LABEL COLLAPSE IN DIFFUSION LANGUAGE MODEL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) applies language models to new tasks by prepending a few labeled demonstrations to a query. Diffusion language models (dLLMs) are entering deployment, yet their ICL failure modes remain largely uncharacterized. We define label collapse as the failure mode in which the model generates query-independent labels, and show that it is widespread: across six dLLMs from three model families, collapse persists even under class-balanced alternating prompts, its frequency is stratified by model family rather than by instruction tuning, and some label word pairs collapse in every model tested. The failure hides beneath an accuracy that sits only modestly above chance, so accuracy-only evaluation reports a healthy model. As a diagnostic, we measure separation, the log-odds of the two label words at the query position, which exposes collapse in a single forward pass. Building on this diagnostic, we introduce demo-anchored calibration (DAC), which anchors on the demonstrations themselves: a paired midpoint over per-demonstration readings cancels what each demonstration's own label word contributes, and the recovered bias estimate repairs most collapsed conditions without access to the test batch and transfers to free generation. Overall, our results show that label collapse is invisible to standard evaluation, diagnosable in a single forward pass, and repairable from the demonstrations already in the prompt.

1 INTRODUCTION

In-context learning (ICL), the ability to perform new tasks from demonstrations in the prompt without weight updates, is a foundational capability of modern language models (Brown et al., 2020; Dong et al., 2024). Diffusion language models (dLLMs), which generate text by iterative denoising rather than left-to-right factorization, now match autoregressive (AR) models on a range of benchmarks and are entering deployment (Nie et al., 2025; Ye et al., 2025). Their ICL capability has been demonstrated (Nie et al., 2025), but their ICL failure modes remain largely uncharacterized: prior work has studied how demonstration ordering affects autoregressive ICL (Lu et al., 2022; Zhao et al., 2021), yet the interaction between label arrangement, label composition, and model bias has not been systematically examined for this architecture class.

We define *label collapse* as a failure mode in which the model generates query-independent labels: regardless of the query input, the model emits the same label word (Figure 1). The failure is not an artifact of imbalanced or adversarial prompts: it arises across six dLLMs from three model families (LLaDA, iLLaDA, and Dream, in base and instruct variants), even under class-balanced alternating demonstrations. Figure 1 shows how easily it is triggered: swapping the label words assigned to the demonstrations, with everything else held fixed, collapses a model that performs the task almost perfectly under its natural label words. To expose the failure, we measure *separation*, the log-odds of the two label words at the query position, which is the quantity that label-bias calibration methods operate on (Zhao et al., 2021; Holtzman et al., 2021; Zhou et al., 2024). Separation traces the failure to the geometry of the label space rather than to lost query information: a condition-level bias shifts every query's log-odds alike, so the queries with the weakest separation cross the decision boundary first, and the model emits one label regardless of input only after nearly all of them have crossed. A single forward pass at the query position predicts whether generation will collapse, at a fraction of the cost of iterative decoding, guarded by a label-extraction protocol that handles tokenizer and formatting artifacts.

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.

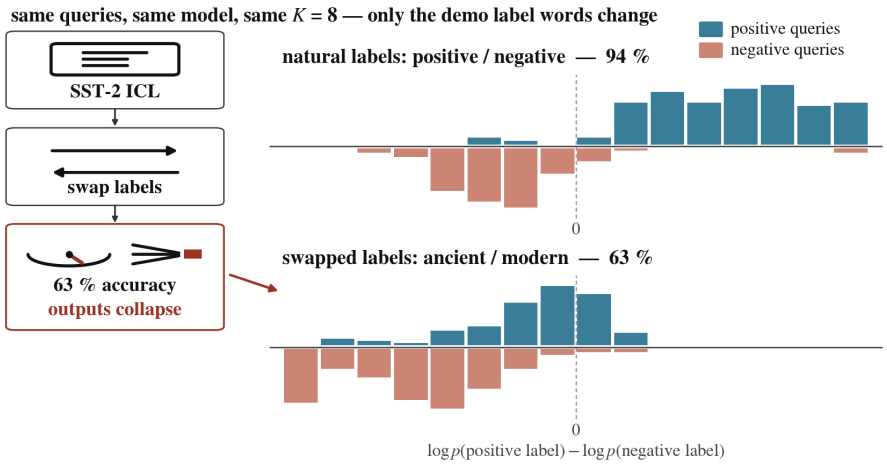


Figure 1: **Label collapse, hidden inside in-context learning.** Per-query log-odds of the two label words on the same queries: natural (top) vs. swapped (bottom) label words.

This geometric account yields a direct repair. Reading the log-odds with each demonstration's own text in the query slot yields a per-demonstration reading in which the condition-level bias appears as a shared offset; averaging within label groups and taking the paired midpoint cancels the parroting component—the part of each reading driven by the demonstration's own label word—and recovers the bias estimate $\hat{b}$. Subtracting $\hat{b}$ from the query reading defines *demo-anchored calibration* (DAC), which needs no test batch, no generated samples, and no extra forward passes beyond the demonstrations already in the prompt. DAC repairs most collapsed conditions across the six models, more than twice as many as the strongest calibration baseline that likewise requires no test batch, and its bias estimate transfers to free generation, predicting the degree of collapse across word pairs. We position DAC as the interventional leg of the causal chain (the bias estimate locates the cause, its removal repairs the failure) rather than as a competing calibration method; batch calibration, which sees the test batch, remains the upper bound and is compared in Section 4.

Two boundaries qualify the repair. The bias estimate is only moderately stable across demonstration counts, which we report openly, and how iterative denoising interacts with label biases remains open. Our contributions are:

- **Phenomenology.** We define label collapse and establish it as a recurring failure mode of dLLM ICL across six models, 20 word pairs, and four arrangements, with some word pairs collapsing in every model, collapse frequency stratified by model family, and arrangement-driven direction reversals (Section 3).
- **Invisibility under accuracy.** Collapse leaves accuracy at or only modestly above chance, so the failure is undetectable without model-internal diagnostics (Section 3).
- **Diagnostic.** We adopt separation, the label log-odds at the query position, as a model-agnostic single-pass diagnostic with a tokenizer-guarded extraction protocol, and show that the bias structure it exposes predicts which collapsed conditions calibration can and cannot repair (Section 4).
- **Causal repair.** We introduce demo-anchored calibration, which anchors on the demonstrations themselves, repairs most collapsed conditions without test-batch access, and whose bias estimate transfers to free generation (Section 4).

A full discussion of related work, covering label biases and demonstration order, ICL mechanisms, probing and logit lens, and diffusion language models, is in Appendix A.

2 EXPERIMENTAL SETUP

Models. We evaluate six diffusion language models: LLaDA-8B-Base (Nie et al., 2025), LLaDA-8B-Instruct, iLLaDA-8B-Base (Nie et al., 2026), iLLaDA-8B-Instruct, Dream-7B-Base (Ye et al., 2025), and Dream-7B-Instruct. The six models form three model families—LLaDA, iLLaDA, and Dream—each contributing a base variant and an instruction-tuned variant, so that family-level effects and instruction-tuning effects can be examined separately. iLLaDA is not an SFT (supervised fine-tuning) derivative of LLaDA, and Dream is trained with a masked-diffusion recipe independent of LLaDA's, so the three families differ in pre-training procedure rather than in architecture class. All models use `gen_length=3, steps=8, block_length=3, temperature=0`, under a unified decoding protocol (Appendix D). For Dream, iterative decoding rarely produces either label word (invalid-output rate ≥ 0.995 across all generation conditions; Appendix B), so its readable channel is the single pass.

Tasks and Label Word Pairs. We use two classification datasets: SST-2 (Socher et al., 2013) for sentiment classification and XSTest (Röttger et al., 2024) for safety classification; each word pair draws its queries and demonstrations from one of the two pools (SST-2 for 17 pairs, XSTest for three), and the gold label of a query is its dataset label mapped to the pair's positive or negative word—the same mapping that labels the demonstrations. We define 20 label word pairs across four semantic domains. **(i)** Sentiment: (1) positive/negative, **(ii)** Safety: (2) clean/toxic and (3) safe/unsafe, **(iii)** Geometric: (4) circle/square, **(iv)** Adjective and Abstract Domains: (5) true/false, (6) yes/no, (7) cat/dog, (8) hot/cold, (9) correct/wrong, (10) right/left, (11) easy/hard, (12) fast/slow, (13) strong/weak, (14) large/small, (15) light/dark, (16) ancient/modern, (17) kind/cruel, (18) quiet/loud, (19) rich/poor, and (20) wet/dry. Pairs (1)–(10) are *task pairs* and (11)–(20) *antonym pairs*. All prompts use a fixed template in which each demonstration is a sentence followed by its label word after "Answer:", and the prompt ends with "Answer:" after the query, with no trailing space. A complete prompt example is given in Appendix D.

Evaluation Metrics. We evaluate two modes: (1) **Generation (gen)**: the model generates the label token; we measure it by `pos_rate` (fraction of outputs matching the positive label) and by accuracy (an output is correct only if it matches the query's gold label word; invalid outputs, matching neither label, are incorrect); (2) **Single-pass (sp)**: from one forward pass we read the logits of the two label words w_+ and w_- at the masked (answer) position; their difference is the *separation*,

$$\mathrm{sep}(x) \;=\; \mathrm{logit}(w_+ \mid x) - \mathrm{logit}(w_- \mid x) \;=\; \log \frac{P(w_+ \mid x)}{P(w_- \mid x)}, \tag{1}$$

abbreviated *sep* in formulas (x: the query together with its demonstrations); the second equality is exact because the softmax denominators cancel. The single-pass `pos_rate` is the fraction of queries with $\mathrm{sep}(x) > 0$, and single-pass accuracy is the fraction whose argmax label matches the query's gold label word. Separation is thus the label log-odds, the quantity that label-bias calibration methods correct against reference points (Zhao et al., 2021; Holtzman et al., 2021; Zhou et al., 2024); we use it, without correction, as a diagnostic (Section 4). We use four severity categories with explicit membership conditions, all stated in terms of `pos_rate` and accuracy on the class-balanced query sets. (1) *Hard collapse*: `pos_rate` ≤ 0.05 or ≥ 0.95 *and* accuracy at chance (≤ 0.55), observed in every tested condition of a model–pair combination; the model emits one label word for virtually every query. (2) *Collapse*: `pos_rate` ≤ 0.20 or ≥ 0.80 with accuracy near chance (≤ 0.60). (3) *Drift*: `pos_rate` deviating from balance by at least 0.15 (i.e., ≤ 0.35 or ≥ 0.65) while accuracy remains well above chance (≥ 0.70); drift is *functional* when accuracy lies within 0.05 of the ceiling $0.5 + \min(p, 1-p)$ imposed by its own skew (p denotes `pos_rate`). (4) *Healthy*: `pos_rate` strictly within ± 0.15 of balance and accuracy ≥ 0.85. In later sections, the *balance band* means `pos_rate` strictly inside $(0.2, 0.8)$, the complement of the category-2 thresholds.

The thresholds are anchored to the observed distribution rather than tuned (hard collapse's 0.05 boundary sits just above the largest `pos_rate` of any hard-collapse condition, 0.03) and are stable under perturbation: shifting a single boundary by ± 0.05 reclassifies at most 7.0% of conditions, leaving the median grade unchanged.

Demonstration Configurations. We use four arrangements of the same demonstration pool: *alternating* (PNPN), *grouped* (PPNN), *blocked* (PPPNNN), and *shuffled* (a random permutation

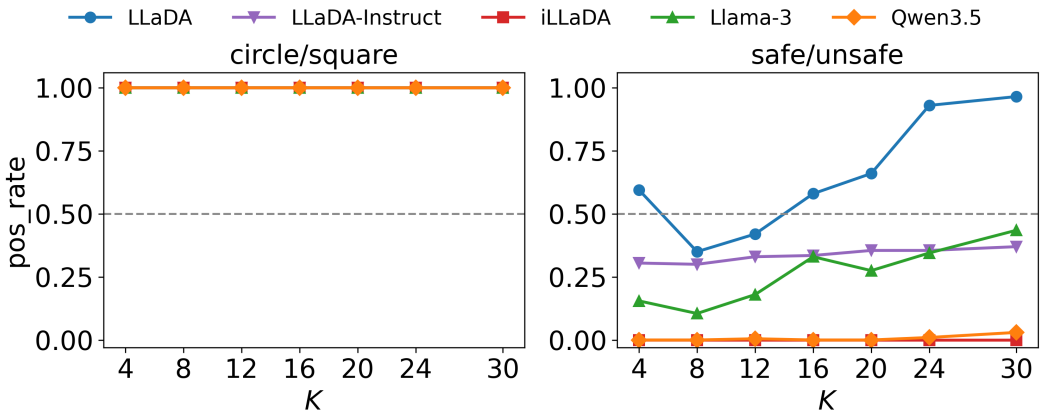


Figure 2: Label collapse across demonstration counts: generation `pos_rate` on the two pairs with alternating arrangement; remaining pairs in Appendix D.

redrawn independently for each query, from a fixed seed), where P and N denote the positive and negative labels; all four order the same label multiset, $\lceil K/2 \rceil$ positives and $\lfloor K/2 \rfloor$ negatives at each K (K: the number of demonstrations; constructions in Appendix D). The main grid sweeps all six models over alternating, grouped, and blocked at $K \in \{4, 8, 12, 16, 20, 24, 30\}$, the even-$K$ grid on which every arrangement is exactly class-balanced, with all 20 word pairs and both channels per condition. The shuffled arrangement is run in both evaluation channels for all six models (per-query random draws). All six models are additionally swept at the odd $K \in \{5, 7, 9, 11, 13\}$ under alternating in the single-pass channel, and three of them (LLaDA-Base, LLaDA-Instruct, iLLaDA-Base) extend to every odd K from 5 to 29 in both channels (supporting the parity analyses, Appendix G). As an unconditional baseline, K=0 (template + query only) is evaluated for all 20 word pairs in both modes. All cross-model snapshots use K=8 alternating. Each condition uses 200 query sentences.

