoints (OLMo-1B-SFT, Qwen3-0.6B) the floors rise to median 0.0065 (p95 0.014) at n=238 and median 0.012 (p95 0.025) at n=139; for the K=90 probe the floor is median 0.023, p95 0.031. All floors are *conditional* on the measured sequence distribution, so sampling-JS cells are read against their own floors. OLMo-1B-SFT's nominal-count JS of 0.010

Table 9: **Ground-truth probe audit** (KL in bits against the analytically known target; branch-sum sequence-level readout, Eq. (1)). **post**: post-trained checkpoint, chat template (cov ≥ 0.99 except Mistral 0.988); **base**: raw prompting, conditional on a candidate emission (cov 0.004–0.34); parentheses in the base `uniform10` cell: unconditional bookkeeping deficit (Appendix C, Table 15). $\kappa = \mathrm{KL}/\log_2 K$ and modal token on `uniform10`; verdict bands: collapse $\kappa \geq 0.6$, moderate $0.3 \leq \kappa < 0.6$. Alphabet, Qwen `uniform20`, and Mistral post cells use the corrected recompute (Appendix I; alphabet bases largely non-compliant, in-set mass dominated by "I", conditional only); ^{cal}free-running calibration exception; ^{a}cov 0.37 (deterministic on "1"); ^{u}band-boundary verdict (paraphrase interval crosses below the moderate band). Protocol detail and paraphrase spreads: Table 12 (Appendix A). Cell shading on post rows encodes KL magnitude (green <0.3; orange 0.3–1; deepening red ≥ 1); κ/verdict shaded by band. The main battery uses the branch-sum estimand; the terminator-complete strict-event recomputation of the 18 validation cells is reported in Table 19 (Appendix F).

Lineage (chat)	U(1–10)	U(1–20)	d6	coin	A–Z	wkday	biased d6	κ	modal	verdict
Llama-3.1-8B	**2.297**	3.460	0.835	0.126	1.008	0.397	0.682	0.69	7 (75%)	collapse
base (raw)	0.63 (3.25)	1.01	1.02	0.47	3.81	0.61	0.21			
Mistral-7B-v0.3	**1.325**	0.838	0.115	0.197	0.493	0.207	1.000^{a}	0.40	7 (48%)	moderate^{u}
base (raw)	1.15 (3.28)	1.52	1.01	0.33	4.08	0.18	0.22			
Qwen2.5-7B	**3.321**	2.298	1.602	1.000	2.333	0.800	1.000	1.00	7 ($\geq$99.8%)	collapse
base (raw)	0.37 (2.45)	0.91	1.15	0.60	2.99	0.73	0.11			
Qwen3-8B	**3.303**	2.365^{cal}	2.050	0.942	2.689	1.427	0.357	0.99	7 (99.8%)	collapse
base (raw)	0.64 (2.42)	1.18	0.53	0.04	2.36	0.65	0.29			
Qwen3-4B	**3.321**	3.273	2.572	1.000	3.029	0.527	2.462	1.00	7 ($\geq$99.99%)	collapse
base (raw)	0.07 (2.52)	0.69	0.06	0.00	2.48	0.61	0.12			
Qwen3-1.7B	**1.517**	3.539	0.245	1.000	4.571	1.473	0.516	0.46	5 (38%)	moderate
base (raw)	0.16 (2.85)	0.93	0.09	0.62	1.65	0.34	0.13			

Table 10: **OLMo-2 stage trends**, each cell *value/cov*: 1124-7B base→SFT→DPO (left) and 0425-1B →RLVR1 (right), the only open 4-stage pipeline. All rows except biased d6 are conditional KL [bits]; biased d6 is the entropy deficit $\mathbb{H}(q) - \mathbb{H}(\hat{p})$ (negative = model entropy above target). Stage cells chat protocol; base cells raw. †corrected values (Appendix I). Unconditional bookkeeping deficit: Table 16 (Appendix C). Monotone-worsening chains: `uniform10`, `uniform20`, and dice on 1124-7B; none on 0425-1B. Cell shading encodes KL magnitude as in Table 9 (biased d6 row unshaded: entropy-deficit scale).

	OLMo-2-1124-7B			OLMo-2-0425-1B			
Probe	base	SFT	DPO	base	SFT	DPO	RLVR
Unif(1–10)	0.014/.02	0.283/.80	0.569/.91	0.463/.02	0.346/.63	0.403/.39	0.493/.56
Unif(1–20)	0.114/.02	0.631/.89	1.135/.96	0.896/.02	0.574/.58	0.520/.42	0.695/.61
d6	0.213/.01	0.281/.52	0.460/.57	0.367/.03	0.981/.09	1.016/.16	0.922/.37
coin	0.371/.02	0.034/.93	0.528/.98	0.359/.01	0.122/.61	0.066/.85	0.094/.92
Unif(A–Z)†	3.002/.08	0.403/.99	0.748/1.00	2.277/.06	1.402/.70	0.892/.39	0.855/.32
weekday	0.228/.01	0.224/.98	0.410/1.00	0.399/.01	0.707/.98	0.561/.97	0.736/.98
biased d6	−0.092/.00	−0.106/.67	−0.101/.64	+0.423/.02	−0.077/.68	−0.018/.38	+0.216/.57

lies within its n=238 floor, while Qwen3-0.6B's 0.027 sits at the ∼97th percentile of its n=139 floor. Both checkpoints were therefore revalidated at n=1024: the JS values tighten to 0.0049 for Qwen3-0.6B (392 valid samples) and 0.0029 for OLMo-1B-SFT (668 valid), both well inside their conditional floors. The apparent tension was small-sample fluctuation, not estimator drift.

C Unconditional Deficit Decomposition for Base Models

For uniform targets, write $\tilde{q}$ for the model's candidate distribution augmented with an out-of-set class of mass $1 - \mathrm{cov}$, π for the conditional readout on the candidate set, and U_K for the uniform law on K candidates. The augmented entropy identity below is descriptive bookkeeping, not a quality score:

$$D_{\mathrm{full}} \;=\; \log_2 K - H(\tilde{q}) \;=\; \underbrace{(1-\mathrm{cov})\log_2 K - \mathbb{H}_2(\mathrm{cov})}_{\text{compliance}} \;+\; \underbrace{\mathrm{cov}\cdot \mathrm{KL}(\pi \| U_K)}_{\text{fidelity}} . \qquad (6)$$

This identity can be negative because it is not a divergence; adding out-of-set mass can lower it. Table 15 tabulates Eq. (6) for the seven non-OLMo bases × five uniform-target probes; the alphabet probe is omitted (its base cells are reported under the corrected measurement, Appendix I), and the OLMo-2 base cells are decomposed per stage in Table 16. Two readings follow. First, the *compliance* term dominates everywhere: bases leave 66–99.995% of mass outside the candidate set

Table 11: **Readout diagnostic summary.** Lens rows record the layer at which candidate-normalized entropy declines. Steering rows compare the tested base-minus-instruct direction with random controls (Lℓ_as = layer ℓ at strength α=s). The few-shot row reports 0-shot→5-shot KL. Patch rows report base-state replacement, with a comparable shuffled control. ^{p}pre-correction load; the Qwen3 steering row uses the corrected readout.

Model	Exp.	Result
llama31 / qwen25_7 / qwen3_8 / olmo2	lens	onset L20 / L23 / L18 / L23
olmo1b (RLVR)	lens	late-layer decline only (0.29 bits, no onset)
llama31	steer	L12_a10: anti $+1.00$ vs rand $+0.59$
qwen25_7	steer	L10_a20: anti $+0.00$ vs rand $+0.00$
qwen3_8	steer	L14_a20: anti $+0.09$ vs rand $+0.06^{p}$
olmo2	steer	L12_a10: anti $+1.00$ vs rand $+0.03$
olmo1b (RLVR)	steer	L6_a10: anti $+1.00$ vs rand $+0.45$
llama31 / qwen25_7 / qwen3_8	fewshot	0-shot KL 2.30/3.32/3.30 → 5-shot 1.12/3.09/2.19
llama31 / qwen25_7 / qwen3_8 / olmo2	patch	$\geq$90% entropy at 32/32, 28/28, 36/36, 32/32 layers
olmo1b (RLVR)	patch	$\geq$90% at 13/16 layers (99.9% from L2 on)

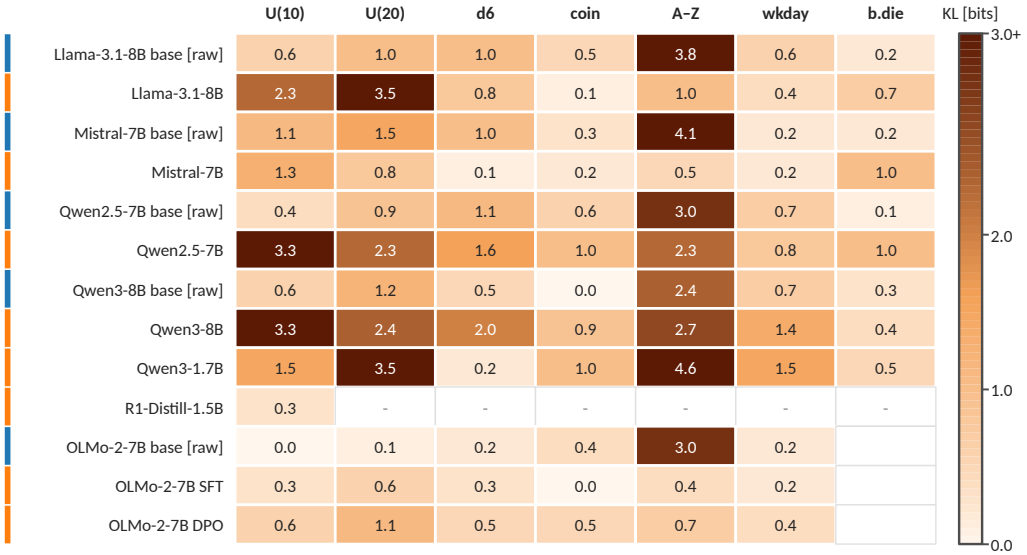


Figure 4: KL panorama across lineages and probe families. Each row is a checkpoint, base models under raw prompting and post-trained counterparts under chat templates; columns are the seven probe families. Thirteen of the audited checkpoints are shown: the endpoint pairs of the four chat families plus the Qwen3-1.7B endpoint, R1-Distill, and the OLMo-2-7B chain; Qwen3-4B, Qwen3-0.6B, the Qwen3-1.7B base, and the OLMo-2-1B stages are omitted for space and appear in Tables 9 and 10. Cell color encodes the finite-path sequence-level KL in bits against the analytically known target (light ≈ 0, saturated ≥ 3.3 bits). The R1-Distill row reports only `uniform10`: elsewhere its coverage is near zero and no conditional KL exists. OLMo-2 biased-die cells are blank here because that probe is scored as an entropy deficit on a different scale (Table 10). Numeric values for every printed cell are in Tables 9 and 12.

under bare prompts. Their unconditional deficits span 2.4–3.3 bits on the primary `uniform10` probe and reach 4.3 bits on `uniform20` (lower on the higher-coverage coin and dice cells, down to 0.4 bits), bounded by each target's $\log_2 K$ (dice $\approx$2.58 to uniform20 4.32). The deficit is dominated by non-answering rather than by maldistribution of answers. Second, the *fidelity* term is small wherever the base complies substantially (Qwen3-4B dice: coverage 0.34, fidelity 0.02 bits), and conditional KLs are only interpretable jointly with coverage. The quantity a pipeline-stage comparison wants is D_{full}, and the same forward passes already determine it.

