

Alignment Sharpens Elicited Label Distributions under Logit and Sampled Elicitation

Christy Jacob Tharakan
BITS Pilani

Kohsheen Tikun
New York University

Abstract

Alignment may encourage a model to commit to one answer even when human judgments differ. We study this effect by comparing base and aligned checkpoints on ChaosNLI and Measuring Hate Speech using logit, sampled, and verbalized elicitation. Across eight model pairs from 7B to 32B, including a mixture-of-experts model, alignment lowers the entropy of logit-elicited label distributions. Sampled elicitation agrees on the three pairs tested. The effect survives prompt-format, answer-order, word-only, and task-compliance controls. Its implications for human fit are less straightforward: Jensen–Shannon distance worsens on ChaosNLI, but changes in both directions on the native MHS hate-speech task. Six of seven aligned checkpoints in the main sweep fit ChaosNLI annotations worse than a uniform predictor, yet temperature scaling to base mean entropy removes much of the base–aligned gap. Majority-label agreement also changes in both directions. Verbalized probabilities show a different pattern, with entropy increases and decreases across three families and substantial sensitivity to the prompt’s numeric example in some base checkpoints. Staged comparisons of two Ai2 lineages further show that the contributions of SFT and DPO depend on task and prompt format. These results support a consistent sharpening effect under logit and sampled elicitation, while showing why entropy alone cannot establish whether alignment reduces a model’s ability to represent human disagreement.

1 Introduction

People often disagree about whether one sentence entails another or whether a comment is hateful. Such disagreement can reflect legitimate differences in interpretation rather than noise around a single ground truth (Plank, 2022; Nie et al., 2020; Sap et al., 2022). In settings where these differences matter, a model serving a diverse population

may need to represent more than the majority view (Sorensen et al., 2024).

Instruction tuning and preference optimization make language models more useful (Ouyang et al., 2022; Rafailov et al., 2023), but may also narrow the judgments they express. Prior work reports reduced output diversity and uneven representation of human opinions after alignment (Kirk et al., 2024; Santurkar et al., 2023; Durmus et al., 2023). One way to study this narrowing is to elicit a distribution over task labels and measure its entropy. The difficulty is that the resulting distribution also depends on how we ask: prompt templates and answer surface forms can affect the probabilities assigned to labels (Holtzman et al., 2021; Zhao et al., 2021).

We examine this problem through matched base and aligned checkpoints from Qwen2.5, Qwen3, OLMo-2, and Ministral-3, using ChaosNLI and Measuring Hate Speech (MHS). Our study compares three ways of eliciting label distributions, tests sensitivity to prompt format and response compliance, and separates distributional spread from fit to human annotations and majority-label agreement. Staged comparisons of OLMo-2 and OLMoE examine where changes arise during training.

Alignment consistently sharpens logit-elicited distributions, and sampled elicitation confirms the direction on the tested pairs. This result holds across formats, answer orders, and label-word scoring conventions, including a completion format in which base models almost always provide valid answers. The interpretation is more limited than the entropy change alone suggests. Much of the ChaosNLI fit gap disappears after matching mean entropy, fit changes on native MHS depend on the model and format, and verbalized probabilities respond differently to alignment. The contribution is therefore a controlled account of when sharpening occurs and what it does, and does not, establish about the representation of human disagreement.

2 Related Work

Human disagreement and pluralism. Work on annotation disagreement argues for retaining variation in human judgments rather than reducing it to a single label (Plank, 2022; Sap et al., 2022; Uma et al., 2021). ChaosNLI makes this variation measurable through 100 annotations per item (Nie et al., 2020). The pluralism taxonomy of Sorensen et al. (2024) distinguishes goals that require representing a distribution of views from those that require acknowledging a range of reasonable answers. Since standard calibration measures can be misleading when annotators disagree (Baan et al., 2022), we compare model and annotator distributions directly.

Alignment, diversity, and opinion. Kirk et al. (2024) find that RLHF reduces output diversity relative to base and supervised models, while Padmakumar and He (2024) observe reduced diversity in human-LLM co-writing. The GPT-4 report also describes worse calibration after RLHF (OpenAI, 2023). Aligned models represent human opinions unevenly (Santurkar et al., 2023; Durmus et al., 2023), and can describe opinion distributions more accurately than they simulate them (Meister et al., 2025). Lake et al. (2024) interpret alignment as moving models from distributional toward Overton pluralism. Proposed training approaches address these limitations: Yao et al. (2025) introduce a group-distributional objective after finding that DPO unlearns minority preferences, and Zhang et al. (2024) show that producing diffuse distributions on request can require dedicated fine-tuning. Sociodemographic prompting alone does not reliably reproduce subjective human judgments (Sun et al., 2025).

Elicitation. Surface-form competition and content-free prompt biases complicate the interpretation of next-token answer probabilities (Holtzman et al., 2021; Zhao et al., 2021). Our study builds on these observations by comparing three elicitation methods across matched checkpoints, with controls for format, answer order, label wording, and compliance. We ask whether sharpening persists under these controls and how much of the accompanying fit change can be explained by sharpness.

3 Setup

Data. We use two datasets. ChaosNLI (Nie et al., 2020) provides approximately 100 human labels

per item for three-class NLI ($H_{\max} \approx 1.585$ bits). We use the MNLI-matched subset of the official v1.0 release (1,599 items), with label order entailment, neutral, contradiction. MHS (Sachdeva et al., 2022) supports two distinct probes. For binary offensiveness ($H_{\max}=1$ bit), we group ratings by comment, binarize at ≥ 2 , and retain comments with at least three annotators. We use this probe only for entropy because offensiveness differs from the corpus’s native hate-speech construct. Human fit on MHS is evaluated separately using its native three-way question (no/unclear/yes), with 400 comments sampled after annotator deduplication and a requirement of at least three valid ratings (Section 4.5).

Models. The main logit sweep contains seven base-aligned pairs: Qwen2.5 at 7B, 14B, and 32B (Qwen Team, 2024); Qwen3 dense 14B and mixture-of-experts 30B-A3B (Qwen Team, 2025); and OLMo-2 at 13B and 32B (OLMo Team, 2025). We compare all three elicitation methods on Qwen2.5-14B, OLMo-2-13B, and Ministral-3-8B. Staged analyses use the base, SFT, and DPO checkpoints of OLMo-2-7B and OLMoE-1B-7B-0125. Throughout, “aligned” denotes the family’s released instruction-tuned or preference-optimized checkpoint.

Elicitation. The *logit* probe presents lettered answer options and normalizes their next-token logits into a label distribution. The *sampled* probe estimates label frequencies from N stochastic generations. The *verbalized* probe asks for the probability that a random human annotator would choose each label, summing to one. It requests a JSON object, includes a numeric example, and uses greedy decoding. This last probe measures a stated distribution of human judgments, which can differ from the model’s own answer behavior (Meister et al., 2025).

The default logit condition applies the tokenizer’s chat template when available and otherwise uses a plain prompt ending in “Answer:”. All Qwen base and aligned checkpoints ship chat templates, so each Qwen pair uses the same format. OLMo-2 base checkpoints do not, making their default comparison raw-base versus chat-aligned. We test both checkpoints with templates disabled, reverse answer-letter order, and score label words under a shared raw prompt (Section 4.2). These are separate controls, not a full format $\times$ order factorial. The staged analyses also compare raw and

injected-template conditions (Section 4.7).

Items and decoding. Logit elicitation uses $n=400$ items per dataset. Sampled elicitation uses $n=150$ for Qwen2.5-14B and OLMo-2-13B and $n=400$ for Ministral-3-8B; verbalized elicitation uses $n=400$ for all three. Diagnostic and example-variant controls use $n=150$. For ChaosNLI and binary MHS, we take the first n items in fixed release order, after filtering MHS. These sets are identical across checkpoints, and the shared 150 items are exact prefixes of the 400-item sets, verified by item ID. Native three-way MHS instead uses a fixed-seed sample from sorted eligible comment IDs. Appendix G reports common-item comparisons and subset sensitivity.

Sampled elicitation uses $N=20$, temperature 0.7, top- p 0.95, and decoding seed 0, with independent-seed, decoding-grid, and sample-count controls in Appendix J. The logit probe requires no response parsing and uses a uniform fallback only if no label token exists. Sampled distributions use parsed answers, with a uniform fallback if all N fail; the compliance probe’s stricter handling is specified in Appendix L. Verbalized responses are parsed as JSON, with a label:number regex fallback and a uniform distribution on whole-response failure.

Metrics and uncertainty. We report mean Shannon entropy in bits, Jensen–Shannon distance (JSD, the metric form of the divergence; Lin, 1991), and, where indicated, total variation. Entropy measures spread, JSD measures fit to observed annotations, and majority agreement compares the model’s argmax with the human plurality label. Unless stated otherwise, 95% intervals use a percentile bootstrap over evaluation items with 10,000 resamples; paired differences resample items jointly and follow the aligned–base convention. Independent decoding seeds assess generation variability separately. Seed repetitions do not provide independent replications of deterministic logit or greedy verbalized outputs.

4 Results

4.1 Sharpening across families and sizes

Under the default logit probe, alignment reduces mean entropy on binary MHS from 0.42–0.86 bits to 0.01–0.12 across all seven pairs (Table 1); every paired 95% interval excludes zero (Appendix E). Qwen distributions become nearly one-hot ($\leq$

0.019 bits). OLMo-2 retains more entropy (0.109–0.124 bits), but still drops by 0.63–0.73 bits. Sharpening occurs throughout the tested size range and in the MoE model. We report absolute paired differences because the controls below show that base entropy, and hence any collapse ratio, depends on the probe.

On ChaosNLI, entropy falls from 1.06–1.39 bits to 0.08–0.15 for Qwen and 0.38–0.83 for OLMo-2. JSD worsens for all seven pairs (0.21–0.38 $\rightarrow$ 0.37–0.51; paired intervals in Appendix E). Section 4.4 examines how much of this fit gap remains after adjusting sharpness.

4.2 Prompt-format and verbalizer controls

We test whether sharpening depends on prompt format or answer-letter assignment (Table 3). Disabling the chat template lowers Qwen2.5-14B base entropy from 0.608 to 0.424 bits on MHS and from 1.389 to 1.072 on ChaosNLI (Appendix A). Reversing answer order raises Qwen2.5-7B base MHS entropy from 0.493 to 0.932. Despite this sensitivity, the MHS controls leave aligned entropy at 0.017–0.035 bits and base entropy at 0.42–0.93, a gap of at least 0.39 bits in every comparison. The broader controls in Appendix D agree in direction.

Reversing letters cannot by itself rule out a letter prior, since entropy is invariant to a pure permutation. We therefore score label words without answer letters, using the same raw prompt for base and aligned checkpoints (Appendix H). Entropy decreases in all four model–dataset cells under summed and length-normalized sequence scoring, both with and without content-free PMI calibration (Zhao et al., 2021); every 95% interval excludes zero.

Aligned JSD varies less across these formats: Qwen2.5-7B scores 0.507, 0.469, and 0.477 on ChaosNLI under chat, raw, and reversed conditions. This stability is checkpoint-specific; injecting a template into the OLMo-2-7B base changes its fit substantially (Section 4.7).

4.3 Task compliance under a shared completion format

High base entropy could also arise from failure to follow the task. We test this with an identical-prefix completion probe, tokenized directly without chat templates. Demonstrations are selected on held-out development items by maximum-minimum parse coverage across checkpoints, without using entropy or JSD, and frozen before evaluation. A strict

Model	MHS ($H_{\max}=1$)		ChaosNLI ($H_{\max}\approx 1.585$)	
	Base	Aligned	Base	Aligned
Qwen2.5-7B	0.493	0.019	1.348	0.094
Qwen2.5-14B	0.608	0.017	1.389	0.115
Qwen2.5-32B	0.420	0.013	1.059	0.081
Qwen3-14B	0.772	0.015	1.282	0.107
Qwen3-30B-A3B (MoE)	0.834	0.016	1.085	0.151
OLMo-2-13B	0.857	0.124	1.229	0.829
OLMo-2-32B	0.740	0.109	1.153	0.377

Table 1: Mean entropy (bits) under the default logit probe, $n=400$ per cell. Qwen pairs use chat templates on both sides; OLMo-2 uses raw base and chat aligned prompts. MHS is the binary offensiveness probe; ChaosNLI is three-class NLI. Aligned entropies are bold to highlight the sharpening effect; bold does not indicate better human fit. Values come from released per-item outputs. Every paired entropy difference is negative; bootstrap intervals are in Appendix E.

parser checks whether the response begins with an unambiguous answer, independently of its correctness (Appendix L).

Base checkpoints achieve 99.6–100% zero-shot parse coverage (Tables 26, 27). The only checkpoint below 99% is Qwen2.5-14B-instruct on ChaosNLI, at 86% zero-shot and 97% few-shot. Sharpening persists: zero-shot logit ΔH is -0.631 $[-0.664, -0.596]$ for Qwen and -0.321 $[-0.350, -0.294]$ for OLMo-2 on ChaosNLI, and -0.387 – -0.361 on MHS. Sampled ChaosNLI differences are -0.484 and -0.459 . Logit JSD still worsens ($+0.177$ for Qwen; $+0.014$ $[+0.001, +0.028]$ for OLMo-2), although the entropy gaps are smaller than under chat prompts.