3 Label Collapse Phenomenon

We now characterize the failure on the main grid of Section 2. At the K=8 alternating snapshot, 67 of 120 model–pair conditions collapse (Table 1); two word pairs collapse in all six models, and collapse frequency is stratified by family (LLaDA 28 of 40 conditions, Dream 28 of 40, iLLaDA 11 of 40); four AR baselines (Llama-3, Qwen3.5; base and instruct) (Grattafiori et al., 2024; Qwen Team, 2026) collapse at intermediate rates (Appendix C). The trigger is minimal (Figure 1): on the same 200 queries, swapping the label words assigned to the demonstrations of a healthy condition drops single-pass accuracy from 94% to 63%, with one word covering 170 of the 200 readings (LLaDA-Base, SST-2, alternating, K=8). Figure 2 and Table 1 resolve the conditions into three groups.

Three groups. The label words govern whether a condition collapses at all; the model family governs how often it does. (i) *Universal.* circle/square collapses in all six models and stays collapsed nearly everywhere: in generation, above 0.80 in all 84 conditions of the four models with readable generation (the three balanced arrangements, K=4–30); in the single pass, Dream-Base stays above 0.97 throughout, and Dream-Instruct leaves the collapsed band only at four large-K non-alternating conditions (0.41–0.77). At K=0, the four non-Dream models already lean "circle" (mean $P(\text{circle}) = 0.77$–0.98) and the Dream family leans "square" (0.16 and 0.29), yet demonstrations move all six models onto "circle": the prompt overrides the prior. (ii) *Largely healthy.* correct/wrong stays functional for five of six models (single-pass 0.49–0.59), collapsing only for Dream-Instruct. (iii) *Model-dependent.* safe/unsafe spans both extremes within the dLLMs: the iLLaDA family collapses hard toward "unsafe" (≤ 0.01 at every tested condition), the Dream family collapses hard toward "safe" (0.98–1.00), and the LLaDA family stays balanced under alternating (0.25–0.96).

Table 1: Single-pass `pos_rate` at the $K=8$ snapshot (alternating; 200 queries per cell; ✗ = collapse, △ = drift, unmarked = neither; –B/–I = base/instruct; Section 2).

Word pair	LLaDA-B	LLaDA-I	iLLaDA-B	iLLaDA-I	Dream-B	Dream-I
Collapsed in all six models						
ancient/modern	0.98 ✗	0.98 ✗	0.81 ✗	0.96 ✗	1.00 ✗	1.00 ✗
circle/square	1.00 ✗	1.00 ✗	1.00 ✗	1.00 ✗	1.00 ✗	1.00 ✗
Model-dependent						
large/small	1.00 ✗	0.96 ✗	0.81 ✗	0.46	1.00 ✗	1.00 ✗
cat/dog	0.55	1.00 ✗	0.60	0.85 ✗	1.00 ✗	1.00 ✗
clean/toxic	1.00 ✗	1.00 ✗	0.42	0.30 △	0.00 ✗	0.03 ✗
easy/hard	0.86 ✗	0.61	0.99 ✗	0.58	0.00 ✗	0.00 ✗
fast/slow	1.00 ✗	1.00 ✗	0.51	0.54	1.00 ✗	1.00 ✗
hot/cold	1.00 ✗	1.00 ✗	0.49	0.47	0.83 ✗	1.00 ✗
safe/unsafe	0.41	0.26 △	0.00 ✗	0.00 ✗	1.00 ✗	1.00 ✗
kind/cruel	1.00 ✗	1.00 ✗	0.47	0.51	0.00 ✗	0.40
light/dark	1.00 ✗	1.00 ✗	0.49	0.49	1.00 ✗	0.80 △
quiet/loud	1.00 ✗	1.00 ✗	0.71 △	0.56	1.00 ✗	0.54
true/false	0.44	0.42	0.00 ✗	0.00 ✗	0.99 ✗	0.56
wet/dry	1.00 ✗	1.00 ✗	0.48	0.52	1.00 ✗	0.78 △
positive/negative	1.00 ✗	1.00 ✗	0.51	0.49	0.64	0.37
rich/poor	1.00 ✗	1.00 ✗	0.54	0.53	0.21 △	0.78 △
right/left	1.00 ✗	1.00 ✗	0.60	0.52	0.51	0.29 △
strong/weak	0.50	0.43	0.51	0.48	0.85 ✗	1.00 ✗
yes/no	0.48	0.53	0.49	0.51	0.97 ✗	0.03 ✗
Healthy in five of six						
correct/wrong	0.49	0.57	0.52	0.58	0.59	1.00 ✗

Family divergence on antonym pairs. The antonym pairs separate the families. At $K=12$ under blocked, LLaDA-Base and LLaDA-Instruct each collapse on 8 of 10 antonym pairs, all toward the positive word except ancient/modern, which flips to "modern"; the iLLaDA family collapses on 1 of 10 (ancient/modern, both variants); the Dream family collapses most often in the single pass (Dream-Base 9 of 10) but in mixed directions. The alternating arrangement shows the same contrast already at $K=4$ (LLaDA-Base 7 of 10) and it widens with K (10 of 10 by $K=12$).

Inversion on the task's own pair. SST-2's own label pair, positive/negative, splits the field in the opposite direction: the LLaDA family collapses toward "positive" (21 of 21 and 19 of 21 generation conditions, accuracy 0.48–0.52), the iLLaDA family stays healthy (0 of 21, accuracy 0.92–0.96), and the Dream family drifts mildly in the single pass at $K=8$ (0.37–0.64, Dream-Instruct toward "negative", consistent with its $K=0$ prior of 0.33; grid-wide ranges in Section 5). The prior does not explain the split: at $K=0$, iLLaDA leans "positive" at 0.97, more strongly than LLaDA's 0.90, yet only LLaDA collapses; LLaDA-Instruct sits near balance at $K=0$ (0.56) and collapses by $K=4$.

Hard collapse is arrangement-indifferent. When collapse is hard, reordering does not rescue it. On circle/square, every model stays at 0.96–1.00 (collapse toward circle) across all arrangements and K, with four Dream-Instruct cells at large K dropping toward balance (0.41–0.77). The arrangement instead modulates the direction of weaker pairs (Section 6).

Effect of demonstration count. Demonstration count is not a reliable cure. On safe/unsafe, iLLaDA stays collapsed at or below 0.01 at every K and arrangement in both channels, LLaDA-Base moves toward balance as K grows and then flips its sign at large K (0.35 at $K=8$ to 0.96 at $K=30$, alternating), and the Dream family stays collapsed on the single pass throughout. Contrary to the monotonic gains reported for many-shot ICL (Agarwal et al., 2024), more demonstrations can harden rather than heal the bias (Appendix G for the parity analyses on the odd-K sweep of the six models).

4 MECHANISM: SEPARATION AS DIAGNOSTIC

This section localizes the bias inside the network, traces it to token priors that predate the demonstrations, asks when the single pass faithfully predicts generation, and turns the diagnostic into a repair.

Where the bias lives. In all six models the bias is present at or before the final projection layer; the intermediate layers do not repair it. Applying the logit lens (projecting each layer's hidden state through the output vocabulary) to every layer of all six models (33 layers for the LLaDA and iLLaDA families, 29 for Dream; layers are indexed L0, L1, …, with L0 the embedding layer; six word pairs, alternating, $K \in \{8, 30\}$), we find two emergence patterns that split by family. The LLaDA family shows label separation already at layer 0 (-3.4 to $+2.3$), consistent with a token-embedding-level bias that the contextual layers then carry to the answer position. The iLLaDA and Dream families stay near zero through the intermediate layers ($|\text{sep}| \leq 0.35$ at L0) and emerge only in the final layers (peak layers 28–32 for iLLaDA and 25–28 for Dream), so for them the bias is assembled late rather than inherited from the embedding. Under both patterns, the last projection step does not undo the offset, and no intermediate layer reduces it.

Token-level prior. Roughly seven in ten collapsed conditions point in the same direction as their zero-shot prior. We measure the prior with no demonstrations in the prompt ($K{=}0$): the separation at the answer position of the bare query already picks a side for most pairs and models, and its direction agrees with the $K{=}8$ collapsed reading in 46 of 67 cases (69%, single-pass channel). The agreement is a shared offset, not destiny: demonstrations can reverse the inherited direction, as on positive/negative, where iLLaDA's prior leans "positive" at 0.97 yet it stays healthy under demonstrations, and the Dream family's circle/square prior leans "square" yet every model collapses toward "circle" (Section 3). Three pairs constructed after these analyses, matched in token length, replicate the pattern, and the prior-direction exceptions again trace to a single family (Appendix I).

Calibration baselines. The fact that a single shared offset per condition restores balance is the formal signature of query-independent bias. On the $K{=}8$ snapshot (120 model–pair conditions, 67 collapsed, single-pass channel), standard calibrations that subtract a reference point repair only a few collapsed conditions: prior calibration 16 (Holtzman et al., 2021), in-context calibration 20 (He et al., 2024), contextual calibration 22 (Zhao et al., 2021), and domain calibration 24 (Fei et al., 2023). Batch calibration (Zhou et al., 2024), which subtracts the test-batch mean, returns all 67 collapsed conditions to balance—but it requires access to the test batch at deployment, and its success is exactly the statement that a per-condition offset is the failure.

Demo-anchored calibration. Anchoring on the demonstrations themselves recovers that offset without the test batch and repairs most collapsed conditions. Reading the log-odds with each demonstration's own text in the query slot yields a per-demonstration reading in which the condition-level bias appears as a shared offset; averaging within label groups and taking the paired midpoint cancels the parroting component—the part of each reading driven by the demonstration's own label word—and recovers the bias estimate $\hat{b} = \frac{1}{2}(\bar{s}_P + \bar{s}_N)$, with $\bar{s}_P$ and $\bar{s}_N$ the mean single-pass log-odds over the positive- and negative-labeled demonstrations. Subtracting $\hat{b}$ from the query reading defines *demo-anchored calibration* (DAC), which needs no test batch, no generated samples, and no extra forward passes beyond the demonstrations already in the prompt. DAC repairs 53 of the 67 collapsed conditions (a trimmed variant, dropping the largest |readout| per side, 54), more than twice the strongest baseline above (domain calibration); every condition repaired by any baseline is repaired by batch calibration, so the baselines are subsumed by the single-offset account rather than complementary to it. The repair is causal, not cosmetic: the same estimate transfers to free generation, where $|\hat{b}|$ predicts the degree of collapse across word pairs ($r{=}0.80$; 74 usable pairs, conditions with unreadable generation excluded). Together with the failed baselines and the batch-calibration upper bound, this completes the interventional leg of the causal argument.

Stability of the estimate. The estimate is only moderately stable across demonstration counts. Calibrated at $K{=}4$, DAC retains 90% of its repair power; across $K \in \{4, \ldots, 30\}$ the per-condition

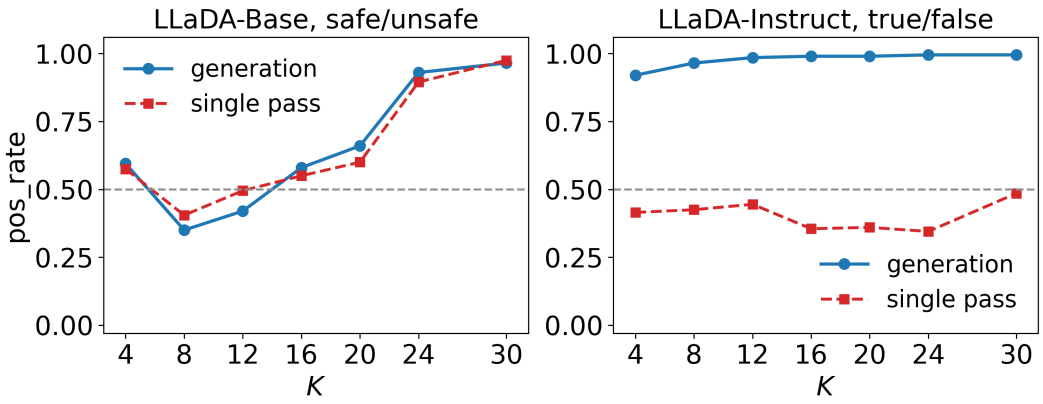


Figure 3: Generation vs. single-pass `pos_rate` on the same queries (alternating, K=4–30): (a) the two channels track each other; (b) generation collapses while the single pass does not.

coefficient of variation of $\hat{b}$ has median 0.29, which we report openly, and how iterative denoising interacts with label biases remains open.

Does the single pass predict generation? It predicts generation collapse with high precision, and its failures are informative. Across the 1,680 conditions with readable generation (four models, three balanced arrangements, K=4–30; generation readings follow the label-extraction protocol of Appendix F), generation and single-pass `pos_rate` agree within 0.10 in 84% of conditions and within 0.30 in 91%, and the criterion of extreme single-pass imbalance (`pos_rate` ≤ 0.05 or ≥ 0.95) predicts generation collapse with 93.7% precision (84.3% recall). The exceptions concentrate in two directions: a single pass that anticipates a collapse the iterative process never commits to, and generation that has flipped while the single-pass bias lags. The largest sustained exception is a decoupling condition (Figure 3): LLaDA-Instruct on true/false, where generation collapses to "true" while the single pass stays near balance yet retains classification accuracy (0.81–0.94)—the single pass keeps query dependence that iterative denoising erases.

5 Model Family and SFT Effects

Training can move label bias in opposite directions, and the six models provide two kinds of comparison. *Same-family SFT*: each family contributes a base and an instruction-tuned variant that share pre-training weights, so each family isolates the effect of chat-style SFT (chat SFT for short), and the three families replicate each other. *Cross-family*: iLLaDA is pre-trained from scratch on a different corpus (12T tokens vs. LLaDA's 2.3T) (Nie et al., 2026), so iLLaDA-vs-LLaDA differences confound pre-training data, scale, and procedure, and we attribute them to the family as a whole. This section examines three effects: (i) repair across families, (ii) the family dependence of majority conformity under imbalanced demonstrations, and (iii) the family and SFT dependence of the gen–sp decoupling; a fourth, a stable third-word answer introduced by chat SFT, is documented in Appendix E.