Table 16 extends the same decomposition to every cell of the OLMo-2 stage battery (T2), so base-vs-stage comparisons can be read at a matched estimand instead of comparing conditional KLs across coverage gaps.

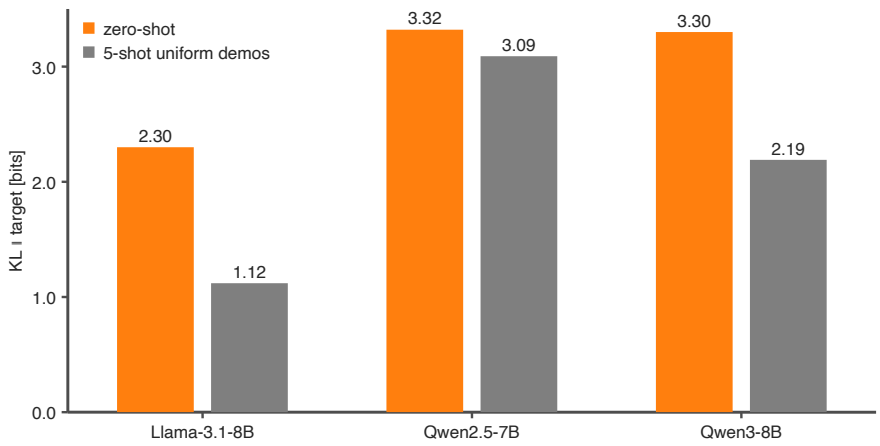


Figure 5: Five uniform demonstrations only partially restore entropy on the three tested instruct models (0-shot→5-shot conditional KL in Table 11); Mistral-7B and the base checkpoints were not run in this experiment.

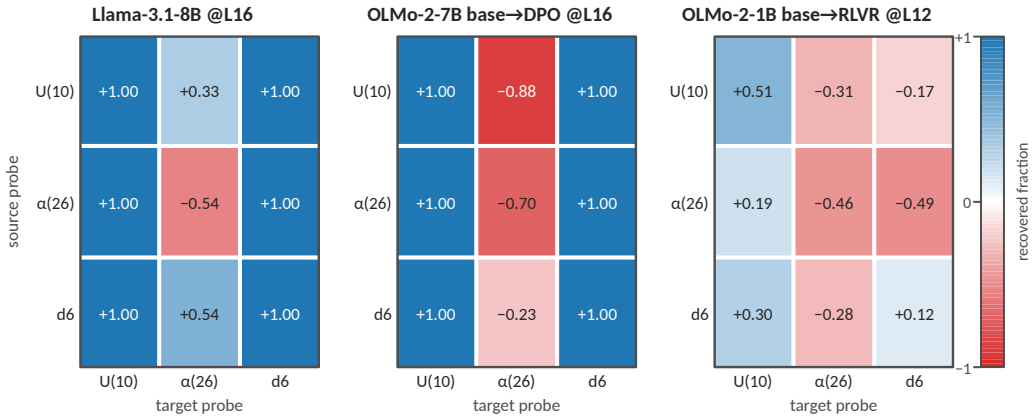


Figure 6: Transfer of anti-collapse directions between probes (α=10, last-token hidden state). Cell (i, j) is the fraction of the entropy gap to uniform recovered when the base-minus-chat direction fitted on probe i is injected into probe j (1=full recovery; negative =further concentration). Panels: Llama-3.1-8B at layer 16 (left), OLMo-2-7B base→DPO at layer 16 (center), OLMo-2-1B base→RLVR1 at layer 12 (right).

D Experiment Lineage and Raw-Record Recompute

Every number in this paper is script-generated from stored per-cell raw records, never hand-copied. Table 17 records, for each experiment family, the estimator/readout version it uses and the status of its revalidation against those raw records. Record-batch identifiers r3–r21 and b2/b4 name the released raw-record batches. Three estimator versions were used over the course of the study. v1 is first-token marginalization with leading-space pooling, retired except as the demonstrated artifact. v2 is the naive path-marginalized branch sum, without cross-candidate prefix correction. v3 adds prefix subtraction for shared-leading-token candidates (the branch readout of Eq. (1)) and the terminator-mass strict event; it is defined in the released answer-language specification and implemented with machine-checked invariants. The estimator test suite pins the uniform-target identity KL = deficit, the two decomposition identities $D_{\text{full}} - c \cdot \text{KL} = (1-c)\log_2 K - \mathbb{H}_2(c)$ and, subtracting the unweighted KL, $D_{\text{full}} - \text{KL} = (1-c)H(\pi) - \mathbb{H}_2(c)$, and an exhaustive hand-computable path-enumeration model. Both decomposition identities are machine-checked, with $|D_{\text{full}} - c \cdot \text{KL}| \leq 0.081$ bits at c=0.99 (the ≈0.08-bit bound of interpretation note (i), Section 2.4). The mechanism experiments (lens, steering, patching, transfer, coverage, calibration, and strength–response readouts) read the next-token distribution at the final position over the enumerated candidate variants; they are variant-enumeration-sensitive but not path-marginalization-sensitive. We recomputed every post-steering cell (14 pre/steered distribution pairs) from its stored raw distribution under the unified conditional-KL

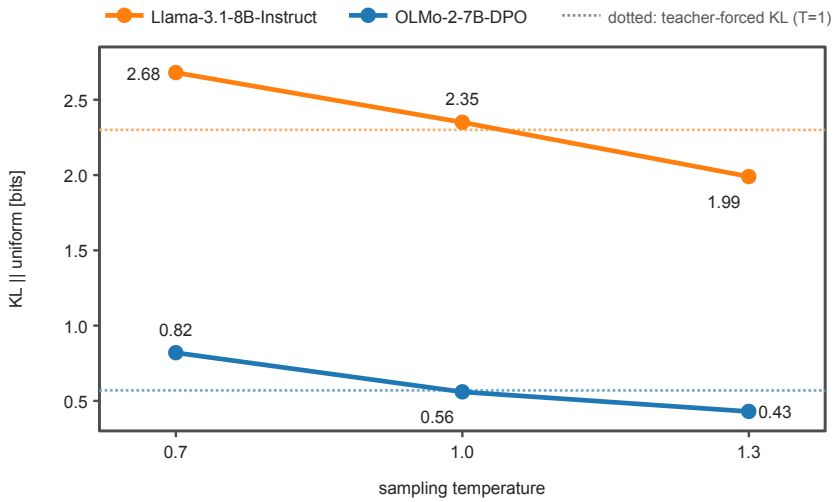


Figure 7: Empirical temperature sweep. The three tested temperatures change but do not eliminate Llama-3.1's deviation. The dotted line is the teacher-forced branch KL at temperature 1 and is not used as a closed-form reference for other temperatures.

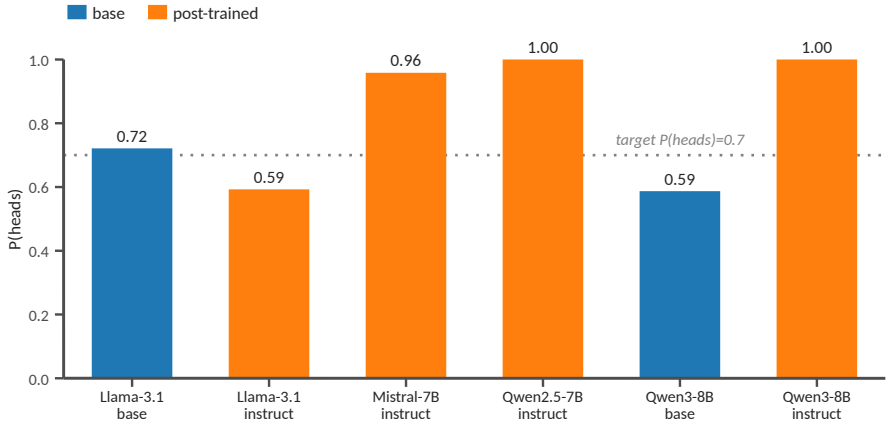


Figure 8: Post-trained models degrade tracking of stated biases, from roughly p-insensitive Heads priors in the base models (0.71 for Llama-3.1 at p=0.7, $\sim$0.75 across p in the biased-coin sweep; 0.58 for Qwen3-8B) to dilution to deterministic mode-locking (label swaps split this endpoint into label-anchoring vs. majority-locking vs. probability-sensitive-but-miscalibrated tracking).

metric: all cells match to $\leq 10^{-3}$ bits (the stored distributions are rounded to four decimals). The "exactly uniform" steered cells verify at the same $\leq 10^{-3}$ tolerance, with entropy $= \log_2 K$ to the same precision. The full raw pre/steered distributions for the calibration and strength–response experiments are included in the supplementary material.

E Authoritative Readings

This appendix states the current authoritative readings (Table 18): one row per headline quantity, the value the paper stands behind, and where it lives.

