CAUSAL INDUCTIVE BIASES FOR IN-CONTEXT LEARNING IN DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Why does next-token pretraining produce in-context learners? Autoregressive (AR) training directly supervises prediction from the observed prefix, whereas masked and diffusion objectives distribute supervision across many observation patterns. We connect AR pretraining to Bayes-optimal episodic prediction and quantify the uncertainty reduction obtained by observing the future. These results motivate a supervision-allocation hypothesis: training more often on earlier context and later targets should promote sequential contextual adaptation. We test this hypothesis with matched AR, MLM and diffusion transformers, then introduce a minimal position- dependent corruption process. This intervention improves few-shot adaptation across diverse algorithmic tasks, survives native iterative generation under multiple diffusion objectives, and transfers to LLaDA-8B through continued pretraining. Together, these results identify training-objective factorisation as a central inductive bias for the emergence of in-context learning.

1 INTRODUCTION

In-context learning (ICL)—the ability of a language model to adapt to examples or instructions supplied in its prompt without updating its parameters—is one of the most striking capabilities to emerge from large-scale language modelling (Brown et al., 2020). A substantial literature has studied *how* transformers may implement ICL, including Bayesian inference over latent tasks (Xie et al., 2022; Akyürek et al., 2023), implicit optimisation (von Oswald et al., 2023) and induction-head circuits (Olsson et al., 2022). Yet a fundamental question remains unresolved: why does ordinary language-model pretraining produce an in-context learner in the first place?

The autoregressive objective contains a suggestive asymmetry. At every position, the model is trained to predict an unknown continuation from the prefix already observed. This is also the basic interface of ICL: demonstrations arrive in context, parameters remain fixed, and the model must change its prediction for what comes next. Autoregressive pretraining therefore optimises exactly the form of prefix-conditioned prediction that ICL later requires at inference.

Non-causal language-model objectives let us separate that alignment from architecture and representational capacity. Masked language models and discrete diffusion language models can be trained on the same underlying sequence distributions, but factorise prediction through bidirectional conditionals rather than left-to-right next-token prediction.(Devlin et al., 2019; Austin et al., 2021; Sahoo et al., 2024; von Rütte et al., 2025). While they may represent the same joint distribution as an AR model, their training objective does not directly optimise the prefix-conditioned prediction that sequential adaptation requires. This gives us a controlled question: does ICL emerge merely from learning the data distribution, or does the factorisation of the training objective provide a crucial inductive bias?

Sequential meta-learning makes the distinction precise. Under a latent task distribution and population log loss, the optimal next-token predictor is the Bayesian posterior predictive (Ortega et al., 2019). Thus AR directly trains the conditional required for sequential adaptation. A non-causal objective may represent the same conditional, but recovering it from future-conditioned predictions requires marginalising unknown futures; we show that the associated uncertainty gap is exactly a conditional mutual information.

This account suggests a direct causal test. Discrete diffusion language models are especially useful here because they retain the same basic modelling goal and transformer architecture while replacing

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.

left-to-right prediction with non-causal denoising. We first ask whether the predicted difference appears without intervention: across natural-text context-use probes and matched AR, MLM and diffusion transformers, the training factorisation tracks whether models exploit additional context. We then alter only the diffusion corruption process, increasing the prevalence of prediction problems in which earlier context is available for later targets.

The resulting intervention changes ICL in the predicted direction. It improves four controlled few-shot tasks spanning retrieval, latent-rule inference, conflicting contextual bindings and sequential state prediction; the effect survives native iterative generation in both MDLM and GIDD; and after 500M tokens of continued pretraining of LLaDA-8B, the advantage grows with contextual interference and follows the imposed positional corruption profile. A loss-weighted control with identical corrupted inputs fails to reproduce either signature, implicating the conditioning structure seen during training rather than simply the amount of supervision assigned to later tokens.

Contributions. **(1) An objective-level theory of ICL.** Building on the sequential-strategy optimality result of Ortega et al. (2019), we show that Bayes-optimal episodic ICL is directly contained in the population AR objective, and derive the future-marginalisation gap faced by non-causal factorisations. Full proofs and scope conditions are in Appendices B–C.
(2) A causal intervention on diffusion training. We introduce a minimal position-dependent corruption process that changes the conditioning configurations presented during training while leaving the architecture and decoder unchanged; its forward-process derivation and supervision controls are in Appendices D–G.
(3) Evidence across objectives, inference procedures and scale. We isolate the objective in matched AR/MLM/diffusion models, improve four controlled ICL tasks, transfer the effect through native MDLM and GIDD generation, and reproduce the predicted intervention fingerprint after continued pretraining of LLaDA-8B.

2 From Sequential Prediction to In-Context Learning

2.1 Autoregressive training directly optimises the ICL predictor

ICL can be written as a sequential prediction problem. Let Z denote a latent task or environment, and let an episode contain demonstrations followed by a query. After observing the corresponding prefix $x_{<t}$, the Bayes-optimal prediction is

$$p(x_t \mid x_{<t}) \;=\; \int p(x_t \mid x_{<t}, z)\, p(z \mid x_{<t})\, dz. \tag{1}$$

The latent task need not be explicitly represented; the object required for ICL is the map from the observed history to this posterior predictive distribution.

The sequential-strategy result of Ortega et al. (2019) now gives a direct connection to pretraining. Under population next-token log loss, every prefix is itself a supervised conditional prediction problem, and its unique optimum is $p(x_t \mid x_{<t})$. When demonstrations and a query are serialised as a prefix, the Bayes-optimal k-shot ICL predictor is therefore not merely compatible with AR training: it is one of the conditional prediction problems that AR training directly optimises.

Appendix B makes this statement exact by defining episodic ICL risk and proving that the population AR optimum simultaneously minimises that risk for every supported k. It also shows that the optimal improvement from demonstrations is exactly their conditional information about the answer. This is a population-optimum statement about training pressure; it does not imply that autoregression is necessary for ICL, or guarantee generalisation outside the training meta-distribution. The novelty for our purposes is not the posterior-predictive optimum itself, but its implication for objective design: AR places the ICL prediction map directly in its training loss, which gives a concrete comparison point for objectives that factorise the same sequence distribution differently.

2.2 The future marginalisation barrier

Masked LMs optimise $\mathbb{E}_{x,S}\big[-\sum_{i\in S}\log p_\theta(x_i \mid x_{\setminus S})\big]$ and discrete diffusion LMs learn denoising conditionals $p_\theta(x^{t-1}_{1:L} \mid x^{t}_{1:L})$ over progressively corrupted sequences. Both are *non-causal*: the

	AR	**MLM**	**MDLM** $\alpha = 0$	**MDLM** $\alpha = 1$
corruption	$x_1\,x_2\,x_3\,x_4\,x_5$ prefix → future	$x_1\,\mathsf{M}\,x_3\,\mathsf{M}\,x_5$	$x_1\,\mathsf{M}\,\mathsf{M}\,x_4\,\mathsf{M}$ $\beta_1 = \cdots = \beta_L$	$x_1\,x_2\,\mathsf{M}\,\mathsf{M}\,\mathsf{M}$ $\beta_1 < \cdots < \beta_L$
mask rate	$\beta_i = \mathbf{1}[i > t]$	$\beta_i = 0.15$ fixed	$\beta_i(t) = \beta(t)$ uniform over position	$\beta_i(t) = \beta(t)\, g_i$ $g_i \propto (i/L)^\alpha$
conditions on	$x_{<i}$	$x_{\setminus S}$	$\tilde{x}^{\,t},\ t$	$\tilde{x}^{\,t},\ t$
objective	$\sum_i \log p(x_i \mid x_{<i})$	$\sum_{i\in S} \log p(x_i \mid x_{\setminus S})$	$\mathbb{E}_t\big[\sum_i \log p(x_i \mid \tilde{x}^{\,t}, t)\big]$	

Figure 1: **Training objectives.** AR directly trains prefix conditionals; MLM and diffusion train under non-causal corruption. The α=1 column is our intervention.

prediction at position i may condition on positions after i. Figure 1 contrasts the resulting training problems.

The difference becomes concrete when the future is informative about the next token. A bidirectional model is rewarded for using that information during training. At test time, however, a sequential prediction must be made before that future has been observed. The prefix conditional is therefore not the conditional systematically presented by the non-causal objective. In general it is the mixture

$$p^\star(x_{k+1} \mid x_{\le k}) \;=\; \sum_{x_{k+2:L}} p^\star(x_{k+1} \mid x_{\le k}, x_{k+2:L})\; p^\star(x_{k+2:L} \mid x_{\le k}), \tag{2}$$

and forming that mixture has a price.

Theorem 2.1 (Future marginalisation barrier)**.** *Let $p^\star$ be a joint distribution over $X_{1:L}$. If $I(X_{k+1}; X_{k+2:L} \mid X_{\le k}) > 0$, then*

$$H(X_{k+1} \mid X_{\le k}) \;>\; \mathbb{E}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})]\,, \qquad X_{k+2:L} \sim p^\star(\cdot \mid X_{\le k}), \tag{3}$$

and the gap equals $I(X_{k+1}; X_{k+2:L} \mid X_{\le k})$ exactly.

The distinction is therefore not what distributions the model can represent, but which prediction problems the training objective repeatedly presents as supervised targets. The proof applies the strict concavity of Shannon entropy (Cover & Thomas, 2006) to the mixture in Eq. 2 (Appendix B). The result does not imply that a joint model cannot produce the correct causal conditional; it says the prediction problem changes. AR places the uncertainty created by the unobserved future inside the training target itself; a non-causal objective trains on the lower-entropy future-conditioned components and must recover their mixture when only a prefix is available. Equality of the represented joint distribution therefore does not imply equality of the inductive bias for sequential adaptation. Appendix C gives a minimal distribution where the joint is captured exactly but the sequential context-compression behaviour differs, and shows how errors compound when the unknown future is replaced by model-generated states.

Evaluation convention. *Our theory concerns the learned prefix conditional, so AR is evaluated by next-token scoring and MLM/diffusion with the same prefix and masked target. This removes decoding-schedule and sampling confounds; Section 6 separately tests transfer to native iterative inference. Noise-time and suffix-mask controls are in Appendix H.*

3 Does the Training Objective Determine ICL?

The theory makes a simple empirical prediction: models trained to predict directly from prefixes should exploit additional context more readily than models trained primarily on bidirectional conditionals. We test this first in natural language, where architecture and scale vary across models, and then in a matched experiment where only the training objective changes.

Natural-text context use. We compare NLL at positions 500 and 50 on OpenWebText (Gokaslan & Cohen, 2019); negative late-minus-early NLL indicates context compression (Kaplan et al., 2020; Olsson et al., 2022). Figure 2(a) shows strong compression for AR (-0.52 and -0.50 nats), near-neutral values for MLM ($+0.08$ and $+0.17$), and the opposite trend for symmetric MDLM and GIDD ($+0.59$ and $+0.65$). Dream-7B (Ye et al., 2025), which is initialised from an autoregressive model

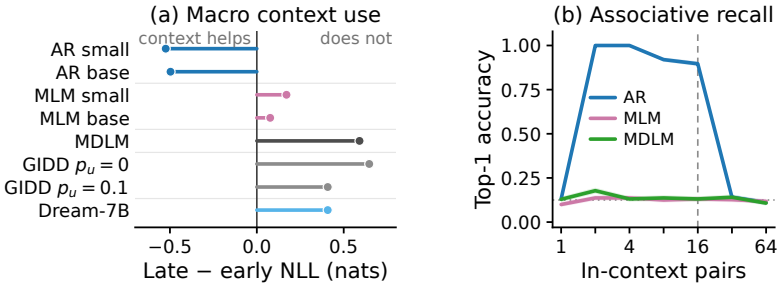


Figure 2: **Objective factorisation predicts context use.** (a) Late-minus-early NLL on natural text; negative values indicate context compression. (b) Matched associative recall: AR learns the mapping while MLM and MDLM remain near chance; the dashed line marks the training context.

and uses context-adaptive token-level noise rescheduling, lies between these regimes; this comparison alone cannot identify the objective as causal, since the models differ in many other ways. Scores for every model, including three not shown in the figure, are in Appendix H.2.

Matched objective control. We train identical transformers on fresh associative-recall episodes, changing only the objective. Figure 2(b) shows that AR learns the in-context mapping while MLM and MDLM remain near chance. Attention/routing and representation-geometry controls are in Appendices J, N and M.

The two experiments address complementary confounds. The natural-text probe shows that the effect is not specific to a synthetic retrieval task, but comparisons across existing models entangle objective, architecture and scale. The matched experiment removes those confounds and recovers the same separation when only the objective changes. Together they motivate treating factorisation—not model family alone—as the variable to intervene on next.

4 Injecting Causal Asymmetry into Diffusion

If weak diffusion ICL arises because prefix-conditioned prediction is underrepresented during training, the direct test is to change that representation while leaving the model itself untouched. We therefore modify only the corruption process so that later targets are more often predicted from relatively cleaner earlier context.

Why later targets? Early in a sequence, little context has been observed, so much of a token's uncertainty cannot yet be reduced from the prefix. Later targets contain more uncertainty that can potentially be removed by exploiting preceding context. Under a fixed corruption budget, shifting denoising targets later therefore increases the proportion of training events in which reducing loss requires using the prefix. Appendix E formalises this intuition in terms of reducible uncertainty; the power law below is a controlled probe, not an assumption that early tokens are unimportant or that this schedule is optimal.

We therefore set

$$g_i \;=\; \frac{(i/L)^{\alpha}}{L^{-1}\sum_j (j/L)^{\alpha}}, \qquad \beta_i(t) \;=\; \mathrm{clip}\big[\beta(t)\, g_i,\, 0,\, 1\big]. \tag{4}$$

At α=0 this is symmetric diffusion; $\alpha > 0$ shifts corruption towards later positions. MDLM and LLaDA use a mean-one profile, so the nominal pre-clipping corruption mass is unchanged; endpoint clipping perturbs this equality slightly and is quantified in Appendix G. GIDD instead uses max-normalisation for numerical stability, giving the causal-bias arm a lower corruption budget than its symmetric baseline. Architecture, attention and decoding are unchanged.

Equivalently, the intervention can be interpreted as tilting the distribution of prediction orders. Absorbing diffusion averages over token prediction orders, while position-specific schedules induce a non-uniform order law (Ou et al., 2025; Garg et al., 2025). Symmetric diffusion spreads training across many conditioning orders; AR concentrates it on the left-to-right interface. Our schedule shifts mass

towards configurations in which earlier tokens are available when later tokens are predicted, without imposing a causal attention mask or collapsing diffusion to AR. Appendix D derives the implemented position-dependent objective and endpoint/clipping caveats; Appendix G gives supervision-budget, convergence and numerical controls.

The prediction is directional: if prefix-conditioned supervision promotes ICL, this tilt should move diffusion models towards stronger contextual adaptation, not necessarily make them equivalent to AR.

5 Causal Asymmetry Improves Few-Shot ICL

We first test the intervention on episodic tasks where the required contextual computation is known. Every evaluation episode defines a fresh latent rule, so the model cannot solve the task by retrieving a fixed association from its parameters: **Key–Value** tests retrieval, **Cipher** latent mapping inference, **OverrideKV** conflicting contextual bindings, and **Binary FSM** sequential rule application. Each point averages 500 episodes.

Figure 3 shows stronger contextual adaptation for causal-bias MDLM across the suite. The precise K-dependence differs by task; the claim is not improvement at every context length, but that changing the supervision structure systematically changes prompt use. Full task generators, additional α values and scored/generative readouts are in Appendix I. Because the scored and generative readouts need not agree, Section 6 tests the intervention under native diffusion inference.

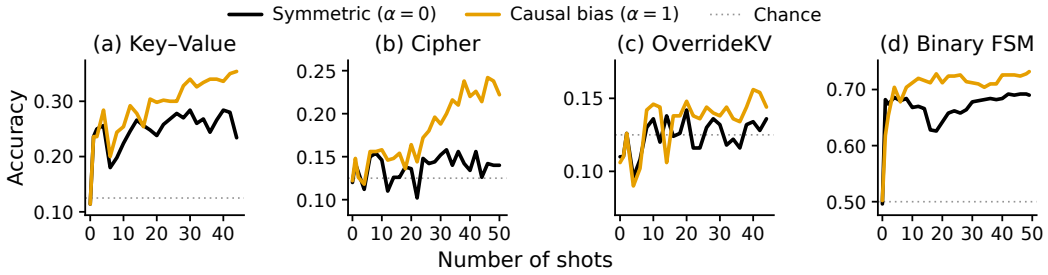


Figure 3: **Causal asymmetry improves few-shot ICL.** Accuracy on four episodic tasks for symmetric (α=0, black) and causal-bias (α=1, gold) MDLM; dotted lines mark chance.

6 From Learned Conditionals to Native Generation

The theory and controlled tasks isolate the conditional predictor learned during training. A diffusion LM, however, generates by repeatedly composing those conditionals through a denoising chain. The next question is therefore whether the training-induced bias survives the model's native inference procedure.