Exhaustive reassignment of invalid samples preserves the entropy-drop sign for OLMo-2, which has no invalid completions, and for Qwen MHS (upper bounds ≤ -0.176). The Qwen ChaosNLI few-shot upper bound is also negative, -0.307 $[-0.391, -0.223]$. Its zero-shot counterpart, -0.009 $[-0.112, +0.095]$, does not certify the sign. For that cell, the logit result provides parse-free evidence of sharpening; the sampled result remains sensitive to missing responses. These bounds quantify reassignment sensitivity rather than recover unobserved answers.

4.4 Human-fit baselines and entropy matching

Two references help interpret ChaosNLI JSD: a uniform predictor scores 0.364, and a human-majority one-hot predictor, which assigns all mass to the observed plurality label, scores 0.441 (Table 2). Six base checkpoints beat uniform (0.21–0.26); Qwen3-30B-A3B is the exception (0.377). Six aligned checkpoints are worse than uniform (0.39–

0.51), while OLMo-2-13B is close to it (0.365).

To separate this gap from mean sharpness, we transform the aligned distribution by renormalizing $p^{1/\tau}$. A fixed split (permutation seed 0) assigns 200 items to calibration and 200 to evaluation. Bisection fits τ to match base mean entropy on calibration items, within 0.02 bits. The search permits $\tau < 1$, although no fit requires it. JSD residuals are measured on held-out items, with intervals conditional on the fitted τ and split.

The residual is positive for Qwen2.5-7B ($+0.021$ $[+0.005, +0.036]$) and OLMo-2 (13B $+0.088$; 32B $+0.021$), indistinguishable from zero for Qwen2.5-14B/32B and Qwen3-14B, and negative for the MoE pair (-0.088 $[-0.125, -0.052]$). Thus, a single temperature adjustment removes much of the observed fit gap. It matches only mean entropy, however, and neither establishes that label rankings are preserved nor isolates information loss. Residual differences may reflect rankings, item-specific confidence, or other distributional changes. This is a diagnostic transform of elicited probabilities, not a decoding intervention.

4.5 Human fit on native MHS annotations

We evaluate human fit on MHS using the native no/unclear/yes hate-speech question. The sample contains 400 seeded comments with at least three valid, deduplicated annotator ratings. Qwen2.5-14B and OLMo-2-13B pairs are evaluated under matched-raw, reversed-label, and default logit formats (Tables 32, 33). Entropy falls in all six comparisons, from 1.07–1.40 to 0.12–0.69 bits (ΔH -0.570 to -1.282 , all intervals excluding zero).

Fit changes depend on the model and format. Qwen JSD improves under reversed labels (-0.075 $[-0.100, -0.050]$) and the default format

Pair	Base _{all}	Aligned _{all}	Al.@ τ_{ho}	Resid. Δ_{ho} [95% CI]
Qwen2.5-7B	0.249	0.506	0.274	+0.021 [+0.005, +0.036]
Qwen2.5-14B	0.257	0.489	0.272	+0.009 [−0.003, +0.022]
Qwen2.5-32B	0.234	0.485	0.232	−0.002 [−0.023, +0.018]
Qwen3-14B	0.251	0.490	0.247	−0.002 [−0.021, +0.018]
Qwen3-30B-A3B	0.377	0.498	0.287	−0.088 [−0.125, −0.052]
OLMo-2-13B	0.220	0.365	0.301	+0.088 [+0.069, +0.107]
OLMo-2-32B	0.207	0.391	0.233	+0.021 [+0.008, +0.034]

Table 2: ChaosNLI JSD under the logit probe ($n=400$). Base/Aligned_{all} are full-set means; Al.@ τ_{ho} is the held-out mean after matching aligned to base mean entropy on 200 calibration items (seed 0; match within 0.02 bits). The residual compares adjusted aligned and base JSD on the same 200 held-out items, conditional on τ and the split. Held-out base means span 0.21–0.38 and are provided in the artifact CSV. Bold residuals have 95% intervals excluding zero. Uniform and human-majority one-hot references score 0.364 and 0.441.

(−0.071), while the raw-format difference is inconclusive (+0.018 [−0.009, +0.045]). OLMo-2 worsens under raw and default formats (+0.072 and +0.115), with no significant change under reversed labels (−0.002 [−0.018, +0.014]).

Most comments have only 3–4 ratings, although a few have hundreds, so the empirical targets are coarse. Checkpoint JSD ranges from 0.49 to 0.65, near the uniform reference (0.555) and far from the plurality reference (0.17, ties split equally). Redrawing each comment’s annotations, with the same draw for both checkpoints, moves Qwen’s raw-format mean Δ JSD from +0.018 to −0.014; OLMo-2’s degradation persists (+0.049). This measures sensitivity to sparse annotations, not a population-level reversal. Brier scores against individual annotations, averaged within comment first, worsen in all six paired comparisons. Held-out entropy matching leaves mixed-sign JSD residuals (Table 34). Sharpening therefore has different implications for fit across tasks, formats, and scoring rules.

4.6 Majority-label agreement

ChaosNLI majority agreement ranges from 0.45 to 0.71 for base checkpoints, above the 1/3 chance rate (Table 9). Their high entropy therefore coexists with task competence. Alignment leaves Qwen2.5 agreement approximately unchanged, lowers OLMo-2-13B agreement significantly (−0.15), and raises MoE agreement significantly (+0.17). No staged OLMo-2-7B transition is significant. Sharpening is thus compatible with gains, losses, or little change in plurality-label accuracy. Native MHS agreement is reported separately in Section 4.5; binary offensiveness has no validated human accuracy reference, so its diffuse base distributions cannot be taken as evidence of human

disagreement.

4.7 Attribution to SFT and DPO

We compare base, SFT, and DPO checkpoints from two Ai2 lineages. For OLMo-2-7B, SFT and DPO use byte-identical chat templates, while the base ships no template. We evaluate the base both raw and with the SFT template injected (Table 29), distinguishing training-stage changes from sensitivity to the base format.

OLMo-2 SFT→DPO lowers entropy on both datasets (ChaosNLI 0.965→0.751; MHS 0.343→0.188) and worsens ChaosNLI JSD by +0.049 [+0.040, +0.058]. Relative to the raw base, JSD changes by +0.005 [−0.009, +0.018] at SFT and +0.054 [+0.038, +0.069] at DPO, concentrating the fit degradation at DPO.

On binary MHS, the larger entropy drop occurs at SFT under either base format: 0.911→0.343 from raw and 0.535→0.343 from the injected template (ΔH −0.192 [−0.221, −0.162]). ChaosNLI depends on the reference format. SFT sharpens relative to the raw base (1.232→0.965), but template injection alone lowers base entropy to 0.718 and raises JSD from 0.288 to 0.482. Relative to that base, SFT increases entropy by +0.247 [+0.204, +0.290]. Since the base has no natural chat format and template injection puts it outside its training format, these checkpoints do not identify a format-independent ChaosNLI base→SFT effect.

For OLMoE-1B-7B-0125, we use the DPO end-point rather than the subsequent RLVR checkpoint and verify identical rendered prompts across all three stages under raw and BOS-free SFT-template formats (Appendix M). Under raw prompts, SFT accounts for nearly all sharpening: ChaosNLI ΔH is −0.289 at SFT and +0.001 at DPO (not

significant); MHS differences are -0.436 and -0.045 . Raw-format ChaosNLI JSD improves at DPO (-0.011 [$-0.017, -0.006$]), opposite to OLMo-2. Under the matched template, however, OLMoE’s MHS base→SFT entropy increases ($+0.072$ [$+0.045, +0.100$]).

Where the base→SFT comparison is robust to format, it sharpens more than SFT→DPO. Beyond this, the attribution depends on lineage, task, and format: each lineage has one task with a format-dependent SFT direction, and their DPO contributions differ. Two lineages from the same developer cannot establish a universal ordering of stage effects.

Item-level entropy correlation. On ChaosNLI, base model entropy correlates only modestly with human entropy across $n=400$ items (Spearman 0.08 – 0.36 ; Appendix E). Alignment lowers this correlation significantly for Qwen2.5-7B and OLMo-2-13B/32B; for 13B it falls from 0.364 to 0.040 ($\Delta -0.323$ [$-0.432, -0.211$]). The MoE correlation instead rises ($+0.231$ [$+0.107, +0.356$]), while the remaining differences are inconclusive (Qwen2.5-14B: $+0.084$ [$-0.018, +0.188$]). High entropy alone therefore says little about whether a model is uncertain on the same items as humans.

4.8 Verbalized probabilities

Verbalized probabilities do not follow a consistent sharpening pattern. Across the four Qwen2.5-14B and OLMo-2-13B cells at $n=150$, entropy increases in two, decreases in one, and has an interval spanning zero in one (Table 15). The latter does not establish equivalence. The 400-item results agree with narrower intervals, including OLMo-2 ChaosNLI $\Delta H +0.810$ [$+0.754, +0.865$]. Verbalized ChaosNLI JSD improves for OLMo-2 ($0.684 \rightarrow 0.328$) but worsens for Qwen ($+0.038$ [$+0.015, +0.062$]), unlike the shared degradation under logit and sampled elicitation.

Ministral-3-8B, evaluated with all three methods on 400 items, sharpens under logit elicitation (ChaosNLI -0.470 ; MHS -0.880), with sampled elicitation agreeing in direction. Its verbalized entropy also decreases on both tasks, but the base probability vector matches the prompt example on 355/400 ChaosNLI and 388/400 MHS items. When the example changes, 137–143/150 responses match the replacement vector; removing it reverses the ChaosNLI entropy-change sign. The

aligned checkpoint is comparatively stable. These are exact vector matches in canonical label order, which do not by themselves establish verbatim text copying (Table 18).

Example sensitivity also appears in OLMo-2 and Qwen. OLMo-2 base reproduces the MHS example in 140/150 responses, and example variants move its mean entropy by up to 0.5 bits. Removing the example reverses Qwen’s MHS entropy increase, while aligned values are more stable (Appendix I).

Per-label verbalized-logit Spearman correlations are strongest mainly on dominant labels of example-robust checkpoints (Table 19). Example-sensitive outputs can still correlate with logits: Ministral base ChaosNLI has significant correlations of 0.32 and 0.35 on two labels. Its MHS correlations and those of OLMo-2 base are compatible with zero. Top-label disagreement also occurs after alignment: OLMo-2-instruct agrees with its logit prediction on only 30% of MHS items. Together, these diagnostics show heterogeneous verbalized behavior across three families, with some effects dependent on the numeric example; they do not identify a mechanism such as learned hedging.

5 Discussion

The distinction between sharpening and representing disagreement is central to these results. Alignment reduces entropy under logit and sampled elicitation, including on items where human annotators disagree strongly (Appendix F). Yet lower entropy does not reliably imply worse majority agreement, and higher entropy does not ensure that a model tracks human uncertainty. The sharpness adjustment and native MHS results further show that the ChaosNLI fit gap cannot support a general claim that alignment worsens fit to human judgments.

Elicitation matters because the three probes ask different questions. Logits and samples measure label-selection behavior; verbalized probabilities ask the model to describe human responses. Their divergence is consistent with the describe/simulate distinction of Meister et al. (2025), while the example controls show that stated probabilities can also inherit the prompt’s values. Evaluations of pluralism should therefore specify the task, elicitation method, and target annotator distribution, and report fit alongside spread and competence.

Inference-time interventions offer limited improvements (Appendix K). Higher decoding temperature improves sampled ChaosNLI fit for

OLMo-2-13B (-0.072 JSD at $T=1.3$), but not significantly for Qwen2.5-14B. Asking for all reasonable labels (Lake et al., 2024) improves JSD under a fixed parser-constructed estimator, without establishing recovery of a human population distribution. These findings motivate training objectives that address distributional fit directly (Yao et al., 2025; Zhang et al., 2024).

6 Conclusion

Across eight base-aligned pairs from three developers, alignment sharpens logit-elicited label distributions; sampled elicitation agrees on the three pairs tested. Format, verbalizer, and compliance controls support this direction, while staged comparisons show that its attribution to SFT and DPO varies with lineage, task, and format. Sharpening has no single implication for human fit: mean-entropy matching removes much of the ChaosNLI JSD gap, native MHS changes are mixed, and verbalized probabilities show heterogeneous, sometimes example-sensitive behavior. Assessing pluralism therefore requires direct comparisons with human distributions under the intended task and elicitation method, rather than entropy alone.

7 Limitations

Each probe estimates a distribution from model behavior, rather than exposing an internal ground-truth distribution. Our format and answer-order controls are not a full factorial design, and absolute entropy remains probe-dependent. The verbalized results cover three families, with diagnostics that constrain possible explanations without identifying a mechanism.