Repair across families. The positive/negative repair is a pre-training effect, and chat SFT does not reproduce it. On positive/negative, the iLLaDA family is healthy (0 of 21 and 0 of 21 generation conditions, accuracy 0.92–0.96) while the LLaDA family collapses toward "positive" (21 of 21 and 19 of 21, accuracy 0.48–0.52); the Dream family sits in between, with Dream-Base only drifting (0.53–0.86 across its 28 single-pass conditions, 2 leaving the balance band) and Dream-Instruct collapsing toward "negative" (20 of 28). All three base variants are pre-trained without chat-style fine-tuning, so the ordering iLLaDA-healthy / Dream-intermediate / LLaDA-collapsed reflects pre-training procedures rather than fine-tuning: the collapse-prone label bias is acquired during pre-training and can be avoided or softened by a different recipe. The contrast extends to the antonym pairs: under alternating at K=8, the iLLaDA family collapses on 3 and 1 of 10 antonym

pairs, versus 9 and 8 for the LLaDA family and 9 and 5 for Dream. Chat SFT does not reproduce the repair: LLaDA-Instruct collapses toward "positive" at every tested K (0.87–1.00).

SFT effects are bidirectional and pair-dependent. Chat SFT adds failure modes in one family while removing some in another. On positive/negative, Dream-Instruct collapses where Dream-Base only drifts (20 of 28 single-pass conditions versus 2, toward "negative"), whereas on the antonym pairs under alternating it collapses less than Dream-Base (5 versus 9 of 10). LLaDA-Instruct sits on the collapsed side of its base on positive/negative (19 of 21) and answers outside the label pair with a single stable third word ("neutral") that more demonstrations gradually retire (Appendix E).

Conformity under imbalance is a family property. Which imbalance makes a model follow the majority is set by the model family, not by the architecture. Replacing the class-balanced alternating arrangement with imbalanced ones (majority share 66.7–75%; PPN and PPPN constructions, Appendix K) splits the six models along family lines. The LLaDA and Dream families follow the majority label almost completely already at these shares (median `pos_rate` 0.95–1.00), whereas the iLLaDA family stays near balance (median 0.48–0.50; only 4–8 of the 20 pairs collapse outright). The iLLaDA resistance is not invulnerability—its antonym pairs collapse under blocked arrangements (Section 3)—but within this grid no architecture effect separates the three masked-diffusion families: two conform, one resists.

gen–sp decoupling is not created by chat SFT. Persistent decoupling occurs in a base model and a chat-tuned model alike, so iterative denoising, not SFT, erases the query dependence. Two persistent cases survive extraction screening (both pairs whose label words are single tokenizer tokens): LLaDA-Instruct on true/false (generation 0.92–0.99 versus single pass 0.34–0.48, at every K) and iLLaDA-Base on right/left (generation 0.20–0.39, accuracy 0.23–0.52, versus single pass 0.50–0.67, accuracy 0.79–0.93, at every K). Since the second case is a base model, the decoupling is a property of iterative denoising itself—one that family and word pair gate—rather than an artifact of chat tuning; where it occurs, the single pass retains the query dependence that generation loses, which is what makes separation usable as a diagnostic. The largest raw gen–sp gaps trace to largely unparseable outputs on the multi-token pair ancient/modern—extraction losses (Appendix F), not behavioral decoupling.

Synthesis. Which label pairs are dangerous is a property of the training procedure, not of the architecture. Chat SFT adds failure modes (a stable third-word answer, direction shifts on positive/negative) without repairing the positive/negative collapse, while a different pre-training procedure repairs positive/negative and most antonym pairs but leaves safe/unsafe hard-collapsed, where LLaDA-Base retains partial function.

6 Arrangement Effects

This section asks whether rearranging the same class-balanced demonstrations can prevent collapse. We find that (i) no arrangement is uniformly safest or most harmful, (ii) collapse driven by a strong token prior is indifferent to arrangement, and (iii) on weaker pairs arrangement can reverse *which* label the model collapses to.

Arrangement sensitivity is model and pair dependent. No arrangement is uniformly safest, and the default alternating protocol is, on average, slightly the most severe. At $K=8$ (single-pass channel), alternating is the most severe arrangement on average for five of the six models and within 0.01 of the most severe for the sixth (Dream-Instruct, where shuffled edges it), over the 20 word pairs, though the margins are small (alternating 0.15–0.41 mean deviation from balance, vs. 0.11–0.40 grouped, 0.12–0.40 blocked, and 0.11–0.41 shuffled, with the Dream family's margins within 0.01 throughout), and at the individual-pair level both directions occur: grouped is more severe than alternating for 1–9 of the 20 pairs per model, alternating more severe for 4–10. The diagnostic value is the contrast itself: arrangement-sensitive pairs compute their bias partly from demonstration order, while arrangement-indifferent pairs read it out of the model regardless of order—the two groups below. Under shuffled, the two evaluation channels agree as closely as on the balanced arrangements (83%/91% within 0.10/0.30, 560 conditions).

Strong-bias collapse is arrangement-indifferent. When the token prior is strong, no arrangement averts the collapse. On safe/unsafe, the iLLaDA family emits the negative word at rate 1.00 under every arrangement at every K, on both channels; on ancient/modern, both Dream models emit the positive word at 0.98–1.00 under all four arrangements, including shuffled; on circle/square, every model stays collapsed across arrangements and K (Section 3). Rearranging changes what the model sees first, last, and interleaved; when none of it matters, the bias does not pass through demonstration processing at all.

Arrangement reverses the collapsed label on weaker pairs. The same class-balanced demonstrations can force opposite collapsed labels, differing only in order. Figure 13 (Appendix K) traces both cases across the four arrangements: what changes from arrangement to arrangement is not only the distance from the dashed balance line but which side of it. On ancient/modern, the LLaDA and iLLaDA families collapse to the positive word under alternating (0.81–1.00, median 1.00) but fall back toward balance under grouped, blocked, and shuffled (medians 0.54–0.60), with LLaDA-Instruct crossing to the negative word under blocked (down to 0.01). On cat/dog, LLaDA-Instruct collapses to "cat" under alternating (0.68–1.00) but leans to "dog" under every other arrangement in the single pass (grouped 0.00–0.49, blocked 0.00–0.27 reaching 0.00 from $K{=}12$ on, shuffled 0.06–0.28); in the generation channel the alternating collapse persists (0.83 at $K{=}8$) while the other two arrangements stay near balance (0.50 grouped, 0.42 blocked; Table 5). Arrangement is thus a latent variable in every reported ICL accuracy: a fixed test set can yield opposite systematic answers depending only on demonstration order. Block position adds a second lever: collapse follows the trailing block rather than the majority count, and the effect of K depends on whether a minority exists (Appendix K).

7 Conclusion and Discussion

We presented a systematic study of in-context learning failure modes in diffusion language models: six models from three families, 20 label word pairs, four arrangements, and two evaluation channels. Label collapse is architecturally universal but graded: every model collapses on some pairs, some pairs collapse in every model, and the frequency stratifies by model family rather than by instruction tuning. Balanced-set accuracy conceals the failure, and separation exposes it in a single forward pass.

Mechanistically, which label pairs are dangerous is a property of the training procedure rather than the architecture, and the bias is largely readable at the output head before any demonstration is seen. A demonstration set with a single label (majority share 100%, Appendix K) overrides the query entirely, while a minority restores enough structure that the effect of demonstration count becomes arrangement-dependent. The failure is repairable from the demonstrations themselves, without test-batch access, and the repair transfers to free generation—with one systematic exception, where generation collapses although the single pass remains discriminative.

These findings translate into protocol guidance: audit label tokens before running an evaluation, report demonstration order alongside accuracy, and include a single-pass diagnostic in the pipeline.

Which ingredient of training makes a given label pair dangerous, whether label collapse extends beyond binary classification under a fixed template, and why iterative denoising discards the query dependence that the single pass retains all remain open; we develop these questions, further implications, and future extensions in Appendices L and M.

Limitations

Our study covers six dLLMs in the 7B–8B range (four 8B–9B AR baselines enter only the census appendix), two English classification datasets, and one fixed binary-choice template, so collapse at other scales, in other languages, and in other task formats is unmeasured. Our mechanistic account stops at the output layer: intermediate-layer interventions did not yield reliable repairs, so the failure is corrected rather than explained from inside the network. The γ-sweep and block-order imbalance grids (Appendix K) cover the five models available when they were run; only the PPN/PPPN constructions include Dream (single-pass channel).

REPRODUCIBILITY STATEMENT

All source code and anonymized summary data are available at `https://anonymous.4open.science/r/code_release-53E1`. The evaluation uses publicly available models and datasets; full hyperparameters appear in Section 2 and Appendix N.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2404.11018. arXiv:2404.11018.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1107–1128, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Yuxin Dong, Jiachen Jiang, Zhihui Zhu, and Xia Ning. Understanding task vectors in in-context learning: Emergence, functionality, and limitations. *arXiv preprint arXiv:2506.09048*, 2025.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2023. URL https://arxiv.org/abs/2305.19148. arXiv:2305.19148.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. How robust are LLMs to in-context majority label bias? In *AAAI 2024 Workshop on Responsible Language Modeling*, 2024. URL https://arxiv.org/abs/2312.16549. arXiv:2312.16549.

Kang He, Yinghan Long, and Kaushik Roy. Prompt-based bias calibration for better zero/few-shot learning of language models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 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)*. Association for Computational Linguistics, 2021. URL https://arxiv.org/abs/2104.08315. arXiv:2104.08315.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. *arXiv preprint arXiv:2502.08009*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2022. URL https://arxiv.org/abs/2104.08786. arXiv:2104.08786.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2502.09992. Oral. arXiv:2502.09992.

Shen Nie, Qiyang Min, Shaoxuan Xu, Zihao Huang, Yuxuan Song, Yong Shan, Yankai Lin, Wayne Xin Zhao, Chongxuan Li, and Ji-Rong Wen. Improved large language diffusion models. *arXiv preprint arXiv:2606.25331*, 2026. URL https://arxiv.org/abs/2606.25331.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8nwDdndpJiDw/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://arxiv.org/abs/2209.11895. arXiv:2209.11895.

Qwen Team. Qwen3.5 model series. Model cards, Hugging Face, 2026. URL https://huggingface.co/Qwen. Initial release 2026-02-16.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.07524. arXiv:2406.07524.

Mario Sanz-Guerrero and Katharina von der Wense. Mitigating label length bias in large language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP*, 2025.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. On the loss of context-awareness in general instruction fine-tuning. *arXiv preprint arXiv:2411.02688*, 2024. URL https://arxiv.org/abs/2411.02688.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025. URL https://arxiv.org/abs/2508.15487.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 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*. PMLR, 2021. URL https://arxiv.org/abs/2102.09690. arXiv:2102.09690.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

A Related Work

We review label biases and demonstration order in ICL, mechanistic accounts, and diffusion language models, closing with the interpretability tools adopted in our analysis.

Label Biases and Demonstration Order in ICL. Lu et al. (Lu et al., 2022) showed that permuting the same demonstrations can swing ICL accuracy from near-random to near-optimal, establishing order sensitivity as a central concern. Min et al. (Min et al., 2022) found that randomly replacing demonstration labels barely hurts performance: demonstrations primarily convey the label space, the input distribution, and the response format. Our results expose the flip side: while the *truth* of demonstration labels may be optional, *which words occupy the label space* is consequential; switching the label pair from positive/negative to circle/square induces query-independent collapse across all six dLLMs. Zhao et al. (Zhao et al., 2021) identified majority label bias, recency bias, and common token bias (Holtzman et al., 2021) and proposed contextual calibration; later work measured robustness to the majority-label bias induced by skewed demonstration distributions, finding robustness boundaries that vary widely across models and tasks (Gupta et al., 2024), and defined a typology of vanilla, context, and domain label biases (Fei et al., 2023). All of these methods correct the label log-odds (the quantity we measure as separation) against a chosen reference point; Zhou et al. (Zhou et al., 2024) unify the family as subtracting a context bias and show that a batch-mean reference dominates. Our use is orthogonal: rather than correcting predictions, we read the joint distribution of the shared bias and the query-dependent component, which turns the family itself into a probe of where the bias lives (Appendix H). The domain-label bias of Fei et al. (Fei et al., 2023) (task accuracy pinned at chance by the label words themselves) anticipates the label-word sensitivity we isolate; we extend it to strictly class-balanced arrangements, across architectures, and with the bias decomposed at the output head. Our collapse arises under strictly class-balanced arrangements, placing it outside the majority-label-bias setting; label words have been varied only jointly with wholesale template changes (Zhao et al., 2021), never isolated under a fixed template, and collapse has not been characterized as a query-independent failure across architectures. This is the gap our paper addresses.

Induction Heads and ICL Mechanisms. Induction heads (Olsson et al., 2022) are a hypothesized mechanism for much of in-context learning, raising the question of whether label collapse is simply an induction failure. The evidence in Section 4 argues that it is not: collapse is label-word-specific rather than pattern-specific, its direction is arrangement-invariant for strong-bias pairs, and demonstration and query information remain decodable even when the model is collapsed. Induction heads explain how models copy patterns, but not *which* label token dominates when the copied signal and the output projection disagree; our separation diagnostic measures exactly this disagreement. Complementing this mechanistic line, task-vector analyses locate the demonstration-induced adaptation in low-dimensional representation shifts (Dong et al., 2025), and geometric studies contrast prompt-induced adaptation with the mechanism that transforms base representations during instruction tuning (Kirsanov et al., 2025); both operate at the representation level, whereas our diagnostic reads the label decision directly at the output projection.

Diffusion Language Models and ICL. Discrete diffusion LMs generate text by iteratively denoising a masked sequence rather than decoding left to right (Lou et al., 2024): MDLM (Sahoo et al., 2024) showed that simple masked diffusion with a simplified objective is more performant than previously thought, LLaDA (Nie et al., 2025) scaled it to 8B parameters, Dream (Ye et al., 2025) demonstrated competitive reasoning and planning, and iLLaDA (Nie et al., 2026) was trained from scratch with fully bidirectional attention (12T tokens), releasing Base and SFT variants. SFT is known to degrade pre-trained capabilities in AR models (Wang et al., 2024); we extend this observation to dLLMs, where it can alter label bias directions between Base and Instruct variants. The ICL failure modes of dLLMs remain uncharacterized; we provide the first systematic evaluation of them, showing that dLLMs share the AR label-collapse phenotype while raising new questions about how iterative denoising interacts with label biases.