6.1 Context-conditioned generation across diffusion objectives

We test this on OverrideKV using native 64-step generation: demonstrations and query are clamped, the answer begins fully masked, and each family generates under a decoder held fixed across its two arms (Appendix L.1). Because inference-time choices—the masking pattern presented to the model and the order in which tokens are unmasked—are themselves known to change how diffusion language models use context (Piskorz et al., 2025a), that procedure is held fixed across arms. Alongside MDLM we train GIDD (von Rütte et al., 2025), whose forward process includes a distinct uniform-noise component, so this is simultaneously a cross-objective transfer test.

Figure 4 shows transfer across both inference procedure and diffusion objective: at K=32, MDLM lift is 0.510 versus 0.332, while GIDD retains 0.425 versus 0.005 for its symmetric arm. Transfer in both families closes a different gap than the few-shot results of Section 5: the effect is present not only in the local readout we analyse theoretically, but in the models' actual iterative inference

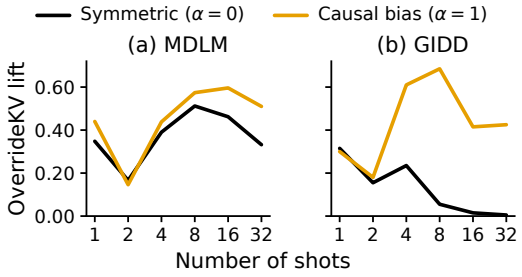


Figure 4: **Native-generation transfer.** At larger context sizes, causal-bias training retains stronger OverrideKV generation under native 64-step MDLM and GIDD inference.

procedure. We use p_{uniform}=0.1 and the remask/clean protocol as the closest common evaluation across the two families (n=500 and n=200 episodes); the p_{uniform}=0.2 arms, teacher-forced and free variants and the discrepant scored readouts are in Appendix L.

6.2 Unrestricted natural-language generation

The contextual task establishes transfer to native inference, but it leaves open whether the same intervention simply trades contextual behaviour against ordinary generation quality. We therefore evaluate unrestricted continuations on OpenWebText, WikiText-103, TinyStories and Sherlock across denoising schedules and sampling temperatures (Table 1). Generated continuations are compared against held-out continuations by unigram precision, trigram precision and longest matching span, with a held-out judge NLL for GIDD; all comparisons are paired on the same prefixes and Holm-corrected jointly.

Table 1: **Native-generation quality.** Paired, Holm-corrected comparisons of causal-bias versus symmetric generation. "Directional" is the raw win count at the reference temperature; the protocol and per-metric counts are in Appendix L.

Family	Metric	Holm causal / sym.	Directional
MDLM	Longest shared span	12 / 0	17/20
MDLM	Unigram precision	12 / 0	19/20
MDLM	Trigram precision	3 / 0	17/20
GIDD (p_u=0.1)	Judge NLL	9 / 0	11/12
GIDD (p_u=0.1)	Longest shared span	6 / 0	7/12
GIDD (p_u=0.1)	Trigram precision	1 / 0	8/12
GIDD (p_u=0.1)	Unigram precision	1 / 2	7/12

Table 1 is therefore a control on generative competence, not another ICL metric. MDLM shows broad improvement—every Holm-significant comparison across the three continuation-overlap measures favours causal bias—while GIDD is mixed but shows no systematic degradation: judge NLL and longest-span metrics favour the intervention, and unigram precision contains two corrected losses. The main cross-family conclusion is consequently narrower and stronger: the ICL intervention survives native generation without being explained by a collapse in ordinary generation quality.

The MDLM advantage is absent at one fully parallel step, emerges under iterative refinement—though its magnitude is not monotone in the number of steps—and survives repetition filtering. The evaluation protocol, Holm-corrected counts for every metric, the refinement-budget control and the repetition control are in Appendix L.

We also test an orthogonal implication of the marginalisation view: whether the model's direct conditional agrees with the same conditional after integrating out an unknown neighbouring token. The discrepancy improves substantially for MDLM, while GIDD stays substantially inconsistent in both arms. We therefore treat self-consistency as model-specific supporting evidence rather than a central consequence of the intervention; the definition, top-M controls and superseded GIDD protocol are in Appendix K.

7 SCALING THE INTERVENTION BY CONTINUED PRETRAINING

Finally, we ask whether the intervention can move an already trained large diffusion LM. Starting from LLaDA-8B (Nie et al., 2025), we continue-pretrain SYM and FORWARD for 500M tokens on identical data, optimiser schedule and random draws; only the corruption profile differs. The base checkpoint separates ordinary continued pretraining from the intervention, and because both arms see identical data and compute, paired SYM–FORWARD differences isolate the positional intervention far more tightly than comparisons across independently trained model families.

Figure 5(a) uses key–value retrieval with one relevant binding and $K-1$ competing for the same key; we call these distractor bindings rather than shots, since the extra demonstrations carry no further evidence about the answer. All models degrade with interference, but FORWARD degrades more slowly: at K=16, accuracy is $0.354/0.422/0.458$ for base/SYM/FORWARD. The paired margin

$$\Delta R \;=\; R_{\text{FORWARD}} - R_{\text{SYM}}, \qquad R \;=\; z(\text{target}) - z(\text{distractor}), \tag{5}$$

a within-example logit margin at a fixed query position (q=2048), is essentially null without interference ($\Delta R = 0.023$ at K=1) and grows as bindings compete (0.142, 0.204, 0.334 and 0.289 at $K = 2, 4, 8, 16$), with paired bootstrap intervals excluding zero for $K \geq 2$.

Figure 5(b) provides the stronger fingerprint. Moving the query through positions with training gain $g(q) \in \{0.25, 0.5, 1, 1.5\}$ leaves K=1 near zero, but with interference the effect changes sign with the imposed profile: at K=16, ΔR moves from -0.403 to $+0.646$. The intervention is worse where training allocated less relative supervision and better where it allocated more. A uniform improvement could still reflect generic continued-training drift; the sign reversal is stronger because the model becomes worse where the intervention reduced relative supervision and better where it increased it.

Target weighting is not sufficient. Figure 5(c) compares FORWARD with a loss-weighted SYM arm whose corrupted inputs are bit-identical to SYM but whose gradient weight increases with position. It reproduces neither signature: at q=2048 its margin stays non-positive as K grows, and its K=16 positional slope is -0.042 versus $+0.857$ for FORWARD. The control therefore rules out positional target weighting as a sufficient explanation and points instead to the conditioning geometry—what context is visible when each target is learned. Together, panels (b) and (c) identify not simply where the intervention acts, but what part of the intervention matters: changing the contexts from which targets are learned, rather than merely weighting later targets more heavily. Full margins and the clipping-aware control analysis are in Appendix G.1.

The language-modelling controls at 8B are a clean null: symmetric-mask cross-entropy for SYM versus FORWARD is $0.9024/0.9041$ at t=0.2, $1.8203/1.8242$ at t=0.5 and $4.1673/4.1711$ at t=0.8—identical to two decimals and within 0.004 nats throughout. Training details and LM controls are in Appendices F–G, key–value accuracies in Appendix I, and the margin, position-sweep and control tables in Appendix G.1.

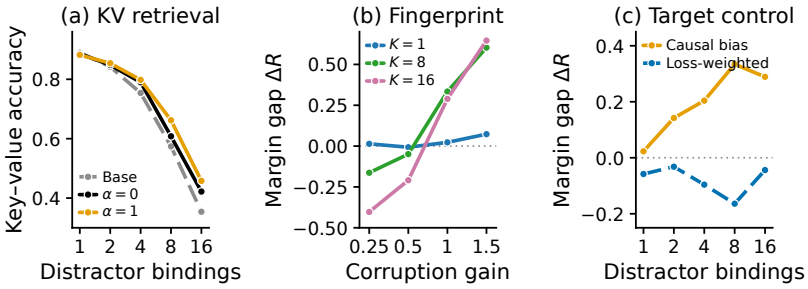


Figure 5: **Scaling to LLaDA-8B.** (a) KV retrieval after 500M-token continued pretraining. (b) FORWARD−SYM margin follows the queried position's training gain $g(q)$. (c) Loss-weighted SYM with identical corrupted inputs does not reproduce the effect, separating conditioning geometry from target weighting.

8 DISCUSSION

Taken together, the results support a specific objective-level account of ICL. Autoregressive training directly optimises the problem of predicting the next token from the observed prefix. Symmetric masked diffusion instead spreads supervision across conditionals defined by many different observation patterns. When we shift that training distribution towards configurations in which earlier context is available for predicting later targets, contextual adaptation improves in the predicted direction—across controlled few-shot tasks, native iterative generation, and continued pretraining at 8B scale. The relevant distinction is therefore not what joint distribution the model can represent, but which conditional prediction problems its training objective repeatedly asks it to solve.

Autoregression is not necessary for ICL, but it gives sequential context use a privileged place in the training objective. Masked and diffusion models can represent the same joint distribution and can themselves exhibit contextual adaptation. Our position-biased diffusion models also do not recover the full AR solution in the matched associative-recall experiment (Appendix J), so the intervention should not be understood as turning diffusion into autoregression. The distinction is one of inductive pressure: AR trains the model to predict from the observed prefix at every position, whereas non-causal objectives distribute supervision across many different conditioning structures.

This distinction explains why we separate local conditional evaluation from native generation. The theory concerns $p(x_t \mid x_{<t})$, so one-step scoring is the cleanest diagnostic of the conditional map the objective has learned, while a diffusion sampler repeatedly composes such conditionals under its own masking schedule. Transfer to iterative OverrideKV generation in both MDLM and GIDD—and to broad natural-language generation for MDLM—is therefore additional evidence rather than an assumption built into the theory.

The intervention should likewise be viewed as a probe of the hypothesis, not an optimal recipe. A fixed power-law schedule lets us change the distribution of conditioning configurations while leaving architecture and decoding fixed.

The LLaDA target-weighting control sharpens this interpretation. Loss-weighted SYM receives the same corrupted inputs as SYM and changes only how strongly later targets contribute to the loss. It reproduces neither of FORWARD's defining signatures: its advantage does not grow with contextual interference, and it does not develop the corresponding positional dependence. The effect therefore does not come from weighting later targets more heavily. It comes from changing which context is visible when those targets are learned.

More broadly, the sequential-strategy view suggests a design principle beyond this schedule: training should place sufficient mass on prediction problems for which useful information must be extracted from the available context. AR does this automatically through factorisation; diffusion permits a continuum between symmetric and strongly directional conditioning structures.

Limitations. The controlled from-scratch experiments use small and mid-scale models and synthetic tasks because these provide exact objective control and fresh task generators. Our 8B result is a 500M-token continued-pretraining intervention on one public base model rather than a from-scratch scale replication. The power-law corruption profile is only one possible schedule, and the scored and native-generation readouts do not always agree. Finally, the Bayesian/sequential-strategy result is relative to the training meta-distribution and does not imply arbitrary out-of-distribution task generalisation. Appendix O gives the full limitations and statistical details.

9 RELATED WORK

Causal factorisation is not necessary for ICL. Masked models can exhibit generative ICL (Samuel, 2024), and recent theory obtains comparable excess-risk rates for causal and masked objectives under suitable assumptions (Liu et al., 2026). Our claim is instead about inductive bias: left-to-right AR makes the sequential posterior predictive a direct target at every prefix, while symmetric masked objectives distribute supervision over other conditioning patterns.

Theories of ICL. ICL has been studied as Bayesian inference over latent tasks (Xie et al., 2022; Akyürek et al., 2023), implicit optimisation (von Oswald et al., 2023), algorithm learning (Garg

et al., 2022; Li et al., 2023; Lin et al., 2024) and induction-head computation (Olsson et al., 2022), with complementary work on the data properties that promote ICL (Chan et al., 2022; Min et al., 2022; Wei et al., 2022). Closest to our starting point is memory-based sequential meta-learning: Ortega et al. (2019) show that sequential log-loss meta-training has the posterior predictive as its optimal history-to-prediction strategy, and Lampinen et al. (2024) explicitly place language-model ICL within this broader family of meta-learned sequential adaptation, while prequential accounts connect next-token loss to meta-learned generalisation (Elmoznino et al., 2025). We build on this line but ask a different question: what happens when the same sequence distribution is learned under a different factorisation? Proposition B.1 makes episodic ICL a conditional subproblem of AR log loss; Theorem 2.1 then isolates the extra marginalisation required by future-conditioned objectives. To our knowledge, prior work has not used this sequential-strategy equivalence to derive a factorisation-level prediction for non-causal language-model objectives and test it through a controlled diffusion intervention.

Masked and diffusion language models. Discrete diffusion was formalised for categorical data by Austin et al. (2021) and Hoogeboom et al. (2021) following Sohl-Dickstein et al. (2015), and developed through MDLM (Sahoo et al., 2024), GIDD (von Rütte et al., 2025) and related formulations (Lou et al., 2024; Campbell et al., 2022). Absorbing diffusion can be interpreted as expected any-order autoregression (Ou et al., 2025), and multivariate noise schedules as nonuniform distributions over decoding orders (Garg et al., 2025); state-dependent masking schedules have also been developed independently (Shi et al., 2024). These results complement our formulation: our contribution is not the existence of position-dependent schedules, but a fixed directional tilt motivated by sequential ICL, together with a controlled causal test across MDLM, GIDD and LLaDA-8B. Block and semi-autoregressive models interpolate between the two paradigms (Arriola et al., 2025; Han et al., 2023; Stern et al., 2019; Li et al., 2022), and hybrids reintroduce partial causal structure (Ye et al., 2025; Nie et al., 2025). Recent diffusion-ICL studies report strong masked-model ICL, bidirectional induction circuits and a marked sensitivity to the masking pattern presented at inference (Piskorz et al., 2025b;a; Catruna & Radoi, 2026). Closest to our 8B result, Li et al. (2026) show that query position strongly affects diffusion-LLM ICL under fixed models and attribute the effect to decoding dynamics; we instead manipulate the training-time positional corruption profile and obtain a within-checkpoint sign reversal that follows that profile, separating a learned training bias from a generic inference-time position effect. Masked LMs (Devlin et al., 2019) likewise model $p(x_i \mid x_{\setminus i})$ rather than the sequential conditional. Our intervention is complementary: it changes the training distribution of conditionals and measures the resulting positional fingerprint while decoding is held fixed.

10 Conclusion

We asked why next-token pretraining so reliably produces models that adapt in context. We show that the sequential-strategy optimality principle of Ortega et al. (2019) has a direct consequence for ICL: under a latent task distribution, Bayes-optimal episodic ICL is itself a conditional subproblem of population autoregressive training. AR therefore directly supervises the posterior-predictive map required for sequential adaptation. Non-causal objectives can represent the same joint distribution, but they distribute training over different conditionals and must recover the prefix predictor through future marginalisation. The experiments identify this objective structure as a major inductive bias for ICL. Matched AR, MLM and diffusion models diverge when only the training factorisation changes; tilting diffusion corruption towards the sequential interface improves four distinct forms of contextual adaptation—associative retrieval, latent-rule inference, overriding a conflicting binding and sequential rule application; the effect survives native MDLM and GIDD generation; and at 8B scale it grows with contextual interference, changes sign with the imposed positional profile, and cannot be reproduced by positional loss weighting alone. ICL is therefore not determined only by what distribution a language model can represent. It depends critically on how that distribution is factorised.

AI USE STATEMENT

We used large language models (ChatGPT and Claude) to assist in improving the clarity and readability of the manuscript, including grammar and minor wording improvements. The LLMs did not contribute to research ideation, experimental design, analysis, or interpretation of results. All scientific content, analyses, and conclusions are the work of the authors.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Marianne Arriola, Aaron Gokaslan, Justin T. Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Aaron van den Oord. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279, 2022.

Andy Catruna and Emilian Radoi. Induction in both directions: A mechanistic analysis of in-context learning in masked diffusion language models. *arXiv preprint arXiv:2607.15893*, 2026.

Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35: 18878–18891, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Eric Elmoznino, Tom Marty, Tejas Kasetty, Léo Gagnon, Sarthak Mittal, Mahan Fathi, Dhanya Sridhar, and Guillaume Lajoie. In-context learning and occam's razor. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. arXiv:2410.14086.

Prateek Garg, Bhavya Kohli, and Sunita Sarawagi. Masked diffusion models are secretly learned-order autoregressive models. *arXiv preprint arXiv:2511.19152*, 2025. EurIPS 2025 Workshop on Principles of Generative Modeling.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. `https://github.com/jcpeterson/openwebtext`, 2019.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11575–11596, 2023.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Andrew K. Lampinen, Stephanie C. Y. Chan, Aaditya K. Singh, and Murray Shanahan. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4328–4343, 2022.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594, 2023.

Zhengheng Li, Panrui Li, Xuyang Liu, and Puzhi Xia. Where to place the query? unveiling and mitigating positional bias in in-context learning for diffusion llms via decoding dynamics. *arXiv preprint arXiv:2606.19349*, 2026.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, 2024.