F Free-Running Validation and Tokenizer Boundary Audit

Design. The strict answer-event estimator of Section 2 evaluates the declared finite token-path inventory; it is the final v3 estimator in the lineage of Appendix D (variant enumeration and terminator mass; cross-candidate prefix subtraction applies to the companion branch readout). The *empirical* side of every sampling-validation closure rests on the surface-form-only parser plus an end-of-turn requirement: the generation's final token must be an EOS/turn-terminator, and a truncated

Table 12: **Sequence-level `uniform10` readout (primary).** KL [bits] from the path-marginalized sequence-level estimator; $\kappa = \mathrm{KL}/\log_2 10$; cov = probability mass on the candidate set (raw-protocol KL is conditional on a candidate emission). *Spread* = max−min KL across the three prompt paraphrases (p1 reproduces the seq(chat) column to within 0.001 bits, a built-in consistency check). JS = divergence between 384-sample free-running draws (fewer for the two added checkpoints) and the sequence-level estimate ([†]raw-protocol JS; per-cell noise floors in Appendix B); qwen3_0.6B's 0.027 is an n=139 value revalidated at 0.0049 in the 1024-sample rerun. Mistral's first-token uniformity is a shared-prefix tokenizer artifact unmasked here (1.32 bits); qwen25_7 retains collapse under raw prompts (cov 0.975; strict terminator-complete event 0.860 vs. strict-parser valid rate 0.861, Table 19), while the remaining families leave the candidate set (cov ≤0.15). Raw-protocol κ is candidate-conditional at the stated cov and not comparable to cov≈1 chat cells; the r1distill, qwen3_0.6B (chat cov 0.35), and olmo1b-sft chat cells are excluded from cross-model κ rankings for the same reason. [*]corrected re-run; immediate-answer coverage ≈0. [‡]low-coverage paraphrases conflate fidelity with answer-rate shifts (Appendix F). OLMo-2 stage cells: Table 10.

Model	seq (chat)	κ	seq (raw)	κ	cov (raw)	base (raw)	top-1 (chat)	spread	JS
llama31-8B	2.297	0.69	0.547	0.16	0.013	0.626	7 (75%)	0.27	0.003[†]
mistral7	1.325	0.40	0.279	0.08	0.029	1.150	7 (48%)	0.82	0.001
qwen25_7	3.321	1.00	**3.317**	**1.00**	**0.975**	0.373	7 (≥99.8%)	0.01	0.000
qwen3_8	3.303	0.99	0.630	0.19	0.054	0.644	7 (99.8%)	0.02	0.001
qwen3_4	3.321	1.00	0.366	0.11	0.142	0.069	7 (≥99.99%)	0.93	0.011
qwen3_1.7B	1.517	0.46	0.521	0.16	0.063	0.163	5 (38%)	2.38[‡]	0.004
qwen3_0.6B	1.284	0.39	1.025	0.31	0.006	0.714	1 (44%)	0.41[‡]	0.027
r1distill	0.326[*]	0.10	—	—	<0.01	1.104	5 (20%)	—	0.003
olmo2-7B sft	—	—	0.464	0.14	0.013	—	—	0.48	—
olmo2-7B dpo	—	—	0.450	0.14	0.053	0.014	—	0.42	—
olmo1b sft	0.346	0.10	0.517	0.16	<0.01	0.463	7 (21%)	0.14	0.010
olmo1b dpo	—	—	0.640	0.19	<0.01	—	—	0.20	—
olmo1b rlvr	—	—	0.426	0.13	<0.01	—	—	0.10	—

Table 13: Free-running calibration of the deterministic branch and strict-event readouts. Comparisons use matched answer events and cell-specific sampling-noise bands. Sixteen of 18 validation cells pass; Qwen3 `uniform20` and Llama3.1 biased-dice remain explicit exceptions. The temperature sweep is interpreted empirically because token-level temperature does not generally induce the answer-level transformation $\pi^{1/\tau}/Z_\tau$ for multi-token paths. The paraphrase rows show the collapse verdicts are stable under prompt rewording; all ranges use the corrected sequence-level estimator. On the smallest models the low-coverage paraphrases conflate fidelity with answer-rate (qwen3_1.7B p3 at cov 0.001, qwen3_0.6B p2 at cov 0.001); see the spread column of Table 12.

Model	Check	Empirical vs. readout [bits]	JS
llama31/instruct	384-sample validation	emp KL 2.22 vs TF 2.30[†]	JS 0.002
mistral7/instruct	384-sample validation	emp KL 1.39 vs seq 1.32	JS 0.001
qwen25_7/instruct	384-sample validation	emp KL 3.32 vs TF 3.32	JS 0.000
qwen3_8/instruct	384-sample validation	emp KL 3.30 vs TF 3.30	JS 0.001
olmo2/sft	384-sample validation	emp KL 0.33 vs TF 0.28	JS 0.003
olmo2/dpo	384-sample validation	emp KL 0.58 vs TF 0.57	JS 0.005
olmo1b/rlvr	384-sample validation	emp KL 0.50 vs TF 0.49	JS 0.004
llama31/instruct	temperature 0.7	emp KL 2.68 vs TF 2.30	JS 0.018
llama31/instruct	temperature 1.0	emp KL 2.35 vs TF 2.30[†]	JS 0.001
llama31/instruct	temperature 1.3	emp KL 1.99 vs TF 2.30	JS 0.009
olmo2/dpo	temperature 0.7	emp KL 0.82 vs TF 0.57	JS 0.012
olmo2/dpo	temperature 1.0	emp KL 0.56 vs TF 0.57	JS 0.002
olmo2/dpo	temperature 1.3	emp KL 0.43 vs TF 0.57	JS 0.008
llama31/instruct	3 prompt paraphrases	KL range [2.02, 2.30]	spread 0.27
mistral7/instruct	3 prompt paraphrases	KL range [0.51, 1.32][§]	spread 0.82
qwen25_7/instruct	3 prompt paraphrases	KL range [3.31, 3.32]	spread 0.01
qwen3_8/instruct	3 prompt paraphrases	KL range [3.30, 3.32]	spread 0.02
qwen3_4/instruct	3 prompt paraphrases	KL range [2.39, 3.32]	spread 0.93
qwen3_1_7/instruct	3 prompt paraphrases	KL range [0.68, 3.06][‡]	spread 2.38
qwen3_0_6/instruct	3 prompt paraphrases	KL range [1.28, 1.69][‡]	spread 0.41
r1distill/instruct	3 prompt paraphrases	cov = 0 on all paraphrases[*]	—

[†]the 384-sample and temperature (T=1.0) checks are independent sampling runs of the same policy; both agree with the teacher-forced 2.30 at JS 0.002/0.001, below the conditional JS floor of interpretation note (iii), Section 2.4: their KL differences (−0.08/+0.05 bits) are sampling noise, not estimator discrepancy.
[*]corrected re-run: immediate-answer coverage ≈0, so no in-set KL is defined.
[§]mistral7 κ range 0.15–0.40: the 0.51-bit paraphrase falls *below* the moderate band.
[‡]low end of the range computed at cov 0.001 (residual mass); not a fidelity estimate.

continuation is out-of-set even if its decoded text matches a candidate. To test whether the parser and the estimator describe the same event on these cells, we ran 18 validation cells, each with N=2048 unfiltered free-running generations (temperature 1.0, top-p 1.0, max 32 new tokens) and raw token IDs plus decoded text stored for every sample. The cells cross four core models (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Qwen3-8B-Instruct, OLMo-2-0425-1B-RLVR1) × three probes (uniform10, uniform20, biased-dice) in chat mode, plus six high-risk cells (Mistral-7B-Instruct uniform20; OLMo-2-0425-1B-SFT uniform10; Qwen3-0.6B uniform10; both alphabet probes for Llama-3.1 and Qwen3; Qwen2.5-7B uniform10 in *raw* mode). For each cell we report the three coverage quantities *separately*: the branch sum (variant-summed, pre-terminator), the finite-inventory strict-event coverage, and the observed strict-parser valid rate. We also report their two distinct gaps (branch−strict and observed−strict), the empirical-vs-readout JS divergence, and a per-cell calibration verdict. JS is compared against a parametric bootstrap (4000 multinomial draws of size N_{valid} from the finite-inventory strict conditional; 99% band), with the χ^2 statistic $8N_{\text{valid}} \ln 2 \cdot \text{JS}$ against $\text{df} = |\text{support}|-1$ as a cross-check (Table 19; full per-cell bands in the supplementary material).

Table 14: **Final claims and their scope conditions.** One-glance summary; authoritative per-claim readings in Table 18. Evidence codes: T1 = Table 9, T2 = Table 10, T3 = the intervention battery (Table 11; Appendix F).

Claim (quantitative content)	Scope conditions	Evidence
Post-trained models violate requested sampling distributions: 4/6 comparable families collapse ($\kappa \geq 0.6$, up to 3.32 bits; $\geq 99.8\%$ on one token in three), Mistral moderate (κ=0.40).	Primary probe, chat protocol, cov$\geq$0.99; verdicts paraphrase-stable, small-model magnitudes not (Table 12).	T1
The failure is protocol-conditioned but family-dependent: bare prompting reverts five endpoint models (0.28–0.63 bits; qwen3_0.6B reverts to 1.025 bits at cov 0.006); qwen25_7 stays fully collapsed (374/374 = "7").	Bare vs. chat on `uniform10`; not universal.	T1, Table 12
Collapse is range-specific anchoring: attractor moves "7"→"12"/"14"→"42"/"57" as K: 10→20→90.	Enumerated numeric probes with known uniform targets.	T1, Table 18
Stage attribution: SFT→DPO worsens 6/6 KL-scored families (sign test p=0.031 across probes on a single pipeline, direction-consistency only); RLVR raises KL on 4/6 KL-scored probes and the biased-die entropy deficit, while the endpoint stays linearly reversible.	OLMo-2 chains only; one checkpoint per stage; no second RLVR pipeline.	T2
Soft/hard reversibility boundary: hard models (near-zero base entropy) resist steering and contextual rescue; soft models respond—at a capability cost (question-answering/arithmetic degradation).	Fixed family of last-position diff-in-means interventions at tested layers/strengths.	T3
Failure modes factorially separable: label-identity anchoring, position-modulated argmax-following, miscalibrated tracking.	Biased-coin swaps + 24×24 label×position×probability sweep.	Fig. 8; Table 23
Closed-set probe distortion and open-ended diversity require separate reporting: pair-level coupling is exploratory, survives Holm only on the rank readouts, and is null over six lineages; semantic metrics show no stable coupling.	Coupling on distinct-2, 12 pairs, retained-record batches only; semantic axis reported separately.	App. G; Table 18
The sequence-level path-marginalized estimator is the readout of record for enumerated-candidate probes; first-token and Monte-Carlo baselines fail in characterized regimes (16/18 validation cells within bootstrap bands, 2 flagged).	Candidates sharing leading tokens; tested immediate-termination regime.	Table 13; App. B