Probing and Logit Lens. Linear probing (Belinkov, 2022) trains classifiers on hidden states to test what information is encoded; the logit lens (nostalgebraist, 2020) decodes intermediate layers through the output vocabulary. We apply the logit lens to dLLMs, reading label-token logits at the

Table 2: Dream variants at $K=8$ alternating (single pass, 20 word pairs): collapse counts by tier, median `pos_rate`, and direction–prior agreement ($\equiv$ marks agreement); column definitions in the text.

Model	hard	any	median	$K=0$ med.	dir. $\equiv$ prior
Dream-7B-Base	14/20	16/20	0.985	0.750	16/20
Dream-7B-Instruct	12/20	12/20	0.785	0.672	15/20

masked answer position across all transformer layers (six word pairs spanning the severity range—circle/square, ancient/modern, safe/unsafe, positive/negative, strong/weak, easy/hard; Section 4), to locate where the label bias is set; extending trained probes and denoising-step trajectories to dLLMs remains open.

B The Dream Lineage

Dream-7B (Ye et al., 2025) is a masked diffusion language model trained independently of the LLaDA recipe; we evaluate its Base and Instruct variants under the main-grid protocol (20 word pairs, all four arrangements plus the zero-shot baseline, both channels, 200 queries per condition; 2,320 records). Iterative decoding never enters the label space: across all 1,160 generation conditions, the invalid rate never falls below 0.995—the label words are essentially unreachable by free denoising, so the readable channel for Dream is the single pass, which is exactly the channel the separation diagnostic targets. The single-pass behavior matches the main phenomenon (Table 2): at $K=8$ alternating, 28 of 40 model–pair conditions sit outside the balance band (≥ 0.30; Dream-Base 16/20, Dream-Instruct 12/20; hard tier ≥ 0.45: 14 and 12), with median `pos_rate` 0.985 and 0.785. The collapse direction is set before any demonstration: it agrees with each variant's own $K=0$ single-pass prior in 16/20 (Base) and 15/20 (Instruct) of conditions, and the $K=0$ medians (0.750, 0.672) sit inside the same positive-leaning range as the other primary dLLMs. The priors themselves do not transfer across lineages: Dream's $K=0$ directions agree with iLLaDA's in only 9/20 (Base) and 8/20 (Instruct) of pairs, consistent with the model-internal reading of the label prior (Section 4).

C Ten-Model Census

Under a uniform snapshot protocol (the 20 word pairs, $K=8$ alternating, both channels, 200 queries per condition), the six primary dLLMs of the main text join four supplementary AR baselines—two families in base and instruct variants (Llama-3-8B, Qwen3.5-9B) (Grattafiori et al., 2024; Qwen Team, 2026)—in a ten-model census; the Dream variants are analyzed in Appendix B, and iLLaDA-Instruct is a main-grid model. Table 3 reports single-pass collapse counts in the two tiers used throughout. Within every model family (lineage), the two variants differ by at most three collapsed conditions (hard tier: Dream 14 vs. 12, LLaDA 13 vs. 14, iLLaDA 4 vs. 4, Llama-3 12 vs. 10, Qwen3.5 5 vs. 2), while across lineages the counts span 2–14: the LLaDA and Dream lineages sit at the heavy end, iLLaDA at the robust end, and the two AR lineages spread across the middle (Qwen at 2–5, Llama-3 at 10–12); the uniform-snapshot counts may differ by one or two conditions from the main-grid statistics of Section 3, which apply the parseability flags of Appendix F (e.g., Llama-3 shows 12 here vs. 13 under the re-parsed readings). The census is a coverage claim, not an SFT claim: the SFT-specific conclusions of Section 5 rest on the three within-family Base/Instruct contrasts.

Supplementary K-sweep, calibration, and arrangement materials of five-model vintage (the LLaDA and iLLaDA families and two AR baselines) predate the Dream extension; they are retained for the per-pair diagnostic detail they add, and their captions should be read in that frame.

D Extended Experimental Details

Prompt template and arrangement examples. Each demonstration is a sentence drawn from its pair's demonstration pool (SST-2 movie reviews or XSTest statements; Section 2) followed by its label on the next line, and the query is followed by "Answer:" with no trailing space. The following

Table 3: Single-pass collapse counts at $K=8$ alternating over the 20 word pairs, ten models, uniform snapshot; tiers as defined in Section 2.

Lineage	Model	hard	any
LLaDA	LLaDA-8B-Base	13/20	14/20
	LLaDA-8B-Instruct	14/20	14/20
iLLaDA	iLLaDA-8B-Base	4/20	6/20
	iLLaDA-8B-Instruct	4/20	5/20
Dream	Dream-7B-Base	14/20	16/20
	Dream-7B-Instruct	12/20	12/20
Llama	Llama-3-8B-Base	12/20	15/20
	Llama-3-8B-Instruct	10/20	12/20
Qwen	Qwen3.5-9B-Base	5/20	7/20
	Qwen3.5-9B-Instruct	2/20	5/20

concrete example shows the alternating (PNPN) arrangement at $K=4$, using real demonstrations from our SST-2 pool:

> exciting and direct , with ghost imagery that shows just enough to keep us on our toes .
> **Answer:** positive
> too slow , too long and too little happens .
> **Answer:** negative
> it 's worth seeing just on the basis of the wisdom , and at times , the startling optimism , of the children .
> **Answer:** positive
> makes for a pretty unpleasant viewing experience .
> **Answer:** negative
> *light years / several warp speeds / levels and levels of dilithium crystals better than star trek xi* .
> **Answer:**

The four arrangements differ only in the ordering of the same demonstration pool: at every K, each orders the same label multiset, $\lceil K/2 \rceil$ positives and $\lfloor K/2 \rfloor$ negatives. Alternating prefixes the label cycle (P, N, P, N, ...), so its prompts are nested across K; grouped concatenates the two classes (P, P, N, N at $K=4$); blocked alternates P- and N-blocks of up to three, shortening the tail block when K is not divisible by six (PPPNNNPN at $K=8$); shuffled permutes the multiset at random, drawn independently at each K (fixed seed). Each query draws an independent class-balanced permutation; the draw sequence is fixed by the seed and K alone, so it is shared across models and word pairs at the same K and unrelated across different K; unlike the alternating prefixes, shuffled does not form a nested family across K. All four are exactly class-balanced at even K; at odd K each carries one extra positive, and alternating additionally ends on a positive demonstration (Appendix G). An earlier variant of this template used a trailing space after the final colon (`Answer: `); the two variants differ in the tokenization of the label position and in bias direction, and all results in this paper use the no-space variant. The template itself was fixed before the main experiments and selected for label-output validity: candidate templates with weaker answer-position cues leave the generation channel largely unreadable (invalid-output rates up to 0.95 on a leading-space variant, including numeric outputs), while the selected template yields single-token label outputs for 0.92–1.00 of queries on the anchor pairs (Appendix F).

K grids. Table 4 lists the exact K values per model and arrangement.

Condition counts. The main grid comprises 5,040 conditions: 2,520 generation and 2,520 single-pass, each with 200 queries (six dLLMs, three balanced arrangements, seven K). Generation is fully readable for the four LLaDA- and iLLaDA-family models; Dream's iterative decoding rarely enters the label space (Appendix B), so its generation conditions are reported but excluded from paired channel analyses. The analyses of Section 4 use the 1,680 paired conditions that remain (four models, three balanced arrangements, $K \geq 4$; generation readings follow the re-parsed values of Appendix F). The shuffled arrangement adds 1,680 conditions under the same protocol (six dLLMs,

Table 4: Demonstration-count coverage per model and arrangement (seven even K, 20 word pairs, both channels). Six dLLMs also sweep the odd $K \in \{5, 7, 9, 11, 13\}$ under alternating in the single-pass channel, and three of them (LLaDA-Base, LLaDA-Instruct, iLLaDA-Base) cover every odd K from 5 to 29 in both channels (Appendix G); the AR rows inherit their balanced-arrangement sweeps from the five-model supplementary grids and join the census snapshot (Appendix C); shuffled was not run for the AR models.

Model	alternating	grouped	blocked	shuffled
LLaDA-8B-Base	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
iLLaDA-8B-Base	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
LLaDA-8B-Instruct	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
iLLaDA-8B-Instruct	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
Dream-7B-Base	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
Dream-7B-Instruct	4–30 (7)	4–30 (7)	4–30 (7)	4–30 (7)
Llama-3-8B (base/instruct)	4–30 (7)	4–30 (7)	4–30 (7)	—
Qwen3.5-9B (base/instruct)	4–30 (7)	4–30 (7)	4–30 (7)	—

both channels), and the $K{=}0$ baseline 240 (six models, both channels where readable; zero-shot generation outputs are largely unparseable, $47/50$ on the original five-model grid, Appendix H). The three post-hoc equal-length pairs of Appendix I (420 single-pass conditions, 405 with $K \geq 4$, and 405 generation conditions), and their Dream single-pass (42-condition) and iLLaDA-Instruct (162-condition) extensions, are likewise evaluated outside the main grid and excluded from all counts above. A small set of exploratory re-runs, consisting of 12 generation and 12 single-pass conditions using demonstration pools from other domains (fact statements for true/false, math solutions for correct/wrong, at $K \in \{4, 8\}$ under alternating), and 39 logit-lens conditions are excluded from all counts and analyses. Three supplementary grids are likewise evaluated outside the main grid and excluded from all counts above: the ten-model snapshot and Dream lineage of Appendices C and B (1,440 records per non-Dream model, 2,320 for Dream), the imbalance and block-order grids of Appendix K (5,600 paired conditions), and the layer-level interventions of Appendix J (10 conditions $\times$ 10 layers $\times$ two channels).

Model-specific configurations. LLaDA-8B-Base/iLLaDA use `transformers==4.46.3`; Qwen3.5 uses `transformers>=5.15`. All models use `torch_dtype=torch.bfloat16`. Exact environment pins are recorded in the released code.

Compute resources. All experiments ran on a single rented NVIDIA A100-40GB GPU (bf16, inference only, no training). Total usage across all experimental versions is estimated at roughly 1,000 GPU-hours ($\pm$30%), based on output-file timestamp ranges over the 67-day project period.

Asset licenses. Licenses, verified via the Hugging Face API: LLaDA-8B weights under MIT; iLLaDA-8B weights under Apache-2.0; Llama-3 (accessed via the NousResearch mirror of Meta-Llama-3-8B-Instruct) under the Meta Llama 3 Community License; Qwen3.5-9B-Base and Dream-7B under Apache-2.0 and MIT respectively; XSTest under CC-BY-4.0; SST-2 carries no unified license (GLUE defers to the terms of the original dataset, which we cite).

Decoding parameters. For dLLMs, `block_length=3` constrains remasking to contiguous blocks of three generated tokens: with `gen_length=3`, the entire answer (three positions) forms a single block, refined over `steps=8` denoising iterations with low-confidence remasking. In the released code, the single-pass (sp) mode is named `ppl` for historical reasons.

On positive/negative the other four models lean toward "positive" while iLLaDA stays balanced; on clean/toxic LLaDA-Base and LLaDA-Instruct collapse toward "clean" while iLLaDA and Qwen3.5 stay near balance (Figure 4).

The continuous single-pass criterion $|\text{sep}| > 3$, the log-odds counterpart of the hard-collapse margin, is used only as a model-internal predictor, not as a severity definition.

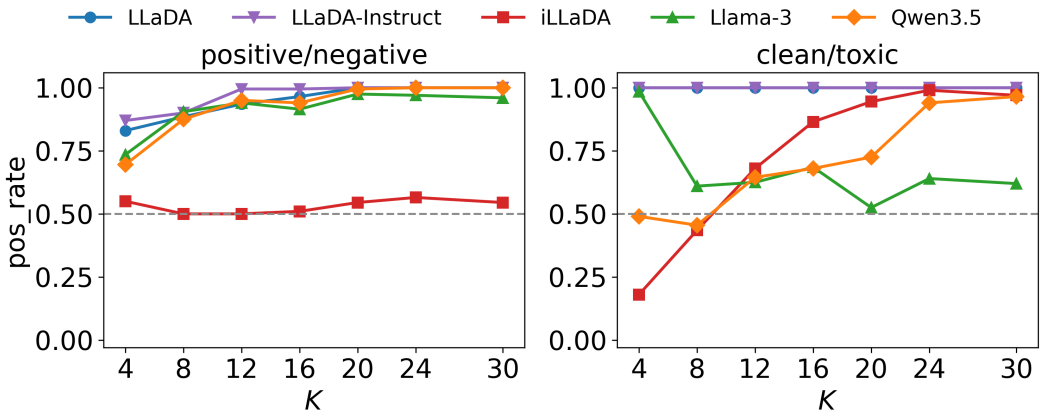


Figure 4: K-sweep generation `pos_rate` for the remaining two representative pairs (positive/negative, clean/toxic; alternating; gray dashed line = 0.5), complementing Figure 2.

Cluster-bootstrap confidence intervals. The 1,680 paired conditions are nested (20 word pairs, four models, three arrangements), so confidence intervals that treat conditions as independent understate the uncertainty. Resampling word-pair clusters with replacement (2,000 resamples) gives the extreme single-pass imbalance criterion (`pos_rate` ≤ 0.05 or ≥ 0.95) a 95% CI of $[88.8\%, 97.2\%]$ for precision and $[70.8\%, 95.1\%]$ for recall; the lower bound of the precision CI stays within 88.8–90.6% under model-, model–pair-, or model–pair–arrangement-level clustering, whereas the recall CI widens under model-level clustering ($[74.1\%, 88.3\%]$), which is why the main text leans on precision.

E Neutral Third-Word Answers under Chat SFT

Chat SFT introduces a stable third-word answer that demonstrations gradually retire. LLaDA-Instruct sometimes answers with a word outside the label pair, and the word is almost always the same one: "neutral". The small-K alternating runs with $1 \leq K < 4$ (outside the main grids), used only in this analysis, support a finer view of this effect: from K=3 through K=8, "neutral" accounts for 91–97% of all out-of-pair outputs, at rates of 0.09–0.19 per condition, before declining to zero by K=24 (at K=2 the share is 77%, the remainder being truncated spellings of "negative"). Two regimes thus coexist in the chat-tuned model. Without demonstrations, it reaches for whatever answer words it knows: in zero-shot conditions, the out-of-pair outputs are scattered across the label vocabularies of other tasks (numbers, "yes"/"no", star ratings), with no "neutral" at all. With even a few demonstrations, it converges on this single stable third word, which further demonstrations gradually retire. LLaDA-Base never produces out-of-pair answers at any K, so the effect is specific to the chat-tuned variant.