Chenrui Liu, Chuanlong Xie, Falong Tan, Yicheng Zeng, and Lixing Zhu. A unified framework for in-context learning with causal and masked language models. *arXiv preprint arXiv:2607.04081*, 2026.

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*, volume 235 of *Proceedings of Machine Learning Research*, 2024. arXiv:2310.16834.

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 *Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

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*, volume 38, 2025. arXiv:2502.09992.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Pedro A. Ortega, Jane X. Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, Siddhant M. Jayakumar, Tom McGrath, Kevin Miller, Mohammad Azar, Ian Osband, Neil Rabinowitz, András György, Silvia Chiappa, Simon Osindero, Yee Whye Teh, Hado van Hasselt, Nando de Freitas, Matthew Botvinick, and Shane Legg. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, 2025. arXiv:2406.03736.

Julianna Piskorz, Cristina Pinneri, Alvaro Correia, Motasem Alfarra, Risheek Garrepalli, and Christos Louizos. Masks can be distracting: On context comprehension in diffusion language models. *arXiv preprint arXiv:2511.21338*, 2025a.

Julianna Piskorz, Cristina Pinneri, Alvaro Correia, Dana Kianfar, Motasem Alfarra, and Christos Louizos. In-context learning in diffusion language models. In *NeurIPS 2025 Workshop: What Can('t) Transformers Do?*, 2025b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M. Rush, Yair Schiff, Justin T. Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Samuel. Berts are generative in-context learners. *Advances in Neural Information Processing Systems*, 37, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2406.04329.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. Insertion transformer: Flexible sequence generation via insertion operations. In *International Conference on Machine Learning*, pp. 5976–5985. PMLR, 2019.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174, 2023.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61810–61843, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

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.

A Appendix Overview and Notation

Appendices B–C establish the objective-level theory and minimal construction; D–G derive the intervention and document training and supervision controls; H–L give evaluation robustness and complete task and generation results; M–N test alternative representation and circuit explanations; O collects limitations. Table 2 maps these components to the main text, and Table 3 fixes notation and model names used throughout.

Table 2: **Map of the appendix.**

	Appendix	Contents
theory	B Theory	Proofs of Theorem 2.1, Proposition B.1, Corollary B.2 and Propositions B.3–B.4; relation to sequential meta-learning and Bayesian accounts; scope of the claim.
	C Minimal construction	A distribution with a strict future-marginalisation gap despite exact joint modelling; numerical instantiation; iterative-denoising amplification.
	D Forward process and objective	Position-dependent corruption, interior change-of-time weighting, and endpoint/clipping caveats.
setup	E Why positional supervision	The reducible-uncertainty argument for reallocating corruption towards later positions.
	F Models and training	Every model in the paper: parameters, depth, width, context, steps, optimiser; held-out loss curves.
	G Supervision controls	Nominal-budget normalisation, supervision-matched checkpoints, numerical controls, α sweep, and target-weighting versus conditioning-geometry control.
protocol	H One-step robustness	Noise-time invariance of the denoiser; the suffix-mask (distraction) probe at both scales.
	I Tasks and battery	Episode format and generation for all synthetic tasks; scoring; where scored and generated readouts disagree.
results	J Matched-objective extras	Associative recall with both diffusion arms, and the attention-by-relative-position maps.
	K Conditional coherence	Marginalisation self-consistency: MDLM improvement, corrected GIDD protocol, superseded fill-tail protocol.
	L Native generation	Full generation protocol, Holm-corrected comparisons, repetition control.
controls	M Representation geometry	Toy-models-of-superposition control: can non-causal objectives learn structured geometry at all?
	N Routing and circuits	QK induction and OV copy diagnostics; the OverrideKV haystack and recency-last decomposition.
	O Limitations	Scale, schedule, readouts, statistical power and scope.

Table 3: **Notation and model names.** Names in the left column are used consistently in every figure, table and section.

Symbol / name	Meaning
$x_{<t}$, $x_{\leq k}$	Prefix of a sequence; the conditioning set of the sequential predictor.
Z, $p(z \mid x_{<t})$	Latent task and its posterior after a prefix; the predictor of interest is the posterior predictive of Eq. 1.
α	Strength of the position-dependent corruption; α=0 is the unmodified objective.
g_i, $\beta_i(t)$, t_i	Position gain, per-position masking probability and per-position effective diffusion time (Eq. 15).
p_u	GIDD uniform-noise mixing probability (0.1 or 0.2 here).
K	Number of in-context demonstrations, or distractor bindings where stated.
S	Number of `[MASK]` tokens appended after the query in the distraction probe.
ΔR	Paired margin difference $R_{\text{FORWARD}} - R_{\text{SYM}}$ with $R = z(\text{target}) - z(\text{distractor})$.
QK gap, OV gap	Attention difference towards the recent binding; logit contribution of the attended value to the gold token.
AR, MLM, MDLM	Autoregressive, masked-LM and masked-diffusion objectives.
Symmetric (α=0)	Diffusion arm with position-independent corruption; black in every figure.
Causal bias (α=1)	Diffusion arm with the position-dependent corruption of Section 4; gold in every figure.
SYM, FORWARD	The two LLaDA-8B continued-training arms, symmetric and position-biased.
Haystack, recency-last	OverrideKV variants testing retrieval across distance and recency against frequency (Appendix N).

B Theory, Proofs and Relation to Sequential Meta-Learning

B.1 ICL as a conditional subproblem of autoregressive pretraining

We first make precise what an ICL episode is, so that the statement is about ICL and not only about conditional maximum likelihood. An episode draws a latent task and then demonstrations and a query from it, $Z \sim p(z)$ and $(D_k, U_\star, V_\star) \sim p(\cdot \mid Z)$, where D_k is a sequence of k demonstrations and $(U_\star, V_\star)$ is the query and its answer. A few-shot learner is judged on the episodic ICL risk

$$\mathcal{R}_k^{\mathrm{ICL}}(q) \;=\; \mathbb{E}\big[-\log q\big(V_\star \mid U_\star, D_k\big)\big], \tag{6}$$

and we serialise the episode so that the prefix immediately before the answer is $H_k = \mathrm{serialize}(D_k, U_\star)$.

Proposition B.1 (AR pretraining contains Bayes-optimal ICL)**.** *Let sequences be generated by drawing $Z \sim p(z)$ and then a trajectory conditionally on Z, and let $\mathcal{L}_{\mathrm{AR}}(q) = \mathbb{E}\big[-\sum_t \log q(X_t \mid X_{<t})\big]$. Then $\mathcal{L}_{\mathrm{AR}}$ is uniquely minimised, pointwise at every prefix of positive probability, by*

$$q^\star(x_t \mid x_{<t}) \;=\; p(x_t \mid x_{<t}) \;=\; \int p(x_t \mid x_{<t}, z)\, p(z \mid x_{<t})\, dz, \tag{7}$$

and consequently, at every serialised query prefix H_k, $q^\star(V_\star \mid U_\star, D_k)$ equals the Bayes-optimal k-shot ICL predictor $\int p(V_\star \mid U_\star, D_k, z)\, p(z \mid U_\star, D_k)\, dz$, so that

$$q^\star \in \arg\min_q \mathcal{L}_{\mathrm{AR}}(q) \implies q^\star \in \arg\min_q \mathcal{R}_k^{\mathrm{ICL}}(q) \quad \textit{for every } k, \tag{8}$$

for every query prefix supported by the training meta-distribution.

Proof. Write the population AR loss as a sum of per-prefix cross-entropies, $\mathcal{L}_{\mathrm{AR}}(q) = \sum_t \mathbb{E}_{x_{<t}}\big[\mathrm{CE}\big(p(\cdot \mid x_{<t}),\, q(\cdot \mid x_{<t})\big)\big]$. Each term depends on q only through $q(\cdot \mid x_{<t})$ for that prefix, and cross-entropy is minimised over distributions uniquely at the true conditional, giving the first equality of Eq. 7; the second is the law of total probability under the latent-task model, with $p(z \mid x_{<t}) \propto p(z)\, p(x_{<t} \mid z)$. Taking $x_{<t} = H_k$, so that $X_t = V_\star$, gives the posterior predictive over the latent task given the demonstrations and query. For the risk statement, $\mathcal{R}_k^{\mathrm{ICL}}(q) = \mathbb{E}_{H_k}\big[\mathrm{CE}(p(\cdot \mid H_k), q(\cdot \mid H_k))\big] + \mathrm{const}$ is again minimised pointwise by the true conditional, which $q^\star$ matches at every prefix in the support; prefixes of probability zero are unconstrained. □

At the population optimum, the Bayes-optimal ICL conditional is therefore a direct regression target of the AR objective rather than an additional objective imposed at test time. Finite models, optimisation and out-of-support tasks remain separate questions, discussed in Appendix B.7.

B.2 In-context gain equals predictive information

Corollary B.2 (In-context gain)**.** *Let $\mathcal{R}_k^\star = H(V_\star \mid U_\star, D_k)$ be the optimal k-shot risk. Then*

$$\mathcal{R}_0^\star - \mathcal{R}_k^\star \;=\; I\big(V_\star; D_k \mid U_\star\big), \qquad \mathcal{R}_{k-1}^\star - \mathcal{R}_k^\star \;=\; I\big(V_\star; D_k \mid U_\star, D_{<k}\big). \tag{9}$$

Proof. By Proposition B.1 the optimal k-shot risk is the conditional entropy $H(V_\star \mid U_\star, D_k)$, and $\mathcal{R}_0^\star = H(V_\star \mid U_\star)$; the definition of conditional mutual information gives the first identity, and the chain rule applied with $D_{<k}$ in the conditioning set gives the second. □

Under the Bayes-optimal sequential strategy, demonstrations improve log loss exactly to the extent that they carry conditional information about the answer. This is what licenses the context-use diagnostics of Section 3: a model that fails to improve with context is failing to realise an information-theoretic gain available to the optimal predictor of the same distribution. Note that this quantity, $I(V_\star; D_k \mid U_\star)$, is a different conditional mutual information from the marginalisation gap $I(X_{k+1}; X_{k+2:L} \mid X_{\le k})$ of Theorem 2.1: the first is over demonstrations and the answer, the second over the next token and the unknown future. They are related in form, not in value.

B.3 Proof of Theorem 2.1

Proof. By the law of total probability, $p^\star(x_{k+1} \mid x_{\le k}) = \sum_{x_{k+2:L}} w_{x_{k+2:L}}\, p^\star(x_{k+1} \mid x_{\le k}, x_{k+2:L})$ with weights $w_{x_{k+2:L}} = p^\star(x_{k+2:L} \mid x_{\le k})$. Shannon entropy is strictly concave on the simplex, so whenever the constituent distributions are not all equal, $H\big(\sum_j w_j q_j\big) > \sum_j w_j H(q_j)$. Taking expectations over $x_{\le k}$ gives the claim, and the chain rule for mutual information identifies the gap as $I(X_{k+1}; X_{k+2:L} \mid X_{\le k})$, which is positive exactly when the constituent distributions differ. □

B.4 Causal boundaries are rare under symmetric masking

Proposition B.1 concerns supervision of the form "clean prefix, unknown future". Define the AR-aligned corruption event at boundary k, $E_k = \{X_{1:k} \text{ visible},\ X_{k+1:L} \text{ masked}\}$.

Proposition B.3 (AR-aligned supervision is rare under symmetric diffusion)**.** *Under position-independent masking at rate β with $\beta \sim \mathcal{U}(0,1)$,*

$$\Pr(E_k) \;=\; \int_0^1 (1-\beta)^k \beta^{L-k}\, d\beta \;=\; B(k+1, L-k+1) \;=\; \frac{1}{(L+1)\binom{L}{k}}. \tag{10}$$

Proposition B.4 (Monotone corruption tilts supervision towards causal boundaries)**.** *Let positions be masked independently with rates $\beta_1, \dots, \beta_L$ and let $i < j$ with $\beta_i \le \beta_j$. Then $\Pr(i \text{ clean}, j \text{ masked}) / \Pr(i \text{ masked}, j \text{ clean}) = (1-\beta_i)\beta_j / \big(\beta_i(1-\beta_j)\big) \ge 1$, with equality iff $\beta_i = \beta_j$. Any monotone non-decreasing profile therefore places at least as much mass on every prefix-clean/suffix-masked pattern as on its position-swapped counterpart.*

Proposition B.3. Conditional on β, positions are masked independently, so $\Pr(E_k \mid \beta) = (1-\beta)^k \beta^{L-k}$. Integrating over $\beta \sim \mathcal{U}(0,1)$ gives the Beta integral $B(k+1, L-k+1) = k!\,(L-k)!/(L+1)! = 1/\big((L+1)\binom{L}{k}\big)$. □

Proposition B.4. Fix the configuration of all positions other than i and j and let C be its probability, common to both events by independence. Then the two probabilities are $C(1-\beta_i)\beta_j$ and $C\beta_i(1-\beta_j)$, so the ratio is as stated; $\beta \mapsto \beta/(1-\beta)$ is strictly increasing on $(0,1)$, so $\beta_i \le \beta_j$ gives a ratio of at least one. Summing over the remaining configurations preserves the inequality. □

For an interior boundary Proposition B.3 is combinatorially tiny: at L=512, k=256 it is below 10^{-150}. An autoregressive model receives prefix-conditioned supervision at every token of every sequence; symmetric diffusion assigns combinatorially small probability to the exact clean-prefix/masked-suffix configuration. Two remarks. The uniform prior on β is the standard continuous-time masked-diffusion schedule at a uniformly sampled time; any schedule with full support on $(0,1)$ gives the same conclusion with a different constant. And Proposition B.4 is a statement about the corruption distribution alone—it holds for any monotone profile, including but not limited to the power law of Eq. 4. Neither proposition says that the exact event E_k is the only useful form of causal supervision; it is the strictest one, and the tilt result applies to the whole family of prefix-clean patterns.

This view is consistent with complementary analyses of absorbing diffusion as any-order autoregression: symmetric masking distributes training across many observation orders, while multivariate schedules induce nonuniform weights over those orders (Ou et al., 2025; Garg et al., 2025). From this perspective, AR places all mass on the left-to-right order, whereas our monotone positional schedule tilts the diffusion order distribution towards that interface without collapsing it to a single autoregressive order. We use the order view as an interpretation of the directional bias rather than as a proof about the implemented objective: the exact order decomposition assumes endpoint-complete position schedules, which our linear rescaling does not satisfy for $g_i < 1$.

B.5 Relation to sequential meta-learning

Existing Bayesian accounts establish that ICL can arise through posterior inference over latent structure in the pretraining distribution (Xie et al., 2022), while sequential-strategy work shows more generally that memory-based predictors trained under sequential proper losses converge to posterior-predictive strategies; recent perspectives explicitly connect this family of memory-based

meta-learning to language-model ICL (Lampinen et al., 2024), and prequential accounts connect next-token loss to meta-learned generalisation (Elmoznino et al., 2025). The sequential-strategy literature makes a structural point: training a memory-based predictor under sequential log loss on trajectories drawn from a task distribution makes the optimum implement the posterior predictive, with the Bayesian update amortised in the model's state rather than computed explicitly (Ortega et al., 2019). The Monte-Carlo objective marginalises the latent task implicitly, because the loss depends on the trajectory and not on the parameter that generated it. Proposition B.1 is the version of that statement we need: it identifies the ICL risk of Eq. 6 as a conditional subproblem of the sequential log loss, so the claim is not that ICL *resembles* Bayesian inference but that the factorisation of the training objective determines whether the posterior-predictive map is the object being optimised.

B.6 Relation to Bayesian accounts of ICL

Xie et al. (2022) studies conditions under which ICL emerges from a latent-concept pretraining distribution, and Akyürek et al. (2023) and von Oswald et al. (2023) show that transformers can implement learning algorithms in their forward pass. Proposition B.1 isolates the part needed here: for serialised ICL episodes, the episodic query risk is literally a conditional subproblem of population AR log loss. We then ask what changes when the training factorisation no longer privileges that conditional—a question about objectives rather than about the computation a given model runs, and one that predicts a difference between models trained on identical data with different factorisations, which is what Section 3 tests.

B.7 Scope: support, out-of-distribution tasks and non-sequential problems

Two caveats bound the claim. First, Bayes-optimality is with respect to the meta-distribution on which the sequential log loss is defined. A novel in-context task is handled to the extent that it is inferable under that learned meta-distribution; genuinely out-of-support tasks require additional assumptions, and nothing here proves arbitrary out-of-distribution generalisation. Second, although any task can be serialised, the theorem concerns adaptation problems whose evidence arrives as a history. It does not say that every useful computation is intrinsically sequential.

C A Minimal Construction Where the Barrier Bites

This appendix gives a data distribution on which the direct future-conditioned objective faces a strict sequential marginalisation gap even when the joint is modelled exactly, together with a numerical instantiation and an analysis of what iterative denoising adds on top.