The study covers two English datasets with two- or three-label tasks. ChaosNLI and binary MHS use release-order prefixes rather than probability samples of the corpora. Random-subset checks support stable paired differences within these pools (Appendix G); native three-way MHS instead uses a seeded sample. Its targets usually contain only 3–4 ratings per comment, so JSD measures fit to coarse observed annotations. The binary offensiveness probe has no validated human-fit target.

Response validity does not establish comprehension. The compliance probe’s 99% coverage threshold is operational, and its bounds address reassignment of invalid responses, not every form of label-dependent missingness. Raw completions are available for the compliance probe, verbalized

diagnostics, and Ministral generation runs. The archived Qwen/OLMo sampled, independent-seed, and decoding-grid runs retain only parsed labels and counts, preventing raw-text revalidation; their bounds rely on stored sample classifications.

Temperature scaling matches mean entropy without controlling item-specific spread, and its intervals are conditional on the fitted temperature and calibration split. It is a diagnostic rather than a deployment remedy. The Overton estimator constructs a distribution from parsed labels; its prompt and weight comparisons do not validate that distribution as a population estimate.

Stage attribution covers two lineages from one developer. Their SFT and DPO effects differ, and the base→SFT direction depends on prompt format for one task in each lineage. Cross-family comparisons also combine released checkpoints whose training recipes and data are not always disclosed. Neither stage ordering nor effect magnitude should be assumed to generalize beyond these conditions.

Ethical Considerations

We use publicly released datasets and open checkpoints, with no new human annotations. MHS contains offensive and hateful text, which we process only to elicit label distributions and neither redistribute nor generate. Our findings are descriptive: higher entropy is not inherently desirable, and agreement with an annotator distribution does not establish fairness or population representativeness. Whether a system should reflect diverse judgments or commit to a consensus depends on its application and on whose perspectives the target distribution represents.

References

- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1892–1915. Association for Computational Linguistics.
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#). *arXiv preprint arXiv:2306.16388*.

- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. [Surface form competition: Why the highest probability answer isn't always right](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7038–7051. Association for Computational Linguistics.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. [Understanding the effects of rlhf on llm generalisation and diversity](#). In *The Twelfth International Conference on Learning Representations (ICLR)*. ArXiv:2310.06452.
- Thom Lake, Eunsol Choi, and Greg Durrett. 2024. [From distributional to overton pluralism: Investigating large language model alignment](#). *arXiv preprint arXiv:2406.17692*.
- Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. [Benchmarking distributional alignment of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1: Long Papers*, pages 24–49. Association for Computational Linguistics. ArXiv:2411.05403.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. [What can we learn from collective human opinions on natural language inference data?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143. Association for Computational Linguistics.
- OLMo Team. 2025. [2 OLMo 2 furious](#). *arXiv preprint arXiv:2501.00656*.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2203.02155.
- Vishakh Padmakumar and He He. 2024. [Does writing with language models reduce content diversity?](#) In *The Twelfth International Conference on Learning Representations (ICLR)*. ArXiv:2309.05196.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10671–10682. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2305.18290.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP (NLPerspectives) at LREC 2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. ArXiv:2303.17548.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. [A roadmap to pluralistic alignment](#). In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. ArXiv:2402.05070.
- Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. [The effect of sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*. ArXiv:2311.09730.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2025. [No preference left behind: Group distributional preference optimization](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv:2412.20299.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J. Zico Kolter, and Daphne Ippolito. 2024. [Forcing diffuse distributions out of language models](#). In *First Conference on Language Modeling (COLM)*. ArXiv:2404.10859.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Appendix

The following sections provide the detailed measurements, uncertainty estimates, and controls supporting the main analysis. We group results by the question they address: prompt sensitivity, elicitation and parsing, human disagreement, decoding, compliance, training stages, and native MHS fit. Table captions specify the estimator, sample size, and interpretation of each comparison.

A Additional entropy values

Tables 3 and 4 compare the default, raw, and reversed-answer conditions for binary MHS and ChaosNLI. They show how prompt choices affect entropy while preserving the base-aligned direction.

Condition	Qwen2.5-7B	Qwen2.5-14B
base + chat (default)	0.493	0.608
base + raw (forced)	0.534	0.424
base + reversed	0.932	0.759
aligned + chat (default)	0.019	0.017
aligned + raw (forced)	0.035	0.034
aligned + reversed	0.020	0.023

Table 3: Mean entropy on binary MHS under format and answer-order controls ($n=400$ per cell; not a full factorial). Default uses chat templates on both sides of each Qwen pair; forced-raw disables them. Aligned entropy remains lower throughout, although base entropy varies from 0.42 to 0.93 bits.

Table 4 reports ChaosNLI entropy ($H_{\max} \approx 1.585$ bits) under the format and answer-order controls used for MHS. Raw conditions disable the template for both checkpoints. Aligned entropy remains below base in every comparison.

Condition	Qwen2.5-7B	Qwen2.5-14B
base + chat (default)	1.348	1.389
base + raw (forced)	0.884	1.072
base + reversed	1.399	1.320
aligned + raw	0.211	0.371
aligned + reversed	0.157	0.062

Table 4: ChaosNLI (three-class, $H_{\max} \approx 1.585$ bits) entropy under the prompt-format and verbalizer-order controls ($n=400$ per cell). Base entropy is format-sensitive but stays far above aligned in every condition; the collapse is not specific to the binary MHS task.

B Three-elicitation comparison

Table 5 compares Qwen2.5-14B and OLMo-2-13B using logit ($n=400$), sampled ($n=150$, $N=20$), and verbalized ($n=150$) elicitation. Table 6 restricts all three methods to the shared first 150

items, verified by item ID. The same pattern holds on these common items: alignment lowers entropy on both datasets and worsens ChaosNLI JSD under logit and sampled elicitation, whereas verbalized changes vary across cells. Binary MHS is reported only for entropy.

C Sampled parse rates and missingness

Sampled elicitation drops responses that cannot be parsed and uses a uniform distribution only when all N responses fail. This can bias estimates if missingness depends on the label. The logit probe has no per-response parsing step, so its sharpening result is independent of this issue, but it does not establish that sampled estimates are unbiased.

Table 8 measures parse coverage over 3,000 completions per cell ($n=150$, $N=20$, temperature 0.7, top- p 0.95). Both aligned checkpoints and OLMo-2-13B base parse at 100%, with effective $N=20.0$. Qwen2.5-14B base parses at 86.9% on ChaosNLI (effective $N=17.4$) and 95.4% on MHS (19.1). No item in any cell uses the all-fail uniform fallback.

Aligned parse failure therefore cannot explain near-zero aligned entropy. Smaller effective sample counts ordinarily bias the plug-in entropy estimate downward, which would reduce the observed gap by understating base entropy. Label-dependent missingness could nevertheless shift the base estimate in either direction. We retain that qualification for sampled results and treat the logit results as separate, parse-free evidence.

D Additional robustness and majority agreement

The MHS controls also cover OLMo-2-13B and Qwen2.5-32B. Under reversed answer order, base and aligned entropy are 0.683 and 0.103 bits for OLMo-2, and 0.524 and 0.011 for Qwen. Raw prompting gives aligned entropy of 0.301 and 0.072 bits, respectively. Aligned entropy remains below base in each condition, although the gap varies with the probe.

Table 9 reports majority-label agreement on

Metric	Model / Dataset	Base $\rightarrow$ Aligned		
		Logit	Sampled	Verbalized
H	Qwen2.5-14B ChaosNLI	1.389 $\rightarrow$ 0.115	1.022 $\rightarrow$ 0.083	1.046 $\rightarrow$ 1.043
	Qwen2.5-14B MHS	0.608 $\rightarrow$ 0.017	0.381 $\rightarrow$ 0.007	0.437 $\rightarrow$ 0.612
	OLMo-2-13B ChaosNLI	1.229 $\rightarrow$ 0.829	0.860 $\rightarrow$ 0.486	0.270 $\rightarrow$ 1.142
	OLMo-2-13B MHS	0.857 $\rightarrow$ 0.124	0.712 $\rightarrow$ 0.104	0.937 $\rightarrow$ 0.383
JSD (ChaosNLI)	Qwen2.5-14B ChaosNLI	0.257 $\rightarrow$ 0.489	0.288 $\rightarrow$ 0.503	0.273 $\rightarrow$ 0.311
	OLMo-2-13B ChaosNLI	0.220 $\rightarrow$ 0.365	0.290 $\rightarrow$ 0.453	0.684 $\rightarrow$ 0.328

Table 5: Base $\rightarrow$ aligned comparisons for Qwen2.5-14B and OLMo-2-13B: logit ($n=400$), sampled ($n=150$, $N=20$), and verbalized ($n=150$). The 150-item sets are exact prefixes of the logit sets; Table 6 compares all methods on these shared items. Entropy is in bits. JSD is reported only for ChaosNLI because binary MHS measures offensiveness. Logit and sampled elicitation show lower entropy and worse ChaosNLI fit after alignment; verbalized changes differ across cells.

Pair	Dataset	Method	H : B $\rightarrow$ A	JSD: B $\rightarrow$ A
Qwen2.5-14B	ChaosNLI	logit	1.393 $\rightarrow$ 0.118	0.245 $\rightarrow$ 0.498
		sampled	1.022 $\rightarrow$ 0.083	0.288 $\rightarrow$ 0.503
		verbalized	1.046 $\rightarrow$ 1.043	0.273 $\rightarrow$ 0.311
Qwen2.5-14B	MHS	logit	0.607 $\rightarrow$ 0.012	—
		sampled	0.381 $\rightarrow$ 0.007	—
		verbalized	0.437 $\rightarrow$ 0.612	—
OLMo-2-13B	ChaosNLI	logit	1.236 $\rightarrow$ 0.810	0.217 $\rightarrow$ 0.374
		sampled	0.860 $\rightarrow$ 0.486	0.290 $\rightarrow$ 0.453
		verbalized	0.270 $\rightarrow$ 1.142	0.684 $\rightarrow$ 0.328
OLMo-2-13B	MHS	logit	0.863 $\rightarrow$ 0.149	—
		sampled	0.712 $\rightarrow$ 0.104	—
		verbalized	0.937 $\rightarrow$ 0.383	—

Table 6: The three elicitation methods computed on the *identical* shared 150 items (the common prefix). The cross-method pattern of Table 5 is unchanged, so the elicitation differences are not an item-sampling artifact. MHS is an entropy probe; its human-reference columns are omitted.

Pair	Dataset	logit	sampled	verbalized
Qwen2.5-14B	ChaosNLI	-1.2742 [-1.3016, -1.2465]	-0.9388 [-1.0105, -0.8635]	-0.0036 [-0.0405, +0.0341]
OLMo-2-13B	ChaosNLI	-0.4000 [-0.4389, -0.3610]	-0.3746 [-0.4668, -0.2840]	+0.8727 [+0.7881, +0.9546]
Ministral-3-8B	ChaosNLI	-0.4621 [-0.5076, -0.4179]	-0.6536 [-0.7179, -0.5902]	-0.3041 [-0.3627, -0.2401]
Qwen2.5-14B	MHS	-0.5904 [-0.6149, -0.5658]	-0.3744 [-0.4341, -0.3134]	+0.1745 [+0.1336, +0.2140]
OLMo-2-13B	MHS	-0.7336 [-0.7574, -0.7088]	-0.6085 [-0.6628, -0.5539]	-0.5542 [-0.5888, -0.5195]
Ministral-3-8B	MHS	-0.8722 [-0.9103, -0.8294]	-0.8690 [-0.9066, -0.8286]	-0.6352 [-0.6596, -0.6101]

Table 7: Three families compared across elicitation methods: paired ΔH (aligned–base, bits, 95% intervals). Qwen and OLMo use 150 items for sampled and verbalized estimates and 400 for logits. Ministral-3-8B uses the matched 150-item subset here; its full-400 results are in the unified artifact table.

Model	Ckpt	ChaosNLI (%/eff. N)	MHS (%/eff. N)
Qwen2.5-14B	base	86.9 / 17.4	95.4 / 19.1
	aligned	100 / 20.0	100 / 20.0
OLMo-2-13B	base	100 / 20.0	100 / 20.0
	aligned	100 / 20.0	100 / 20.0

Table 8: Sampled parse rate and mean effective N (valid responses per item), with $n=150$ and $N=20$ (3,000 completions per cell). There are no all-fail items. Qwen2.5-14B base is the only checkpoint below 100% coverage; label-dependent missingness may affect its estimate in either direction. Aligned checkpoints discard no samples.

Pair	ChaosNLI	
	B→A	Δ [95% CI]
Qwen2.5-7B	.63→.62	-.01 [-.06, +.04]
Qwen2.5-14B	.63→.63	-.01 [-.05, +.04]
Qwen2.5-32B	.64→.69	+.05 [-.01, +.10]
Qwen3-14B	.71→.66	-.05 [-.10, +.01]
Qwen3-30B-A3B	.45→.62	+.17 [+ .10, +.24]
OLMo-2-13B	.71→.56	-.15 [-.21, -.10]
OLMo-2-32B	.71→.70	-.01 [-.05, +.03]

Table 9: ChaosNLI majority-label agreement under the default logit probe ($n=400$, paired bootstrap intervals; chance 1/3). Staged OLMo-2-7B agreement is $0.580 \rightarrow 0.585 \rightarrow 0.560$, with no significant transition. Binary MHS agreement is omitted because its annotations do not validate the offensiveness construct; those values remain in the archived CSVs.