F Parsing Protocol and Artifact Audit

Generation outputs are read at the text level: the first content token of the output (skipping leading newlines and whitespace) is matched, case-insensitively, against the two label words of the pair; `pos_rate` is the fraction of the 200 queries whose output matches the positive label, and outputs matching neither word count as invalid. Three artifact classes initially confounded this reading; cells where more than 30% of outputs remain invalid even after correction are noted below:

(i) **Tokenizer splits.** A label word may tokenize into multiple pieces, so first-token matching fails: "ancient" splits in all three dLLM vocabularies ("an"+"cient"), "toxic" in Qwen3.5 ("to"+"xic"), and "clean", "kind", "cruel", "quiet", "rich", and "wet" split in various models. On ancient/modern under alternating, LLaDA-Base outputs "ancient" for 200/200 queries at K=12 while the first-token parser records a parse error and `pos_rate` 0.00; text-level matching restores the correct reading.

(ii) **Format shifts.** Llama-3 capitalizes its answers ("TRUE"/"FALSE" on true/false), and Qwen3.5 prefixes answers with a newline (or "Answer:") followed by free-form continuations. Case-insensitive matching and newline skipping restore these.

(iii) **Genuine out-of-pair answers.** These are not artifacts: Qwen3.5 answers "wrong" or "middle" on right/left at small K (Section 3), and such outputs remain invalid by design.

We re-parsed all 1,624 recorded generation conditions (the 1,612 main grid plus the 12 exploratory) under this protocol; per condition we retain whichever reading (original or corrected) yields the lower invalid rate. We hand-verified three anchors: LLaDA-Base ancient/modern, alternating $K{=}12$ (`pos_rate` $0.00 \rightarrow 1.00$; 200/200 outputs "ancient"); Llama-3 true/false, alternating $K{=}30$ ($0.25 \rightarrow 0.85$); iLLaDA ancient/modern, alternating $K{=}4$ ($0.00 \rightarrow 0.91$). Three cells remain in that regime (iLLaDA right/left, Llama-3 right/left, Llama-3 wet/dry): their outputs leave the label vocabulary under both readings.

Extraction audit and template ablation. Two audits support the generation readings used in the main text. First, under the selected template, parse errors are dominated by multi-token label words rather than noise: the whole-word subclass accounts for roughly 70–90% of parse errors across the four non-Dream models, and numeric outputs vanish, so generation `pos_rate` is conservative but not distorted. A looser whole-string reading (counting any output whose full text equals a label word) flips the collapse verdict of 7 of 120 $K{=}8$ conditions, all in the iLLaDA family and all on multi-token pairs; every condition quoted in the main text involves single-token pairs and is unchanged under the two readings. Second, a five-template ablation on four single-token pairs shows that the generation–single-pass decoupling of Section 4 is stable in direction across the three variants that preserve the answer cue, while failing under two degenerate variants for attributable reasons: the leading-space variant collapses the extraction floor itself (label validity 0.37), and a no-cue variant that deletes "Answer:" removes the answer-position semantics rather than the diagnostic. Dream is excluded from generation-side audits (Section 2).

G Parity Structure of the Alternating Arrangement

The alternating arrangement has a structural property that interacts with the demonstration count K: an alternating block of even length ends on a negative demonstration and is exactly class-balanced, whereas a block of odd length ends on a positive demonstration and carries one extra positive. Parity of K therefore changes both the terminal label and the class balance of the prompt. Under alternating we observe parity-structured behavior on several model–pair combinations: on safe/unsafe, the reversal of LLaDA-Base appears only on even $K \geq 24$, its single pass staying in the collapsed band at every odd K (0.08–0.19, no reversal), while LLaDA-Instruct sits on the negative side of balance throughout with a weak parity echo (0.20–0.25 at odd K vs. 0.26–0.39 at even) and never collapses outright; on clean/toxic, iLLaDA stays near hard collapse on odd K (`pos_rate` ≤ 0.07) while rising through 0.23–0.98 on even K. The direction of these oscillations is consistently opposed to what recency bias and majority label bias predict (the model favors the label demonstrated at the prompt's end less, not more). We therefore treat alternating parity as an arrangement–count interaction to be characterized, not as a collapse criterion. Figure 5 shows the safe/unsafe example: the two channels track each other, LLaDA-Base's single pass stays in the collapsed band at odd K (0.08–0.19) while rising on even K toward the even-only reversal, LLaDA-Instruct stays on the negative side at every K with a weak parity echo, and iLLaDA sits at hard collapse throughout.

Each shuffled condition averages its 200 queries over independently drawn class-balanced permutations, so no reading is specific to a single permutation.

The gen–sp correspondence of Section 4 extends beyond its two spotlighted cells: Figure 6 shows the distribution of per-condition gen–sp gaps over all 1,680 paired conditions (four models, three balanced arrangements), 84% of the conditions lie below 0.10 and 91% below 0.30, and the tail concentrates in LLaDA-Instruct (71 of its 420 conditions at or above 0.30, versus 24–37 for each other model); Figure 7 traces the K-sweep of one cell per kind of above-threshold behavior, making the taxonomy directly visible.

The four panels read as follows. *(a)* Channel flip (Llama-3 right/left): at even K the two channels point to opposite labels (gap up to 0.92) while the single pass itself has lost query dependence. *(b)* Transient

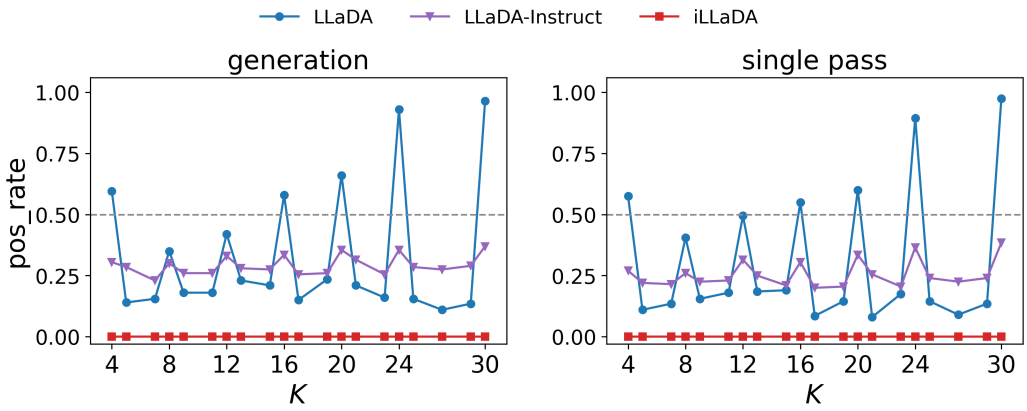


Figure 5: safe/unsafe under alternating, per model, against K: generation `pos_rate` (left) and single-pass `pos_rate` (right); every integer K from 4 to 30 (even K from the main grid, odd K per Table 4); the dashed line marks balance.

Table 5: Cross-arrangement generation `pos_rate` at K=8 (arrangement codes as in Section 2; ✗ = collapsed, △ = drift).

Model	Word pair	PNPN	PPNN	PPPNNN	Pattern
LLaDA-Base	safe/unsafe	0.35 △	0.25 △	0.30 △	All drift
iLLaDA-Base	safe/unsafe	0.00 ✗	0.00 ✗	0.00 ✗	All collapse
LLaDA-Inst.	true/false	0.97 ✗	0.97 ✗	0.84 ✗	Always "true"-biased
LLaDA-Inst.	cat/dog	0.83 ✗	0.50	0.42	Alternating collapse
Llama-3	cat/dog	1.00 ✗	0.95 ✗	0.40	Alternating and grouped collapse
Qwen3.5-9B	safe/unsafe	0.00 ✗	0.00 ✗	0.00 ✗	All collapse

episode (Qwen3.5 true/false): generation stays near balance through K=16 and then drifts to the positive word (0.86–0.87 at K=20–30), while the single pass tracks it closely throughout—a late divergence, not a hard collapse. *(c)* Persistent exception (LLaDA-Instruct true/false): generation at 0.92–0.99 at every K while the single pass (0.81–0.94 accurate) stays mid-range. *(d)* Off-label outputs (iLLaDA right/left): outputs matching neither label word reach 48–76% of queries; among the parseable remainder, generation accuracy is 0.90–1.00, so its low `pos_rate` reflects off-label emission rather than a query-blind label choice.

H Calibration Baselines

We evaluate two standard label-bias calibrations, on the scoring channel and, for the dLLMs, inside the decoding loop. *Contextual calibration (CC)* (Zhao et al., 2021): W is the label-word probability averaged over three content-free queries ("N/A", "[MASK]", and the empty string) appended after the same demonstrations, and the calibrated score is $\mathrm{sep} - \log(W_{\mathrm{pos}}/W_{\mathrm{neg}})$; in the generation channel the same correction is injected symmetrically into the two label logits before each remasking step (the injection is defined inside the diffusion remasking loop, which autoregressive decoding lacks, so the AR models are calibrated at the scoring level only). *PMI* (pointwise mutual information) (Holtzman et al., 2021): we use a query-conditional denominator (the separation of the same query at K=0), a stronger normalization than the domain-conditional premise of the original formulation. Coverage: the scoring-channel calibration comprises 190 conditions ($K \in \{4, \ldots, 30\}$, the original five models; 86 of the 100 model–pair cells at K=8, the remainder lacking historical raw outputs), and the decoding injection comprises 114 conditions (three dLLMs; 20 word pairs at K=8 plus the K sweeps of positive/negative, safe/unsafe, and true/false); ancient/modern is excluded from cross-version comparisons because its outputs vary across experimental versions (successive template and code iterations), a version-drift effect distinct from the run-to-run variance that deterministic decoding eliminates.

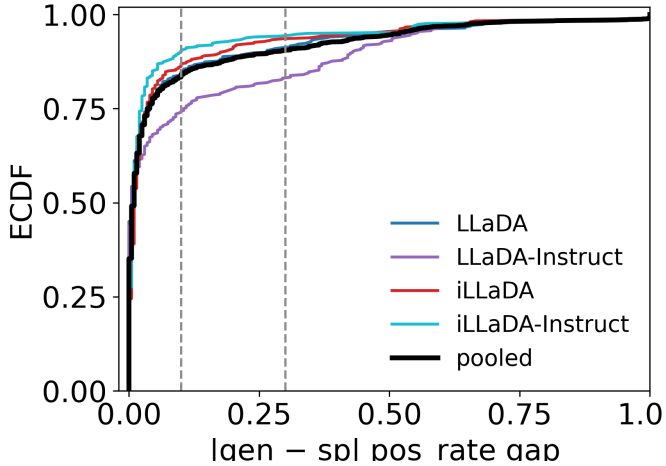


Figure 6: Empirical CDF of per-condition gen–sp `pos_rate` gaps (four generation-capable dLLMs, three balanced arrangements; thin curves per model, thick black pooled; dashed guides at 0.10/0.30).

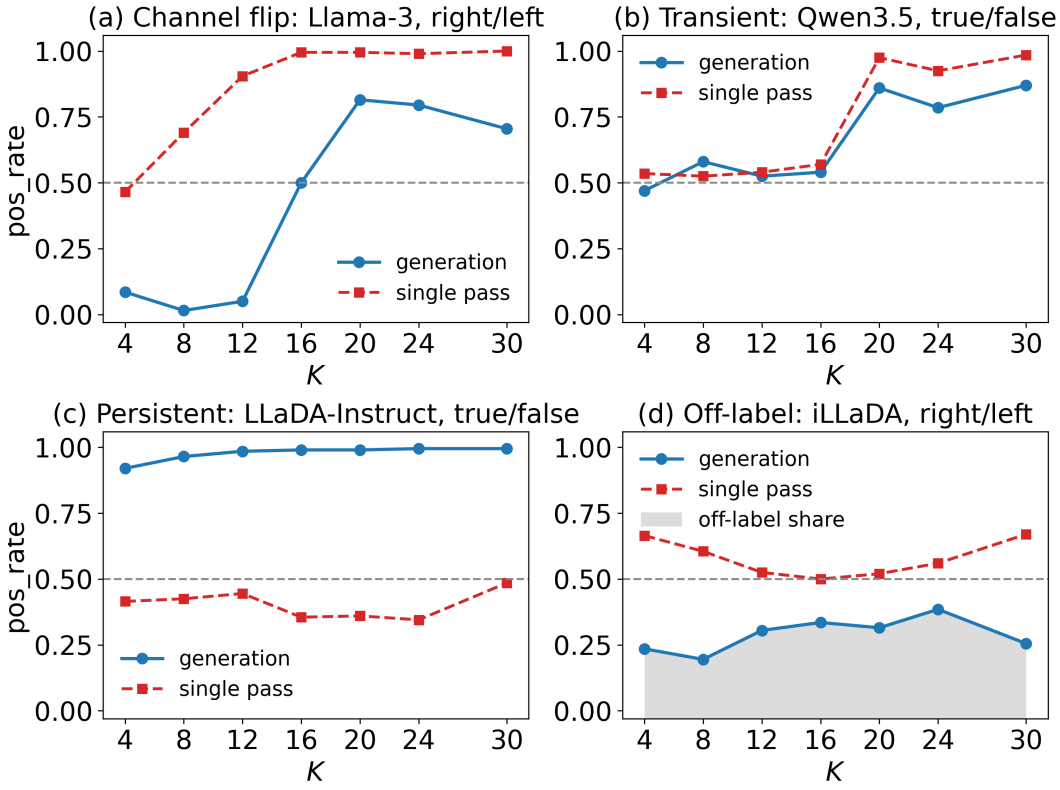


Figure 7: K-trajectories of one alternating cell per kind of above-threshold gen–sp behavior (blue solid: generation; red dashed: single pass; gray dashed line: balance; gray band in (d): outputs matching neither label word).