C.1 Construction

Let $h \in \{1, \dots, H\}$ be a latent hypothesis with prior $\pi(h)$. A sequence $x_{1:L}$ is generated in two stages:

1. the prefix $x_{1:k}$ is drawn from a hypothesis-dependent distribution $p_{\mathrm{pre}}(x_{1:k} \mid h)$ that is informative but non-degenerate about h, so that $p(h \mid x_{1:k})$ is not a point mass;
2. conditioned on h, each suffix token x_i, $i > k$, is drawn i.i.d. from a hypothesis-specific distribution p_h.

We assume the $\{p_h\}$ are not all identical. Everything below is a statement about $p^\star$ itself, so it holds at the population optimum of any objective and cannot be fixed by more capacity or more data.

Proposition C.1 (Autoregressive behaviour)**.** *For the construction above,*

$$p^\star(x_{k+1} \mid x_{\le k}) \;=\; \sum_h p(h \mid x_{\le k})\, p_h(x_{k+1}), \tag{11}$$

and, for a fixed target Y and nested observed histories C_m, conditional entropy satisfies $H(Y \mid C_{m+1}) \le H(Y \mid C_m)$. This is an expectation over histories, not a claim that predictive entropy decreases after every individual observation.

An autoregressive model trained to the population optimum matches the conditional of Proposition C.1 exactly (Proposition B.1), so context compression is not something AR has to discover: it is what the objective optimises.

Proposition C.2 (Joint objectives face a strict gap)**.** *Let $\{p_h\}$ be non-identical. Then*

$$H(X_{k+1} \mid X_{\le k}) \;>\; \mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})]\,. \tag{12}$$

Proof. Fix a prefix $x_{\le k}$. For a realised suffix $f = x_{k+2:L}$, Bayes' rule gives $p(h \mid x_{\le k}, f) \propto p(f \mid h)\, p(h \mid x_{\le k})$, so

$$q_f(x_{k+1}) \;:=\; p^\star(x_{k+1} \mid x_{\le k}, f) \;=\; \sum_h p(h \mid x_{\le k}, f)\, p_h(x_{k+1}). \tag{13}$$

Because the suffix likelihoods $p(f \mid h)$ differ across hypotheses, observing f sharpens the posterior over h in a direction that depends on f; since the $\{p_h\}$ are not all identical, the resulting $\{q_f\}$ are not all identical either. The law of total probability writes the causal conditional as the mixture $q = \sum_f w_f q_f$ with $w_f = p^\star(f \mid x_{\le k})$, and strict concavity of Shannon entropy gives $H(\sum_f w_f q_f) > \sum_f w_f H(q_f)$. The left-hand side is $H(X_{k+1} \mid X_{\le k})$ and the right-hand side is $\mathbb{E}_{X_{k+2:L}}[H(X_{k+1} \mid X_{\le k}, X_{k+2:L})]$. □

The two propositions together isolate the phenomenon: the joint distribution is the same object in both cases, but the conditional that AR optimises is the marginalised one, while a joint objective is rewarded for the sharper future-conditioned family and must recover the marginalised conditional at inference.

C.2 Numerical instantiation

We instantiate the construction with $H{=}3$ hypotheses over a vocabulary of 10 token values, with $h{=}1$ favouring low values, $h{=}2$ middle values and $h{=}3$ high values, so that different futures genuinely imply different predictions. The prefix induces a posterior concentrated on $h{=}2$ (mass 0.5) while retaining mass on the other two hypotheses.

Figure 6 shows the resulting quantities. The causal conditional (panel c) is the posterior-weighted mixture of the three hypothesis distributions and is diffuse, with $H = 2.254$ nats. Each future-conditioned distribution (panels d–i) sharpens in a different direction, with entropies between 1.920 and 2.128 nats and an expectation of ≈ 2.077 nats. The gap

$$\Delta \;=\; 2.254 - 2.077 \;=\; 0.177 \text{ nats} \tag{14}$$

is the cost of marginalisation, and confirms Proposition C.2 numerically. Diffusion training observes futures and learns the sharp conditionals of panels (d)–(i); at inference the future is unknown and the diffuse conditional of panel (c) is required. AR never has access to the future and so never faces the mismatch.

C.3 What iterative denoising adds

Theorem 2.1 concerns prefix-only prediction with a perfectly specified joint. Iterative denoising changes the problem in one respect: the model conditions not on a future drawn from $p^\star(\cdot \mid x_{\le k})$ but on a future $\hat{x}_{k+2:L}$ produced by its own dynamics, with the prefix coordinates clamped. Two reference quantities frame what follows: conditioning on ground-truth futures (the training-like signal) gives ≈ 1.87 nats in our instantiation, and the marginalised causal conditional gives ≈ 2.19 nats, larger by Theorem 2.1.

Three error modes (Figure 7). *(A) Miscalibration.* If the model's implicit posterior $p_\theta(h \mid x_{\le k})$ puts too much mass on the wrong hypothesis, sampled futures are consistent with that hypothesis and the induced conditional is sharp but wrong: entropy falls while accuracy degrades—confident incorrectness rather than amplified uncertainty. *(B) Uninformative futures.* If $\hat{x}_{k+2:L}$ carries little information about h, the effective posterior stays diffuse and the induced conditional lands at or just above the marginalised baseline: iterative conditioning buys nothing. *(C) Systematic bias.* If the denoising dynamics favour futures consistent with one hypothesis largely independently of the prefix

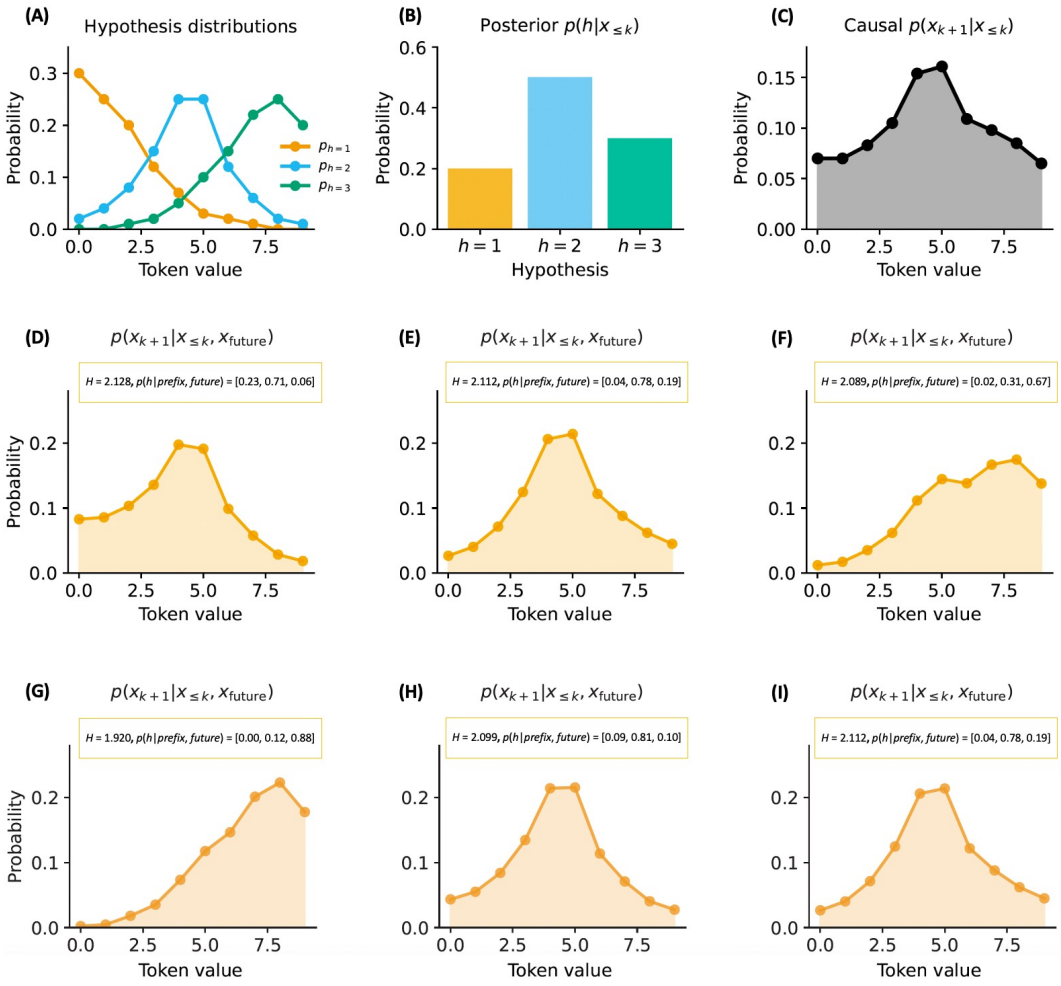


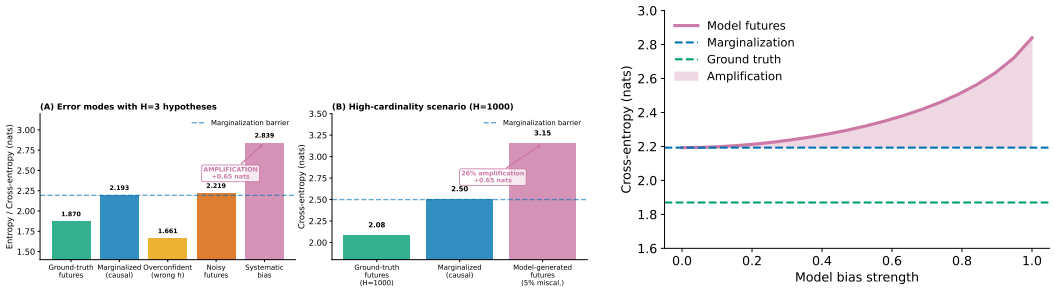
Figure 6: **Numerical verification of Theorem 2.1.** Panel (a): hypothesis distributions $\{p_h\}$. Panel (b): prefix posterior, concentrated on h=2 but not degenerate. Panel (c): causal conditional, diffuse ($H = 2.254$ nats). Panels (d)–(i): future-conditioned distributions for six realisations of $x_{k+2:L}$, with $H \approx 1.92$–2.13 nats. Marginalisation gap $\Delta = 0.177$ nats.

evidence, the model frequently conditions on badly mismatched futures. Always conditioning as if $h{=}1$ gives $\mathbb{E}_h[\mathrm{CE}(p_h, p_1)] \approx 2.84$ nats, well above the 2.19-nat baseline. This is the amplification effect: model-generated futures can make prefix-conditioned prediction strictly worse than the intrinsic barrier.

Bias parameterisation. Writing $p_\theta(h \mid x_{\leq k}) = (1-\beta)\, p^\star(h \mid x_{\leq k}) + \beta\,[1,0,0]$ interpolates between calibration ($\beta{=}0$) and total bias ($\beta{=}1$). Expected cross-entropy increases monotonically in β (Figure 7b), and the region above the dashed baseline is amplification beyond Theorem 2.1.

High-cardinality settings are more brittle. The three-hypothesis case understates the few-shot regime, where many plausible latent tasks can have nearly disjoint label supports. With $H{=}1000$, $V{=}10{,}000$, a true posterior of 0.90 on $h^\star$ and a model that samples $h^\star$ with probability 0.85, wrong hypotheses placing roughly $1/V$ on the correct token give $\mathbb{E}[\mathrm{NLL}] \approx 0.85 \cdot 2.08 + 0.15 \cdot 9.21 \approx 3.15$ nats. Small calibration errors in the induced future distribution have outsized effects once the hypothesis space is large and wrong hypotheses are strongly incompatible with the correct label.

Contrast with autoregressive evaluation. Under teacher forcing, an AR model is evaluated at every position conditioned on the ground-truth prefix, so there is no analogous dependence on model-generated futures. Prefix-conditioned evaluation of a diffusion model necessarily involves unknown futures, and iterative denoising replaces them with model-imputed ones, adding a distribution shift on top of the marginalisation barrier itself.



(a) Error modes against the barrier.

(b) Cross-entropy versus bias strength.

Figure 7: **Iterative denoising amplification.** (a) Ground-truth future conditioning (green) is the training-like target; the marginalised causal conditional (blue dashed) is the Theorem 2.1 barrier. Scenario A (orange): miscalibration. Scenario B (red): uninformative futures. Scenario C (purple): systematic bias, exceeding the marginalisation baseline. The right panel of (a) shows the same mechanism in the high-cardinality setting. (b) Expected cross-entropy increases monotonically with bias strength β; the dashed line is the marginalised baseline.

D The Position-Dependent Forward Process and Its Objective

The intervention is not merely a positional reweighting of a fixed denoising loss: it changes the forward corruption distribution itself. We define the implemented position-dependent process and derive the corresponding interior change-of-time weighting. Because clipping modifies the hazard near the endpoints and the linear rescaling $t_i = t g_i$ is not endpoint-complete for coordinates with $g_i < 1$, the implemented objective should not be described as an exact terminal ELBO without an additional boundary term. These caveats do not affect the controlled comparison at $\alpha{=}0$, nor the distinction between changing the corrupted inputs and reweighting an unchanged input distribution.

D.1 Forward process

Let α_t be the schedule of the unmodified masked-diffusion process, i.e. the probability that a token survives uncorrupted at time t, with $\alpha_0 = 1$, $\alpha_1 = 0$ and α_t strictly decreasing and differentiable. Sample one global time $t \sim \mathcal{U}(\epsilon, 1-\epsilon)$. With L the sequence length, an offset o and a strength

$\alpha \geq 0$, define the position gain and the per-position effective time

$$g_i \;=\; \frac{\big((i+o)/(L-1+o)\big)^{\alpha}}{Z}, \qquad t_i \;=\; \mathrm{clip}\big(t\, g_i,\; \epsilon,\; 1-\epsilon\big), \tag{15}$$

and corrupt position i with the *same* kernel as the unmodified process, but at time t_i:

$$q\big(z_t \mid x\big) \;=\; \prod_{i=1}^{L} q_i\big(z_t^i \mid x^i;\, t_i\big), \qquad q_i(z_t^i = \texttt{[M]} \mid x^i) = 1-\alpha_{t_i}, \quad q_i(z_t^i = x^i \mid x^i) = \alpha_{t_i}. \tag{16}$$

Only the time at which a position is corrupted depends on i; the kernel, the state space and the absorbing structure are unchanged.

D.2 The bound

Define the per-position schedule $\alpha_t^{(i)} := \alpha_{t g_i}$ (clipped as above). Each $\alpha^{(i)}$ is differentiable and non-increasing in t, because $t \mapsto t g_i$ is an increasing reparameterisation of time with $g_i > 0$. It does not in general reach the absorbing state at $t = 1-\epsilon$: for $g_i < 1$ the coordinate terminates at $\alpha_{(1-\epsilon)g_i} > 0$, which is the endpoint incompleteness discussed below. The forward process is therefore a product of L independent absorbing processes, each with its own schedule.

For a single absorbing-state process with schedule α_t, the continuous-time tokenwise bound of Sahoo et al. (2024) reads

$$-\log p_\theta(x) \;\leq\; \mathbb{E}_{t\sim\mathcal{U}(0,1)}\, \mathbb{E}_{z_t\sim q(\cdot\mid x,t)} \left[\frac{-\alpha_t'}{1-\alpha_t} \sum_{i:\, z_t^i = \texttt{[M]}} -\log p_\theta\big(x^i \mid z_t\big) \right], \tag{17}$$

where $\alpha_t' = \mathrm{d}\alpha_t/\mathrm{d}t$. The derivation of Eq. 17 treats positions independently given t: the forward kernel factorises, the reverse posterior factorises, and the KL between them is a sum of per-position terms. That is the only structural property it uses, and our modification preserves it. Substituting $\alpha^{(i)}$ for α at position i and applying the chain rule, $\mathrm{d}\alpha_t^{(i)}/\mathrm{d}t = g_i\, \alpha'(t g_i)$, gives

$$-\log p_\theta(x) \;\leq\; \mathbb{E}_t\, \mathbb{E}_{z_t} \left[\sum_{i=1}^{L} \underbrace{\frac{-g_i\, \alpha'(t g_i)}{1-\alpha_{t g_i}}}_{w_i(t)}\; \mathbf{1}\big[z_t^i = \texttt{[M]}\big]\; \big(-\log p_\theta(x^i \mid z_t, t)\big) \right]. \tag{18}$$

For the mask-only linear schedule used by MDLM and LLaDA the weight collapses. With $\alpha_u = 1-u$, i.e. masking probability $m_i(t) = t g_i$, we have $\alpha'(u) = -1$ and $1 - \alpha_{t g_i} = t g_i$, so

$$w_i(t) \;=\; \frac{-g_i\, \alpha'(t g_i)}{1-\alpha_{t g_i}} \;=\; \frac{g_i}{t\, g_i} \;=\; \frac{1}{t} \qquad \text{for every position } i. \tag{19}$$

The per-position hazard ratio is the baseline weight, independent of g_i. The intervention therefore changes the corruption distribution and nothing else: the loss weight is exactly the one the symmetric arm already uses, which is the sharpest form of the claim that this is not a reweighted cross-entropy.

Three observations make the status of Eq. 18 clear.

It is a bound for the process we actually sample from. The expectation over z_t is taken under the position-dependent kernel, and the weight $w_i(t)$ is the one implied by that position's own schedule. Both sides refer to the same forward process, so on the interior the inequality holds in the same sense as Eq. 17.