Pair	Dataset	H base	H aligned	ΔH [95% CI]	JSD base	JSD aligned	ΔJSD [95% CI]
Qwen2.5-7B	ChaosNLI	1.348	0.094	-1.254 [-1.280, -1.228]	0.249	0.506	0.257 [0.238, 0.276]
Qwen2.5-7B	MHS	0.493	0.019	-0.475 [-0.498, -0.452]	—	—	—
Qwen2.5-14B	ChaosNLI	1.389	0.115	-1.274 [-1.301, -1.246]	0.257	0.489	0.232 [0.213, 0.251]
Qwen2.5-14B	MHS	0.608	0.017	-0.590 [-0.615, -0.566]	—	—	—
Qwen2.5-32B	ChaosNLI	1.059	0.081	-0.979 [-1.005, -0.952]	0.234	0.485	0.251 [0.233, 0.268]
Qwen2.5-32B	MHS	0.420	0.013	-0.407 [-0.435, -0.381]	—	—	—
Qwen3-14B	ChaosNLI	1.282	0.107	-1.175 [-1.205, -1.143]	0.251	0.490	0.240 [0.221, 0.258]
Qwen3-14B	MHS	0.772	0.015	-0.757 [-0.777, -0.736]	—	—	—
Qwen3-30B-A3B	ChaosNLI	1.085	0.151	-0.933 [-0.971, -0.893]	0.377	0.498	0.120 [0.093, 0.147]
Qwen3-30B-A3B	MHS	0.834	0.016	-0.818 [-0.835, -0.799]	—	—	—
OLMo-2-13B	ChaosNLI	1.229	0.829	-0.400 [-0.439, -0.361]	0.220	0.365	0.145 [0.129, 0.160]
OLMo-2-13B	MHS	0.857	0.124	-0.734 [-0.757, -0.709]	—	—	—
OLMo-2-32B	ChaosNLI	1.153	0.377	-0.776 [-0.807, -0.743]	0.207	0.391	0.184 [0.170, 0.198]
OLMo-2-32B	MHS	0.740	0.109	-0.631 [-0.662, -0.601]	—	—	—

Table 10: Paired bootstrap 95% intervals for default logit estimates ($n=400$; 10,000 joint item resamples). Every ΔH interval excludes zero, as does every ChaosNLI ΔJSD interval. Binary MHS measures offensiveness, so human-reference columns are omitted; the archived CSVs retain them.

Pair	n	ρ base	ρ aligned	aligned-base [95% CI]
Qwen2.5-7B	400	0.244 [0.152, 0.337]	0.031 [-0.060, 0.130]	-0.213 [-0.336, -0.077]
Qwen2.5-14B	400	0.221 [0.119, 0.317]	0.305 [0.210, 0.388]	+0.084 [-0.018, +0.188]
Qwen2.5-32B	400	0.252 [0.153, 0.343]	0.279 [0.186, 0.370]	+0.028 [-0.089, +0.144]
Qwen3-14B	400	0.258 [0.161, 0.350]	0.326 [0.230, 0.411]	+0.068 [-0.035, +0.172]
Qwen3-30B-A3B	400	0.083 [-0.017, 0.185]	0.314 [0.220, 0.401]	+0.231 [+0.107, +0.356]
OLMo-2-13B	400	0.364 [0.274, 0.453]	0.040 [-0.060, 0.141]	-0.323 [-0.432, -0.211]
OLMo-2-32B	400	0.317 [0.219, 0.406]	0.209 [0.110, 0.302]	-0.107 [-0.185, -0.023]

Table 11: Spearman correlation between per-item model entropy and human entropy on ChaosNLI, with bootstrap 95% CIs (2,000 resamples) and the CI of the aligned-base *difference* on paired resamples. The alignment effect on the correlation is significantly negative for OLMo-2-13B/32B and Qwen2.5-7B, significantly positive for Qwen3-30B-A3B, and not distinguishable from zero for the remaining pairs.

ChaosNLI, complementing the distributional comparisons in Section 4.6.

E Bootstrap intervals and entropy correlations

Table 10 gives paired bootstrap 95% intervals for each default-probe estimate and difference. Table 11 reports model-human entropy correlations, sample sizes, and paired difference intervals. Tables are generated from saved per-item measurements; each generated table’s source CSV is recorded in its TeX file.

F Deltas stratified by human disagreement

Table 12 stratifies the ChaosNLI base→aligned deltas by the human-agreement bin. Sharpening occurs on highly disputed items just as strongly as on high-agreement items (the ΔH rows are near-constant within each pair), while the JSD degradation concentrates on the contested strata.

G Item-subset sensitivity and matched-item comparison

Table 13 quantifies how the paired deltas vary across random 150-of-400 item subsets of the logit pool, and locates the actual first-150 prefix within

Pair	Bin	n	H : B→A	ΔH [95% CI]	ΔJSD [95% CI]	ΔAcc [95% CI]
Qwen2.5-7B	low	76	1.294→0.095	-1.199 [-1.258, -1.135]	+0.050 [+0.005, +0.098]	+0.026 [-0.079, +0.132]
	med	182	1.342→0.074	-1.268 [-1.300, -1.234]	+0.273 [+0.249, +0.296]	-0.027 [-0.099, +0.044]
	high	142	1.385→0.118	-1.267 [-1.314, -1.216]	+0.348 [+0.325, +0.370]	+0.000 [-0.085, +0.085]
Qwen2.5-14B	low	76	1.317→0.032	-1.285 [-1.324, -1.243]	+0.061 [+0.021, +0.103]	-0.026 [-0.105, +0.039]
	med	182	1.393→0.119	-1.274 [-1.314, -1.230]	+0.229 [+0.205, +0.252]	-0.033 [-0.104, +0.033]
	high	142	1.423→0.154	-1.269 [-1.318, -1.215]	+0.327 [+0.301, +0.354]	+0.042 [-0.042, +0.120]
Qwen2.5-32B	low	76	0.965→0.037	-0.928 [-0.989, -0.862]	+0.118 [+0.077, +0.158]	+0.118 [+0.026, +0.211]
	med	182	1.058→0.070	-0.988 [-1.025, -0.950]	+0.270 [+0.244, +0.297]	-0.005 [-0.082, +0.071]
	high	142	1.112→0.118	-0.994 [-1.037, -0.949]	+0.297 [+0.274, +0.319]	+0.070 [-0.014, +0.155]
Qwen3-14B	low	76	1.144→0.060	-1.084 [-1.149, -1.018]	+0.116 [+0.071, +0.163]	-0.118 [-0.197, -0.039]
	med	182	1.293→0.114	-1.179 [-1.223, -1.133]	+0.242 [+0.217, +0.267]	-0.060 [-0.132, +0.011]
	high	142	1.341→0.123	-1.218 [-1.270, -1.165]	+0.303 [+0.274, +0.331]	+0.007 [-0.092, +0.106]
Qwen3-30B-A3B	low	76	1.020→0.073	-0.947 [-1.026, -0.864]	+0.065 [-0.010, +0.140]	+0.158 [+0.013, +0.303]
	med	182	1.098→0.161	-0.937 [-0.997, -0.875]	+0.127 [+0.087, +0.168]	+0.148 [+0.044, +0.247]
	high	142	1.102→0.181	-0.921 [-0.978, -0.859]	+0.142 [+0.106, +0.178]	+0.204 [+0.085, +0.324]
OLMo-2-13B	low	76	1.099→0.826	-0.273 [-0.368, -0.180]	+0.138 [+0.100, +0.175]	-0.289 [-0.395, -0.184]
	med	182	1.232→0.804	-0.428 [-0.484, -0.370]	+0.139 [+0.117, +0.161]	-0.121 [-0.192, -0.049]
	high	142	1.296→0.864	-0.432 [-0.494, -0.370]	+0.156 [+0.130, +0.182]	-0.120 [-0.218, -0.021]
OLMo-2-32B	low	76	1.003→0.279	-0.724 [-0.777, -0.669]	+0.117 [+0.082, +0.149]	-0.079 [-0.158, -0.013]
	med	182	1.146→0.342	-0.804 [-0.851, -0.755]	+0.184 [+0.163, +0.204]	+0.000 [-0.049, +0.049]
	high	142	1.242→0.475	-0.767 [-0.822, -0.711]	+0.220 [+0.198, +0.242]	+0.014 [-0.056, +0.085]

Table 12: ChaosNLI aligned—base differences by human disagreement, defined by plurality vote share: low ≥ 0.8 , medium $[0.6, 0.8)$, high < 0.6 . Intervals use paired item bootstraps. Entropy falls by similar amounts across strata, while JSD degradation grows with disagreement and majority-agreement changes vary across pairs.

Pair	Dataset	Metric	Full 400	First 150	ptile	Subsample mean $\pm$ sd	[2.5%, 97.5%]
Qwen2.5-14B	ChaosNLI	dH	-1.2742	-1.2749	48	-1.2743 $\pm$ 0.0190	[-1.3099, -1.2373]
Qwen2.5-14B	ChaosNLI	dJSD	+0.2319	+0.2526	95	+0.2327 $\pm$ 0.0122	[+0.2109, +0.2560]
Qwen2.5-14B	MHS	dH	-0.5904	-0.5958	36	-0.5896 $\pm$ 0.0163	[-0.6217, -0.5574]
OLMo-2-13B	ChaosNLI	dH	-0.4000	-0.4262	16	-0.3994 $\pm$ 0.0265	[-0.4506, -0.3462]
OLMo-2-13B	ChaosNLI	dJSD	+0.1450	+0.1571	89	+0.1448 $\pm$ 0.0104	[+0.1235, +0.1645]
OLMo-2-13B	MHS	dH	-0.7336	-0.7141	91	-0.7345 $\pm$ 0.0154	[-0.7655, -0.7036]

Table 13: Subset sensitivity from 1,000 seeded, paired 150-of-400 logit subsamples. The first-150 prefix used for sampled and verbalized comparisons is shown with its percentile in this distribution. This assesses item-subset variability, not independent replication. Directions are stable; the largest prefix effect is Qwen2.5-14B ChaosNLI ΔJSD (95th percentile), a +0.02 shift. Binary MHS JSD is omitted but retained in the archived CSV.

that distribution; Table 6 recomputes the three-elicitation comparison on the identical shared 150 items.

H Semantic-verbalizer control

The semantic-verbalizer control removes answer letters and uses the same raw prompt for base and aligned checkpoints. It lists “entailment, neutral, contradiction” for ChaosNLI or “offensive, not offensive” for binary MHS, asks “Answer with exactly one label”, and contains neither letter options nor the word “letter”. We score candidate label words by full-sequence log-probability and normalize across labels. Scores are either summed or normalized by token length, each with or without content-free PMI calibration (Zhao et al., 2021). Calibration subtracts the score from a prompt containing the instructions and label list but no item content (equations in Appendix N). Table 14 reports length-normalized results for Qwen2.5-14B and OLMo-2-13B on both datasets, using $n=400$ and 10,000 paired percentile-bootstrap resamples.

Appendix N reports all four scoring conventions.

Every model–dataset cell sharpens under all four scoring conventions, with all 95% intervals excluding zero. Length-normalized ChaosNLI entropy falls from 0.958 to 0.312 for Qwen (ΔH -0.646 [-0.688, -0.604]) and from 0.861 to 0.511 for OLMo-2 (-0.350 [-0.380, -0.319]). PMI calibration yields drops of -0.702 and -0.385, respectively, making a simple content-free label-prior explanation unlikely. Summed scores agree in direction. ChaosNLI JSD increases by +0.08 to +0.18, and binary MHS entropy decreases by -0.13 to -0.38, with all intervals excluding zero. Sharpening therefore persists when scoring label words directly, after matching formats and accounting for token length.

I Verbalized-output diagnostics

The raw-response diagnostics use greedy generations whose parsed distributions exactly reproduce the archived per-item vectors (maximum L_∞ difference 0.0000 across all eight cells). Table 16

Family	Dataset	entropy (bits)		ΔH raw	ΔH content-free	ΔJSD raw
		base	aligned	(aligned–base) [95% CI]	(aligned–base) [95% CI]	(aligned–base) [95% CI]
Qwen2.5-14B	ChaosNLI	0.958	0.312	−0.646 [−0.688, −0.604]	−0.702 [−0.738, −0.666]	+0.181 [+0.164, +0.197]
	MHS	0.481	0.149	−0.331 [−0.356, −0.307]	−0.378 [−0.408, −0.348]	—
OLMo-2-13B	ChaosNLI	0.861	0.511	−0.350 [−0.380, −0.319]	−0.385 [−0.415, −0.356]	+0.083 [+0.071, +0.095]
	MHS	0.865	0.739	−0.126 [−0.144, −0.108]	−0.272 [−0.291, −0.254]	—

Table 14: Word-only semantic-verbalizer control under a shared raw format. Values use length-normalized full-sequence log-probabilities ($n=400$; 10,000 paired bootstrap resamples; 95% intervals). Entropy decreases in every cell with and without PMI calibration; summed scoring agrees (Table 31). ChaosNLI JSD worsens for both families. Binary MHS is an entropy-only probe, so its JSD is omitted.