Repair rates. On the scoring channel at $K{=}8$ (67 collapsed, 53 healthy conditions, six dLLMs), reference-point calibrations repair little: contextual calibration 22/67 (33%; 17 of 61 hard-tier and 5 of 6 between-margin conditions, the tiers defined by $|\texttt{pos_rate} - 0.5| \geq 0.45$ and $0.30 \leq |\texttt{pos_rate} - 0.5| < 0.45$), prior calibration 16 (14 hard-tier), in-context calibration 20, domain calibration 24, and contextual calibration breaks 6/53 healthy ones. Per model (collapsed / CC-

repaired): LLaDA 14/2, LLaDA-Instruct 14/3, iLLaDA 6/2, iLLaDA-Instruct 5/2, Dream-Base 16/5, Dream-Instruct 12/8.

A unified view, and three further calibrations. All methods above correct the same quantity: writing $\mathrm{sep}(x)$ for the label log-odds of query x (Eq. 1), every calibration predicts positive iff $\mathrm{sep}(x) > c$ for a reference point c given by the content-free log-odds for CC, the zero-shot log-odds of the same query for our PMI variant, the in-domain random-word log-odds for domain-context calibration (DC) (Fei et al., 2023), the null-prompt log-odds for IntrCal (He et al., 2024), and the test-batch mean $\bar{\mathrm{sep}} = \frac{1}{n}\sum_i \mathrm{sep}(x_i)$ for batch calibration (BC) (Zhou et al., 2024). For two labels the BC decision reduces to a sign test of centered separations, $\hat{y}(x) = \mathbf{1}[\mathrm{sep}(x) - \bar{\mathrm{sep}} > 0]$, the log-space formulation of Zhou et al. (Zhou et al., 2024) with softmax denominators canceled. On the same 67 collapsed conditions: domain calibration repairs 24 (overlapping contextual calibration in 20 and repairing four conditions it does not), in-context calibration repairs 20, and batch calibration returns the `pos_rate` of all 67 to the balance band. The finer residual analysis was established on the original five-model grid (LLaDA and iLLaDA families, two AR baselines, 45 collapsed): nine conditions carried no query information in the centered log-odds, and the analyses below (small-batch simulation, treatment boundary, scoring-channel sweeps, and the associated figures) retain that five-model frame, with the Dream rows entering through the six-model counts above. Length-normalized whole-label scoring (NCC; scoring both label words' full token sequences) (Sanz-Guerrero & von der Wense, 2025) was evaluated on the cells selected for it: the 18 cells whose two label words differ in token count under the model's own tokenizer (six word pairs: clean/toxic, kind/cruel, ancient/modern, quiet/loud, rich/poor, wet/dry) plus ten length-symmetric controls (positive/negative and safe/unsafe, five models each), all on the original five-model grid. Of these 28 cells, 15 are collapsed in the uncalibrated single pass, and NCC repairs 0 of the 15: length normalization does not move the collapsed pole. For the dLLMs, which have no left-to-right factorization of the label probability, the score is computed by prefix teacher forcing (label tokens $1, \ldots, i{-}1$ are filled into the answer slot and the logit of token i is read at its position, one forward pass per position), so it reflects the forced-conditioning distribution of the scoring channel rather than free iterative denoising; the AR models use the standard full-sequence log-probability. Both label words are scored under the same scheme, so the forcing shapes the absolute scores, not which label wins. BC is transductive: it requires access to the test batch, and its authors list this batch dependence as their primary limitation and leave calibration of *generative* tasks as open future work (Zhou et al., 2024), a question our decoding-injection results answer directly (the readout-style correction inside the loop returns 8 of 34 collapsed $K{=}8$ conditions to the band, three of them at a parsing cost, and only LLaDA-Instruct on positive/negative holds across the K sweep). We therefore read BC as a probe of where the bias lives rather than a deployable fix. The signature transfers out of the original word pairs: on three fresh equal-length pairs (Appendix I), BC returns all 298 collapsed single-pass conditions, across the LLaDA and iLLaDA families and two AR baselines, to balance (accuracy restored in 207) and damages none of the healthy ones.

Small-batch simulation. The transductive cost of BC is in practice mild. Simulating batches offline on the stored per-query separations (86 $K{=}8$ conditions on the original five-model grid, 200 queries each), we estimate the batch offset from m sampled queries, subtract it in log space, and evaluate on the disjoint remainder (the streaming deployment pattern: each batch's offset estimate is applied to queries it does not contain); a leave-one-out variant, where each query is centered by the other members of its own batch, behaves the same. With $m{=}4$, all 45 collapsed conditions return to the balance band (as with the full batch), 30 of 45 recover accuracy at least 0.7 (full batch: $36/45$), and the 41 healthy conditions keep their accuracy (median 0.92 before and after). The simulation draws batches from a single condition's query pool, an i.i.d. assumption; batches drawn from shifted or mixed distributions remain an open deployment question.

DAC variants and a global-calibration baseline. DAC is the method this paper contributes: to our knowledge, no prior calibration anchors on the demonstrations themselves (Table 6 lists the anchor families; the closest, generative calibration, samples model-generated pseudo-inputs instead). Table 6 completes the method comparison on the same $K{=}8$ scoring-channel conditions (67 collapsed, 53 healthy). DAC-basic places each labeled demonstration in the query slot, reads its separation, and subtracts the midpoint $\hat{b}$ of the positive- and negative-demo means at test time; the leave-one-out (LOO) variant estimates $\hat{b}$ from the demonstrations other than the one being read

Table 6: Repair/harm comparison on the $K=8$ scoring channel (67 collapsed, 53 healthy; six dLLMs). "Reference point": source of the subtracted constant; DAC variants are demo-anchored and inductive, BC uses the test batch (transductive).

Method	Reference point	Repaired	Broken
prior calibration	$K=0$ reading of the same query	16/67	13/53
in-context calibration	null-prompt log-odds	20/67	5/53
contextual calibration	content-free prompt	22/67	6/53
domain calibration	in-domain random words	24/67	6/53
DAC-basic	demonstrations (query-slot readout)	53/67	3/53
DAC-trimmed	demonstrations, trimmed	54/67	2/53
batch calibration	test batch	67/67	0/53

(repairing 51/67), and the trimmed variant drops the single largest |readout| per side before averaging. A global-calibration baseline (GC, our reproduction of **?**), established on the original five-model grid, repairs 21/45 while breaking none, but stalls in the hard tier; every DAC variant repairs a subset of the BC-repaired conditions (nested, not competing), and most DAC-basic failures are conservative under-corrections that trade accuracy for safety rather than overshooting.

The prior share draws the treatment boundary. The boundary analysis was established on the original five-model grid; on its 38 hard-tier conditions we define the prior share as $|\log(W_{\text{pos}}/W_{\text{neg}})|$ divided by the mean separation recomputed from the raw single-pass outputs. A condition counts as *repaired* here if contextual calibration returns it to the balance band ($|\texttt{pos_rate} - 0.5| < 0.3$ after calibration); the null-prompt baseline repairs the identical set, and the decoding channel is immune on the hard tier regardless of share, so the boundary below is drawn on the scoring channel (batch calibration is excluded from this notion: it re-centers all 45, leaving no immune conditions). Repaired conditions have prior share 0.63–0.95 (median 0.74); immune conditions 0.06–0.8 (median 0.46). No condition with share below 0.55 is ever repaired ($0/18$); four of the five conditions at or above 0.8 are repaired (the exception is Qwen3.5 circle/square, share 0.80, which stays at 1.0); and the 0.55–0.8 band is mixed. The prior's direction is not discriminative (every repaired and every immune condition aligns with it), so amplitude, not direction, separates the two groups. The boundary is a property of the model–pair combination rather than of the word pair: circle/square carries prior shares of 0.95 (Llama-3; repaired, $1.00 \rightarrow 0.49$), 0.83 (LLaDA-Instruct; repaired to drift, 0.68), 0.69 and 0.61 (LLaDA-Base, iLLaDA; stay at 1.0), and 0.80 (Qwen3.5; stays at 1.0). Repairs can also decay with K: LLaDA positive/negative remains repaired through $K=16$ (0.56–0.61) but returns to 1.0 by $K=24$, as the demonstration-induced component outgrows the fixed subtraction.

Scoring-channel sweeps. The sweeps and figures of this and the next two paragraphs show the original five-model grid. Figure 8 shows the word pairs that respond (solid before, dashed after). On positive/negative, the two AR baselines are repaired at every K (0.49–0.62) and iLLaDA is balanced throughout, while the two LLaDA-family models are repaired only transiently: LLaDA-Base holds the band through $K=16$ and then re-collapses monotonically toward the positive word (0.90 at $K=20$ to 1.00 at $K=30$), and LLaDA-Instruct holds through $K=20$ before drifting out (0.84 and 0.95 at $K=24$ and 30). On safe/unsafe, Llama-3 sits in the band at every K (0.28–0.58; its uncalibrated sweep re-enters the band by itself from $K=12$), Qwen3.5 is lifted from the floor into the band from $K=12$ and stays (0.30–0.48), while iLLaDA stays pinned and LLaDA overshoots below balance at small K and misses the pair's late re-collapse at $K=30$ (0.87). Figure 9 shows the word pairs that barely move: circle/square stays at 1.0 for LLaDA, iLLaDA, and Qwen3.5 at every K, Llama-3 recovers only on part of the grid, and LLaDA-Instruct is lifted at $K=8$ alone (0.68) and re-pins elsewhere; on true/false, iLLaDA stays at 0.0 and Llama-3's collapse is untouched (0.89–0.98), Qwen3.5—balanced through $K=16$—collapses from $K=20$ and is only partially held (0.72 at $K=24$; 0.80 and 0.90 at $K=20$ and 30 remain out), and the two balanced LLaDA-family models stay balanced (LLaDA-Instruct drifting just outside the band, 0.81 at $K=30$).

Decoding-channel injection. Figure 10 shows the injected decoding sweep for the two anchor word pairs (three dLLMs; solid before, dashed after injection). At $K=8$, the injection returns 8 of the 34 collapsed conditions to the band, five of them with recovered accuracy (LLaDA and LLaDA-Instruct

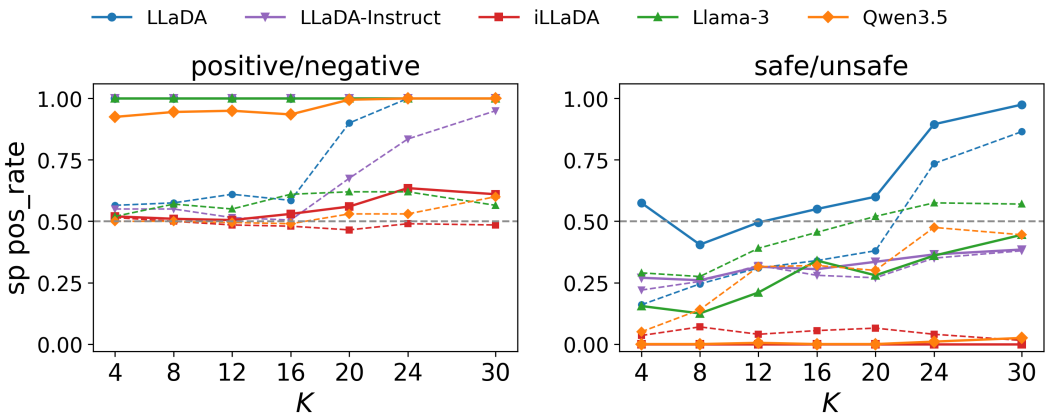


Figure 8: Single-pass `pos_rate` before (solid) and after (dashed) contextual calibration (alternating; panels follow Figure 2): responsive word pairs.

on positive/negative, LLaDA-Instruct on cat/dog and light/dark, iLLaDA on easy/hard) and three at a parsing cost (e.g., light/dark, below); on the sweeps, only LLaDA-Instruct positive/negative holds across the whole range (0.87–1.0 $\rightarrow$ 0.49–0.51), LLaDA positive/negative holds through K=24 and re-collapses at K=30 (1.0 $\rightarrow$ 0.905), and on safe/unsafe the injection overshoots below balance at small K (LLaDA, 0.595 $\rightarrow$ 0.225 at K=4), returns LLaDA's late collapse to the band only at K=24 (0.93 $\rightarrow$ 0.705) before losing it at K=30 (0.87), and lifts iLLaDA off the floor only transiently (0.255 peak at K=12, back to 0.045 at K=30). The hard tier is immune in this channel as well: circle/square remains at 0.99–1.0 for all three dLLMs, iLLaDA true/false remains at 0.0, and on true/false the injection leaves decoding essentially unchanged throughout; the balanced conditions (iLLaDA positive/negative, LLaDA-Instruct safe/unsafe) stay balanced under injection. The residual failure mode is parsing cost, where calibrated outputs leave the label vocabulary (LLaDA light/dark: invalid rate 0.01 $\rightarrow$ 0.23). The scoring and decoding verdicts agree on 18/26 collapsed conditions; the eight disagreements concentrate on conditions where the correction is small relative to decoding dynamics.

Zero-shot prior statistics. Over the 67 collapsed model–pair combinations (six dLLMs), the zero-shot separation agrees with the collapse direction in 46/67; the unigram analyses below were established on the original five-model grid, where restricting to combinations with a meaningful zero-shot prior ($|\text{sep}| > 1$) raises the agreement to 24/29 (83%). The unigram logits, defined on the 37 collapsed combinations whose two label words are single tokens, agree in 28/37, rising to 18/21 (86%) among those with $|\text{logit diff}| > 1$. The two prior measurements cross-validate each other over all 82 single-token model–pair combinations, healthy or collapsed (sign agreement 73%, r=0.68). Zero-shot generation produces neither label in 47/50 conditions (invalid rate $\geq$ 0.98): with no demonstrations almost no model enters the label vocabulary, so the readable quantity at K=0 is the prior rather than behavior. Excluding circle/square, whose collapse trivially follows its token prior, changes the full-set agreement from 33/45 and 28/37 to 28/40 and 23/32, qualitatively unchanged.