Endpoint caveats. Two things happen at the ends of the schedule and we state both. A coordinate with $g_i > 1$ saturates at $t = 1/g_i$; beyond that point $\mathrm{d}m_i/\mathrm{d}t = 0$, the coordinate accrues no further hazard, and its time integral covers $[0, 1/g_i]$ rather than the full interval. Coordinates with $g_i < 1$ satisfy $m_i(1-\epsilon) = (1-\epsilon) g_i < 1$, so they do not reach the absorbing state by the end of the schedule so a terminal KL term would be required for a fully exact bound. We therefore describe the implemented objective as a position-dependent denoising objective carrying the baseline hazard weight of Eq. 19 on the interior, with a truncated time integral for saturating positions and an unaccounted terminal term for non-absorbing ones. We do not claim endpoint exactness.

It is not a reweighted cross-entropy. A reweighting would keep q fixed and multiply the per-position loss by g_i. Here g_i enters twice—once through the masking probability $1 - \alpha_{tg_i}$ that determines *which* positions are supervised, and once through the derivative $\alpha'(tg_i)$ in the weight—and the two are tied together by the schedule. Writing $u = tg_i$, the weight is exactly the unmodified weight evaluated at u, scaled by $\mathrm{d}u/\mathrm{d}t = g_i$, which is what a change of time variable produces.

It reduces to the baseline. At $\alpha = 0$ we have $g_i \equiv 1$, $t_i = t$, $w_i(t) = -\alpha'_t/(1 - \alpha_t)$ for every i, and Eq. 18 is Eq. 17. The α=0 arm is therefore the unmodified objective, not an approximation of it, which is what makes the comparison in the main text a controlled one.

The same argument applies to the hybrid GIDD process, whose ELBO is likewise a sum of per-position terms with weights determined by that position's diffusion time; substituting t_i for t per position leaves the bound valid. The position-specific effective time t_i determines the forward corruption and the corresponding per-position ELBO weight. The denoiser itself receives the same global time t as in the baseline; the positional profile is not exposed as an additional conditioning signal. The resulting objective remains a valid variational bound because the forward process and its per-position weights are matched to the position-dependent schedules; validity does not require the network to observe t_i explicitly. This is also the stricter control: the model must infer the positional bias from the supervision distribution rather than being told g_i.

D.3 Normalisation, clipping and endpoints

Two implementation choices matter and are not cosmetic. For MDLM we renormalise the raw profile to mean one, $Z = L^{-1} \sum_j ((j + o)/(L - 1 + o))^\alpha$, which preserves the *nominal* pre-clipping corruption mass: $L^{-1} \sum_i g_i = 1$, so $\sum_i tg_i = tL$ before clipping. Clipping perturbs the realised mask count slightly, so the arms are nominally rather than exactly budget-matched. For GIDD we instead normalise so that $\max_i g_i = 1$. The reason is the GIDD weight, which behaves as $1/((1-t)t)$: mean-renormalisation scales the profile up to $\max_i g_i \approx 1.996$, drives t_i above $1 - \epsilon$ for late positions and gives those tokens weights of order 10^4. Empirically this placed 15.3% of tokens at the clamp (against 0.01% at α=0) and made the first α=1 runs diverge near 200k steps. With $\max_i g_i = 1$ we have $t_i \leq t$ by construction and the upper clamp is structurally unreachable (measured exposure 0.000%); the cost is that the expected budget is then strictly *lower* for α=1 than for α=0, so the GIDD results are obtained with less corruption, not more. We use $o = 1$ rather than $o = 0$ so that $g_1 = 1/L$ rather than 0, which reduces lower-clamp exposure from 0.263% to 0.068% and leaves the first position with roughly $3\times$ more training signal.

E Why Positional Supervision?

The intervention has a simple reading in terms of where reducible uncertainty lives. Decompose the predictive entropy at position t as

$$H(X_t \mid X_{<t}) \;=\; H(X_t) \;-\; I(X_t; X_{<t}), \tag{20}$$

where the mutual-information term is exactly the part of the loss that can be reduced by exploiting the prefix. Early in a sequence little context has been observed, so most of the uncertainty is irreducible from the prefix and a denoising target placed there is solved mainly by learning marginal statistics. Later in the sequence substantially more conditional information is available, and a target placed there can be solved by using the prefix. Under a fixed corruption budget, uniform masking spends equal supervision on these qualitatively different prediction problems. Position-dependent corruption reallocates training events towards positions where contextual information exists, increasing the fraction of supervision for which reducing the loss requires using the prefix.

We do not claim that early tokens are unimportant, or that a power law is the optimal profile. It is a minimal one-parameter intervention that reallocates a fixed budget towards conditionally informative positions, chosen because it adds no learned components that would confound the comparison.

F Models, Data and Training

Table 4 lists every model used in the paper. The 170M MDLMs use feedforward dimension 3072, conditional embedding dimension 128, dropout 0.0 and 169.7M total parameters (92.1M non-embedding);

Table 4: **Models used in this paper.** The matched toy suite differs only in the objective; the 170M MDLM and the GIDD small models share the same backbone and parameter count; LLaDA-8B is a public checkpoint we continue-train for 500M tokens per arm. "Params" is total trainable parameters, with non-embedding parameters in brackets where they differ materially. "Ctx" is the maximum sequence length seen in training.

Model	Objective	Params	Layers	Width	Heads	Ctx	Steps	Optimiser
Matched toy suite (5 seeds each; vocabulary 9 incl. `[M]`, learned positional embeddings)								
AR-small	AR	42.5k [33.7k]	2	64	4	32	15k	Adam, lr 3×10^{-4}
MLM-small	MLM (15%)	42.5k [33.7k]	2	64	4	32	15k	Adam, lr 3×10^{-4}
MDLM-small	MDLM (toy)	42.5k [33.7k]	2	64	4	32	15k	Adam, lr 3×10^{-4}
Symmetric (α=0)	MDLM (faithful)	42.5k [33.7k]	2	64	4	32	15k	Adam, lr 3×10^{-4}
Causal bias (α=1)	MDLM + pos. bias	42.5k [33.7k]	2	64	4	32	15k	Adam, lr 3×10^{-4}
TMS control (Appendix M; vocabulary 2^5+1, attention-only, no MLP blocks)								
TMS suite (5 objectives)	AR / MLM / MDLM / α=0, 1	56.6k	2	64	4	16	8k	Adam, lr 3×10^{-4}
OpenWebText (Gokaslan & Cohen, 2019), GPT-2 tokenizer (Radford et al., 2019), sequences packed to 512 tokens; Adam (Kingma & Ba, 2015)								
MDLM α=0	MDLM	169.7M [92.1M]	12	768	12	512	1M	Adam
MDLM α=1	MDLM + pos. bias	169.7M [92.1M]	12	768	12	512	1M	Adam
GIDD p_u=0.1, α=0	GIDD	169.7M [92.1M]	12	768	12	512	200k	Adam, lr 5×10^{-4}
GIDD p_u=0.1, α=1	GIDD + pos. bias	169.7M [92.1M]	12	768	12	512	200k	Adam, lr 5×10^{-4}
GIDD p_u=0.2, α=0	GIDD	169.7M [92.1M]	12	768	12	512	200k	Adam, lr 5×10^{-4}
GIDD p_u=0.2, α=1	GIDD + pos. bias	169.7M [92.1M]	12	768	12	512	200k	Adam, lr 5×10^{-4}
Public checkpoint, continued training (500M tokens per arm)								
LLaDA-8B base	masked diffusion	8.0B	32	4096	32	4096	—	—
LLaDA SYM	symmetric	8.0B	32	4096	32	4096	500M tok	Adam
LLaDA FORWARD	+ pos. bias	8.0B	32	4096	32	4096	500M tok	Adam

sequences shorter than 512 tokens are filtered and longer ones truncated, with no padding. The GIDD models use the upstream `small` configuration (hidden 768, conditional dimension 128, 12 blocks, 12 heads, RoPE, softmax attention) with batch size 64 across 8 devices, weight decay 0.02 and 10k warmup steps. All four GIDD arms are trained for 200k steps, and every GIDD number reported in this paper—ICL tasks, native OverrideKV generation, common-objective perplexity and the macro diagnostics—is measured at that matched 200k checkpoint, so no comparison benefits from one arm having seen more optimisation than another. Training runs were continued past 200k for monitoring only; the two α=0 arms were resumed once after a non-finite-gradient abort. The matched toy suite trains on 32-token sequences with batch size 64 for 15k steps, five seeds per objective.

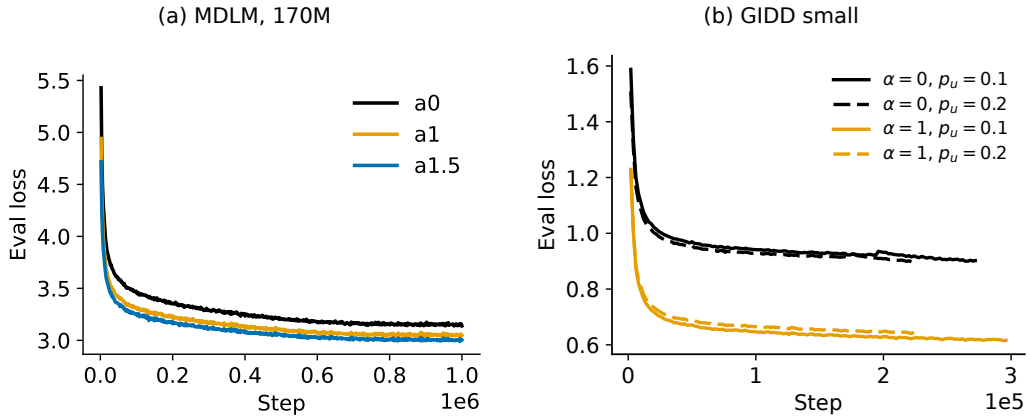


Figure 8: **Training convergence.** All MDLM (a, where `a0`/`a1`/`a1.5` denote $\alpha \in \{0, 1, 1.5\}$) and GIDD (b) arms converge stably. Raw vertical offsets across different α values are objective-specific, because α changes the corruption distribution the ELBO is taken over, and should not be read as like-for-like language-model quality; common-objective and native-generation controls are reported separately.

G SUPERVISION-BUDGET CONTROLS

The obvious alternative explanation is that α=1 simply receives more or easier supervision. The following controls address it.

- **Nominal versus realised budget.** For MDLM and LLaDA, $L^{-1}\sum_i g_i = 1$, so before clipping $\sum_i tg_i = tL$ exactly. Clipping makes the realised mask count smaller: in the large-L limit, with g uniform on $[0, 2]$ and $t \sim \mathcal{U}(0, 1)$, the mean realised masking rate at α=1 is $\mathbb{E}[\min(tg, 1)] = 5/8 - \frac{1}{4}\log 2 \approx 0.452$, against 0.5 for the symmetric arm—about 10% less corruption, not more. We therefore call the arms nominally budget-matched rather than realised-corruption matched, and note that the discrepancy runs against the intervention. The supervision-matched and target-weighting controls below test the corresponding alternative explanations directly.
- **Max-normalised variant (GIDD).** With $\max_i g_i = 1$ the expected budget is strictly *lower* for α=1 than for α=0, so the GIDD gains in Section 6 are obtained with less, not more, corruption.
- **Clamp exposure.** Measured upper-clamp exposure is 0.000% under the max-normalised profile and lower-clamp exposure 0.068% with offset o=1 (versus 0.263% at o=0).
- **Supervision-matched checkpoints.** Beyond the 200k checkpoints used throughout, a second matched pair at ≈219–220k gives the same ordering, so equalising total supervised-token count does not explain the effect.
- **Numerical controls.** A bf16 round-trip of the α=0 arm and alternative t-sampling modes (flat and MLE variants) reproduce the sweep results within measurement noise.
- α **sweep and convergence.** All arms converge stably for $\alpha \in \{0, 1, 1.5\}$ (Figure 8); the intervention does not slow convergence. Raw held-out losses are not comparable across α, since α changes the corruption distribution, so we rely on the common-objective and generation controls instead.

G.1 TARGET ALLOCATION VERSUS CONTEXT ALLOCATION

The positional intervention changes two things at once, and this control separates them:

$$\underbrace{\text{FORWARD}}_{\text{target allocation + context allocation}} \quad \text{versus} \quad \underbrace{\text{loss-weighted SYM}}_{\text{target allocation only}}.$$

Construction. All three arms draw one global time $t \sim \mathcal{U}(10^{-4}, 1 - 10^{-4})$ per sequence and use the profile $g_i \propto ((i + 1)/L)^{\alpha}$ with α=1 and L=4096, normalised so that $\mathrm{mean}(g) = 1$ exactly; g then ranges from 0.00049 at the first position to 1.9995 at the last.

Arm	Corrupted input	Position-dependent loss weight
SYM	$m_i(t) = t$	$1/t$
FORWARD	$m_i(t) = \mathrm{clip}(t\, g_i)$	$1/t$
Loss-weighted SYM	$m_i(t) = t$	g_i/t

The loss-weighted arm draws the same global time and the same per-position uniforms as SYM, so its mask patterns are bit-identical to SYM's; only the gradient weight differs, ramping from 0.10 to 1.90 across position deciles.

Why the two are matched. Away from endpoint clipping the two interventions have exactly matched expected loss mass at every position and time: conditioned on t, FORWARD contributes $tg_i \cdot (1/t) = g_i$ at position i, and the loss-weighted arm contributes $t \cdot (g_i/t) = g_i$. Clipping creates a residual mismatch for late positions at large t, where FORWARD contributes $m_i(t)/t < g_i$; a clipping-aware matched weight would be $w_i^{\mathrm{match}}(t) = m_i^{\mathrm{FWD}}(t)/t^2$. We therefore describe the control as matched on the interior rather than mathematically exact, and note that the residual mismatch reduces FORWARD's target mass rather than increasing it.

Margins across interference, q=2048**.**

K	SYM R	loss-weighted R	FORWARD R	lossw − SYM	FWD − SYM
1	12.083	12.025	12.105	-0.058	$+0.023$
2	10.208	10.176	10.350	-0.032	$+0.142$
4	8.729	8.633	8.933	-0.096	$+0.204$
8	7.077	6.913	7.411	-0.164	$+0.334$
16	5.036	4.992	5.325	-0.044	$+0.289$

Margins across query position. Each cell is loss-weighted − SYM / FORWARD − SYM.

q ($g(q)$)	K=1	K=8	K=16
512 (0.25)	-0.077 / $+0.014$	-0.064 / -0.163	$+0.120$ / -0.403
1024 (0.50)	-0.063 / -0.007	-0.104 / -0.049	-0.032 / -0.209
2048 (1.00)	-0.058 / $+0.023$	-0.164 / $+0.334$	-0.044 / $+0.289$
3072 (1.50)	-0.055 / $+0.073$	-0.176 / $+0.603$	$+0.039$ / $+0.646$

Slopes against $g(q)$.

K	Loss-weighted	FORWARD
1	$+0.016$	$+0.052$
8	-0.095	$+0.633$
16	-0.042	$+0.857$

The control reproduces neither signature: it does not grow with K at fixed position and it has no positional slope, while FORWARD has both. If anything the loss-weighted arm drifts slightly negative, most visibly at K=8 where it runs -0.06 to -0.18 across positions. Since the two arms match on expected supervision per position (away from clipping) and differ only in which tokens are visible while each target is predicted, the effect lives in the second channel: target allocation is not the cause, context allocation is.

Caveat. The loss-weighted columns above are unpaired condition means read from evaluation logs, unlike the FORWARD − SYM contrast of Section 7, which is computed per example. Paired bootstrap intervals for the loss-weighted arm, together with its self-consistency and generation numbers, are not yet available; we therefore report its separation from FORWARD as a qualitative discriminator and attach no significance statement to it.

H ONE-STEP EVALUATION: ROBUSTNESS

Because one-step conditional evaluation is our primary diagnostic (Section 2.2), we report two probes of its validity.

H.1 NOISE-TIME CONDITIONING

This is a confirmation rather than a discovery. Sahoo et al. (2024) show that for absorbing-state diffusion the reverse posterior depends on the corrupted sequence only through which positions are masked: given the mask pattern, the denoiser's prediction is independent of the diffusion time, which is precisely what licenses their time-independent (SUBS) parameterisation. Our probe checks that the trained models inherit this property, because our evaluation protocol relies on it: if the supplied t mattered, one-step scoring would be an arbitrary choice among many.