Findings. (i) For every cell the strict-parser valid rate matches the finite-inventory strict-event coverage to within the sampling noise floor ($|\Delta_o| \leq 0.0193$), so the branch-vs-strict coincidence defined analytically in Section 2 holds *empirically* on free-running generations, not only on forced token slices. (ii) The largest Δ_o occur exactly where coverage is lowest (OLMo-1B-RLVR1 ∼0.55–0.60, Qwen3-0.6B 0.34–0.36), consistent with multinomial noise on a support-truncated event rather than a systematic parser bias. (iii) One cell isolates the branch-vs-strict distinction directly: without the chat template, Qwen2.5-7B `uniform10` has branch coverage 0.977 but strict coverage 0.8604, with strict-parser valid rate 0.8613. The ∼11.7-point shortfall is the measured cost of requiring a terminator and excluding surface forms outside the strict event, and the earlier branch-sum estimator would have overstated in-set mass by exactly this amount on raw-template prompts ([†]flagged: raw protocol, outside the tested immediate-termination regime). (iv) **Two cells fail the JS calibration and are reported as exceptions, not validated.** *Qwen3 uniform20 (chat)*: JS = 0.0106 vs. a bootstrap 99% band of $[0.0001, 0.0014]$ ($\chi^2 = 120.5$, df = 18 over the strict-event support, bootstrap $p = 0.0003$). The deviation concentrates on two candidates ("12": 944 observed vs. 799 expected, "13": 298 vs. 484, $z \approx +5.1$ and -8.5), so it is a statistically detectable mismatch between the sampled generations and the strict-event readout for this cell, not small-sample noise. The deterministic readout *understates* qwen3's concentration on this probe (the sampled distribution is more peaked on "12"): substituting the sampled distribution for the deterministic readout would *raise* the tabulated KL (2.365 bits, κ=0.55, moderate band). The tabulated verdict is therefore the conservative of the two readings, no verdict changes under either, and no family-level verdict in this paper rests on the direction of this discrepancy; the cell is nevertheless flagged as a calibration exception wherever it appears. *Llama3.1 biased-dice (chat)*: marginal fail (JS = 0.0015, band

Table 15: Unconditional deficit decomposition for base models under raw prompting on the uniform-target probes (Eq. (6), uniform targets only): $D_{\text{full}} = \underbrace{(1-\text{cov})\log_2 K - \mathbb{H}_2(\text{cov})}_{\text{compliance}} + \underbrace{\text{cov}\cdot\text{KL}}_{\text{fidelity}}$ on the numeric probes. Bases carry their deficit almost entirely in the compliance term; their rare candidate emissions are near-uniform (small fidelity term). D_{full} is an exact bookkeeping identity (not a divergence; can in principle be negative); $\text{JS}(\tilde{q}, (U_K, 0))$ is the nonnegative companion summary. The conditional KLs of the two R1-Distill rows are computed on residual in-set branches at unrounded coverage 3.8×10^{-4} / 5×10^{-5} (printed in full, since a cell with exactly zero coverage reports no conditional KL anywhere in this paper).

Model	Probe	cov	KL (cond.)	compl.	fid.	D_{full}
Llama-3.1-8B	Unif(1–10)	0.006	0.626	3.249	0.004	3.252
Llama-3.1-8B	Unif(1–20)	0.006	1.013	4.241	0.006	4.247
Llama-3.1-8B	Dice	0.009	1.022	2.483	0.010	2.493
Llama-3.1-8B	Coin	0.005	0.466	0.948	0.002	0.951
Llama-3.1-8B	Weekday	0.013	0.605	2.670	0.008	2.678
Qwen2.5-7B	Unif(1–10)	0.119	0.373	2.401	0.044	2.446
Qwen2.5-7B	Unif(1–20)	0.072	0.910	3.637	0.066	3.703
Qwen2.5-7B	Dice	0.027	1.153	2.339	0.031	2.370
Qwen2.5-7B	Coin	0.022	0.604	0.825	0.013	0.838
Qwen2.5-7B	Weekday	0.018	0.728	2.630	0.013	2.643
Qwen3-8B	Unif(1–10)	0.129	0.644	2.336	0.083	2.420
Qwen3-8B	Unif(1–20)	0.173	1.177	2.908	0.204	3.112
Qwen3-8B	Dice	0.041	0.532	2.231	0.022	2.253
Qwen3-8B	Coin	0.063	0.036	0.600	0.002	0.602
Qwen3-8B	Weekday	0.132	0.645	1.872	0.085	1.957
Mistral-7B	Unif(1–10)	0.004	1.150	3.275	0.004	3.279
Mistral-7B	Unif(1–20)	0.004	1.523	4.267	0.006	4.273
Mistral-7B	Dice	0.006	1.010	2.520	0.006	2.526
Mistral-7B	Coin	0.006	0.330	0.940	0.002	0.943
Mistral-7B	Weekday	0.003	0.179	2.769	0.001	2.769
Qwen3-4B	Unif(1–10)	0.101	0.069	2.514	0.007	2.521
Qwen3-4B	Unif(1–20)	0.111	0.687	3.337	0.077	3.413
Qwen3-4B	Dice	0.342	0.064	0.775	0.022	0.797
Qwen3-4B	Coin	0.068	0.004	0.574	0.000	0.575
Qwen3-4B	Weekday	0.037	0.610	2.473	0.023	2.495
Qwen3-1.7B	Unif(1–10)	0.054	0.163	2.841	0.009	2.850
Qwen3-1.7B	Unif(1–20)	0.056	0.928	3.768	0.052	3.820
Qwen3-1.7B	Dice	0.053	0.088	2.146	0.005	2.151
Qwen3-1.7B	Coin	0.132	0.620	0.306	0.082	0.388
Qwen3-1.7B	Weekday	0.032	0.339	2.516	0.011	2.527
R1-Distill-1.5B	Unif(1–10)	0.004	1.104	3.276	0.004	3.280
R1-Distill-1.5B	Unif(1–20)	0.004	1.451	4.273	0.005	4.278
R1-Distill-1.5B	Dice	0.002	0.337	2.561	0.001	2.562
R1-Distill-1.5B	Coin	3.8×10^{-4}	0.420	0.995	0.000	0.995
R1-Distill-1.5B	Weekday	5×10^{-5}	0.458	2.806	0.000	2.806

Table 16: Unconditional deficits D_{full} (bits) for every cell of Table 10, computed with the same bookkeeping identity as Table 15 (uniform targets only; the biased die is excluded). Base cells are compliance-dominated: the D_{full} column restores the deficit their small conditional KLs understate (e.g. 7B base Unif(1–10): 0.014 conditional vs. 3.13 unconditional). Read each column jointly with the coverage column of Table 10; cell-level reading notes are in the table footnotes.

	OLMo-2-1124-7B			OLMo-2-0425-1B			
Probe	base	SFT	DPO	base	SFT	DPO	RLVR
Unif(1–10)	3.13	0.17	0.38	3.11	0.49	1.23	0.74
Unif(1–20)	4.08	0.54	1.02	4.08	1.17	1.74	1.14
d6	2.52	0.39	0.39	2.35	1.98	1.70	1.02
coin	0.88	−0.25	0.42	0.89	−0.50	−0.41	−0.24
Unif(A–Z)†	4.19	0.36	0.74	4.21	1.50	2.26	2.58
weekday	2.66	0.14	0.40	2.67	0.60	0.43	0.63

D_{full} is an exact bookkeeping identity, not a divergence, and can be negative for a fully compliant near-uniform stage ($\log_2(K+1) > \log_2 K$ in the $K+1$-class augmented distribution): the negative cells (7B/1B coin SFT–RLVR) are exactly those near-perfectly-compliant near-uniform stages, a bookkeeping sign artifact, not models "more uniform than uniform". At coverage ≥ 0.92 conditional and unconditional readouts differ by at most $\approx$0.33 bits; the low-coverage 1B alphabet and dice cells understate the unconditional deficit by up to $\approx$1.7 bits.
†alphabet uses the corrected measurement (case-folded A–Z branches, uniform(26) target).

$[0.0000, 0.0015]$, $\chi^2 = 17.0$, df $= 5$, bootstrap $p = 0.0037$). The remaining sixteen cells sit inside their bootstrap bands (per-cell bands and χ^2 statistics in the supplementary material). Small JS alone is not evidence of agreement; each cell must clear its own noise floor.

Mechanism of the two exceptions (forward-path attribution). Re-running both exception cells in a single process with three matched readouts isolates the source. The strict-event estimator is exactly

Table 17: Experiment lineage: readout version per experiment family and raw-record recompute status. v1/v2/v3 are the estimator versions defined in Appendix D; "HS" marks hidden-state readouts (single forward pass at the final position, variant-enumeration-sensitive only).

Experiment family	Readout	Raw-record recompute
Probe battery	seq v2 (v3 strict on the 18 validation cells)	chat cells from the r5b/r5c sequence battery; base cells recomputed in the unified r8a/r8e battery; OLMo stage cells recomputed (App. I); paraphrase cells re-measured with the corrected estimator
Lens / steer / patch / transfer	HS (v2 variants)	not path-estimator-sensitive
Few-shot rescue	seq v2	distributions recorded; reproduced
Coverage readout	candidate mass (v2/v3)	distributions recorded; reproduced
Post-steering distributions	cond. KL (v2 variants)	14/14 cells recomputed $\leq 10^{-3}$
Raw protocol	seq v2 + cond. reporting	distributions + JS recorded
Alphabet (post-correction readout)	sequence-level, case-folded branches	distributions recorded; parser pinned by tests
Label sweeps	seq v3 final	recomputed from raw distributions (measured rerun)
Disentangling sweep	seq v3 finite-path	3×576 units; label/pos. R^2 recomputed
Steering calibration / strength	cond. KL + abs. mass	revalidated; raw distributions re-dumped in rerun
Sampling validations	seq v2/v3 vs. empirical	JS + valid counts recorded
Strict validation (18 cells)	strict v3 finite-path terminal enumeration	r19 re-derivation from stored prompts: 18/18 cells reproduced in coverage (Table 20)
Semantic readouts	pooled cosine + Vendi (pinned embedder)	r19 same-text recomputation: b2/b4 reproduce to within rounding; batch-3 archived signs confirmed (App. I)
Interventions	harness readouts	r19 free-running replication: baselines replicated; steered/patched conditions as summarized (App. F)

reproducible across environments (cross-environment JS $\leq 4.1\times10^{-11}$ bits on both cells), ruling out estimator or inventory drift. The mismatch arises instead from the bf16 forward path: generation scores tokens with incremental KV-cache updates in batches of 32, while the deterministic readout scores every candidate path with a full-prefix teacher-forced forward. Measured in-process on Qwen3 `uniform20`, the full-prefix versus incremental second-step distributions differ by JS $= 0.0072$ bits against an archived estimator-vs-sampling gap of 0.0086, while the incremental forward and the sampled frequencies agree to JS $= 0.0004$. These measurements establish sensitivity to numerical execution, not an additive decomposition of the archived gap. Batch size alone also shifts the distribution (batch-1 vs. batch-32), and the deviation direction matches the archive ("12" up, "13" down): systematic numerics, not noise. For Llama3.1 biased-dice the same replay lands inside its bootstrap band (percentile 0.9635), so the archived marginal FAIL was a boundary event. The two cells remain flagged wherever they appear; one exception is mechanistically *forward-path numerics* (bf16 incremental batched decoding versus teacher-forced scoring), the other is a bootstrap boundary event, and neither is an estimator or inventory error. An exact-agreement variant would score candidates through the same incremental path as generation, or evaluate in fp32.