Model	Dataset	H : B→A	ΔH [95% CI]
Qwen2.5-14B	MHS	0.437→0.612	+0.175 [+0.135, +0.214]
Qwen2.5-14B	ChaosNLI	1.046→1.043	−0.004 [−0.041, +0.035]
OLMo-2-13B	ChaosNLI	0.270→1.142	+0.873 [+0.790, +0.955]
OLMo-2-13B	MHS	0.937→0.383	−0.554 [−0.589, −0.519]

Table 15: Verbalized probabilities for Qwen2.5-14B and OLMo-2-13B ($n=150$; paired bootstrap intervals). Entropy increases in two cells, decreases in one, and has an interval spanning zero in one.

Model	Dataset	full/part/fail	distinct	top vec	ex. copies	agree _{logit}	H all	H full-only
olmo2-13b-base	ChaosNLI	140/10/0	5	73%	5	31%	0.2695	0.2673
olmo2-13b-base	MHS	150/0/0	2	93%	140	18%	0.9375	0.9375
olmo2-13b-instruct	ChaosNLI	150/0/0	16	34%	0	69%	1.1422	1.1422
olmo2-13b-instruct	MHS	150/0/0	7	45%	0	30%	0.3833	0.3833
qwen2.5-14b-base	ChaosNLI	150/0/0	13	41%	2	69%	1.0464	1.0464
qwen2.5-14b-base	MHS	139/0/11	9	40%	0	82%	0.4371	0.3926
qwen2.5-14b-instruct	ChaosNLI	150/0/0	15	21%	0	74%	1.0428	1.0428
qwen2.5-14b-instruct	MHS	150/0/0	11	35%	0	97%	0.6117	0.6117

Table 16: Verbalized-output diagnostics on the stored raw completions ($n=150$ /cell, greedy). “full/part/fail”: complete JSON parses with all labels / partial parses (regex or missing labels) / failures that fall back to uniform. “distinct”: number of distinct probability vectors emitted; “top vec”: share of the most frequent vector; “ex. copies”: outputs identical to the numeric example embedded in the prompt; “agree_{logit}”: top-label agreement with the same checkpoint’s logit distribution on the same items; the last columns give mean entropy over all items vs. complete parses only (exclusion sensitivity). Nearly all vectors are multiples of 0.05.

records complete, partial, and failed parses; the number and frequency of distinct vectors; matches to the prompt example; top-label agreement with the same checkpoint’s logit distribution; and entropy sensitivity to excluding incomplete parses.

Complete-parse coverage is at least 92.7%. The largest number of uniform fallbacks is 11/150, for Qwen2.5-14B base MHS, and excluding incomplete parses changes mean entropy by at most 0.044 bits. Each cell emits only 2–16 distinct probability vectors across 150 items, almost all on a 0.05 grid.

OLMo-2-13B base returns the prompt example verbatim on 140/150 MHS items and places 73% of ChaosNLI outputs on one stock vector. Aligned checkpoints in this comparison do not copy the example. Changing or removing it (Table 17) shifts base outputs substantially: OLMo-2 base ChaosNLI entropy moves from 0.270 to 0.792 un-

der the alternative example, with 77/150 copies of that example, and to 0.567 with no example. Qwen base MHS rises from 0.437 to 0.641 without an example. Aligned OLMo-2 changes are smaller (ChaosNLI 1.142 → 1.097; MHS 0.383 → 0.377). Both OLMo-2 entropy-change directions persist, but Qwen’s MHS increase reverses without an example (0.641 → 0.557).

Top-label agreement with the logit probe is 18–31% for OLMo-2 base and 69–97% for Qwen and OLMo-2-instruct on ChaosNLI. OLMo-2-instruct agrees on only 30% of MHS items, so disagreement between stated and answer distributions is not confined to base checkpoints. These observations identify example sensitivity and differences between elicitation methods, without establishing their causal mechanism.

Model	Dataset	variant	n	H	JSD	Agree	Parse
olmo2-13b-base	ChaosNLI	example2	150	0.792	0.551	0.300	—
olmo2-13b-base	ChaosNLI	noexample	150	0.567	0.597	0.327	—
olmo2-13b-base	MHS	example2	150	0.920	—	—	—
olmo2-13b-base	MHS	noexample	150	0.930	—	—	—
olmo2-13b-instruct	ChaosNLI	example2	150	1.097	0.380	0.333	—
olmo2-13b-instruct	ChaosNLI	noexample	150	1.203	0.355	0.373	—
olmo2-13b-instruct	MHS	example2	150	0.377	—	—	—
olmo2-13b-instruct	MHS	noexample	150	0.446	—	—	—
qwen2.5-14b-base	ChaosNLI	example2	150	1.013	0.297	0.607	—
qwen2.5-14b-base	ChaosNLI	noexample	150	1.121	0.252	0.647	—
qwen2.5-14b-base	MHS	example2	150	0.414	—	—	—
qwen2.5-14b-base	MHS	noexample	150	0.641	—	—	—
qwen2.5-14b-instruct	ChaosNLI	example2	150	1.024	0.343	0.473	—
qwen2.5-14b-instruct	ChaosNLI	noexample	150	1.073	0.293	0.547	—
qwen2.5-14b-instruct	MHS	example2	150	0.586	—	—	—
qwen2.5-14b-instruct	MHS	noexample	150	0.557	—	—	—

Table 17: Greedy verbalized elicitation with a changed numeric example (example2, different values and argmax) or no example (noexample), $n=150$. Default-example means appear in Table 15. Whole-response parse failures use the same uniform fallback as the main protocol.

Model	Dataset	Variant	n	f/p/inv	unif	dist.	top	ex.match	agr _{lg}	H
qwen2.5-14b-base	ChaosNLI	orig_example	150	150/0/0	0	13	0.407	2	0.693	1.0464
qwen2.5-14b-base	ChaosNLI	ext250	250	250/0/0	0	15	0.372	0	0.700	0.9912
qwen2.5-14b-base	MHS	orig_example	150	139/0/11	0	9	0.396	0	0.878	0.3926
qwen2.5-14b-base	MHS	ext250	250	233/0/17	0	8	0.378	0	0.893	0.3568
qwen2.5-14b-instruct	ChaosNLI	orig_example	150	150/0/0	0	15	0.213	0	0.740	1.0428
qwen2.5-14b-instruct	ChaosNLI	ext250	250	250/0/0	0	19	0.224	0	0.692	1.0017
qwen2.5-14b-instruct	MHS	orig_example	150	150/0/0	0	11	0.353	0	0.973	0.6117
qwen2.5-14b-instruct	MHS	ext250	250	250/0/0	0	10	0.388	0	0.948	0.5981
olmo2-13b-base	ChaosNLI	orig_example	150	140/10/0	0	5	0.733	5	0.313	0.2695
olmo2-13b-base	ChaosNLI	ext250	250	238/12/0	0	6	0.668	12	0.268	0.3433
olmo2-13b-base	MHS	orig_example	150	150/0/0	0	2	0.933	140	0.180	0.9375
olmo2-13b-base	MHS	ext250	250	250/0/0	0	2	0.940	235	0.212	0.9408
olmo2-13b-instruct	ChaosNLI	orig_example	150	150/0/0	0	16	0.340	0	0.687	1.1422
olmo2-13b-instruct	ChaosNLI	ext250	250	250/0/0	0	18	0.332	0	0.588	1.1149
olmo2-13b-instruct	MHS	orig_example	150	150/0/0	0	7	0.447	0	0.300	0.3833
olmo2-13b-instruct	MHS	ext250	250	250/0/0	0	10	0.488	0	0.332	0.3700
ministral3-8b-base	ChaosNLI	orig_example	400	400/0/0	4	9	0.887	355	0.037	1.3559
ministral3-8b-base	ChaosNLI	example2	150	149/0/1	0	5	0.919	137	0.919	1.3844
ministral3-8b-base	ChaosNLI	noexample	150	141/8/1	26	7	0.383	NA	0.396	0.9149
ministral3-8b-base	MHS	orig_example	400	391/0/9	0	2	0.992	388	0.450	0.9671
ministral3-8b-base	MHS	example2	150	143/0/7	0	1	1.000	143	0.531	0.8813
ministral3-8b-base	MHS	noexample	150	137/0/13	24	5	0.540	NA	0.745	0.3214
ministral3-8b-instruct	ChaosNLI	orig_example	400	400/0/0	0	21	0.603	0	0.573	1.0824
ministral3-8b-instruct	ChaosNLI	example2	150	150/0/0	0	18	0.540	0	0.627	1.0515
ministral3-8b-instruct	ChaosNLI	noexample	150	150/0/0	0	18	0.540	NA	0.573	1.0588
ministral3-8b-instruct	MHS	orig_example	400	400/0/0	0	11	0.502	0	0.970	0.3285
ministral3-8b-instruct	MHS	example2	150	150/0/0	0	9	0.520	0	0.953	0.3314
ministral3-8b-instruct	MHS	noexample	150	150/0/0	0	10	0.433	NA	0.960	0.2726

Table 18: Verbalized-output audit recomputed from raw responses (three families): full/partial/invalid parses (a well-formed uniform answer counts as a FULL parse, tallied separately), distinct probability vectors, dominant-vector share, exact matches of the prompt’s numeric example vector (canonical label order; N/A when the prompt has no example; vector equality alone does not establish verbatim text copying), top-label agreement with the same checkpoint’s logit distribution (valid parses only; a diagnostic, not proof of invalidity), and mean entropy over valid parses.

Model	Dataset	Label	n	ρ [95% CI]
qwen2.5-14b-base	ChaosNLI	entailment	150	0.814 [0.759, 0.855]
qwen2.5-14b-base	ChaosNLI	neutral	150	0.344 [0.191, 0.491]
qwen2.5-14b-base	ChaosNLI	contradiction	150	0.343 [0.150, 0.522]
qwen2.5-14b-base	MHS	not_offensive	150	0.706 [0.589, 0.799]
qwen2.5-14b-base	MHS	offensive	150	0.706 [0.584, 0.801]
qwen2.5-14b-instruct	ChaosNLI	entailment	150	0.891 [0.848, 0.919]
qwen2.5-14b-instruct	ChaosNLI	neutral	150	0.289 [0.125, 0.438]
qwen2.5-14b-instruct	ChaosNLI	contradiction	150	-0.171 [-0.349, 0.006]
qwen2.5-14b-instruct	MHS	not_offensive	150	0.831 [0.752, 0.890]
qwen2.5-14b-instruct	MHS	offensive	150	0.843 [0.768, 0.899]
olmo2-13b-base	ChaosNLI	entailment	150	0.004 [-0.148, 0.159]
olmo2-13b-base	ChaosNLI	neutral	150	0.057 [-0.096, 0.210]
olmo2-13b-base	ChaosNLI	contradiction	150	0.054 [-0.102, 0.208]
olmo2-13b-base	MHS	not_offensive	150	0.044 [-0.125, 0.204]
olmo2-13b-base	MHS	offensive	150	0.044 [-0.122, 0.199]
olmo2-13b-instruct	ChaosNLI	entailment	150	0.721 [0.610, 0.807]
olmo2-13b-instruct	ChaosNLI	neutral	150	-0.052 [-0.242, 0.148]
olmo2-13b-instruct	ChaosNLI	contradiction	150	-0.056 [-0.225, 0.113]
olmo2-13b-instruct	MHS	not_offensive	150	0.330 [0.161, 0.496]
olmo2-13b-instruct	MHS	offensive	150	0.330 [0.154, 0.495]
minstral3-8b-base	ChaosNLI	entailment	400	0.321 [0.236, 0.400]
minstral3-8b-base	ChaosNLI	neutral	400	0.005 [-0.108, 0.115]
minstral3-8b-base	ChaosNLI	contradiction	400	0.349 [0.258, 0.426]
minstral3-8b-base	MHS	not_offensive	400	-0.000 [-0.098, 0.103]
minstral3-8b-base	MHS	offensive	400	-0.000 [-0.097, 0.102]
minstral3-8b-instruct	ChaosNLI	entailment	400	0.625 [0.558, 0.683]
minstral3-8b-instruct	ChaosNLI	neutral	400	0.282 [0.180, 0.380]
minstral3-8b-instruct	ChaosNLI	contradiction	400	0.540 [0.455, 0.619]
minstral3-8b-instruct	MHS	not_offensive	400	0.824 [0.784, 0.857]
minstral3-8b-instruct	MHS	offensive	400	0.824 [0.784, 0.857]

Table 19: Per-label Spearman correlation between verbalized and logit probabilities on matched items (item-bootstrap 95% CIs; constant vectors reported as undefined, not zero).