The probe presents clean demonstrations and query, replaces only the answer token with `[MASK]`, fixes the mask pattern, and sweeps the supplied noise level t. Predictions are invariant to t to four decimal places:

Model	Task	t=0.0	t=0.05	t=0.15
MDLM (toy)	ICL recall	0.1211	0.1211	0.1211
MDLM (toy)	MQAR	0.2973	0.2973	0.2973
Symmetric (α=0)	ICL recall	0.1281	0.1281	0.1281
Symmetric (α=0)	MQAR	0.2953	0.2953	0.2953
Causal bias (α=1)	ICL recall	0.1246	0.1246	0.1246
Causal bias (α=1)	MQAR	0.3043	0.3043	0.3043

At scale the same holds: sweeping $t_{\text{eff}} \in \{0.01, 0.05, 0.1, 0.2\}$ at a fixed mask pattern has no measurable effect on the 170M MDLM arms. Once the mask pattern is fixed, the supplied time does not modulate the one-step prediction. Two consequences follow. First, MDLM masked-token scoring, MLM masked-token scoring and AR next-token scoring can be compared on the same footing, because none of them has a free inference-time parameter to tune. Second, none of the results in this paper can be explained by a favourable choice of t at evaluation, for either arm.

H.2 Full macro context-use results

Table 5 lists the macro context-use score for every model evaluated, including the three arms omitted from Figure 2(a) for space. The score is the late-minus-early NLL difference on matched tokens, in nats; negative values indicate that the later occurrence is easier. The stored evaluation artefact records this difference rather than the two absolute NLLs separately, so we report the difference alone.

Table 5: **Macro context use, all models.** Late-minus-early NLL (nats) on matched OpenWebText tokens at positions 500 versus 50. Negative = context compression.

Model	Score	Model	Score
AR small	-0.524	MDLM α=0	$+0.591$
AR base	-0.497	MDLM α=1	$+0.353$
LLaDA-8B	-0.057	GIDD p_u=0	$+0.647$
MLM base	$+0.077$	GIDD p_u=0.1	$+0.408$
MLM small	$+0.171$	GIDD p_u=0.2	$+0.181$
Dream-7B	$+0.408$		

Two entries are worth noting. The causal-bias arm (MDLM α=1, $+0.353$) sits between symmetric MDLM and the AR models, i.e. the intervention moves the macro diagnostic in the predicted direction without closing the gap to AR. LLaDA-8B is close to zero, which is consistent with the Section 7 finding that its base checkpoint neither compresses context in the AR sense nor shows the strong anti-compression of the small symmetric diffusion models.

Figure 2(a) summarises the macro diagnostic; the underlying measurements are the per-model early and late NLL on matched tokens (positions 50 and 500) over OpenWebText, with the score defined as their difference. For the diffusion arms the same protocol is applied at the supplied noise level with the answer position masked, and the t-sweep of Appendix H.1 confirms the values are insensitive to that choice. Per-model early/late values, the full position-wise loss curves from which the two-point score is taken, and the exact corpora and position grids are collected here.

H.3 Suffix-mask distractors

Piskorz et al. (2025a) show that appended `[MASK]` tokens can distract masked diffusion models, and argue that this compromises context comprehension. We replicate the probe, find the mechanism real but the sign task-dependent, and report a result in the opposite direction on exactly the tasks this paper is about.

Toy suite. We append $S \in \{0, 4, 8, 16, 32\}$ mask tokens after the clean sequence and re-measure answer accuracy, over 5 seeds.

Task	Model	S=0	S=4	S=8	S=16	S=32	Δ (sd over seeds)
Induction	AR	0.172	0.172	0.172	0.172	0.172	+0.000 (0.000)
	MLM	0.103	0.102	0.098	0.098	0.098	−0.004 (0.009)
	MDLM (toy)	0.033	0.031	0.030	0.027	0.025	−0.008 (0.008)
	Symmetric (α=0)	0.071	0.069	0.070	0.068	0.063	−0.008 (0.013)
	Causal bias (α=1)	0.089	0.088	0.091	0.093	0.089	+0.000 (0.017)
MQAR	AR	0.312	0.312	0.312	0.312	0.312	+0.000 (0.000)
	MLM	0.301	0.306	0.304	0.304	0.279	−0.023 (0.053)
	MDLM (toy)	0.309	0.279	0.246	0.204	0.140	−0.169 (0.158)
	Symmetric (α=0)	0.309	0.307	0.304	0.266	0.236	−0.073 (0.133)
	Causal bias (α=1)	0.307	0.312	0.313	0.311	0.264	−0.043 (0.084)
Directional	AR	0.140	0.140	0.140	0.140	0.140	+0.000 (0.000)
	MLM	0.072	0.066	0.067	0.069	0.062	−0.009 (0.016)
	MDLM (toy)	0.036	0.032	0.031	0.030	0.030	−0.007 (0.010)
	Symmetric (α=0)	0.041	0.039	0.043	0.043	0.042	+0.001 (0.009)
	Causal bias (α=1)	0.054	0.048	0.048	0.042	0.041	−0.013 (0.020)
ICL recall	AR	0.416	0.416	0.416	0.416	0.416	+0.000 (0.000)
	MLM	0.116	0.116	0.116	0.116	0.116	+0.000 (0.000)
	MDLM (toy)	0.129	0.129	0.128	0.128	0.128	−0.001 (0.001)
	Symmetric (α=0)	0.128	0.123	0.123	0.123	0.123	−0.005 (0.011)
	Causal bias (α=1)	0.124	0.124	0.124	0.124	0.124	+0.000 (0.000)

Δ is accuracy at S=32 minus accuracy at S=0, with the standard deviation of the per-seed difference in brackets (n=5, so the standard error is that number divided by $\sqrt{5}$).

Only one cell is above noise. Read against the seed spread, the table is mostly null. The single effect that clears its own noise is the simplified toy MDLM on MQAR: -0.169 with a per-seed standard deviation of 0.158, i.e. about 2.4 standard errors, and even that is driven by a subset of seeds—at S=32 the mean accuracy is 0.140 with a standard deviation of 0.141, so some seeds collapse to near zero while others barely move. The faithful implementations on the same task (-0.073 ± 0.133 and -0.043 ± 0.084) are within one and a half standard errors of zero and we do not claim them as effects; the same is true of every entry on induction, directional prediction and ICL recall, where no $|\Delta|$ exceeds 0.013. AR is exactly invariant, as it must be: appended masks are not in its conditioning set. We therefore report the catastrophic-collapse phenomenon as real but narrow—one task, one simplified implementation, and not stable across seeds.

At scale, suffix masks help, and they help most on the ICL tasks. At 170M parameters on the synthetic ICL battery the sign reverses and the magnitudes are large.

Model	Task	S=0	Δ at S=32	Δ at S=128
Symmetric (α=0)	OverrideKV	0.340	+0.324	+0.445
	Minimal override	0.143	+0.379	+0.402
	Random mapping change	0.138	+0.345	+0.385
	Binary FSM	0.659	+0.038	+0.039
Causal bias (α=1)	OverrideKV	0.124	+0.573	+0.714
	Minimal override	0.132	+0.329	+0.380
	Random mapping change	0.110	+0.399	+0.463
	Binary FSM	0.496	+0.205	+0.205

The pattern is systematic. The three tasks that gain most—OverrideKV, minimal override and random mapping change—are precisely the tasks whose solution requires deciding *which* of several competing in-context bindings governs the answer, i.e. the ones that need strong recency-weighted attention over the demonstrations. Binary FSM, which requires applying a rule rather than resolving a binding, gains an order of magnitude less. On OverrideKV the causal-bias arm goes from 0.124 to 0.838 with 128 appended masks. Our reading is that appended masks are not neutral padding and not simply distractors either: they are additional denoising targets that compete for attention mass, and the outcome depends on what that mass is taken *from*. When it is taken from an uninformative region, attention concentrates on the demonstrations and the query, and recency-sensitive retrieval improves; when it is taken from the evidence itself, as in toy MQAR, performance degrades.

Relation to the marginalisation barrier. These two phenomena are at different levels and should not be conflated. Mask distraction is an inference-time property of a particular model and prompt format: it can be measured only by adding masks, and it can go in either direction. The future-marginalisation barrier is a property of the training objective that is present under clean-prefix evaluation with no appended masks at all, which is the setting of every result in the main text. Our replication therefore neither contradicts Piskorz et al. (2025a) nor is explained by them.

I Synthetic Task Generation and the Full Few-Shot Battery

Few-shot episode format. For each task we generate 500 evaluation episodes with $k \in \{0, 1, 2, 4, 6, 8, 10, 12, \ldots, 48, 50\}$. Each episode uses a fresh random seed and a pool of 60 demonstrations from which the first k are selected. A prompt is a flat token sequence containing demonstrations followed by a query. Demonstrations use the same surface template within a task, but the latent rule is resampled per episode. Thus memorising a global mapping cannot solve the evaluation; the model must infer the episode-specific rule from in-context examples. The query answer token is the only prediction target.

Key–Value retrieval. Keys are sampled as `wug`*NNNN* for *NNNN*$\in [0000, 9999]$, yielding 10,000 possible key strings. Each episode samples a fresh mapping from keys to one of 8 label tokens. Demonstrations present key–label pairs. The query repeats one of the demonstrated keys and asks for its label. This is the cleanest associative-memory task: the correct answer is present in the prefix, and the challenge is retrieving the matching key–value binding rather than learning a compositional rule.

Cipher. Cipher uses a fresh random permutation per episode. Input symbols are the first 8 lowercase letters (a–h), and labels are 8 fixed output tokens. Demonstrations reveal pairs from the permutation; the query asks for the output of one held-out or repeated input symbol under the same episode permutation. This task requires latent-rule inference from examples rather than direct copying of a previous key string. Increasing K reduces uncertainty about the permutation.

Binary FSM. Binary FSM samples a two-state finite-state machine per episode. Input strings are binary sequences of length 3–12; the FSM has a random 2×2 transition table $T[s][b]$ and a random output flip parameter $\theta \in \{0, 1\}$ that maps final state to label. Demonstrations show input strings and final labels. The query provides a new binary string and asks for its label. This task tests sequential rule application: the model must infer the transition dynamics and apply them across the query string rather than retrieve a single observed pair.

OverrideKV and recency probes. OverrideKV is designed to separate retrieval from recency. In the standard OverrideKV task, keys can appear multiple times with conflicting labels, and the correct answer is the label from the most recent occurrence. This creates a conflict between frequency and recency. In the haystack variant, the queried key is paired with a stale value several times early in the prompt and then with a recent override value at a fixed intermediate position; later demonstrations use unrelated keys. In the recency-last variant, every demonstration uses the same key, with $K-1$ stale values followed by one recent value immediately before the query. Haystack therefore tests whether the model can retrieve a sparse override across distance, while recency-last tests whether the model can overcome frequency pressure from many stale demonstrations.

One-step scoring details. For AR we score the next-token logit under teacher forcing. For MLM and MDLM we replace only the answer token with `[MASK]` and score the answer logit among the candidate labels. Candidate-label restriction is used only for the synthetic classification tasks; macro-ICL NLL uses the full vocabulary. Ties are negligible and are resolved by the deterministic order of candidate labels.

Full few-shot battery. Table 6 gives the complete scored battery behind Figure 3, including the α=1.5 rerun and the two additional cipher variants that do not appear in the main text. Accuracies are means over n independently generated episodes at each K; $K_{\max}$ differs across tasks because episodes are truncated to the 512-token context.

Table 6: **Full few-shot battery (scored readout).** Accuracy against number of demonstrations for symmetric (α=0), causal-bias (α=1) and a stronger tilt (α=1.5). Each entry is a mean over n=500 independently generated episodes; $K_{\max}$ differs by task because episodes are truncated to the 512-token context.

Arm	K=0	K=4	K=12	K=24	K=$K_{\max}$
Key–Value (chance 0.125, $K_{\max} = 44$)					
α=0	0.114	0.256	0.246	0.268	0.234
α=1	0.114	0.284	0.292	0.300	0.354
α=1.5	0.106	0.258	0.240	0.284	0.242
Cipher (chance 0.125, $K_{\max} = 50$)					
α=0	0.120	0.112	0.110	0.148	0.140
α=1	0.122	0.118	0.146	0.172	0.222
α=1.5	0.116	0.134	0.112	0.114	0.138
OverrideKV (chance 0.125, $K_{\max} = 44$)					
α=0	0.110	0.096	0.120	0.116	0.136
α=1	0.106	0.090	0.144	0.136	0.144
α=1.5	0.116	0.094	0.124	0.130	0.154
Binary FSM (chance 0.500, $K_{\max} = 49$)					
α=0	0.496	0.686	0.670	0.662	0.690
α=1	0.502	0.704	0.720	0.724	0.732
α=1.5	0.490	0.682	0.638	0.666	0.700
Composed cipher (chance 0.125, $K_{\max} = 50$)					
α=0	0.118	0.136	0.108	0.120	0.144
α=1	0.136	0.114	0.110	0.104	0.156
α=1.5	0.132	0.102	0.118	0.124	0.142
Last-demo cipher (chance 0.125, $K_{\max} = 50$)					
α=0	0.128	0.152	0.128	0.134	0.140
α=1	0.148	0.144	0.138	0.134	0.142
α=1.5	0.138	0.156	0.120	0.126	0.146

Two observations. First, α=1.5 is not uniformly better than α=1 and is often worse than α=0: the effect is not monotone in the tilt strength, which is why the main text reports α=1 and treats the power law as a probe rather than a tuned schedule. Second, the two additional cipher variants—a two-step rule in which the query asks for $g(f(x))$, and a recency probe in which only the final demonstration block uses the true mapping—stay near chance for every arm, so the intervention does not produce indiscriminate gains on anything labelled "in-context".

LLaDA-8B key–value accuracy. Table 7 gives the accuracies plotted in Figure 5(a). The corresponding within-example margins, the query-position sweep and the loss-weighted control are in Appendix G.1.

Table 7: **LLaDA-8B key–value retrieval.** Accuracy against the number of competing bindings K, for the public base checkpoint and the two 500M-token continued-pretraining arms.

Arm	K=1	K=2	K=4	K=8	K=16
Base	0.890	0.842	0.754	0.574	0.354
SYM (α=0)	0.886	0.848	0.790	0.608	0.422
FORWARD (α=1)	0.882	0.854	0.798	0.662	0.458

Scored and generated readouts can disagree. We flag this explicitly rather than smoothing it over. The scored (one-step) and native-generation readouts are genuinely different measurements, and they do not always move together: in MDLM the scored readout can flatten on a task where generative OverrideKV improves, and in GIDD the mirror image occurs. This is the empirical counterpart of the distinction drawn in Section 2.2: one measures the learned conditional, the other measures what the denoising chain does with it. Where we claim an effect at both levels we show both.

J Position-Biased Arms on Associative Recall

Figure 9 repeats the main-text associative-recall panel with the two position-biased diffusion arms included. Both sit on chance for every context length, which is why they are omitted from Figure 2: at this scale the corruption profile does not by itself produce the recall algorithm, and the intervention's effect appears in the few-shot and generative settings of Sections 5–6 instead. This is the sense in which the intervention is a directional test of the theory rather than a reconstruction of autoregression.

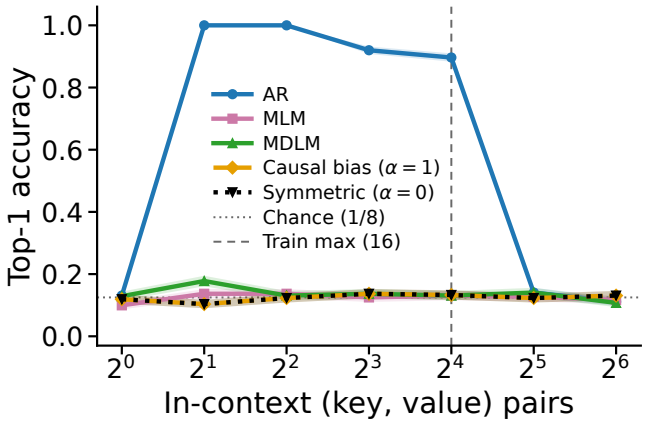


Figure 9: **Associative-recall accuracy, all five objectives.** AR learns the task; MLM, MDLM and both position-biased arms remain at chance.

Figure 10 gives the attention counterpart of the same control. AR concentrates attention in a narrow band immediately before the query and is exactly zero for positive offsets, while MLM and MDLM are broad and near-symmetric about it. We report this as a qualitative correlate of the behavioural split and do not claim that the attention pattern causes it, which is why it is not a main-text panel.

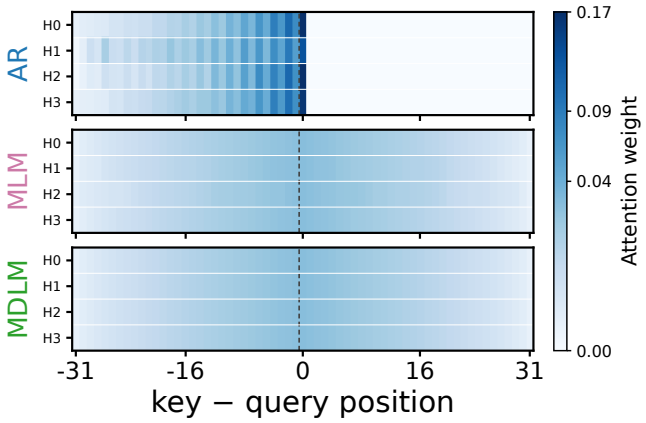


Figure 10: **Attention by relative position, layer 1, four heads per objective, shared colour scale.** AR is concentrated just before the query; MLM and MDLM are diffuse.