Decoder-complete enumeration bound. The released path inventory is built from specified candidate forms; completeness over every token sequence decoding to the same surface is not proven by construction. We therefore enumerate it. For each validation-cell tokenizer, all token sequences of length ≤ 6 whose decoded bytes match a candidate surface (including byte-fallback and non-standard merges) were enumerated exhaustively by byte-level search over the merge table, and the strict-event mass was recomputed on all 18 validation cells. Seventeen of 18 cells are unchanged (JS $= 0$ to printed precision); the single non-trivial correction is OLMo-2-0425-1B-RLVR1 biased-dice, whose coverage rises $0.5590 \rightarrow 0.5647$ ($+0.006$; conditional JS $= 10^{-4}$ bits). Convergence is steep: 99% of the total mass is carried by 1–24 paths per cell out of 24–372 enumerated. The released inventory is thus effectively decoder-complete on these cells, with an empirical error bound of ≤ 0.006 coverage and $\leq 10^{-4}$ bits JS.

Tokenizer boundary audit. For each of five tokenizers (Llama-3.1, Qwen2.5-7B, Qwen3-8B, Mistral SentencePiece v3, OLMo-2-1B) we recorded, per probe and candidate, the token IDs of every surface variant, verified `encode(prompt)+encode(answer) == encode(prompt+answer)` at the seam, and audited the first-token pool of each answer's tokenization. Four tokenizers show *zero* boundary mismatches across all probes (600 audited prompt/candidate/form combinations in total). The Mistral SentencePiece v3 tokenizer shows *60* mismatches (prompt/answer re-tokenization merges or splits bytes across the seam relative to the concatenated encoding), quantifying the byte-level splitting artifact that motivated reporting Mistral results under both encodings elsewhere; all non-Mistral tables are unaffected. Concrete IDs

make the three artifact classes of Section 2 (with the corrected-tokenizer follow-up in Appendix I) inspectable. *(i) Leading-space pooling:* for Llama-3.1 the bare candidate "1" encodes as the single token [16] while its spaced form is [220, 16] ("□"+"1"), so a raw-protocol policy that spreads mass over the shared 220 prefix dilutes every candidate's first-token probability identically. *(ii) Shared-prefix byte splits:* on the Qwen tokenizers at $K{=}20$, candidate "1" is [16] but "10"–"19" are [16, 15], [16, 16], …, [16, 24]. Eleven candidates share leading token 16, exactly the regime where naive branch sums double-count and prefix subtraction is mandatory (subtracted totals 0.68–1.0 for the Qwen families, $\leq$0.001 for Llama-3.1). *(iii) Mistral SentencePiece v3:* every Mistral digit candidate begins with the same shared prefix token 29473 ("□"): "1" is [29473, 29508], "2" is [29473, 29518], "10" is [29473, 29508, 29502]. The first token carries zero candidate information, so first-token marginalization cannot distinguish candidates at all. The boundary mismatches are of the same seam class: for uniform10/chat candidate "1", the standalone encoding is [29473, 29508], but `encode(prompt+"1")` ends […, 29491, 29473, 4, 29473, 29508]: the trailing period of the prompt re-tokenizes across the seam. The Llama-3.1 counterpart verifies exactly ([16] appended unchanged after the chat header tokens [128007, 271]). Full per-candidate inventories (every form's IDs, seam check, and first-token pool) are in the supplementary material.

F.1 Finite-Inventory Strict-Event Re-Derivation

The 18 validation cells are re-derived from their stored prompt-token IDs over the declared finite strict-event inventory, with every path probability recomputed from logits under teacher forcing. The inventory is a disjoint terminal-path enumeration (optional single leading space $\times$ core spelling variants $\times$ optional trailing newline, post-multiplied by the stored terminator set), with no prefix subtraction and no terminator post-multiplication shortcut. Coverage reproduces the archived r11 values in 18/18 cells within $|\Delta| \leq 0.006$ (maximum: OLMo-1B-RLVR1 biased-dice $+0.0057$). The flagged qwen25_7 raw `uniform10` cell reproduces its authoritative strict coverage *exactly* (0.8604), and 14/18 reverse-KLs reproduce to $\leq 10^{-3}$ bits. The remaining four cells split into two groups. Three uniform-20 cells (qwen25_7, qwen3, mistral7) show lower reverse-KL by 0.5–1.2 bits. The re-derivation enumerates the fuller variant set (case variants $\times$ optional lead/newline branches, 2–8 paths per candidate where the archived enumeration used the minimal path set), which flattens the conditional distribution slightly and hence lowers the reverse-KL against the target. Coverage is unchanged to 10^{-4}, so this is a strictly-more-complete-branch effect in the expected direction. We report the recomputed values as the v3 authoritative readings for these cells. The fourth (olmo1b-rlvr biased-dice) shifts by $+0.023$ bits with coverage $+0.0057$, rounding-level, consistent with the archived conditional set being reported at reduced precision. No cell has exactly zero coverage, so every cell reports a finite conditional divergence.

F.2 Free-Running Intervention Replication

The intervention claims are re-tested under continuous free-running generation (2048 samples per cell, temperature 1.0, hooks active at every decoding step) rather than the harness readouts of the post-steering and calibration experiments. Baselines closely reproduce the archived cells (llama31 `uniform10` chat: "7" 75.5%/"8" 21.7%, coverage 1.000; qwen3_8: "7" at 99.95%). The free-running results *strengthen* the paper's capability-cost accounting: at the entropy-restoring strength (L12, $\alpha{=}10$) the steered llama31 emits a strict-valid candidate in only 3.3% of free-running generations (and 59.8% at $\alpha{=}5$). The flatter conditional shape suggested by the harness readout is carried by a residual candidate emission whose absolute mass nearly vanishes in free running, the free-running counterpart of the calibration absolute-mass caveat and the specificity reset reading. On the harness side, the strength–response sweep anchors the capability cost quoted in the main text: arithmetic accuracy runs 1.00 ($\alpha{=}0$) $\rightarrow$ 0.57 ($\alpha{=}5$) $\rightarrow$ 0.00 ($\alpha{=}10$), $n{=}30$, while candidate coverage at the near-flat setting ($\alpha{=}5$, KL 0.0004 bits) is 0.75. In the free-running $\alpha{=}10$ cell the attractor remains modal among the strict-valid generations ("7" at 54.4%, 37/68). The hard-family result is unchanged in kind: at L14/$\alpha{=}20$ the steered qwen3_8 still emits "7" in 98.7% of its 2048 free-running generations (coverage 1.000). The partial conditional-shape recovery at the refined-strength readout ($\sim$21% of the deficit) amounts to a mild conditional reshaping that leaves the free-running policy mode-locked. Finally, last-position base-state patching at layers 12–28 *destroys* strict-valid candidate emission entirely in free running (0/2048 for both the base-state and shuffled-state conditions, versus the readout-layer recovery curves of the patching experiments). State replacement at every decoding step is a far stronger intervention than the single-readout patch the

patching curves evaluate, so those curves do not transfer to free-running policies. Patching evidence is therefore read at the readout layer only, an added scope restriction on the intervention claims.

G Open-Ended Diversity: Per-Pair Batch Detail

This appendix gives the per-pair batch detail behind the open-ended diversity results summarized in Section 4. Each *pair* compares one post-trained checkpoint against its own base on the same prompt pool. *Distinct-2* is the fraction of distinct bigrams in the pooled generations; the Vendi score is an embedding-space diversity index. Both are reported as post−base changes, so negative values mean diversity loss. Table 21 contrasts three independent generation batches.

Both regeneration batches (2 and 3) certify their own precision: batch-3 split-half Spearman–Brown reliability is $\geq$0.94 for every post-trained checkpoint (0.69–0.97 for the noisier bases), and the batch-2 values computed post hoc from the retained raw texts are 0.88–0.99 (mean 0.96). Disattenuating the pair-level correlations against the reliability of the difference score barely moves them ($-0.61\rightarrow -0.61$, $-0.55\rightarrow -0.56$), so the archived batch's null cannot be explained by regeneration noise; the archived batch is excluded from inference solely because its precision is unverifiable. Under a Holm correction across the four statistic$\times$batch readouts, the pair-level coupling between probe KL and distinct-2 survives on both rank readouts (Spearman $0.010 < 0.0125$ and $0.014 < 0.0167$) but not on the Pearson readouts ($0.035 > 0.025$ at the third Holm step). A wild cluster bootstrap over the six lineages gives p=0.17/0.16, and at the lineage level all three batches are null. In sum, the two certifiable batches show a sign-stable pair-level rank association, and the lineage-level effect size remains unresolved at n=6. The qwen3_8 sign flip is localized to the archived batch: in batch 3 all 30 prompts show a diversity loss (per-prompt instruct distinct-2 range $[0.184, 0.400]$ vs. base mean 0.498), and the original-10 and new-20 prompt subsets agree (0.293 vs. 0.288). On the semantic axis, batch-2 Vendi scores fall for all 13 pairs while mean cosine rises in 12/13 pairs in batches 2 and 3 (verified against the archived per-checkpoint values, Appendix I). Neither semantic statistic correlates with probe KL, and the distinct-2 and Vendi changes are themselves rank-uncorrelated (ρ=0.02): the measurement axis matters as much as the batch. A fourth, independent embedding-readout batch (23 checkpoints, 30 prompts $\times$ 64 samples, full text retention) replicates the batch-2 pattern on all three prompt-pool splits: mean cosine rises for 11/12 certified pairs and Vendi falls for 12/12, with the KL-vs-semantic couplings again null.

H Estimator Boundary Narratives: Coverage and the Alphabet Correction

This appendix gives the worked detail behind the two estimator-boundary discussions summarized in the main text: coverage and the alphabet correction.