J Decoding and compliance controls

Independent decoding seeds. Table 20 compares seeds 0, 1, and 2 with the same decoding settings. Mean entropy varies by at most 0.05 bits and JSD by at most 0.02 in all OLMo-2-13B and aligned cells. Qwen2.5-14B base ChaosNLI varies most: 0.11 bits (1.02–1.13) and 0.05 JSD, compared with an alignment entropy gap of approximately 0.95 bits. Deterministic logit and greedy verbalized measurements reproduce per-item outputs exactly on the same hardware; these repetitions are reproducibility checks rather than independent replications (Appendix I).

Temperature/top- p grid. Table 21 varies decoding around (0.7, 0.95). Base distributions generally spread with temperature and sharpen as it falls. Aligned entropy remains below base entropy in every grid cell, so the observed sampled sharpening is not specific to the default configuration.

Sample count. Table 22 compares $N=100$ with $N=20$ on the same first 50 items. OLMo-2-13B and aligned entropy estimates change by at most a few hundredths of a bit. Qwen2.5-14B base ChaosNLI gains +0.22 bits and improves JSD by -0.08 . The plug-in entropy estimator is downward-biased in expectation at small N , especially for diffuse distributions; individual estimates need not be

lower bounds. This bias tends to understate base entropy more than nearly deterministic aligned entropy, making the observed sampled gaps conservative to that extent.

Few-shot demonstrations. Table 23 uses class-balanced demonstrations from positions beyond the evaluation set, with the same demonstrations across checkpoints, two demonstration sets, and one order reversal. The zero-shot logit measurements reproduce the archived per-item results exactly. Few-shot prompting narrows the entropy gap in two logit comparisons (Qwen2.5-14B ChaosNLI and OLMo-2-13B MHS) and widens it in the other two, while aligned entropy remains lower in every variant. Base means vary by up to 0.13 bits across demonstration sets and orders (Qwen base MHS: 0.671–0.798).

Demonstrations raise OLMo-2-instruct’s ChaosNLI majority agreement from 0.56 to 0.68 and leave Qwen-instruct’s approximately unchanged. Base entropy falls in three of four cells; Qwen base MHS is the exception, rising from 0.61 to 0.67–0.80. Sampled few-shot coverage is 100% except for Qwen base (61% ChaosNLI; 76% MHS). Failures cast no vote. These default-format demonstrations therefore do not equalize Qwen’s compliance, and its sampled few-shot estimates retain the missingness qualification in Appendix C.

Model	Dataset	seed	n	H	JSD	Agree	Parse
olmo2-13b-base	ChaosNLI	seed1	150	0.915	0.273	0.667	100.0%
olmo2-13b-base	ChaosNLI	seed2	150	0.888	0.276	0.667	100.0%
olmo2-13b-base	MHS	seed1	150	0.726	—	—	100.0%
olmo2-13b-base	MHS	seed2	150	0.759	—	—	100.0%
olmo2-13b-instruct	ChaosNLI	seed1	150	0.490	0.453	0.547	100.0%
olmo2-13b-instruct	ChaosNLI	seed2	150	0.479	0.458	0.540	100.0%
olmo2-13b-instruct	MHS	seed1	150	0.105	—	—	100.0%
olmo2-13b-instruct	MHS	seed2	150	0.107	—	—	100.0%
qwen2.5-14b-base	ChaosNLI	seed1	150	1.056	0.255	0.700	85.9%
qwen2.5-14b-base	ChaosNLI	seed2	150	1.127	0.238	0.713	88.0%
qwen2.5-14b-base	MHS	seed1	150	0.410	—	—	95.1%
qwen2.5-14b-base	MHS	seed2	150	0.424	—	—	95.6%
qwen2.5-14b-instruct	ChaosNLI	seed1	150	0.073	0.510	0.593	100.0%
qwen2.5-14b-instruct	ChaosNLI	seed2	150	0.077	0.509	0.593	100.0%
qwen2.5-14b-instruct	MHS	seed1	150	0.005	—	—	100.0%
qwen2.5-14b-instruct	MHS	seed2	150	0.005	—	—	100.0%
olmo2-13b-base	ChaosNLI	seed0(orig)	150	0.860	0.290	0.687	—
olmo2-13b-base	MHS	seed0(orig)	150	0.712	—	—	—
olmo2-13b-instruct	ChaosNLI	seed0(orig)	150	0.485	0.453	0.533	—
olmo2-13b-instruct	MHS	seed0(orig)	150	0.104	—	—	—
qwen2.5-14b-base	ChaosNLI	seed0(orig)	150	1.022	0.288	0.673	—
qwen2.5-14b-base	MHS	seed0(orig)	150	0.381	—	—	—
qwen2.5-14b-instruct	ChaosNLI	seed0(orig)	150	0.083	0.503	0.600	—
qwen2.5-14b-instruct	MHS	seed0(orig)	150	0.006	—	—	—

Table 20: Sampled elicitation with independent decoding seeds ($n=150$, $N=20$, $T=0.7$, top- p 0.95). The artifact label 'seed0(orig)' identifies the seed-0 baseline. Variation across seeds is an order of magnitude smaller than the alignment entropy gaps. Deterministic logit and greedy verbalized decoding are not treated as independent seed replications.

Model	Dataset	(T, p)	n	H	JSD	Agree	Parse
olmo2-13b-base	ChaosNLI	T0.3/p0.95	150	0.451	0.369	0.733	100.0%
olmo2-13b-base	ChaosNLI	T0.7/p0.9	150	0.830	0.286	0.707	100.0%
olmo2-13b-base	ChaosNLI	T0.7/p1.0	150	0.947	0.273	0.673	100.0%
olmo2-13b-base	ChaosNLI	T1.0/p0.95	150	1.155	0.254	0.693	100.0%
olmo2-13b-base	ChaosNLI	T1.3/p0.95	150	1.254	0.262	0.620	100.0%
olmo2-13b-base	MHS	T0.3/p0.95	150	0.365	—	—	100.0%
olmo2-13b-base	MHS	T0.7/p0.9	150	0.681	—	—	100.0%
olmo2-13b-base	MHS	T0.7/p1.0	150	0.721	—	—	100.0%
olmo2-13b-base	MHS	T1.0/p0.95	150	0.815	—	—	100.0%
olmo2-13b-base	MHS	T1.3/p0.95	150	0.864	—	—	100.0%
olmo2-13b-instruct	ChaosNLI	T0.3/p0.95	150	0.180	0.521	0.547	100.0%
olmo2-13b-instruct	ChaosNLI	T0.7/p0.9	150	0.389	0.479	0.527	100.0%
olmo2-13b-instruct	ChaosNLI	T0.7/p1.0	150	0.527	0.443	0.547	100.0%
olmo2-13b-instruct	ChaosNLI	T1.0/p0.95	150	0.673	0.417	0.547	100.0%
olmo2-13b-instruct	ChaosNLI	T1.3/p0.95	150	0.843	0.381	0.547	100.0%
olmo2-13b-instruct	MHS	T0.3/p0.95	150	0.048	—	—	100.0%
olmo2-13b-instruct	MHS	T0.7/p0.9	150	0.097	—	—	100.0%
olmo2-13b-instruct	MHS	T0.7/p1.0	150	0.104	—	—	100.0%
olmo2-13b-instruct	MHS	T1.0/p0.95	150	0.129	—	—	100.0%
olmo2-13b-instruct	MHS	T1.3/p0.95	150	0.158	—	—	100.0%
qwen2.5-14b-base	ChaosNLI	T0.3/p0.95	150	0.689	0.322	0.687	76.5%
qwen2.5-14b-base	ChaosNLI	T0.7/p0.9	150	1.070	0.257	0.693	85.6%
qwen2.5-14b-base	ChaosNLI	T0.7/p1.0	150	1.108	0.246	0.713	88.3%
qwen2.5-14b-base	ChaosNLI	T1.0/p0.95	150	1.274	0.249	0.680	91.4%
qwen2.5-14b-base	ChaosNLI	T1.3/p0.95	150	1.330	0.258	0.680	92.6%
qwen2.5-14b-base	MHS	T0.3/p0.95	150	0.152	—	—	98.9%
qwen2.5-14b-base	MHS	T0.7/p0.9	150	0.373	—	—	95.7%
qwen2.5-14b-base	MHS	T0.7/p1.0	150	0.465	—	—	95.9%
qwen2.5-14b-base	MHS	T1.0/p0.95	150	0.642	—	—	95.9%
qwen2.5-14b-base	MHS	T1.3/p0.95	150	0.754	—	—	94.4%
qwen2.5-14b-instruct	ChaosNLI	T0.3/p0.95	150	0.035	0.522	0.580	100.0%
qwen2.5-14b-instruct	ChaosNLI	T0.7/p0.9	150	0.071	0.512	0.580	100.0%
qwen2.5-14b-instruct	ChaosNLI	T0.7/p1.0	150	0.089	0.508	0.580	100.0%
qwen2.5-14b-instruct	ChaosNLI	T1.0/p0.95	150	0.110	0.500	0.580	100.0%
qwen2.5-14b-instruct	ChaosNLI	T1.3/p0.95	150	0.133	0.493	0.580	100.0%
qwen2.5-14b-instruct	MHS	T0.3/p0.95	150	0.000	—	—	100.0%
qwen2.5-14b-instruct	MHS	T0.7/p0.9	150	0.003	—	—	100.0%
qwen2.5-14b-instruct	MHS	T0.7/p1.0	150	0.005	—	—	100.0%
qwen2.5-14b-instruct	MHS	T1.0/p0.95	150	0.009	—	—	100.0%
qwen2.5-14b-instruct	MHS	T1.3/p0.95	150	0.014	—	—	99.9%

Table 21: Sampled elicitation across a decoding grid around the default $T=0.7$, top- p 0.95 setting (seed 0, $n=150$, $N=20$). Base distributions generally spread with temperature; aligned entropy remains lower in every cell.

Model	Dataset	N	n	H	JSD	Agree	Parse
olmo2-13b-base	ChaosNLI	N100	50	0.904	0.273	0.700	100.0%
olmo2-13b-base	ChaosNLI	N20-orig-same50	50	0.856	0.302	0.660	—
olmo2-13b-base	MHS	N100	50	0.699	—	—	100.0%
olmo2-13b-base	MHS	N20-orig-same50	50	0.726	—	—	—
olmo2-13b-instruct	ChaosNLI	N100	50	0.492	0.468	0.460	100.0%
olmo2-13b-instruct	ChaosNLI	N20-orig-same50	50	0.509	0.472	0.460	—
olmo2-13b-instruct	MHS	N100	50	0.115	—	—	100.0%
olmo2-13b-instruct	MHS	N20-orig-same50	50	0.108	—	—	—
qwen2.5-14b-base	ChaosNLI	N100	50	1.222	0.209	0.680	85.9%
qwen2.5-14b-base	ChaosNLI	N20-orig-same50	50	1.007	0.284	0.660	—
qwen2.5-14b-base	MHS	N100	50	0.433	—	—	96.1%
qwen2.5-14b-base	MHS	N20-orig-same50	50	0.357	—	—	—
qwen2.5-14b-instruct	ChaosNLI	N100	50	0.082	0.512	0.500	100.0%
qwen2.5-14b-instruct	ChaosNLI	N20-orig-same50	50	0.073	0.511	0.540	—
qwen2.5-14b-instruct	MHS	N100	50	0.009	—	—	100.0%
qwen2.5-14b-instruct	MHS	N20-orig-same50	50	0.000	—	—	—

Table 22: Sample-count sensitivity: $N=100$ versus $N=20$ on the same first 50 items.