K Conditional Coherence Under Marginalisation

Theorem 2.1 concerns one specific operation: recovering a prefix conditional by marginalising a variable the model was trained to condition on. A model with a coherent joint should be indifferent to the route, so for any intermediate variable y the law of total probability requires

$$p_\theta(x_t \mid h) \;=\; \sum_y p_\theta(x_t \mid h, y)\, p_\theta(y \mid h), \tag{21}$$

and we measure the discrepancy between the two sides directly,

$$\Delta_{\mathrm{SC}} \;=\; \mathrm{TV}\Big[\, p_\theta(x_t \mid h),\; \textstyle\sum_y p_\theta(x_t \mid h, y)\, p_\theta(y \mid h) \,\Big], \tag{22}$$

lower being more self-consistent. Concretely: the left-hand side masks only x_t and asks the model for its prediction given everything else in the prefix. The right-hand side leaves a second position y masked as well, reads off the model's own distribution over y, and averages the model's prediction for x_t over that distribution. The question is therefore whether predicting x_t directly agrees with predicting x_t after integrating out an unknown neighbouring token, using the model's own beliefs about that token throughout. The sum is truncated to the top-M values of y, and we report the top-M sweep so that the truncation is visible.

Why the objectives should differ here is the same reason they differ elsewhere in this paper. A masked language model is trained on a collection of conditionals $p_\theta(x_i \mid x_{\setminus S})$, one per masking pattern, and nothing in the objective requires those conditionals to be the conditionals of any single joint distribution; they are fitted independently and may be mutually incompatible. Masked diffusion is better placed, because its objective is a bound on a joint likelihood, but the model still only ever emits local conditionals under sampled corruption patterns, and coherence between two different patterns is not itself supervised. Causal factorisation is the limiting case in which the chain rule ties the conditionals together by construction. The probe asks how far along that spectrum each trained model actually sits.

This is not another ICL score and not a validation of Theorem 2.1. It is an independent, model-internal test of whether conditionals elicited under different observation patterns compose coherently under the marginalisation operation the theory identifies. It needs no labels, no task and no reference model: the two sides are both the model's own outputs, so a gap is internal incoherence rather than a failure to match some external target.

MDLM. The discrepancy falls from 0.683 for MLM to 0.354 for the symmetric arm and 0.204 for the causal-bias arm, which is the ordering the argument above predicts: independently fitted masked conditionals are the least coherent, a joint-likelihood bound is markedly better, and tilting that bound towards prefix-conditioned supervision is better again. The ordering is stable under the truncation: as M rises to 256 the gap remains approximately 0.32 versus 0.19.

GIDD: a superseded result and its correction. An earlier fill-tail protocol, in which the tail is filled with a blank token rather than left masked, gave a far more dramatic improvement ($0.187 \to 0.064$). Diagnostics showed this to be an artefact: under that protocol the causal-bias arm emits the blank token as argmax in 100 of 100 cases and has extremely low direct entropy, so the two sides of the identity are compared on a degenerate distribution. The corrected masked-tail protocol removes the artefact and is essentially flat: $0.259 \to 0.271$ at p_u=0.1 and $0.255 \to 0.268$ at p_u=0.2, with a fixed-t variant reproducing these within 0.005. We report the corrected numbers and treat the fill-tail result as superseded.

Interpretation. Protocols differ between families and the values should not be compared across rows; the comparison of interest is within a family. On that basis the intervention improves conditional coherence in MDLM, whereas GIDD stays substantially inconsistent in both arms ($\Delta_{\mathrm{SC}} \approx 0.26$–$0.27$) and the intervention does not move it. We therefore treat self-consistency as model-specific supporting evidence rather than a universal consequence of causal asymmetry, and we note that the MDLM protocol is the fill-tail variant, so its comparison with the corrected GIDD numbers is indicative only.

L Native-Generation Evaluation

Protocol. *MDLM.* We evaluate four corpora (OpenWebText, WikiText-103, TinyStories and Sherlock), five refinement budgets $S \in \{1, 4, 16, 64, 256\}$ and four temperatures $\tau \in \{0.7, 0.8, 0.9, 1.0\}$, comparing each generated continuation with the held-out continuation. *GIDD.* A smaller reference-temperature protocol: the same four corpora, $S \in \{4, 16, 64\}$ and $\tau = 1.0$. The metrics are unigram precision, trigram precision and longest matching span; for GIDD we additionally score a held-out judge NLL. Both arms see identical prefixes, so comparisons are paired; significance is a paired sign test, Holm-corrected across the full family of comparisons.

MDLM. Across the resulting 240 metric cells (4 corpora $\times$ 5 budgets $\times$ 4 temperatures $\times$ 3 metrics), every comparison that remains significant after Holm correction favours α=1; none significantly favours α=0. The corrected counts are 12/0 on longest matching span, 12/0 on unigram precision and 3/0 on trigram precision.

The advantage requires iterative refinement. At S=1—a single fully parallel denoising step, i.e. no refinement—there is no systematic advantage for α=1. The advantage appears once refinement is introduced; its magnitude is not monotone in the number of steps. This is the mechanistic control for the claim in Section 6.2: the intervention does not change the marginal token predictions, it changes how they are assembled into a sequence.

Repetition control. Because repeated phrases can inflate n-gram overlap, we repeat the analysis after filtering repetitive generations and matching the comparison pools across arms. The advantage survives: on WikiText-103 the corrected trigram-precision effect rises from $+33\%$ before filtering to $+62.6\%$ after it, so it is not an artefact of repetition.

GIDD. The picture is more mixed and we report it as such. At p_u=0.1, judge NLL gives 9 Holm-corrected wins for α=1 and 0 for α=0 (11/12 directional), and longest shared span gives 6/0 (7/12 directional). Trigram precision gives 1/0 corrected wins, while unigram precision gives 1/2; generation-quality effects in GIDD are therefore metric-dependent.

L.1 Contextual OverrideKV under native generation

This is the protocol behind Figure 4, which Section 6.1 reports.

Episodes and the lift metric. Each episode is an OverrideKV prompt: demonstrations followed by a query whose key has appeared earlier with a conflicting value. Every episode is generated in a *real* condition and a task-matched *control* condition that keeps the query but breaks the demonstrations, and we report

$$\text{lift} \;=\; \text{acc}(\text{real}) - \text{acc}(\text{control}). \tag{23}$$

Differencing against the control removes any part of the score that comes from the model simply favouring a particular answer token. Both conditions use forced choice over the probe's own option tokens, because unconstrained generation places most of its mass outside the candidate set.

Decoding. Generation runs the model's own reverse diffusion over the whole window from an all-`[MASK]` state for 64 steps, with a greedy final step. Two axes are varied. The *prefix mode* is `free` (no correction), `tf` (a prefix position is corrected to its true token once the model unmasks it) or `clean` (the prefix is clamped to the true tokens at every step). The *query mode* is `remask` (the answer position is reset to `[MASK]` after every non-final step, so a single answer token benefits from all 64 passes) or `free`. For MDLM the reverse schedule is additionally either `uniform` or position-`biased`; the MDLM runs use `remask` throughout.

MDLM. n=500 episodes per K and condition, $K \in \{0, 1, 2, 3, 4, 5, 6, 8, 16, 32\}$. *Both* arms are decoded with `uniform_clean`, so the decoding procedure is identical across arms and only training differs. Table 8 gives every variant.

Two readings. First, neither reverse schedule is a neutral choice: `uniform` is the matched reverse process for the α=0 forward process, `biased` for α=1, so decoding both arms with one schedule runs one of them off-spec while matching each arm lets the decoder co-vary with training. We therefore report both. The long-context conclusion is stable across them—the α=1 minus α=0 gap is $+0.134$ (`uniform_clean`) against $+0.110$ (`biased_clean`) at K=16, and $+0.178$ against $+0.192$ at K=32—whereas the mid-range is not: at K=4 the gap is $+0.048$ under `uniform_clean` and -0.014 under `biased_clean`. This is why the claim in Section 6.1 is restricted to large context sizes. The `free` variants are identically zero for both arms: without prefix clamping the model's mass leaves the candidate set entirely, which is why the comparison is made under clamping rather than free generation.

Table 8: **MDLM OverrideKV generation lift** (Eq. 23), all six decode variants, n=500 episodes per cell. The main text reports `uniform_clean`.

Decode variant	Arm	K=1	K=2	K=4	K=8	K=16	K=32
`uniform_clean`	α=0	0.348	0.166	0.390	0.512	0.462	0.332
`uniform_clean`	α=1	0.440	0.146	0.438	0.574	0.596	0.510
`biased_clean`	α=0	0.396	0.212	0.450	0.538	0.458	0.330
`biased_clean`	α=1	0.402	0.176	0.436	0.550	0.568	0.522
`uniform_tf`	α=0	0.214	0.110	0.238	0.272	0.284	0.220
`uniform_tf`	α=1	0.232	0.048	0.152	0.232	0.314	0.308
`biased_tf`	α=0	0.232	0.096	0.252	0.322	0.282	0.202
`biased_tf`	α=1	0.238	0.060	0.170	0.238	0.320	0.374
`uniform_free`	α=0	0.000	0.000	0.000	0.000	0.000	0.000
`uniform_free`	α=1	0.000	0.000	0.000	0.000	0.000	0.000
`biased_free`	α=0	0.000	0.000	0.000	0.000	0.000	0.000
`biased_free`	α=1	0.000	0.000	0.000	0.000	0.000	0.000

GIDD. n=200 episodes per cell at the 200k checkpoint, 64 steps. Table 9 gives both uniform-noise levels under every prefix and query mode.

Table 9: **GIDD OverrideKV generation lift** (Eq. 23), n=200 episodes per cell. The main text reports `remask/clean` at p_u=0.1.

Query/prefix mode	Arm	K=1	K=2	K=4	K=8	K=16	K=32
`remask/clean`	p_u=0.1, α=0	0.315	0.155	0.235	0.055	0.015	0.005
`remask/clean`	p_u=0.1, α=1	0.300	0.180	0.610	0.685	0.415	0.425
`remask/clean`	p_u=0.2, α=0	0.500	0.330	0.715	0.680	0.435	0.180
`remask/clean`	p_u=0.2, α=1	0.340	0.110	0.605	0.520	0.495	0.290
`remask/tf`	p_u=0.1, α=0	0.215	0.070	0.110	0.060	0.020	0.030
`remask/tf`	p_u=0.1, α=1	0.220	0.045	0.270	0.325	0.235	0.245
`remask/tf`	p_u=0.2, α=0	0.350	0.155	0.450	0.495	0.350	0.150
`remask/tf`	p_u=0.2, α=1	0.180	0.050	0.285	0.310	0.305	0.200
`free/clean`	p_u=0.1, α=0	0.245	0.100	0.095	0.045	0.015	0.015
`free/clean`	p_u=0.1, α=1	0.250	0.145	0.565	0.600	0.350	0.445
`free/clean`	p_u=0.2, α=0	0.330	0.185	0.405	0.275	0.110	0.025
`free/clean`	p_u=0.2, α=1	0.115	0.060	0.460	0.355	0.200	0.080
`free/tf`	p_u=0.1, α=0	0.080	0.025	0.060	0.030	0.015	0.005
`free/tf`	p_u=0.1, α=1	0.140	0.020	0.190	0.200	0.155	0.265
`free/tf`	p_u=0.2, α=0	0.135	0.075	0.220	0.140	0.100	0.015
`free/tf`	p_u=0.2, α=1	0.030	0.035	0.155	0.170	0.125	0.075
`remask/free`	p_u=0.1, α=0	0.000	0.000	0.000	0.000	0.000	0.000
`remask/free`	p_u=0.1, α=1	0.000	0.000	0.000	0.000	0.000	0.000
`remask/free`	p_u=0.2, α=0	0.000	0.000	0.000	0.000	0.000	0.000
`remask/free`	p_u=0.2, α=1	0.000	0.000	0.000	0.000	0.000	0.000

The p_u=0.1 result in the main text is stable across prefix and query modes: at every combination except `free`, the causal-bias arm holds a substantial lift at long context while the symmetric arm decays towards zero. The p_u=0.2 arms do *not* reproduce it. There the symmetric arm is the stronger of the two at most K under `remask/clean` (0.715 against 0.605 at K=4, 0.680 against 0.520 at K=8), with the causal-bias arm ahead only at K=16. We therefore restrict the generative GIDD claim to p_u=0.1 and do not present it as a property of the GIDD family as a whole. As in the MDLM runs, the `free` prefix mode collapses to zero lift for every arm.

M Representation-Geometry Control: Toy Models of Superposition

M.1 Why this control exists

A trivial alternative explanation for everything in the main text is that non-causal objectives simply learn worse representations, and that their ICL failure is a symptom of that. This appendix tests the alternative directly: under a data distribution with known latent structure, can a diffusion objective learn coherent global feature geometry at all? If it can, then the sequential failure is specific to directional structure rather than a generic representational deficit.

M.2 Data-generating process

We use a sequential, discrete version of the toy model of superposition. Each token encodes a binary feature vector $f \in \{0,1\}^F$ with $F{=}5$, mapped to one of $2^F{=}32$ semantic tokens (plus a `[MASK]` token), so that AR, MLM and diffusion objectives all have a well-defined cross-entropy on the same vocabulary. Sequences have length 16.

Features are correlated *across* the sequence, which is what gives the model a reason to learn relational structure rather than per-token compression. Each sequence first draws a topic $z \sim \mathcal{U}\{1,\dots,8\}$; each topic carries its own per-feature activation probabilities, and the tokens of a sequence are then drawn i.i.d. from that topic's feature distribution. A model that infers the topic from early tokens predicts later tokens better, which is the discrete analogue of correlated TMS. The activation probability $p \in \{0.05, 0.2, 0.5\}$ is the mean probability that a feature is active, so small p means few active features per token and larger p means *less* sparse data.

M.3 Models and training

All five objectives—AR, MLM, MDLM, and the symmetric ($\alpha{=}0$) and causal-bias ($\alpha{=}1$) faithful-MDLM arms—are trained on identical data with an attention-only 2-layer transformer (no MLP blocks), $d_{\text{model}}{=}64$, 4 heads, learned positional embeddings, batch size 64 and learning rate $3{\times}10^{-4}$. Only the loss and corruption process differ. Runs are repeated across dataset-size caps $N \in \{256, 1024, 4096, \infty\}$; the diagnostics reported here use the uncapped runs. Every diagnostic is computed at $t{=}0$, i.e. on clean inputs, so that it measures what the model has learned about token structure rather than how it handles noise; this convention is identical across objectives.

M.4 What we evaluate

Five diagnostics are computed on the trained models.

Rotation sensitivity. Replace the layer-0 residual stream with $R\,h$ for a random orthogonal R (three rotations, 64 sequences) without compensating downstream reads, and measure the mean KL between the original and perturbed output distributions. A large value means the output distribution is sensitive to the residual basis; a small value means it is not, which can also occur if the predictor is insensitive to residual perturbations in general.

Feature decodability. Fit a linear logistic probe from the layer-0 residual to each of the F binary features (256 sequences) and report mean accuracy. We report no untrained-network baseline, so this measures linear recoverability rather than learned structure.

Feature-direction geometry. Take the normalised probe directions and report their mean off-diagonal inner product. The quantity is signed and averaged, so a value near 0 is consistent with pairwise orthogonality but does not imply it: positive and negative pairs can cancel.

Hamming alignment. Pearson correlation between Hamming distance in feature space and Euclidean distance between the corresponding token embeddings.

Embedding effective rank. Participation ratio of the eigenvalue spectrum of the centred $W_E^\top W_E$, i.e. how many embedding directions are actually used.

Table 10: **TMS geometry diagnostics.** All five objectives on the same data; p is the feature activation probability, so larger p means less sparse data. Rotation KL: sensitivity of the output distribution to random residual rotations. Accuracy: linear decodability of individual features. Inner: mean off-diagonal inner product of feature directions. Hamming: embedding-distance correlation with feature-set overlap. Rank: embedding effective rank.

Objective	p	Rotation KL	Feature acc.	Inner	Hamming r	Eff. rank
AR	0.05	4.21	0.998	+0.057	0.556	7.76
	0.2	6.01	0.990	+0.051	0.659	9.74
	0.5	3.89	0.994	−0.010	0.647	12.17
MLM	0.05	5.38	0.995	+0.048	0.353	8.87
	0.2	6.66	0.984	+0.005	0.382	10.02
	0.5	4.37	0.982	+0.045	0.418	8.58
MDLM	0.05	3.96	0.979	+0.069	0.038	10.35
	0.2	0.74	0.945	+0.031	−0.062	10.18
	0.5	0.55	0.936	+0.021	−0.039	10.16
Symmetric (α=0)	0.05	3.81	0.992	+0.066	0.373	10.31
	0.2	0.92	0.952	+0.046	0.114	10.82
	0.5	0.04	0.922	−0.002	−0.009	9.82
Causal bias (α=1)	0.05	4.00	0.992	+0.066	0.401	9.88
	0.2	0.95	0.952	−0.004	0.128	7.73
	0.5	0.04	0.968	+0.005	0.234	12.51

M.5 Results

Table 10 reports linear feature decodability, sensitivity to residual rotations and embedding-geometry statistics for the trained models, and Figure 11 plots the same values against p. Here p denotes feature activation probability, so larger p means less sparse data. These diagnostics do not establish compressive superposition, identify a privileged coordinate basis, or rule out representation-based explanations of the ICL results. In particular, a small mean signed inner product need not imply pairwise orthogonality, and low rotation KL need not imply a structured rotation-invariant representation. We treat the measurements as descriptive controls only.