Out-of-set mass / coverage. Under the chat protocol, the six directly comparable families place essentially all mass inside the candidate set (sequence-level coverage $\geq$0.99 for five; mistral7 sits at 0.988), so the unconditional and conditional readouts differ by at most the $\mathbb{H}_2$ compliance correction (interpretation note (i), Section 2.4). The OLMo-2 pipeline stages sit lower (0.39–0.91 on `uniform10`, Table 10) and are excluded from that bound. At coverage 0.39 the compliance correction reaches $\approx$1.06 bits, roughly 18$\times$ the 0.057-bit SFT$\rightarrow$DPO conditional-KL difference on 1B `uniform10`; OLMo stage cells therefore always report coverage alongside conditional KL, and the two estimands are never interchanged. Raw-protocol bases leak 0.66–0.99995 of their mass outside the candidate set, so raw-protocol first-token numbers are uninformative without the sequence-level correction.

Alphabet probe, corrected measurement. Two corrections turn the alphabet probe from a collapse measurement into a compliance-and-fidelity decomposition. (i) The original variant enumerated only an A–J candidate head against an A–Z target set; v2 uses all 26 letters with case-folded answer branches and a uniform(26) target. (ii) Even corrected, base models do not treat the prompt as a letter-sampling task: candidate coverage is 0.06–0.11 across the three tested bases, and the dominant *in-set* mass is the pronoun "I" (47–88% of in-set mass) rather than a sampled letter. The earlier reading "base models are already collapsed on alphabet" is retracted: bases are *non-compliant*, not collapsed. The compliance jump at SFT (coverage 0.08$\rightarrow$0.99 at 7B) is itself an alignment-stage effect and should be read separately from distributional fidelity. The numeric probes show the same

distinction more weakly: their low base coverage (0.00–0.34, Table 15) still consists of candidate answers, not one dominant non-answer token.

I Measurement Corrections and Retractions

This appendix consolidates every quantity in this paper that was corrected, replaced, or retracted over the course of the study, with cause and status. Every corrected value is either sampling-validated (JS reported alongside) or analytically derived, and all raw per-sample records are retained.

The corrected `uniform20` anchor inversion. Under first-token marginalization the qwen families appeared to recover on `uniform20` (0.76–0.86 bits) while llama31 appeared to collapse (3.46 bits), the opposite of their `uniform10` severities. The 0.86-bit value is exactly the shared-prefix artifact signature $\log_2(20/11)$. The corrected readout shows the four collapsing families retain high conditional KL at $K{=}20$ (2.30–3.46 bits), joined by qwen3_1.7B (3.54 bits), with llama31 the most concentrated ("14" at 88%). The lesson: what varies across probes is the attractor, not the collapse ("7" $\rightarrow$ "12"/"14" $\rightarrow$ "42"/"57" as the range widens).

The stage-attribution revision (measured base battery). Rows 7–8 change a headline claim. An early version counted *5/6 full-chain monotone-worsening families* on OLMo-2-7B under a base readout that treated the base cell as $\approx$0. The unified sequence-level battery then measured every base cell with coverage: base conditional KLs are not uniformly near zero (on coin the base's 0.371 substantially exceeds SFT's 0.034, on weekday the two nearly tie at 0.228 vs. 0.224; the non-compliant alphabet base improves at SFT). The full-chain count therefore drops to 3/6, while the SFT$\rightarrow$DPO step, which never depended on the base readout, holds on 6/6. Figure 3 was regenerated from the same records, and the unconditional decomposition (Table 16) allows a matched-estimand base-vs-SFT comparison. Seven r3-era stage cells were recomputed from the same records; every difference is $\leq$0.02 bits and none changes an interpretation.

Scope. The three readout-level corrections (rows 1–3) share one root cause: first-token marginalization is invalid whenever candidates share a leading token. That is why the branch-sum readout with prefix subtraction is the reported readout everywhere, with the strict event as the normative quantity (Section 2; Appendix F calibrates both against free-running sampling). Rows 4–6 are target-set and interpretation-level corrections; rows 7–8 record the stage-attribution revision described above.

J Label $\times$ Position $\times$ Probability Disentangling

The two-ordering design behind the factorial sweep in the Results confounds letter identity with display position. This appendix reports the disentangling run: for each of three instruct models (llama31-8b, qwen25_7-7b, qwen3_8-8b) we cross all 24 display orders of $\{A, B, C, D\}$ with all 24 assignments of the stated probabilities $(40, 30, 20, 10)$, 576 units per model plus 24 flat controls $(25/25/25/25)$. Every unit is read out with the finite-inventory strict-event estimator (no sampling noise; per-unit records retain the full four-letter distribution, coverage, and top-1 letter/position). Table 23 reports the mean mass on letters stated at each level, the within-unit correlation between stated probability and mass, the mean conditional KL against the stated profile, $P(\text{top-1} = \text{argmax-stated letter})$, $P(\text{top-1} = \text{first-displayed letter})$, and the one-hot least-squares decomposition of the top-1 letter's mass into label-identity and display-position factors. All summary statistics in this table and in the text below are recomputed directly from the 1800 per-unit raw records under a single verified alignment (the stated probability of each canonical letter is the one attached to it in the display sequence). The identity $\overline{\mathrm{KL}} = \overline{\mathrm{CE}} - \overline{H} = -\sum_p m(p) \log_2 p - \overline{H}$ holds to machine precision for all three models.

Grouped holdout prediction. We expand each unit to four candidate rows in canonical label order and recover each label's displayed position and stated probability from the prompt record. Five ordinary least-squares specifications add, in sequence, stated probability, candidate identity, displayed position, and probability-by-identity/position interactions. Leave-one-display-order-out and leave-one-probability-assignment-out cross-validation each fit on 23 groups and predict the omitted group. The interaction specification reaches $R^2 = 0.777/0.760$ (llama31), $0.510/0.495$

(qwen25_7), and $0.502/0.525$ (qwen3_8), with MAE $0.109/0.114$, $0.220/0.226$, and $0.211/0.204$ respectively (order/assignment). For the order holdout, probability-only R^2 is $0.430/0.336/0.112$; adding position raises qwen25_7 to 0.478, while adding label raises qwen3_8 to 0.382. These tests hold out combinations within one candidate set and prompt template; direct transfer tests fail on all three held-out axes (next paragraph).

Held-out transfer test (new models, templates, labels). We test transport directly on three held-out axes, each measured with the same strict-event estimator on 24 display orders $\times$ 8 stated-probability assignments stratified over the argmax position (192 units per cell, plus flat controls). The axes are (i) *new models*: Mistral-7B-Instruct-v0.3 and Qwen3-4B-Instruct-2507 under the original template; (ii) *new template*: a rephrased instruction (candidate list first, chance phrasing) on the three original models; and (iii) *new candidate labels*: the color set {red, blue, green, yellow} replacing A–D on the three original models. The candidate-level interaction specification of the grouped holdouts, refit on the full 24×24 training grid (pooled in-sample $R^2 = 0.38$; model fixed effects added for the new-model axis and zeroed at prediction), predicts *none* of the held-out axes. Held-out R^2 against the training-mean baseline is negative everywhere (-0.65 to -0.29 template, -0.71 to -0.27 labels, -0.17 to -0.16 new models), and predicted argmax matches the measured argmax only at chance level (0.22–0.33), no better than the stated-probability baseline. The label/position/stated-probability structure is therefore *instance-specific* to each model $\times$ template $\times$ label set, not a transportable law: within-design holdouts interpolate, but the fitted coefficients do not extrapolate. A secondary observation sharpens the template axis: under the rephrased template, Qwen3-8B's strict-event coverage collapses to 1.6×10^{-8}. A wording change alone switches off the bare-letter response format, so the measurement event itself is template-fragile.

Within-unit correlation between stated probability and mass (canonical letter pairing) is 0.660 (llama31), 0.578 (qwen25_7), and 0.325 (qwen3_8); mean conditional KL against the stated profile is 0.63, 1.35, and 1.43 bits respectively (identity-checked: $\overline{\mathrm{KL}} = \overline{\mathrm{CE}} - \overline{H}$ and $\overline{\mathrm{CE}} = -\sum_p m(p)\log_2 p$ hold to 10^{-15}). The fraction of fully deterministic units (top-1 mass > 0.99) is 2.8%, 69.1%, and 41.3%. Per-letter means for qwen3_8-instruct ("A" 0.025, "B" 0.099, "C" 0.317, "D" 0.559) are nearly flat in the stated probability attached to each letter: the "C"-anchor generalizes to a fixed letter-identity preference (C/D), not a position or probability artifact. Flat-control top-1 masses: 0.438 (llama31), 1.000 (qwen25_7, with the chosen letter uniform across the 24 controls), 0.878 (qwen3_8). Raw per-unit records (distribution, coverage, top-1 letter, display order, stated assignment) are retained for all 1800 units. The full per-candidate release contains a 6912-row long table (model $\times$ 576 units $\times$ 4 candidates, each row with display position, stated probability, model mass, and KL contribution), the marginal responses behind Table 23 ($m(p)$, balanced label mass, balanced position mass), a descriptive additive cell-means variance decomposition of the per-candidate mass (main-effect shares of stated-p/letter/position plus residual, per model), and a three-panel summary figure (probability tracking against $y{=}x$; balanced label mass; balanced position mass), all regenerated by a single script from the same verified records (released artifact set `e30_factorial/`). Two interpretation notes on the R^2 columns. First, the balanced 24×24 design makes the label and position factors orthogonal *at the unit level*, but the decomposition is computed on the winner-conditioned subset (each unit contributes its top-1 letter's identity and position), which is *not* balanced. The additivity $R^2_{\text{full}} = R^2_{\text{L}} + R^2_{\text{P}}$ therefore holds only approximately (additivity error $+0.019$, $+0.005$, $+0.016$ for the three models), and we read the decomposition as descriptive rather than as a clean ANOVA. Second, R^2 here is variance of the top-1 *mass value* explained, not top-1 *choice* predictability, and the two diverge at the deterministic boundary. The 69.1% fully deterministic units of qwen25_7 compress the mass variance available to be explained, so its small R^2s coexist with highly predictable choices ($P(t{=}\arg){=}0.69$). For near-deterministic models the factorial R^2 should be read jointly with the choice-level columns, and its low values indicate "no graded mass structure to explain," not "no bias."