Model	Dataset	variant	n	H	JSD	Agree	Parse
olmo2-13b-base	ChaosNLI	fs_setA/logit	400	0.968	0.255	0.685	—
olmo2-13b-base	ChaosNLI	fs_setA/sampling	150	0.578	0.348	0.693	100.0%
olmo2-13b-base	ChaosNLI	fs_setA_rev/logit	400	0.964	0.252	0.682	—
olmo2-13b-base	ChaosNLI	fs_setB/logit	400	0.977	0.253	0.690	—
olmo2-13b-base	ChaosNLI	zs/logit	400	1.229	0.220	0.713	—
olmo2-13b-base	MHS	fs_setA/logit	400	0.606	—	—	—
olmo2-13b-base	MHS	fs_setA/sampling	150	0.404	—	—	100.0%
olmo2-13b-base	MHS	fs_setA_rev/logit	400	0.598	—	—	—
olmo2-13b-base	MHS	fs_setB/logit	400	0.582	—	—	—
olmo2-13b-base	MHS	zs/logit	400	0.857	—	—	—
olmo2-13b-instruct	ChaosNLI	fs_setA/logit	400	0.489	0.372	0.685	—
olmo2-13b-instruct	ChaosNLI	fs_setA/sampling	150	0.235	0.456	0.633	100.0%
olmo2-13b-instruct	ChaosNLI	fs_setA_rev/logit	400	0.497	0.370	0.675	—
olmo2-13b-instruct	ChaosNLI	fs_setB/logit	400	0.485	0.371	0.680	—
olmo2-13b-instruct	ChaosNLI	zs/logit	400	0.829	0.364	0.560	—
olmo2-13b-instruct	MHS	fs_setA/logit	400	0.130	—	—	—
olmo2-13b-instruct	MHS	fs_setA/sampling	150	0.076	—	—	100.0%
olmo2-13b-instruct	MHS	fs_setA_rev/logit	400	0.124	—	—	—
olmo2-13b-instruct	MHS	fs_setB/logit	400	0.079	—	—	—
olmo2-13b-instruct	MHS	zs/logit	400	0.124	—	—	—
qwen2.5-14b-base	ChaosNLI	fs_setA/logit	400	1.153	0.345	0.450	—
qwen2.5-14b-base	ChaosNLI	fs_setA/sampling	150	0.935	0.308	0.727	61.0%
qwen2.5-14b-base	ChaosNLI	fs_setA_rev/logit	400	1.179	0.342	0.450	—
qwen2.5-14b-base	ChaosNLI	fs_setB/logit	400	1.208	0.330	0.455	—
qwen2.5-14b-base	ChaosNLI	zs/logit	400	1.389	0.257	0.632	—
qwen2.5-14b-base	MHS	fs_setA/logit	400	0.745	—	—	—
qwen2.5-14b-base	MHS	fs_setA/sampling	150	0.824	—	—	76.3%
qwen2.5-14b-base	MHS	fs_setA_rev/logit	400	0.798	—	—	—
qwen2.5-14b-base	MHS	fs_setB/logit	400	0.671	—	—	—
qwen2.5-14b-base	MHS	zs/logit	400	0.608	—	—	—
qwen2.5-14b-instruct	ChaosNLI	fs_setA/logit	400	0.106	0.493	0.652	—
qwen2.5-14b-instruct	ChaosNLI	fs_setA/sampling	150	0.049	0.505	0.673	100.0%
qwen2.5-14b-instruct	ChaosNLI	fs_setA_rev/logit	400	0.114	0.504	0.627	—
qwen2.5-14b-instruct	ChaosNLI	fs_setB/logit	400	0.087	0.489	0.670	—
qwen2.5-14b-instruct	ChaosNLI	zs/logit	400	0.115	0.489	0.627	—
qwen2.5-14b-instruct	MHS	fs_setA/logit	400	0.030	—	—	—
qwen2.5-14b-instruct	MHS	fs_setA/sampling	150	0.013	—	—	100.0%
qwen2.5-14b-instruct	MHS	fs_setA_rev/logit	400	0.025	—	—	—
qwen2.5-14b-instruct	MHS	fs_setB/logit	400	0.025	—	—	—
qwen2.5-14b-instruct	MHS	zs/logit	400	0.017	—	—	—

Table 23: Zero-shot and few-shot elicitation. ‘zs’ denotes the zero-shot logit condition; ‘fs_setA/B’ use class-balanced demonstrations from positions 400 onward, disjoint from evaluation items and shared across checkpoints; ‘fs_setA_rev’ reverses their order. ‘sampling’ uses the default sampled settings. Aligned entropy remains below base in every variant. Demonstration choices shift base means by up to approximately 0.13 bits (Qwen2.5-14B MHS).

The shared-prefix compliance experiment is a separate condition.

K Recovery levers

We evaluate two inference-time interventions on aligned checkpoints using ChaosNLI, comparing

each with a baseline measured by the same estimator.

Sampling temperature. We compare decoding temperatures 1.0 and 1.3 with 0.7 on the same 150 items, holding $N=20$, top- p 0.95, and seed 0 fixed. This varies generation-time tempera-

Pair (aligned ckpt)	sampled JSD ($T=0.7$)	sampled JSD ($T=1.0$)	$\Delta T0.7 \rightarrow 1.0$	$\Delta T0.7 \rightarrow 1.3$
Qwen2.5-7B		— (no sampled $T=0.7$ baseline run)		
Qwen2.5-14B	0.5033	0.5005	−0.0028 [−0.0122, +0.0064]	−0.0104 [−0.0222, +0.0008]
Qwen2.5-32B		— (no sampled $T=0.7$ baseline run)		
Qwen3-14B		— (no sampled $T=0.7$ baseline run)		
Qwen3-30B-A3B		— (no sampled $T=0.7$ baseline run)		
OLMo-2-13B	0.4526	0.4169	−0.0357 [−0.0532, −0.0183]	−0.0720 [−0.0914, −0.0521]
OLMo-2-32B		— (no sampled $T=0.7$ baseline run)		

Table 24: ChaosNLI decoding-temperature comparisons on aligned checkpoints, paired by item ID ($n=150$; 95% bootstrap intervals). All conditions use sampled elicitation with $N=20$, top- p 0.95, and seed 0; temperature varies from 0.7 to 1.0 or 1.3. Negative differences indicate improved fit. Checkpoints lacking a compatible sampled baseline are marked unavailable. The Overton estimator is described separately in Section K, with weight sensitivity in Table 25.

ture, separately from the distribution transform in Section 4.4. Compatible sampled baselines are available for Qwen2.5-14B and OLMo-2-13B only; other checkpoints are marked unavailable in Table 24. Qwen’s ChaosNLI JSD changes are small and not significant (−0.003 at $T=1.0$; −0.010 [−0.022, +0.001] at $T=1.3$). OLMo-2 improves significantly (−0.036 [−0.053, −0.018] and −0.072 [−0.091, −0.052]), recovering 0.07 of its 0.16 sampled alignment gap. The effect is therefore checkpoint-dependent. Appendix J reports the surrounding temperature/top- p grid.

Overton list-all prompt. We ask the model to “list ALL labels that a reasonable annotator could justifiably choose,” then end with “Best: ⟨letter⟩”. The single-best comparison prompt is identical except that it forbids alternatives. Both use greedy decoding on the same items and the same parser-constructed estimator. With L the listed labels and b the parsed best label,

$$p(y) \propto \mathbf{1}[y \in L] + w \mathbf{1}[y = b], \quad w = 1, \quad (1)$$

a single parsed label yields a point mass. The list-all prompt improves JSD for all three aligned checkpoints: OLMo-2-7B-DPO 0.560→0.359 (−0.201 [−0.222, −0.179]), Qwen2.5-7B-Instruct 0.619→0.520 (−0.099 [−0.123, −0.075]), and OLMo-2-13B-Instruct 0.627→0.551 (−0.077 [−0.102, −0.052]). Re-scoring the decoded label sets under $w \in \{0.5, 1, 2\}$ changes JSD by at most 0.026 and preserves the direction (Appendix E). These are prompt effects on a constructed distribution, not evidence that the model recovers human population preferences.

L Compliance probe: bounds and details

Run manifests and per-item measurements accompany the analyses. A unified artifact table reports

paired 10,000-resample bootstrap intervals for each included comparison. Raw-response availability is described in Section 7.

The compliance parser requires the first non-empty line to open with an unambiguous decision. Bare or parenthesized labels and letter-then-explanation forms are valid unless another referenced label creates a conflict. Refusals, input restatements, quoted input, and undecided lists are invalid; correctness is not part of the validity test. Demonstrations maximize the minimum parse coverage across checkpoints on development items 600–659. Parser-audit corrections change 189/52,800 stored classifications across 121 items, with no item losing all samples. Reapplying the frozen selection rule selects the executed demonstration condition under each audited parser version.

For each sampled item let c_k be the valid-label counts, m the invalid completions, and $N=20$. We enumerate every allocation a with $\sum_k a_k = m$, score $p_k = (c_k + a_k)/N$, and take per-item extrema of the metric. The paired bounds $L_i = \min f(\text{aligned}_i) - \max f(\text{base}_i)$ and $U_i = \max f(\text{aligned}_i) - \min f(\text{base}_i)$ bracket the aligned–base difference under arbitrary reassignment (Table 28). With $m=0$, enumeration returns the point metric, as verified by the released self-tests. We apply the bounds to compliance runs and archived independent-seed runs; the latter rely on stored per-sample classifications because raw text is unavailable. Inference errors are tracked separately. In the compliance probe, items with no valid sample would be excluded and counted, rather than assigned a uniform vector; no evaluated item requires this exclusion.

Model	$w=0.5$	$w=1$	$w=2$
qwen2.5-7b-instruct	0.513	0.520	0.536
olmo2-7b-dpo	0.358	0.359	0.372
olmo2-13b-instruct	0.535	0.551	0.577

Table 25: Sensitivity of the Overton list-all JSD (ChaosNLI, $n=400$) to the best-choice weight w of Eq. 1: the stored $w=1$ per-item distributions are decoded back to (listed set, best) and re-scored. The improvement direction is unchanged and the magnitude moves by at most 0.026.

Pair	Dataset	Cond.	H B→A	ΔH	JSD B→A	Δ JSD
Qwen2.5-14B	ChaosNLI	zero-shot	0.9738→0.3433	-0.6306 [-0.6642, -0.5964]	0.2796→0.4569	+0.1773 [+0.1606, +0.1935]
Qwen2.5-14B	ChaosNLI	few-shot	0.8882→0.2019	-0.6863 [-0.7228, -0.6489]	0.2338→0.4442	+0.2103 [+0.1961, +0.2252]
Qwen2.5-14B	MHS	zero-shot	0.4372→0.0497	-0.3874 [-0.4139, -0.3607]	—	—
Qwen2.5-14B	MHS	few-shot	0.5070→0.1377	-0.3694 [-0.3936, -0.3450]	—	—
OLMo-2-13B	ChaosNLI	zero-shot	1.5220→1.2008	-0.3212 [-0.3501, -0.2940]	0.2929→0.3071	+0.0141 [+0.0007, +0.0278]
OLMo-2-13B	ChaosNLI	few-shot	1.2443→0.6676	-0.5767 [-0.6056, -0.5480]	0.2371→0.3350	+0.0979 [+0.0837, +0.1117]
OLMo-2-13B	MHS	zero-shot	0.9378→0.5773	-0.3605 [-0.3936, -0.3275]	—	—
OLMo-2-13B	MHS	few-shot	0.8617→0.5725	-0.2892 [-0.3172, -0.2620]	—	—

Table 26: Logit elicitation under an identical raw completion prefix for base and aligned checkpoints ($n=400$; paired bootstrap 95% intervals, aligned–base). JSD is shown only for ChaosNLI; binary MHS is an entropy probe. Few-shot demonstrations maximize minimum parse coverage on development items 600–659 and are frozen before evaluation. Parse coverage measures response compliance, not task competence.

Pair	Dataset	Cond.	n	H B→A	ΔH
Qwen2.5-14B	ChaosNLI	zs/sampled	150	0.6716→0.1877	-0.4839 [-0.5495, -0.4190]
Qwen2.5-14B	ChaosNLI	fs1/sampled	150	0.5527→0.1280	-0.4247 [-0.4975, -0.3505]
Qwen2.5-14B	MHS	zs/sampled	150	0.2326→0.0159	-0.2167 [-0.2677, -0.1685]
Qwen2.5-14B	MHS	fs1/sampled	150	0.2659→0.0576	-0.2084 [-0.2618, -0.1582]
OLMo-2-13B	ChaosNLI	zs/sampled	150	1.3583→0.8994	-0.4589 [-0.5314, -0.3878]
OLMo-2-13B	ChaosNLI	fs1/sampled	150	0.9329→0.3717	-0.5613 [-0.6353, -0.4871]
OLMo-2-13B	MHS	zs/sampled	150	0.8552→0.4496	-0.4056 [-0.4659, -0.3449]
OLMo-2-13B	MHS	fs1/sampled	150	0.7636→0.4401	-0.3234 [-0.3826, -0.2658]

Table 27: Sampled compliance conditions ($N=20$, $T=0.7$, top- p 0.95, seed 0, 128-token cap). Items with no valid samples would be excluded and counted rather than assigned a uniform distribution; none require exclusion.

M Staged training comparisons

Table 29 compares OLMo-2-7B stages under raw and injected-template base conditions (Section 4.7).

Table 30 reports the OLMoE-1B-7B-0125 transitions under the two byte-verified formats of Section 4.7. Verification artifacts (rendered-prompt hashes for all three stages under both formats) ship with the analysis outputs.

N Semantic-verbalizer scoring: definitions and all four conventions

Given a prompt x and a candidate label word $w = (w_1, \dots, w_m)$, we score w by its full-sequence log-probability under the model θ , in two conventions:

$$s_{\text{sum}}(w) = \sum_{t=1}^m \log p_{\theta}(w_t \mid x, w_{<t}), \quad (2)$$

$$s_{\text{len}}(w) = \frac{1}{m} s_{\text{sum}}(w). \quad (3)$$

Content-free (PMI) calibration subtracts the same score under a content-free prompt x_0 (instructions

+ label list, no item content):

$$\tilde{s}(w) = s(w) - s_0(w), \quad s_0(w) \text{ uses } x_0. \quad (4)$$

For each convention we softmax the corresponding raw or calibrated score (s or $\tilde{s}$) over labels to obtain the label distribution $p(w)$. Table 31 gives the paired base→aligned ΔH for all four conventions; every cell is negative with a 95% CI excluding zero, so the entropy collapse does not depend on the scoring choice.