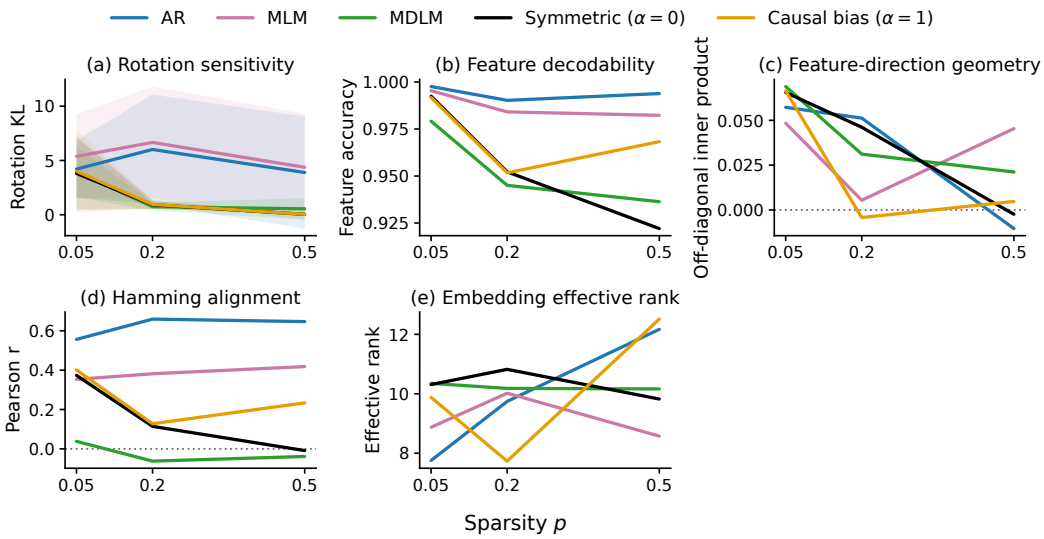


Figure 11: **TMS geometry diagnostics against activation probability.** (a) Sensitivity of the logits to random residual rotations. (b) Linear decodability of individual features. (c) Off-diagonal inner products between feature directions. (d) Pearson correlation between embedding distance and feature-set overlap. (e) Effective rank of the embedding matrix. Values are those of Table 10.

N Attention Routing and Circuit Diagnostics

N.1 What the probes separate

A model can attend to the right position and still fail, because attending is not the same as writing the attended value into the output distribution. We therefore measure two things separately: *QK routing*, whether attention goes to the informative position, and *OV amplification*, whether the value at that position is pushed into the logit of the correct token. Treating them as one "attention" number hides exactly the case that matters—correct routing with a weak value path—which is what distinguishes the large hybrid diffusion models from each other.

N.2 Matched toy suite

On the matched suite of Section 3 we compute three quantities, all on layer 1 of the 2-layer models, averaged over 5 seeds. Because the numbers are few and small, we give them directly rather than as a figure.

QK induction score. On the induction eval set, the average attention from the query position (the second occurrence of a token) to the position holding the first occurrence's continuation, per head. For AR this is the standard induction score; for MLM and MDLM the target is masked and the same question is asked. The uniform reference is $1/T = 1/32 = 0.031$.

OV copy score. For each head, form the token-space OV map $W_E W_V^{(h)} W_O^{(h)} W_U$ under a row-vector convention, take a column-wise softmax, and report the mean of its diagonal—the probability mass a head assigns to copying a token to itself. This map is $V \times V$, which is what makes the uniform reference $1/V = 0.0303$; the residual-space operator $W_V^{(h)} W_O^{(h)}$ alone would not have that reference.

Score–accuracy dissociation. The maximum induction score over heads and layers against final top-1 accuracy, which asks whether task performance is explained by the induction circuit alone.

Table 11: **QK induction score by head, layer 1** (mean ± sd over 5 seeds), with the maximum over all heads and layers and the final induction accuracy. Uniform reference $1/32 = 0.031$.

Model	H0	H1	H2	H3	max	Accuracy
AR	0.062±0.005	0.061±0.005	0.062±0.003	0.060±0.004	0.064±0.004	0.171±0.010
MLM	0.050±0.009	0.047±0.007	0.050±0.005	0.037±0.003	0.058±0.004	0.105±0.007
MDLM	0.031±0.001	0.031±0.000	0.031±0.000	0.031±0.000	0.032±0.000	0.036±0.007
Symmetric (α=0)	0.032±0.001	0.032±0.001	0.032±0.000	0.031±0.000	0.041±0.003	0.006±0.006
Causal bias (α=1)	0.032±0.002	0.032±0.001	0.033±0.002	0.032±0.002	0.052±0.002	0.016±0.009

Table 12: **OV copy score by head, layer 1** (representative seed). Uniform reference $1/V = 0.03030$. No objective departs from the baseline by more than 3×10^{-5}.

Model	H0	H1	H2	H3
AR	0.03032	0.03032	0.03032	0.03033
MLM	0.03030	0.03034	0.03030	0.03032
MDLM	0.03031	0.03029	0.03030	0.03030
Symmetric (α=0)	0.03030	0.03030	0.03030	0.03030
Causal bias (α=1)	0.03030	0.03030	0.03030	0.03030

Tables 11 and 12 give the numbers; three readings follow. First, AR is the only objective with a clearly super-uniform QK induction score on every head (0.060–0.062 against 0.031, with seed spreads of 0.003–0.005); MLM is intermediate (0.037–0.050); the three diffusion arms sit on the uniform reference to within one standard deviation on every head. Second, the maximum column shows that the diffusion arms are not completely devoid of induction structure—the causal-bias arm reaches 0.052 somewhere in the network—but the accuracy column shows this does not translate into task

performance (0.016 versus AR's 0.171). AR itself reaches only about twice uniform, so it solves induction without forming a textbook induction head, and the QK score is not a sufficient statistic for behaviour. Third, the OV copy score is a null for every objective: all twenty head-level values lie within 3×10^{-5} of $1/V$. We report it because a positive OV-copy result would have been the cleanest mechanistic story available, and it is not there; the OV evidence in this paper comes from the OverrideKV decomposition below, not from copy scores.

The fourth diagnostic, attention by relative position, is shown in Appendix J, Figure 10, and is not repeated here.

N.3 OverrideKV: haystack and recency-last

OverrideKV is the probe we use whenever the question is whether a model can decide *which* of several in-context bindings governs a prediction. The generic form presents a key that appears several times with conflicting values; the correct answer is the value from the most recent occurrence, so frequency and recency point in opposite directions and a model that simply counts co-occurrences is wrong by construction. The two variants below stress different halves of that problem, which is why we keep both rather than averaging them.

Haystack: retrieval across distance. The queried key is paired with a *stale* value several times early in the prompt, then once with the *recent* override value at a fixed intermediate position; every demonstration after that uses unrelated keys. The override is a single relevant item buried among many irrelevant ones, and the distance between it and the query grows with K while the stale evidence stays where it is. Failure here means the model cannot locate a sparse, distant, correct binding—an attention-reach failure. Because the override sits at a fixed position, prompts with fewer than $K{=}7$ demonstrations do not contain the intended structure at all: there is no haystack to search. We omit those shots rather than reporting a number that measures something else.

Recency-last: recency against frequency. Every demonstration uses the same key: $K{-}1$ stale values followed by one recent value immediately before the query. Distance is no longer the difficulty—the correct binding is adjacent to the query—but frequency pressure is, and it grows with K: at $K{=}16$ the wrong value has been seen fifteen times and the right one once. Failure here is a weighting failure rather than a reach failure, and a model that has learned "predict the value most often associated with this key" scores near zero. The structure exists from $K{=}2$ onwards.

Together the two variants separate the two ways recency-sensitive retrieval can break, and they are the natural probe for the mechanism this paper proposes: a model trained to denoise later positions from cleaner earlier context has a reason to weight recent bindings, whereas a symmetric objective has none.

Metrics. At the query position we record, averaged over all layers and heads, the *QK gap*

$$\mathrm{QK} \;=\; \frac{1}{LH}\sum_{l,h}\Big(a^{(l,h)}_{q\to\mathrm{recent}} - \overline{a^{(l,h)}_{q\to\mathrm{stale}}}\Big), \tag{24}$$

the attention at the query to the recent label token minus the mean attention to the stale label tokens, and the *OV gap*

$$\mathrm{OV} \;=\; \sum_{l,h} a^{(l,h)}_{q\to k}\big(v^{(l,h)}_{k} W^{(h)}_{O}\big)\cdot W_U[\mathrm{gold}]\Big|_{k=\mathrm{recent}} \;-\; \Big(\text{same at } k=\mathrm{stale}\Big), \tag{25}$$

the logit contribution that the attended position makes to the gold token through the value–output path. The two are deliberately not combined: QK asks where attention goes, OV asks what is written into the logit once it gets there, and a model can pass the first and fail the second.

For the DiT-style diffusion models the OV computation omits the adaLN gate and the final modulation. The stale and recent positions pass through the same gate, but that modulation is featurewise and does not commute with the projection onto $W_U[\mathrm{gold}]$, so it need not preserve the sign or the ordering of the projected difference. We therefore read the OV gap only as a within-model diagnostic, and do not compare OV values across architectures, either in absolute value or as a ratio.

N.4 RESULTS

Table 13: **OverrideKV circuit decomposition, small models.** Median over shots at or above each task's minimum meaningful K (7 for haystack, 2 for recency-last), real context, no suffix masks. QK gap is an attention difference; OV gap is a logit difference and, per Appendix N, is comparable only within an architecture family.

	Haystack			Recency-last		
Model	QK gap	OV gap	Acc.	QK gap	OV gap	Acc.
AR	$+0.0051$	14.00	0.154	$+0.0637$	80.86	0.852
MLM	$+0.0003$	0.10	0.144	$+0.0123$	0.36	0.009
Symmetric (α=0)	$+0.0010$	0.71	0.118	$+0.0055$	2.62	0.664
Causal bias (α=1)	$+0.0018$	1.28	0.147	$+0.0087$	4.35	0.230

Three readings follow from Table 13 and Figure 12.

First, within the two models whose OV path is computed in full—AR and MLM—the OV gap separates them far more than the QK gap does. On recency-last AR's OV gap is 80.9 against 0.36 for MLM, while their QK gaps differ by much less ($+0.064$ versus $+0.012$), so attention routing alone does not explain the behavioural gap. Per Appendix N the diffusion arms are not on this scale and are compared only with one another.

Second, MLM is the instructive failure. On recency-last it has the second-largest QK gap of the small models ($+0.0123$, twice the symmetric diffusion arm) and yet an OV gap more than two orders of magnitude below AR's and an accuracy of 0.009. Attending to the right position is not sufficient; the value path has to carry the binding.

Third, causal bias moves the diffusion arm along the OV axis, not the QK axis: $0.71 \rightarrow 1.28$ on haystack and $2.62 \rightarrow 4.35$ on recency-last, with QK gaps roughly doubling from a near-zero base. The accuracy column shows why we do not over-read this—on recency-last the symmetric arm scores higher (0.664 versus 0.230) despite the weaker circuit, which is one of the scored-readout disagreements flagged in Appendix I.

At scale the same decomposition separates the two public hybrids: LLaDA-8B forms both circuits, whereas Dream-7B forms strong QK routing with a much weaker value path. Both are DiT-style, so we report the ordering rather than the absolute scales, and the behavioural accuracies follow it: at K=7, 0.460 (LLaDA) versus 0.362 (Dream) on haystack and 0.746 versus 0.526 on recency-last. Weak OV in Dream is a candidate explanation for its better-than-MDLM but worse-than-LLaDA macro-ICL behaviour.

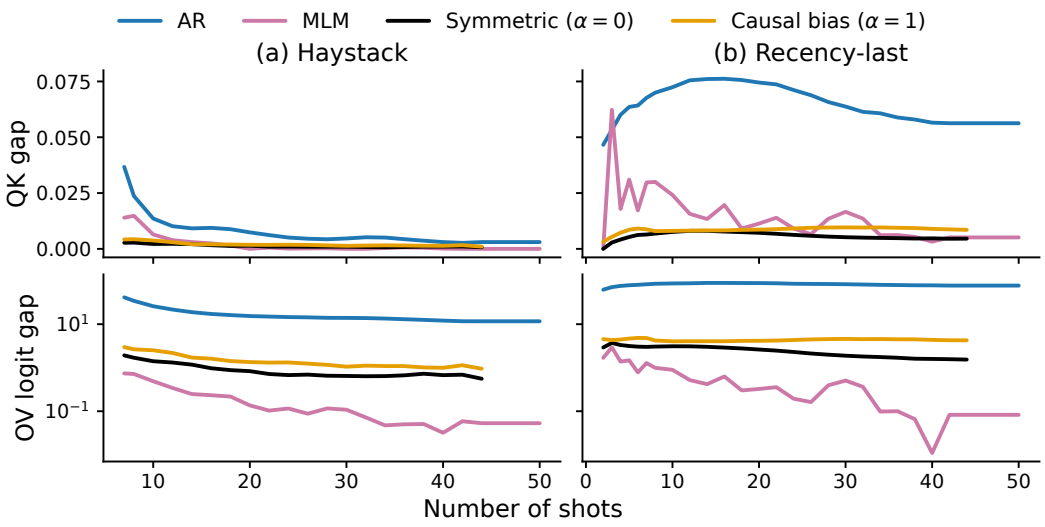


Figure 12: **OverrideKV QK/OV decomposition, small models.** (a) Haystack, (b) recency-last, QK gap above OV logit gap on a shared K axis. Both columns share one axis per row and the OV axis is logarithmic; OV values are comparable only within an architecture family. Shots below each task's minimum meaningful K are omitted.

O LIMITATIONS

Scale. The controlled experiments deliberately use small models and synthetic tasks, because full objective control, clean task generators and circuit probes are what isolate the mechanism, and none of them is available in frontier-scale training. The matched suite is 42.5k parameters, the OpenWebText models are 170M, and the 8B evidence is a 500M-token continued-training study on a single public checkpoint rather than a from-scratch retraining. We therefore claim that the effect exists and follows the predicted positional profile at 8B; we do not claim that its magnitude at 170M transfers, and a from-scratch billion-scale comparison remains the obvious missing experiment.

The intervention is not equivalence with AR. On the matched associative-recall control, only AR learns the algorithm; the position-biased arms remain at chance (Appendix J). Every claim we make about the intervention is directional—it moves behaviour the way the theory predicts—and none of it should be read as position-biased masking recovering autoregressive ICL.

One schedule. We test a single one-parameter power-law profile, with $\alpha \in \{0, 1, 1.5\}$ for MDLM and $\alpha \in \{0, 1\}$ elsewhere. Learned schedules, response-only corruption, block-wise or hybrid AR/diffusion objectives and causal attention constraints are all plausible alternatives we have not compared against, so we cannot claim this profile is optimal—only that it is minimal and that it works.

Readouts disagree. Scored one-step evaluation and native iterative generation are different measurements and do not always agree; we report the disagreements (Appendix I) rather than selecting the favourable readout. The strongest version of our claim concerns native contextual generation and the learned conditionals, not universal superiority under every artificial protocol.

Task realism. The few-shot tasks are synthetic and controlled, which is what lets us infer whether retrieval, recency, rule induction and specific circuits are present, but they are not naturalistic benchmarks, and the macro-ICL probe is a two-point summary rather than a full monotonicity test over positions.

Statistical power. Several appendix results rest on 5 seeds, and we mark those that do not clear their own seed noise—most of the suffix-mask table (Appendix H.3), and the TMS differences between diffusion arms, which are single-seed geometry measurements. Significance claims are made in two

places and with two different procedures: the native-generation comparisons use paired sign tests with Holm correction across the full comparison family (Appendix L), and the LLaDA-8B contrasts use paired bootstrap intervals on within-example margins (Section 7). The loss-weighted control arm is reported as unpaired condition means and is not accompanied by an interval (Appendix G.1). All remaining small-seed controls are treated descriptively.

Scope of the theory. Theorem 2.1 is a statement about conditional entropies at the population optimum. It says nothing about finite-sample behaviour, about what a particular architecture can represent, or about how large the gap is for natural language. Bayes-optimality of the sequential predictor is with respect to the meta-distribution on which the log-loss is defined, so it does not license claims about genuinely out-of-support tasks (Appendix B.7).