Alphabetical monotonicity: "letter identity" is mostly an ordinal prior. Because the 24 stated assignments are balanced over letters by construction, any per-letter mass gradient that survives averaging over assignments cannot be a probability artifact. Averaging each letter's mass over all 576 units, the per-letter means are *strictly monotone in canonical alphabetical order* for llama31 ("A" 0.128, "B" 0.203, "C" 0.268, "D" 0.401; Spearman $\rho{=}1.0$) and qwen3_8 ("A" 0.025, "B" 0.099, "C" 0.317, "D" 0.559; $\rho{=}1.0$), while qwen25_7 is the exception: non-monotone, with its "C" mass dipping below "B" ($\rho{\approx}0.8$). With only $K{=}4$ letters the exact null probability of a perfect rank ordering is $1/24 \approx 0.042$, so we read the monotonicity as suggestive rather than decisive; the

consecutive per-letter gaps themselves grow with alphabetical index for qwen3_8. For two of the three families, the per-letter mass gradient behind Table 23's letter-identity column is therefore not an arbitrary token affinity but an *ordinal prior over the canonical letter sequence*: mass grows with alphabetical index just as it grows with display position for llama31. This is consistent with a shared ordinal prior over whichever ordering (display or canonical) the model treats as primary, though no manipulation isolates a single mechanism. This refines the letter-generality reading: what surfaced there as a fixed "C"-anchor is, at the population level, the upper tail of an alphabetical gradient (for qwen3_8, the last canonical letter "D"). Only qwen25_7's letter preference is idiosyncratic rather than ordinal, consistent with its hard-family profile elsewhere (it retains collapse under raw prompts and shows 0% steering reversal).

K REPRODUCIBILITY

All experiments were run on a single A100-80G (bf16 inference, transformers 4.57.6). The finite-path readouts are deterministic given a checkpoint, and no sampling seeds are involved in any KL/coverage cell. The free-running validation generations and the open-ended batches used temperature 1.0/top-p 1.0 with per-batch fixed seeds recorded in the artifacts. The three open-ended generation batches were produced on the *same* hardware and software stack at different times. Batch 1 (archived) retains only pooled aggregates; its raw generations were not archived, the gap this paper's batch-2/3 protocol was designed to close. Batches 2 and 3 retain every generated text, enabling the post-hoc reliability analysis reported in Appendix G. We can identify no library or configuration change that distinguishes batch 1 from batch 2, so we exclude batch 1 from inference rather than reconcile it. The embedding readouts (mean pairwise cosine and Vendi scores) are computed from the retained batch-2 texts and, for the fourth batch, at generation time, both with a pinned embedder (all-MiniLM-L12-v2). The label-family cells (colors/animals) are finite-path branch-sum readouts of the same kind as the main battery. The probe protocol specification (prompts, candidate sets, targets, tokenizer-variant enumeration rules), the complete per-cell measurement records (every KL/entropy/coverage/sampling-validation cell in this paper, as JSON), the per-sample generation records for batches 2/3, and the analysis code needed to regenerate every table are released as anonymized supplementary material. Two checkpoint-integrity incidents occurred during the audit and were resolved before measurement. A truncated OLMo-2-0425-1B-SFT download was fixed by re-downloading and verifying all shards; a config/architecture field mismatch that can silently misload a module graph was fixed by loading via the explicit class named in `architectures`.

Table 18: **Current authoritative readings (quick reference).** One row per headline quantity, stating the value the paper stands behind and where it lives. All KLs in bits; sequence-level estimator throughout (chat-protocol cells read the branch sum, which coincides with the strict event within the tested regime, Section 2; the flagged raw cell carries both readings, Table 19); full provenance in Appendix D.

Quantity	Current authoritative reading	Where
`uniform10` post KL, chat (primary probe)	llama31 2.297; mistral 1.325; qwen25_7 3.321; qwen3_8 3.303; qwen3_4 3.321; qwen3_1_7 1.517; qwen3_0_6 1.284; r1distill 0.326 (reasoning-conditional readout)	Table 12
K-normalized severity κ	primary-probe post: 0.69 (llama31), 1.00 (qwen25_7), 0.99 (qwen3_8), 1.00 (qwen3_4), 0.46 (qwen3_1_7)	Tables 9, 12
Raw-protocol reversion (chat→bare)	llama31/qwen3/mistral revert to 0.28–0.63 conditional bits (qwen3_0.6B: 1.025; cov $\leq$0.15); qwen25_7 retains 3.317 at branch cov 0.975 (strict-event coverage on the validation wording is lower; reproduces exactly under the r19 re-derivation, Table 20); unconditional deficits stay compliance-dominated	Table 12; T1; Discussion
`uniform20` post KL (corrected)	llama31 3.460; qwen25_7 2.298; qwen3_8 2.365 (chat cell flagged as a calibration exception, Appendix F); qwen3_4 3.273; qwen3_1_7 3.539 (anchors "14"/"12"); the 0.76–0.86-bit values are retracted	Table 9; App. I
90-way probe (`Unif(10--99)`)	qwen3_8 5.73 (89.9% on "42"); qwen25_7 5.05 (70% on "57"); llama31 2.84; collapse re-anchors, does not vanish	Section 4
Steering reversibility	llama31 +100% (L12–20/α=10); qwen3_8 $\leq$9% main grid, 21% at the refined-strength readout; qwen25_7 0%; RLVR1 endpoint +100% (L6/α=10); r19 free-running: steered llama31 α=10 strict-valid cov 0.033 (cost-side confirmation), steered qwen3_8 α=20 still "7" at 98.7% (hard-family confirmation)	T3; Table 11; App. F
Stage attribution (OLMo-2)	7B: SFT→DPO worsens 6/6 KL-scored families, full measured chain (base battery) monotone on 3/6 (biased die: deficit, near-flat); 1B: SFT share family-dependent, chain non-monotone; RLVR raises 4/6 KL probes plus the biased-die entropy deficit; endpoint stays reversible	T2; Table 10
Scale trend (Qwen3)	0.6B→1.7B→4B→8B: 1.28 (cov 0.35)→1.52→3.32→3.30 bits; attractor strength grows with scale (observational)	T1 (scale)
Probe-vs-diversity coupling	moderate pair-level coupling in both regenerated, retained-record batches ($\rho = -0.73/-0.70$); *exploratory*: survives Holm only on the two rank readouts across the four statistic×batch tests (Spearman 0.010/0.014 vs. thresholds 0.0125/0.0167; Pearson 0.035/0.064), wild cluster bootstrap over six lineages p=0.16–0.17 (sign-resolution-limited, 2^6=64); archived batch unverifiable, excluded; dissociation extremes replicate in all three (qwen25_7); Vendi falls 13/13 (b2) and 12/12 (b4, all three prompt-pool splits); cosine rises 12/13 (b2), 12/13 (b3, after correcting a sign-transcription error), 11/12 (b4, p=0.001, all prompt-pool splits); no coupling on either semantic statistic (b2 $p\geq$0.36; b4 $\lvert r\rvert\leq$0.33)	App. G
Format transfer (fair die)	quoted worsens all four families (+0.23–0.63); spelled family-specific; sysuniform helps qwen3_8 (-0.55) but not qwen25_7; die10/colors6/animals6 readouts in the extended battery	released records
Downstream SC bridge (steering screeners)	SC@8 minus greedy per family: -0.011 to $+0.067$, uncorrelated with probe KL ($r = -0.09$, n=7, exploratory); single-gold short-form accuracy is insensitive to support collapse; scope: ceiling-limited screeners	released records
Sampled confidence/width (steering screeners)	Sampled-answer entropy and distinct-count fall directionally with probe KL ($r = -0.43/-0.54$, n=7, ns; permutation p=0.33/0.24); negative in all leave-one-out drops; sharpens to $r = -0.81$ at the 4-lineage level; confidence AUROC does not degrade (LOO range $[0.04, 0.68]$ for agreement — stays positive but approaches zero, weaker evidence than log-p whose LOO range stays clearly positive); scope: ceiling-limited, one family has a single negative item	released records
Paraphrase spread (`uniform10`, chat)	$\leq$0.02 for qwen25_7/qwen3_8; 0.93 (qwen3_4, remains in collapse band); 0.27 (llama31); 0.82 (mistral); up to 2.38 on 1.7B with compliance drop to cov 0.001	Table 12

Table 19: Free-running validation closure with the three coverage objects separated: branch-sum coverage (Br), finite-inventory strict-event coverage (St), and observed strict-parser valid rate (vr) at $N{=}2048$ per cell; Δ_{b}=Br−St, Δ_{o}=vr−St, both computed from unrounded values. cal: JS calibrated against a per-cell parametric bootstrap (4000 multinomial draws, 99% band). Sixteen of eighteen cells pass; the two FAIL cells are reported as calibration exceptions (Findings iv), and the †-marked raw cell carries a protocol flag (outside the tested immediate-termination regime), not a calibration failure. Raw records (token IDs, decoded text, per-sample parser decisions) are retained for every sample.

Cell	Br	St	vr	Δ_{b}	Δ_{o}	JS	cal
Llama3.1 biased-dice (chat)	0.9986	0.9942	0.9922	+0.0044	−0.0020	0.0015	FAIL
Llama3.1 uniform10 (chat)	1.0000	0.9990	0.9995	+0.0010	+0.0005	0.0007	ok
Llama3.1 uniform20 (chat)	1.0000	0.9989	0.9976	+0.0011	−0.0014	0.0017	ok
Qwen2.5 biased-dice (chat)	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	ok
Qwen2.5 uniform10 (chat)	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	ok
Qwen2.5 uniform20 (chat)	1.0000	1.0000	1.0000	0.0000	0.0000	0.0009	ok
Qwen3 biased-dice (chat)	0.9998	0.9998	1.0000	+0.0000	+0.0002	0.0012	ok
Qwen3 uniform10 (chat)	1.0000	0.9998	1.0000	+0.0002	+0.0002	0.0004	ok
Qwen3 uniform20 (chat)	1.0000	0.9994	0.9995	+0.0006	+0.0001	0.0106	FAIL
OLMo-1B-RLVR1 biased-dice (chat)	0.5655	0.5590	0.5776	+0.0066	+0.0187	0.0016	ok
OLMo-1B-RLVR1 uniform10 (chat)	0.5627	0.5528	0.5669	+0.0098	+0.0141	0.0027	ok
OLMo-1B-RLVR1 uniform20 (chat)	0.6126	0.6032	0.5947	+0.0094	−0.0084	0.0027	ok
Llama3.1 alphabet (chat)	0.9997	0.9981	0.9990	+0.0016	+0.0009	0.0029	ok
Qwen3 alphabet (chat)	1.0000	1.0000	1.0000	0.0000	0.0000	0.0015	ok
Mistral-7B uniform20 (chat)	0.9996	0.9731	0.9741	+0.0265	+0.0010	0.0018	ok
OLMo-1B-SFT uniform10 (chat)	0.6326	0.6301	0.6494	+0.0026	+0.0193	0.0016	ok
Qwen2.5 uniform10 (raw)	0.9770	0.8604	0.8613	+0.1166	+0.0009	0.0001	flag†
Qwen3-0.6B uniform10 (chat)	0.3455	0.3438	0.3579	+0.0017	+0.0141	0.0026	ok

Table 20: Strict-event re-derivation of the 18 validation cells from stored prompts (r19) vs. the archived r11 records. u10/u20/dice: `uniform10`/`uniform20`/fair-die probes; "chat" = chat template, "raw" = bare prompt; r11/r19 = archived vs. recomputed. The KL columns archive the r11 validation records' divergence $\mathrm{KL}(q\|P^*)$, with q the target and P^* the model's strict-event conditional (unbounded for concentrated P^*); the main tables report the opposite direction $\mathrm{KL}(P^*\|q) \le \log_2 K$ (for qwen25_7 u10 chat: 3.321 vs. 22.836), so the two are not directly comparable; for the die rows q is the stated biased target. † marks the three uniform-20 cells whose KL shifts under the fuller variant enumeration (see text); the die cell's small shifts are rounding-level. Mode column: top candidate of the recomputed strict event; the qwen25_7 raw uniform10 coverage reproduces exactly.