O Native three-way MHS results

Tables 32 and 33 give checkpoint estimates and paired comparisons for the native no/unclear/yes hate-speech task described in Section 4.5. They include majority agreement, Brier scores, and sensitivity to sparse annotations. Table 34 reports held-out JSD after matching aligned to base mean entropy.

Source	Pair	Dataset	Cond.	f	mean L	mean U	inv.
comp_eval	Qwen2.5-14B	ChaosNLI	zs/sampled	H	-0.5232 [-0.5862, -0.4581]	-0.0088 [-0.1119, +0.0954]	13/432
comp_eval	Qwen2.5-14B	ChaosNLI	zs/sampled	J	-0.0021 [-0.0391, +0.0353]	+0.1815 [+0.1544, +0.2095]	13/432
comp_eval	Qwen2.5-14B	ChaosNLI	fs1/sampled	H	-0.4296 [-0.5029, -0.3555]	-0.3068 [-0.3911, -0.2232]	0/86
comp_eval	Qwen2.5-14B	ChaosNLI	fs1/sampled	J	+0.0906 [+0.0609, +0.1211]	+0.1345 [+0.1087, +0.1617]	0/86
seeds	Qwen2.5-14B	ChaosNLI	1	H	-1.2318 [-1.2834, -1.1766]	-0.9066 [-0.9672, -0.8420]	422/0
seeds	Qwen2.5-14B	ChaosNLI	1	J	+0.1808 [+0.1518, +0.2101]	+0.3348 [+0.3060, +0.3636]	422/0
seeds	Qwen2.5-14B	ChaosNLI	2	H	-1.2408 [-1.2951, -1.1827]	-0.9865 [-1.0460, -0.9240]	359/0
seeds	Qwen2.5-14B	ChaosNLI	2	J	+0.2015 [+0.1716, +0.2304]	+0.3318 [+0.3011, +0.3626]	359/0
comp_eval	Qwen2.5-14B	MHS	zs/sampled	H	-0.2183 [-0.2697, -0.1696]	-0.1891 [-0.2396, -0.1386]	3/24
comp_eval	Qwen2.5-14B	MHS	fs1/sampled	H	-0.2085 [-0.2615, -0.1580]	-0.1795 [-0.2348, -0.1257]	0/17
seeds	Qwen2.5-14B	MHS	1	H	-0.5236 [-0.5741, -0.4708]	-0.3939 [-0.4492, -0.3392]	146/0
seeds	Qwen2.5-14B	MHS	2	H	-0.5200 [-0.5734, -0.4648]	-0.4074 [-0.4605, -0.3534]	131/0
comp_eval	OLMo-2-13B	ChaosNLI	zs/sampled	H	-0.4589 [-0.5308, -0.3880]	-0.4589 [-0.5288, -0.3878]	0/0
comp_eval	OLMo-2-13B	ChaosNLI	zs/sampled	J	+0.0962 [+0.0673, +0.1237]	+0.0962 [+0.0684, +0.1244]	0/0
comp_eval	OLMo-2-13B	ChaosNLI	fs1/sampled	H	-0.5613 [-0.6340, -0.4882]	-0.5613 [-0.6342, -0.4875]	0/0
comp_eval	OLMo-2-13B	ChaosNLI	fs1/sampled	J	+0.1464 [+0.1181, +0.1746]	+0.1464 [+0.1182, +0.1740]	0/0
seeds	OLMo-2-13B	ChaosNLI	1	H	-0.4241 [-0.5126, -0.3336]	-0.4241 [-0.5147, -0.3333]	0/0
seeds	OLMo-2-13B	ChaosNLI	1	J	+0.1798 [+0.1492, +0.2109]	+0.1798 [+0.1496, +0.2100]	0/0
seeds	OLMo-2-13B	ChaosNLI	2	H	-0.4089 [-0.4974, -0.3215]	-0.4089 [-0.4984, -0.3225]	0/0
seeds	OLMo-2-13B	ChaosNLI	2	J	+0.1824 [+0.1513, +0.2133]	+0.1824 [+0.1519, +0.2131]	0/0
comp_eval	OLMo-2-13B	MHS	zs/sampled	H	-0.4056 [-0.4659, -0.3451]	-0.4056 [-0.4672, -0.3448]	0/0
comp_eval	OLMo-2-13B	MHS	fs1/sampled	H	-0.3234 [-0.3802, -0.2667]	-0.3234 [-0.3824, -0.2665]	0/0
seeds	OLMo-2-13B	MHS	1	H	-0.6205 [-0.6697, -0.5707]	-0.6205 [-0.6700, -0.5689]	0/0
seeds	OLMo-2-13B	MHS	2	H	-0.6514 [-0.6993, -0.6013]	-0.6514 [-0.6990, -0.6000]	0/0

Table 28: Worst-case bounds under exhaustive reassignment of invalid responses: paired means of $L_i = \min f(\text{aligned}) - \max f(\text{base})$ and $U_i = \max f(\text{aligned}) - \min f(\text{base})$, with item-bootstrap endpoint intervals. An interval crossing zero does not certify the sign. The comp_eval rows use the audited answer-validity parser on stored raw completions. Seed rows use archived sample classifications because raw text is unavailable. The last column counts invalid completions (base/aligned). These bounds measure reassignment sensitivity rather than recover ground truth.

Dataset	Metric	Base _{raw}	Base _{tmpl}	SFT	DPO
MHS	H	0.911	0.535	0.343	0.188
ChaosNLI	H	1.232	0.718	0.965	0.751
	JSD	0.288	0.482	0.292	0.341

Table 29: OLMo-2-7B staged comparison ($n=400$). Base_{raw} uses a raw prompt; Base_{tmpl} injects the shared SFT/DPO template into the base tokenizer. MHS entropy falls through both stages under either base format. On ChaosNLI, the base→SFT entropy change reverses with format. SFT→DPO uses identical templates, lowers entropy, and worsens ChaosNLI JSD.

Dataset	Format	Transition	ΔH	ΔJSD
ChaosNLI	raw	base→sft	-0.2889 [-0.3097, -0.2683]	+0.0471 [+0.0347, +0.0597]
ChaosNLI	raw	sft→dpo	+0.0014 [-0.0118, +0.0150]	-0.0113 [-0.0171, -0.0055]
ChaosNLI	sft_template_nobos	base→sft	-0.2528 [-0.2801, -0.2249]	+0.0977 [+0.0634, +0.1310]
ChaosNLI	sft_template_nobos	sft→dpo	-0.0167 [-0.0269, -0.0069]	+0.0088 [+0.0041, +0.0134]
MHS	raw	base→sft	-0.4356 [-0.4632, -0.4081]	—
MHS	raw	sft→dpo	-0.0449 [-0.0516, -0.0384]	—
MHS	sft_template_nobos	base→sft	+0.0724 [+0.0452, +0.0996]	—
MHS	sft_template_nobos	sft→dpo	+0.0077 [-0.0014, +0.0172]	—

Table 30: OLMoE-1B-7B-0125 base/SFT/DPO transitions, ending at DPO rather than RLVR. Both conditions use byte-verified identical prompts across stages: a raw prompt or the BOS-free SFT-template string. SFT and DPO templates are identical; removing their tokenizers’ BOS marker matches the base rendering. Binary MHS JSD is omitted because this probe is entropy-only; native three-way MHS is reported in Table 32.

Family	Dataset	ΔH (aligned-base), bits [95% CI]			
		summed, raw	summed, PMI	length-norm, raw	length-norm, PMI
Qwen2.5-14B	ChaosNLI	-0.551 [-0.590, -0.513]	-0.634 [-0.671, -0.596]	-0.646 [-0.688, -0.604]	-0.702 [-0.738, -0.666]
	MHS	-0.140 [-0.170, -0.111]	-0.203 [-0.234, -0.173]	-0.331 [-0.356, -0.307]	-0.378 [-0.408, -0.348]
OLMo-2-13B	ChaosNLI	-0.315 [-0.341, -0.289]	-0.353 [-0.379, -0.326]	-0.350 [-0.380, -0.319]	-0.385 [-0.415, -0.356]
	MHS	-0.273 [-0.296, -0.251]	-0.364 [-0.389, -0.341]	-0.126 [-0.144, -0.108]	-0.272 [-0.291, -0.254]

Table 31: Semantic-verbalizer control, *all four scoring conventions*: paired base→aligned ΔH (word-only prompt, matched raw format; $n=400$ per cell; percentile bootstrap 95% CIs, 10,000 resamples). Every entry is negative with a CI excluding zero, so the entropy collapse is robust to summed vs. length-normalized scoring and to content-free (PMI) calibration.

Model	Format	H	JSD	Agree	ties	Brier	JSD ann-resamp.
qwen2.5-14b-base	raw	1.2815	0.5220	0.4776	21	0.6552	[0.5481,0.5664]
qwen2.5-14b-base	raw_reversed	1.0737	0.5670	0.4090	21	0.7954	[0.5824,0.6025]
qwen2.5-14b-base	default	1.3976	0.6011	0.2876	21	0.7910	[0.6268,0.6445]
qwen2.5-14b-instruct	raw	0.3508	0.5404	0.4565	21	0.9216	[0.5312,0.5580]
qwen2.5-14b-instruct	raw_reversed	0.3912	0.4920	0.5435	21	0.8020	[0.4851,0.5119]
qwen2.5-14b-instruct	default	0.1159	0.5298	0.4908	21	0.9867	[0.5149,0.5423]
olmo2-13b-base	raw	1.3143	0.5305	0.4354	21	0.6736	[0.5564,0.5750]
olmo2-13b-base	raw_reversed	1.2630	0.5686	0.3404	21	0.7608	[0.5907,0.6093]
olmo2-13b-base	default	1.3143	0.5305	0.4354	21	0.6736	[0.5564,0.5750]
olmo2-13b-instruct	raw	0.6801	0.6025	0.3773	21	0.9698	[0.6033,0.6267]
olmo2-13b-instruct	raw_reversed	0.6926	0.5668	0.4116	21	0.8699	[0.5681,0.5918]
olmo2-13b-instruct	default	0.2706	0.6458	0.3562	21	1.1494	[0.6319,0.6596]

Table 32: Native three-way MHS hate-speech task (400 seeded comments, ≥ 3 annotators, no/unclear/yes): logit elicitation under matched-raw, reversed-label, and default formats. Agreement is on unique-plurality items only (tie counts shown). Brier is the mean squared error against individual annotations (within-comment average first). The last column is the 95% interval of mean JSD under within-item annotation resampling (the human distributions rest on few ratings). References (full set): uniform 0.5546; plurality_onehot 0.1715.

Pair	Format	ΔH	ΔJSD	ΔJSD (ann. resamp.)
Qwen2.5-14B	raw	-0.9306 [-0.9635, -0.8969]	+0.0184 [-0.0090, +0.0454]	-0.0136 [-0.0233, -0.0042]
Qwen2.5-14B	raw_reversed	-0.6824 [-0.7213, -0.6419]	-0.0750 [-0.1001, -0.0502]	-0.0938 [-0.1020, -0.0858]
Qwen2.5-14B	default	-1.2817 [-1.3071, -1.2540]	-0.0714 [-0.1092, -0.0340]	-0.1067 [-0.1195, -0.0947]
OLMo-2-13B	raw	-0.6342 [-0.6660, -0.6007]	+0.0719 [+0.0545, +0.0888]	+0.0491 [+0.0413, +0.0566]
OLMo-2-13B	raw_reversed	-0.5704 [-0.6028, -0.5386]	-0.0018 [-0.0180, +0.0144]	-0.0198 [-0.0268, -0.0127]
OLMo-2-13B	default	-1.0437 [-1.0781, -1.0085]	+0.1152 [+0.0895, +0.1396]	+0.0800 [+0.0688, +0.0900]

Table 33: Native three-way MHS: six paired comparisons (2 model pairs $\times$ 3 formats; item-bootstrap 95% intervals). The last column gives mean ΔJSD under within-item annotation resampling, with each comment’s annotations redrawn once per replicate and the same draw applied to both checkpoints. This measures sensitivity to sparse observed annotations, separately from item-bootstrap uncertainty; it is not a population correction.

Pair	Format	τ	Base $_{h_o}$	Aligned $_{h_o}$	AI. $@\tau_{h_o}$	Residual
Qwen2.5-14B	raw	6.44	0.5335	0.5610	0.5368	+0.0033 [-0.0052, +0.0118]
Qwen2.5-14B	raw_reversed	4.48	0.5828	0.5072	0.5153	-0.0675 [-0.0816, -0.0528]
Qwen2.5-14B	default	21.80	0.6145	0.5519	0.5400	-0.0744 [-0.0859, -0.0631]
OLMo-2-13B	raw	2.81	0.5439	0.6225	0.5649	+0.0210 [+0.0142, +0.0278]
OLMo-2-13B	raw_reversed	3.13	0.5805	0.5839	0.5478	-0.0327 [-0.0412, -0.0241]
OLMo-2-13B	default	5.60	0.5439	0.6733	0.5807	+0.0368 [+0.0287, +0.0446]

Table 34: MHS three-way task: aligned distributions temperature-scaled to base mean entropy (200/200 calibration/evaluation split, permutation seed 0; held-out means; interval conditional on the fitted τ and split).