I Equal-Length New Label Pairs

Three fresh word pairs, constructed after the analyses above and evaluated with the identical protocol (XSTest pool, alternating arrangement, K=4–30, 200 queries; both channels for the three LLaDA- and iLLaDA-family models and the two AR baselines, single-pass for the Dream family), test whether token-length asymmetry drives collapse: new/old and valid/invalid are single-token on both sides for all five tokenizers, and moral/immoral is a matched two-token pair (`m+oral` vs. `imm+oral`). Collapse survives equal length: new/old collapses outright for four models (LLaDA, LLaDA-Instruct, Llama-3, Qwen3.5, all toward "old") while iLLaDA's generation alternates with parity over the whole range (even K at "new", odd K at "old"); moral/immoral collapses toward "immoral" for the same four models, while the iLLaDA family dissociates by channel—both single passes stay balanced, but generation does not: iLLaDA-Instruct collapses to 0.00 at every K, and iLLaDA oscillates

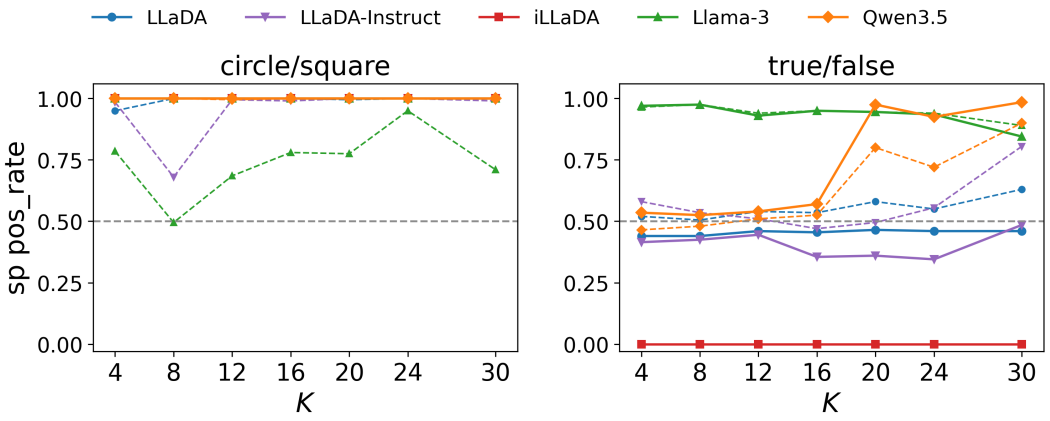


Figure 9: Single-pass `pos_rate` before (solid) and after (dashed) contextual calibration (alternating): immune word pairs.

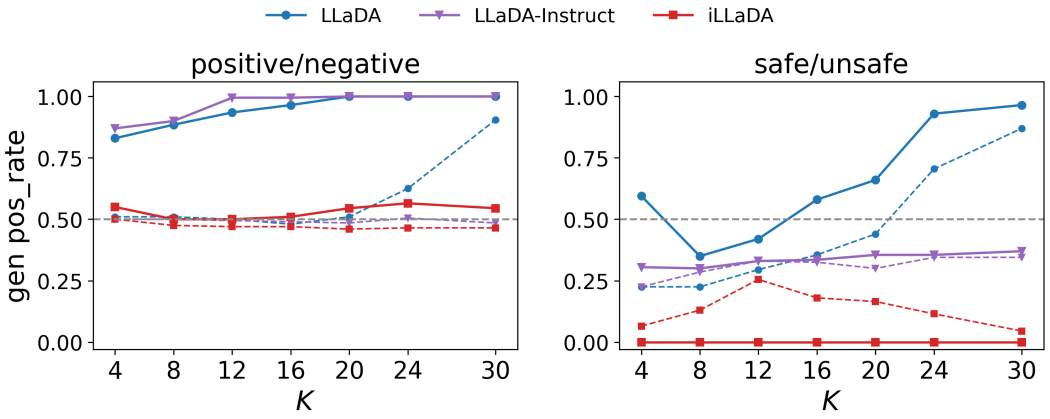


Figure 10: Generation `pos_rate` before (solid) and after (dashed) injecting the contextual-calibration correction into dLLM decoding (alternating; three dLLMs, panels as in Figure 2).

with the same parity as on the other pairs (odd K hard toward "immoral", 0.035 at its deepest, the troughs rising as K grows; even K drifting mildly toward "moral", 0.71 at its peak), a family-wide instance of the gen–sp decoupling of Section 4; valid/invalid collapses slowly rather than never: its generation medians stay balanced (0.21–0.52), but LLaDA-Base and iLLaDA are parity-structured from the lowest K and Qwen3.5 joins from $K{\approx}20$ (reaching 0.84 at $K{=}30$), with even K toward "valid" (LLaDA-Base up to 0.83) and odd K toward "invalid" (LLaDA-Base down to 0.04), while LLaDA-Instruct leans "invalid" throughout and Llama-3 stays near balance (Figure 11). With the iLLaDA-Instruct extension, all six dLLMs are covered. On the single-pass channel, where the Dream family is readable (Appendix B): Dream-Base adds a hard collapse on valid/invalid (≤ 0.015 at every K) and the sole non-monotonic case on new/old (dipping to 0.08 at $K{=}8$–12 and recovering to 0.86 by $K{=}30$); Dream-Instruct pins new/old to "old" (0.00 throughout); and on moral/immoral the family splits—Dream-Base pins the positive word "moral" (0.97–1.00) while Dream-Instruct drifts toward it (0.57–0.98). iLLaDA-Instruct pins new/old to "new" (1.00 at every K, in both channels)—the only model fixed on that side, where iLLaDA oscillates and the rest collapse to "old"—and drifts steadily from "invalid" toward "valid" as K grows ($0.16 \rightarrow 0.70$ single-pass, $0.34 \rightarrow 0.72$ in generation). Equal length thus neither prevents nor guarantees collapse, and the word pair fixes not only whether but also how fast and in what pattern collapse arrives: new/old completes within $K{\leq}8$ for the four pinned models (iLLaDA's alternation persists to $K{=}30$), valid/invalid builds its parity structure from the lowest K and strengthens it gradually, and the alternating arrangement imprints its parity on

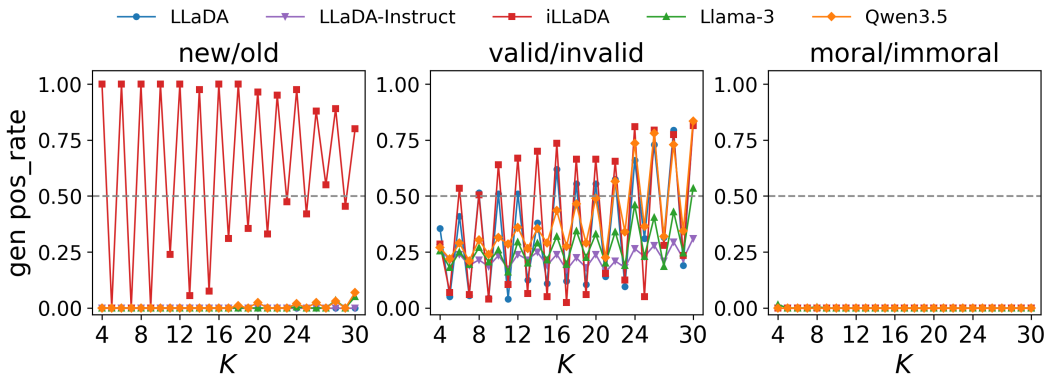


Figure 11: Generation `pos_rate` on the three equal-length pairs (alternating, K=4–30; gray dashed line marks `pos_rate` = 0.5).

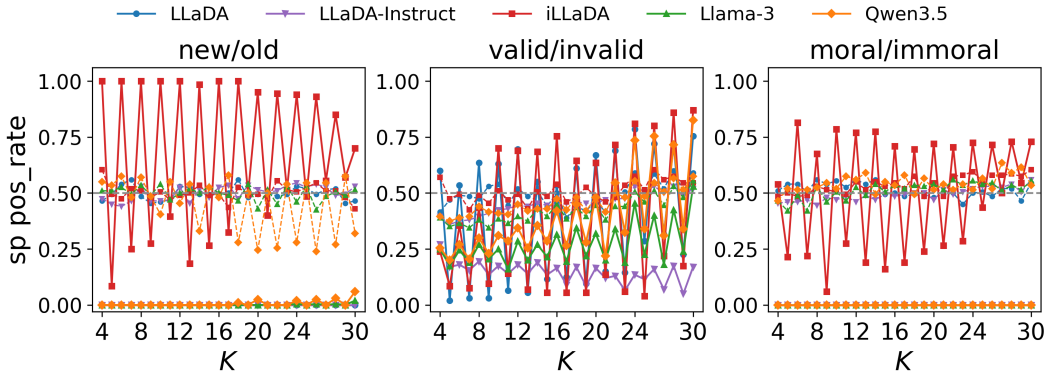


Figure 12: Single-pass `pos_rate` on the three equal-length pairs (alternating; solid = uncalibrated, dashed = batch-calibrated, gray dashed line = balance).

the slow case, with even K (last demonstration "invalid") collapsing toward "valid" and odd K the reverse. Which pairs are dangerous remains a property of the word pair, not of its token lengths.

The direction measurements of Section 4 replicate on the new pairs: among the nine collapsed model–pair combinations, the zero-shot (K=0) separation predicts the collapse direction in 7/9 and empty-context unigram logits in 7/9, the two measurements agreeing with each other in 14/15 combinations overall; the unigram direction matches the corpus-lower-frequency side in 7/9, consistent with the 7/14 rate on the original pairs. The two zero-shot exceptions are a single phenomenon: for LLaDA-Base and LLaDA-Instruct on moral/immoral, both the zero-shot and unigram directions point to "moral" ($+2.05$ and $+1.04$) yet the in-context separation and generation collapse toward "immoral" (LLaDA-Base, sep $= -8.62$ at K=30); the demonstration context can *reverse* the direction inherited from the prior, so direction prediction is probabilistic rather than deterministic. We exclude the slowly drifting valid/invalid combinations from these counts: a parity oscillation around balance has no single collapse direction, and its median flips between channels (LLaDA, 0.38 generation vs. 0.54 single-pass).

Calibration behaves as in Appendix H: contextual calibration repairs only part of the collapsed conditions, and unstably (LLaDA new/old: 0.40 at K=8, back to 0.065 at K=16), whereas batch calibration (Figure 12) returns all 298 collapsed single-pass conditions (of 405 over those five models) to balance and restores accuracy (≥ 0.7) in 207 of them, with no damage to the healthy ones; the parity-drift conditions are included in the 298, and centering flattens their oscillation as well (0.34–0.60). Finally, the multi-token pair reprises the parsing hazard of Appendix F: first-token matching misreads every moral/immoral generation (invalid rate ≈ 1.0, the first tokens being `m`/`imm`), which text-level prefix matching restores (invalid rate 0.00 after re-parsing).

Table 7: Subtraction (treat): generation `pos_rate` per condition and layer, α=1.0; a condition counts as recovered at layer ℓ if the value enters the balance band.

Condition	base	L8	L16	L25	L26	L27	L28	L29	L30	L31	L32
true/false	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060	0.140	0.040
circle/square	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.720	0.000	0.000	0.000
safe/unsafe	0.000	0.000	0.000	0.020	0.000	0.000	0.020	0.020	0.020	0.000	0.000
easy/hard	0.960	1.000	0.980	0.980	0.980	0.960	0.960	0.920	0.640	0.000	0.000
ancient/modern	0.660	0.680	0.640	0.600	0.620	0.800	0.640	0.620	0.520	0.240	0.040
recovered		1	1	1	1	0	1	2	2	1	0

Table 8: Injection (cause): generation `pos_rate` of five healthy conditions with the donor direction injected at α=1.0; only L31 moves any condition.

Condition	baseline	L31	other layers
hot/cold	0.540	0.460 ($\lvert\Delta\rvert$=0.08)	no shift
rich/poor	0.560	0.200 ($\lvert\Delta\rvert$=0.36)	no shift
positive/negative	0.560	0.560 ($\lvert\Delta\rvert$=0.00)	no shift
light/dark	0.560	0.500 ($\lvert\Delta\rvert$=0.06)	no shift
wet/dry	0.540	0.540 ($\lvert\Delta\rvert$=0.00)	no shift

J LAYER-LEVEL INTERVENTIONS

Setup. All interventions run on iLLaDA-8B-Base (L0 = embedding output; L1–L32 = transformer layers). Condition sets were pre-registered: the five collapsed conditions with the largest $|\hat{b}|$ (true/false -9.01, circle/square $+6.14$, safe/unsafe -2.64, easy/hard $+1.47$, ancient/modern $+1.39$) for subtraction, and the five healthy conditions with the largest parroting delta (hot/cold 8.16, rich/poor 7.92, positive/negative 7.89, light/dark 7.58, wet/dry 7.25) for injection. The parroting delta is the query-slot readout: placing a *negative* demonstration in the query slot yields a negative separation in all 41 healthy conditions of the original five-model grid (median -2.72) but flips positive in 32/45 collapsed ones (median $+2.80$)—zero overlap between the groups.

Protocol. Subtraction: $h^{(\ell)} \leftarrow h^{(\ell)} - \Delta^{(\ell)}$ per query, where $\Delta^{(\ell)}$ is the aligned difference of the answer-slot hidden state between the K=8 and K=0 contexts. Injection: $h^{(\ell)} \leftarrow h^{(\ell)} + \bar{\Delta}^{(\ell)}$, the query-averaged difference of the donor condition (true/false), applied to a different healthy condition. $\alpha = 1.0$; layers $\ell \in \{8, 16, 25, 26, 27, 28, 29, 30, 31, 32\}$; the intervention position is the answer slot. Generation is re-run with 50 samples per condition; the pipeline reproduces the unmodified K=8 single-pass readings bitwise, and α=0 reproduces the baseline.

Pre-registered gates. Subtraction passes if some single layer returns $\geq 3/5$ collapsed conditions to the balance band ($|\texttt{pos_rate} - 0.5| < 0.3$ in generation) *and* the early layers (L8, L16) repair $0/5$; injection passes if some layer shifts $\geq 2/5$ healthy conditions by ≥ 0.3.