Cell	cov (r19)	cov (r11)	KL_{19}	KL_{11}	Mode
llama31 u10 chat	0.9990	0.9990	5.880	5.880	7
llama31 u20 chat	0.9989	0.9989	7.037	7.037	14
llama31 dice chat	0.9942	0.9942	0.915	0.915	4
llama31 alphabet chat	0.9981	0.9981	1.756	1.756	K
qwen25_7 u10 chat	1.0000	1.0000	22.836	22.836	7 (>99.9%)
qwen25_7 u20 chat†	1.0000	1.0000	13.010	13.638	14
qwen25_7 dice chat	1.0000	1.0000	10.345	10.345	1
qwen25_7 u10 *raw*	0.8604	0.8604	15.766	15.767	7 (spaced form)
qwen3 u10 chat	0.9998	0.9998	13.868	13.868	7
qwen3 u20 chat†	0.9994	0.9994	13.384	13.899	12
qwen3 dice chat	0.9998	0.9998	0.525	0.525	1
qwen3 alphabet chat	1.0000	1.0000	5.024	5.024	Q
olmo1b rlvr u10 chat	0.5528	0.5528	0.505	0.505	7
olmo1b rlvr u20 chat	0.6032	0.6032	0.648	0.648	7
olmo1b rlvr dice chat	0.5647	0.5590	0.444	0.421	1
mistral7 u20 chat†	0.9731	0.9731	6.597	7.767	14
olmo1b sft u10 chat	0.6301	0.6301	0.433	0.433	7
qwen3_0_6 u10 chat	0.3443	0.3438	1.708	1.707	1

Table 21: Three independent generation batches of open-ended diversity (pooled distinct-2 change, post−base; negative = diversity loss). Batch 1: archived corpus (only pooled aggregates survive; excluded from inference). Batch 2: full regeneration, identical protocol. Batch 3: enlarged protocol (30 prompts × 128 samples). $r_{1/2}$: batch-3 split-half correlation; SB: Spearman–Brown reliability. Vendi Δ (b2): post−base change of the Vendi score on batch-2 texts, negative for all 12 pairs shown and for r1-distill (-34.0); independent batch 4 is negative for 12/12 pairs on all three prompt-pool splits. Pair-level correlations use the 12 pairs (r1-distill excluded); lineage-level rows collapse to six independent clusters.

Pair	batch 1	batch 2	batch 3	btwn-batch SD	$r_{1/2}$/SB (b3)	Regen. (b3)	Vendi Δ (b2)
llama31	−0.316	−0.195	−0.149	0.087	0.990/0.995	—	−30.1
mistral7	−0.241	−0.169	−0.124	0.059	0.986/0.993	—	−34.7
qwen25_7	−0.067	−0.030	−0.013	0.028	0.989/0.994	0.2583/0.2570	−4.6
qwen3_8	+0.109	−0.175	−0.111	0.149	0.988/0.994	0.1708/0.1695	−13.2
qwen3_4	−0.251	−0.199	−0.129	0.061	0.983/0.992	—	−14.2
qwen3_1_7	−0.246	−0.211	−0.137	0.056	0.989/0.994	—	−18.0
qwen3_0_6	−0.161	−0.131	−0.101	0.030	0.968/0.984	—	−16.5
olmo2/sft	−0.108	−0.020	−0.011	0.054	0.986/0.993	—	−7.8
olmo2/dpo	−0.138	−0.033	−0.031	0.061	0.989/0.995	—	−17.9
olmo1b/sft	−0.081	−0.008	−0.005	0.043	0.992/0.996	—	−17.7
olmo1b/dpo	−0.084	+0.010	+0.005	0.053	0.985/0.993	—	−24.0
olmo1b/rlvr	−0.085	+0.007	+0.004	0.052	0.979/0.990	—	−24.5
Pearson r (p)	0.02 (0.95)	−0.61 (0.035)	−0.55 (0.064)				
Spearman ρ (p)	−0.24 (0.44)	−0.73 (0.010)	−0.70 (0.014)				
Lineage-collapsed (6 clusters; pairs averaged within lineage)							
lineage Pearson r (p)	−0.08 (0.88)	−0.37 (0.47)	−0.31 (0.55)				

Batch 1: archived corpus (10 prompts × 32 samples); only pooled aggregates survive, so its reliability cannot be certified and it is excluded from inference.
Batch 2: full regeneration, identical protocol, reliability-certified post hoc from retained raw texts (split-half Spearman–Brown 0.88–0.99, mean 0.96, post-trained checkpoints).
Regen.: same-checkpoint regeneration control, pooled distinct-2 (first vs. repeat). Vendi Δ correlation with probe KL is null (r=0.29, permutation p=0.36, ρ=0.18; batch 4: r=0.18–0.33).
Pair-level p-values are two-sided asymptotic t tests (Pearson, n=12 pairs) and permutation tests (Spearman); lineage rows use n=6 clusters.
Pair-level significance in batches 2/3 is carried partly by within-lineage repetition and does not survive at the lineage level (n=6); full narrative follows the table.

Table 22: Consolidated corrections and retractions. "Replaced" = the original readout is superseded by a valid one on the same data; "Retracted" = the value is withdrawn outright; "Softened" = the interpretation, not the number, changed.

Quantity	Originally reported	Status	Cause	Current value
Qwen families, `uniform20` post KL (first-token readout)	0.76–0.86 bits	Retracted	11/20 candidates share the leading "1" token; pooled-mass artifact (uniform spread over the shared-prefix group alone yields KL = $\log_2(20/11) = 0.86$, top-1 = $1/11$ = 9%, the artifact signature)	Sequence-level 2.30–3.54 bits, qwen3_1.7B at the upper end (Table 9)
Mistral-7B-Instruct, all post readouts (first-token)	Spurious near-uniformity	Replaced	the released SentencePiece v3 tokenizer emits a shared prefix token before every digit and Chinese character; first-token marginalization cannot distinguish candidates	Sequence-level (Table 12); sampling-validated JS $\leq$0.006
Base-model conditional KLs (first-token)	$<10^{-3}$ bits	Replaced	pooled leading-space mass + shared byte-split prefixes overstated base uniformity	Sequence-level conditional 0.01–1.15 bits across all probed bases (Tables 9 and 10)
Alphabet probe target set	v1 uppercase-only	Corrected	case-folded A–Z branches with uniform(26) target; bases are largely non-compliant (cov 5–19% across the six lineages (0.06–0.11 on the three bases under the corrected protocol, Appendix H), dominated by "I")	corrected measurement; conditional on answering, SFT improves fidelity 3.00→0.40 bits
llama31-base biased-coin tracking at p=0.7	"Tracks exactly" (KL 0.0016)	Softened	p-insensitive $\sim$0.75 Heads prior; Heads 0.81–0.82 at $p \in \{0.6, 0.9\}$, so the p=0.7 agreement was coincidental	"$\sim$0.75 Heads prior" wording
qwen3_8 coin mode-locking	"Majority-locking"	Retracted	label swaps: $\geq$99% Heads even when Heads is the stated 30% minority	Label-anchoring on "Heads"
OLMo-2 7B stage monotonicity count (base-cell readout)	5/6 full-chain monotone (first-token base readout $\approx$0)	Replaced	the unified base battery (r8a/r8e) measures base conditional KLs directly: on coin the base's 0.371 substantially exceeds SFT's 0.034 and on weekday the two nearly tie (0.228 vs. 0.224), and the non-compliant alphabet base improves at SFT	full chain monotone-worsening 3/6 (u10/u20/dice); SFT→DPO step 6/6 (T2; Table 10)
Seven r3-era OLMo stage cells	e.g. 1B RLVR `dice` 0.918, `uniform20` 0.694, `weekday` 0.734, biased $-0.016 \rightarrow +0.235$	Replaced	recomputed from the unified r8a/r8e battery; all differences $\leq$0.02 bits, none interpretation-changing	values as printed in Table 10
Batch-3 mean-cosine changes (13 pairs)	e.g. llama31 -0.079, mistral7 -0.118, r1-distill -0.113; "falls 12/13"	Corrected	an earlier draft transcribed every batch-3 Δ with the opposite sign; the archived per-checkpoint cosines (r14) and a same-text identity-corrected recomputation (r19: pooled values reproduced to $\leq$0.0005, identity correction $\leq$0.0025) confirm the corrected signs and magnitudes	batch-3 cosine rises 12/13, sign-consistent with batches 2 and 4
Factorial mean KLs and R^2 decomposition (Table 23)	1.21/1.96/1.78 bits	Corrected	per-unit KLs computed under a display/canonical order mismatch; recomputed from the 1800 raw records under a single verified alignment	0.63/1.35/1.43 bits; identity-checked to machine precision

Table 23: Label × position × probability disentangling (576 units + 24 flat controls per model, finite-inventory strict-event readout). $m(p)$: mean mass on the letter stated at level p (uniform would be p). $P(t{=}\mathrm{arg})$: fraction of units whose top-1 letter carries the highest stated probability (0.25 under no following). $P(t{=}\mathrm{first})$: fraction of units whose top-1 letter is displayed first (0.25 under no position bias). R^2: one-hot least-squares variance decomposition of the top-1 letter's mass into label-identity (L), display-position (P), and joint (L+P) factors.

Model	$m(40)$	$m(30)$	$m(20)$	$m(10)$	$P(t{=}\mathrm{arg})$	$P(t{=}\mathrm{first})$	R^2_{L}	R^2_{P}	$R^2_{\mathrm{L+P}}$
llama31-8b-inst	0.587	0.251	0.104	0.059	0.668	0.104	0.045	0.113	0.178
qwen25_7-inst	0.685	0.215	0.080	0.020	0.688	0.517	0.015	0.021	0.041
qwen3_8-inst	0.476	0.248	0.164	0.114	0.474	0.212	0.035	0.039	0.090