Results. Both gates fail (Table 7, Table 8): the best single-layer subtraction recovers 2/5 collapsed conditions (L29/L30), and cross-condition direction injection moves 1/5 (rich/poor at L31, by 0.36), so the donor direction essentially does not transfer across word pairs. The L8/L16 recoveries trace to ancient/modern, whose generation baseline (0.660) already sits inside the band (its collapse is a single-pass reading, 0.815), so they are baseline continuations, not intervention effects; excluding it, the late layers still recover only circle/square (L29) and easy/hard (L30). Two failure modes coexist: positive-side subtraction *overshoots*—circle/square and easy/hard flip from their collapsed side through the balance band to 0.000 (circle/square at L30: $1.00 \rightarrow 0.00$)—while negative-side collapse (true/false, safe/unsafe) is essentially immovable (≤ 0.14). The position control (Table 9) localizes the effect to the answer slot at late layers; query-slot interventions have zero effect at every layer, and the single-pass shifts (≤ 0.24) are far smaller than the generation flips, so the intervention mainly changes the dynamics of iterative decoding rather than a single forward readout. We read the negative result as follows: the bias is *distributed across layers and bound to the word pair*—there is

Table 9: Position control (exploratory): single-pass `pos_rate` with subtraction applied at the query or the answer slot.

Condition	Position	L8	L16	L25	L30	L31	L32
circle/square	query	1.00	1.00	1.00	1.00	1.00	1.00
	answer	1.00	1.00	1.00	0.98	0.90	0.82
hot/cold	query	0.54	0.54	0.54	0.54	0.54	0.54
	answer	0.54	0.54	0.54	0.40	0.72	0.78

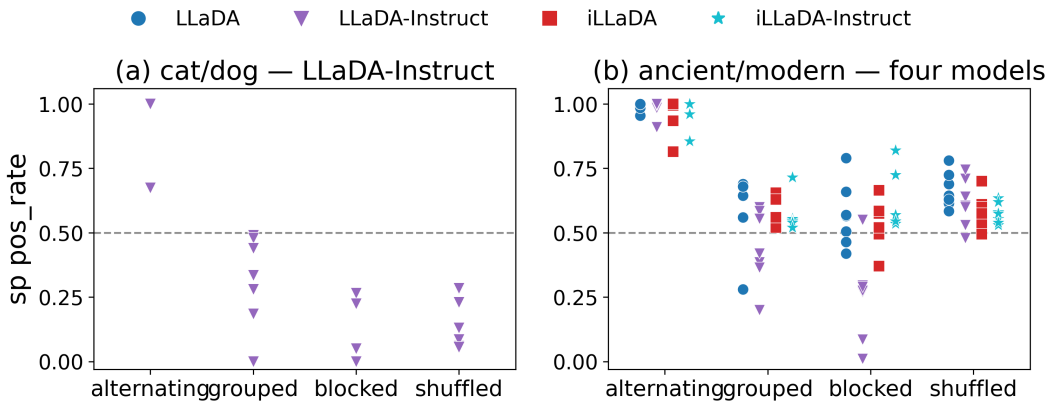


Figure 13: Single-pass `pos_rate` across the four arrangements at K=4–30, one dot per K: (a) cat/dog under LLaDA-Instruct; (b) ancient/modern under the LLaDA and iLLaDA families. The dashed line marks balance. (Referenced from Section 6.)

no single-layer additive bias vector whose removal restores balance. This bounds the "mere threshold shift" reading of the calibration results (a pure threshold shift should be correctable at a single layer, and is not), and motivates the output-level calibration of Appendix H over layer surgery.

K Imbalance and Block-Order Grids

Five models—the LLaDA and iLLaDA families and two AR baselines—are additionally evaluated under eight imbalanced arrangements: shuffled demonstrations with majority share $\gamma \in \{62.5\%, 75\%, 87.5\%, 100\%\}$, and prefix-majority orders with a trailing minority block at $\gamma \in \{75\%, 87.5\%\}$ (a P-majority prefix followed by small N blocks, `blkP`, and its mirror `blkN`); 20 word pairs, $K \in \{4, 8, 12, 16, 20, 24, 30\}$, both channels, 200 queries per condition: 5,600 paired conditions. The gen–sp parity of the diagnostic survives the transfer: of the 3,583 generation-collapsed conditions on these grids, 3,339 (93.2%) are also single-pass collapsed (Pearson $r = 0.930$ across all 5,600 pairs), versus 88.2% on the balanced main grid (the four generation-capable dLLMs). The four phenomena below quantify the conformity thresholds of Section 5 and the block-position effects summarized in Section 6: the dose–response in Table 10 and Figure 14, block recency in Figure 15, the minority dependence of K in Figure 16, and the majority-following family split under PPN and PPPN imbalanced constructions. The PPN and PPPN constructions themselves (majority shares $2/3$ and $3/4$, single-pass channel, all six models) show the split at its sharpest: the LLaDA family pins the majority label (median `pos_rate` 1.00), the Dream family follows it almost completely (0.95–0.99), and the iLLaDA family stays near balance (median 0.48–0.50, with 4–8 of the 20 pairs collapsing outright), and the two AR baselines, evaluated under the same constructions, stay near balance throughout (medians 0.50–0.68).

Collapse follows the trailing block, and the effect of K depends on whether a minority exists. Two further arrangement families sharpen the structural reading (Figure 15). Under a front-loaded majority with a trailing minority block (`blkP75`: a γ=75% P-majority prefix followed by small N blocks), the five models collapse toward the *trailing minority* label, and the collapse deepens with K (iLLaDA `pos_rate` 0.005 at K=12, 0.000 at K=30; Llama-3 0.098 $\rightarrow$ 0.013; Qwen3.5

Table 10: Median generation `pos_rate` toward the majority label under shuffled imbalanced demonstrations, by majority share γ (all tested K pooled).

Model	γ=62.5%	75%	87.5%	100%
LLaDA-Instruct	0.98	0.98	0.99	1.00
LLaDA-Base	0.94	0.95	0.99	1.00
iLLaDA	0.47	0.47	0.55	1.00
Llama-3	0.48	0.50	0.66	0.99
Qwen3.5	0.51	0.52	0.61	1.00

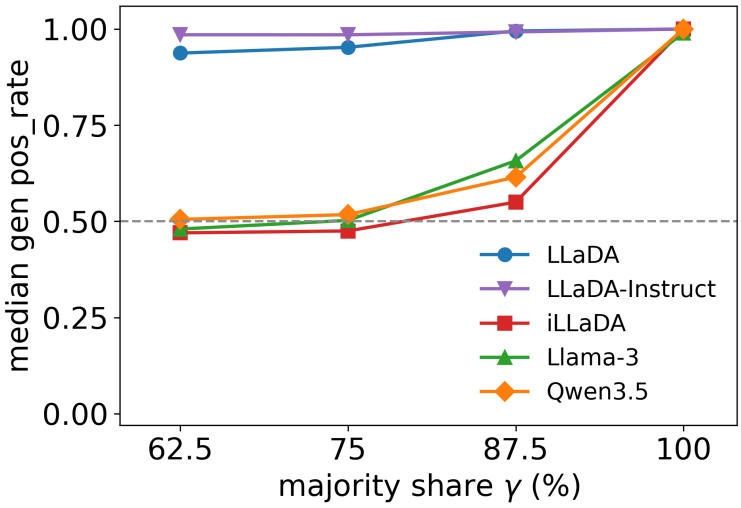


Figure 14: Conformity dose–response: median generation `pos_rate` against the majority share γ under shuffled arrangements (20 pairs, evaluated K pooled).

0.383 $\rightarrow$ 0.190); the mirrored `blkN75` condition collapses toward P in all five models at K=30 (0.80–1.00). The prefix majority is thus not merely discounted—it is overruled by the trailing block, and adding demonstrations amplifies the override. Whether adding demonstrations helps or hurts additionally depends on the presence of a minority. At $\gamma = 87.5\%$ (a 12.5% minority present), iLLaDA, Llama-3, and Qwen3.5 traverse a non-monotone path in K : full collapse at K=4 (0.92–0.99), resistance at K=8 (0.46–0.55), then partial conformity from K=12 on (0.54–0.68 across K=12–30). At $\gamma = 100\%$ this structure vanishes and every K collapses. At γ=87.5% the trailing-block rule holds for the resistant models in both directions (at K=30: `blkP875` medians 0.00–0.13 toward N, `blkN875` 0.86–1.00 toward P), while the LLaDA family is asymmetric: it follows the trailing positive block (0.99–1.00) but only drifts partway toward the trailing negative one (0.57–0.70 at K=30). Neither "more demonstrations are safer" nor "more demonstrations are riskier" holds unconditionally: the effect of K is a function of the demonstration structure, consistent with the parity and reversal patterns of Section 3 and Appendix G.

L Discussion

We distill the findings into practical guidance (what to measure and what to control) and state limitations. Three implications follow from the results above.

Practical implications. Separation is a cheap, model-agnostic pre-deployment check; which words serve as labels is a design decision; and training can move label bias in either direction. One forward pass at the answer slot anticipates behavioral collapse with 93.7% precision, and its disagreements with generation mark exactly the conditions worth auditing; we recommend it as a standard ICL metric, especially for diffusion models, where a single pass is cheap compared with iterative decoding. Label-token choice can systematically disable ICL (circle/square collapses

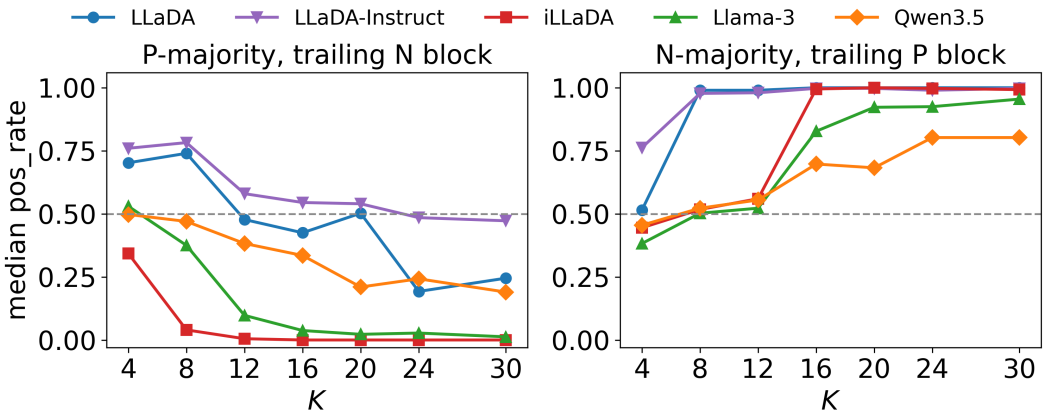


Figure 15: Block recency at γ=75%: median generation `pos_rate` against K; left, P-majority prefix with trailing N block (`blkP75`), right, the mirror (`blkN75`).

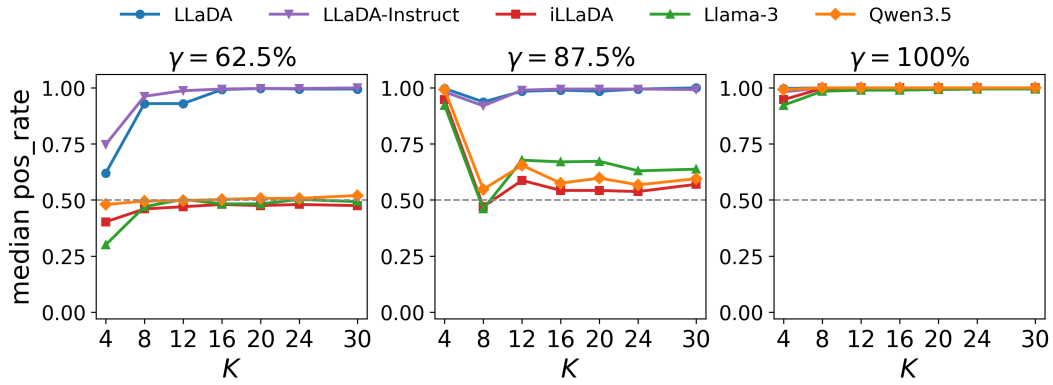


Figure 16: The effect of K depends on the presence of a minority: median generation `pos_rate` per model at γ=62.5% (left), 87.5% (middle), and 100% (right; 20 pairs pooled).

for every architecture tested; the pair can even decide the failure mode, as on ancient/modern), so practitioners should audit label tokens with separation before deployment and report label words alongside ICL accuracy. Because training can strengthen one label pair's bias while leaving another untouched, dangerous pairs cannot be predicted from the training data alone; training data should cover diverse label tokens, as SFT-data composition is known to shape context reliance (Wang et al., 2024), and measuring separation across label pairs after training flags the dangerous ones before deployment.

Limitations. (i) We evaluate 7B–9B models only; larger dLLMs may behave differently. (ii) The 20 word pairs, though diverse, are English, binary, and evaluated under one fixed template; arrangement coverage is uneven across pairs, and the γ-sweep and block-order grids retain the original five-model coverage, with Dream entering only the PPN/PPPN constructions (Appendix K). (iii) Our mechanistic account stops at the output projection; which properties of the training data produce a dangerous pair remains open. (iv) The alternating-parity structure is characterized empirically but not explained; a controlled study is left for future work. (v) Cross-lineage comparisons confound pre-training data, scale, and pre-training procedure; our SFT-specific conclusions rest on the three within-family Base/Instruct contrasts, and the Dream pair's chat tuning differs in provenance from the LLaDA pair's. (vi) The separation quantity itself is standard, the label log-odds that calibration methods operate on (Zhao et al., 2021; Zhou et al., 2024); our contribution is the diagnostic protocol and the phenomena it reveals, not the quantity.

M FUTURE WORK

Several directions remain: (i) extending separation analysis to further dLLM architectures (MDLM and larger Dream variants; Dream enters already in Appendix B); (ii) identifying which training-data properties make a label pair dangerous, and designing data accordingly; (iii) layer-level interventions that intercept collapse before the output projection; (iv) a mechanistic account of gen–sp decoupling in iterative denoising; (v) why collapse follows the trailing block rather than the majority count (Appendix K); (vi) extending the evaluation beyond binary classification under a fixed template, to natural multi-class task distributions and their label-word ecosystems.

N REPRODUCIBILITY

All source code and anonymized summary data are available at `https://anonymous.4open.science/r/code_release-53E1`. The experimental setup uses publicly available models and datasets, with full hyperparameters described in Section 2.