WHAT GETS COMMITTED FIRST SHAPES THE ANSWER IN MASKED DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion language models such as LLaDA and Dream generate text in parallel and in any order, and are now applied to retrieval-augmented question answering. At each step the model predicts many masked positions and the decoder selects which prediction to commit; the commitment is irreversible and becomes context for all later predictions. We ask whether the decoder's choice of which prediction to commit can redirect the final answer while the underlying model state stays fixed. Final accuracy cannot resolve this, since decodings that differ in one commitment differ in every later state. We therefore branch decodings from identical states and vary only the first non-padding commitment. This single change alters 27.3% of final answers. A dominant failure pattern arises when the decoder commits an end-of-turn (EOT) token before any answer content, while answer-supporting predictions are already present elsewhere in the canvas. Token identity governs the effect: committing end-of-sequence (EOS) yields 38.7% exact match at either tested position, and committing EOT yields 0%. Following this finding, EndFirst reweights EOS against non-EOS candidates once, at the earliest step where any uniform EOS reweighting can alter the trajectory, and returns control to the original decoder. With no extra forward pass, EndFirst at 32 passes outperforms a 64-pass retrieval-guided decoder on three QA benchmarks and improves that decoder by 5.3 exact-match points on average across eight LLaDA model-task settings. These results identify commitment selection as a distinct and directly controllable source of error in masked diffusion decoding.

1 INTRODUCTION

Masked diffusion language models such as LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025) generate text by unmasking a fixed-length canvas in parallel and in any order (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024; Gong et al., 2025), now scaled to 8B parameters and applied to knowledge-intensive tasks such as retrieval-augmented question answering (Izacard et al., 2023; Kim & Ye, 2026; Yu et al., 2026a). Their decoding separates two operations. At every step the model predicts tokens for all open positions, and the decoder selects which prediction to commit. A prediction is provisional; a commitment is irreversible and becomes context for every later prediction. Masked diffusion therefore separates what the model currently predicts from which prediction the decoder makes irreversible.

Existing decoding methods improve the predictions or the order in which positions are revealed: retrieval guidance reshapes token distributions with external evidence (Kim & Ye, 2026), and early boundary-token writes are treated as a bias to suppress at every step or to remove by retraining (Huang et al., 2026a; Kim et al., 2026; Liu et al., 2026); Section 6 reviews them. We ask a more specific question: starting from the same decoding state, can changing a single early commitment improve the final answer while leaving all subsequent decoding rules unchanged? Comparing final accuracy across decoders cannot isolate this effect because the decoders may differ at many steps.

To isolate this effect, we branch decoding trajectories from the same intermediate state. At the first non-padding commitment, at step $t^\star$, we replay generation after changing only the executed action, holding the model, retrieved evidence, preceding canvas, and forward-pass budget fixed. Changing

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

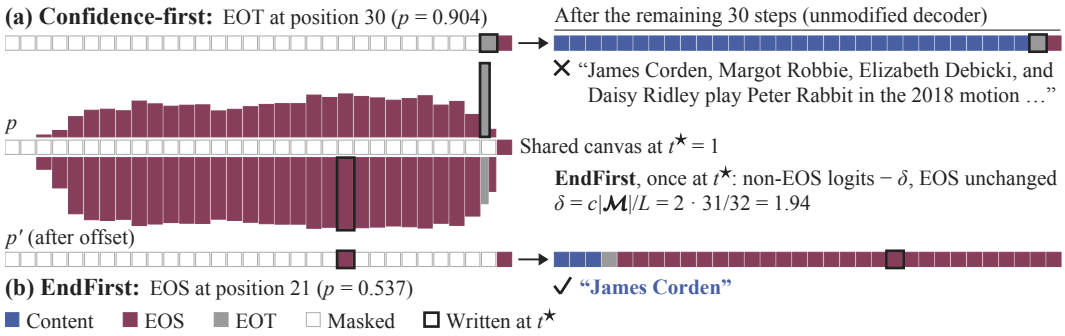


Figure 1: Same evidence, one different write, different answer (LLaDA-1.5, "Who plays Peter Rabbit?"). Both runs share the canvas at the first proposed non-EOS step $t^\star = 1$. Bars show each open position's EOS probability (wine) and the EOT probability at position 30 (gray), before (p) and after (p') EndFirst subtracts δ from non-EOS logits. (a) Confidence-first decoding writes EOT at position 30 and fills the canvas with a long wrong answer. (b) EndFirst writes EOS at position 21, where the model's EOS probability is highest, and the unmodified decoder returns "James Corden".

this single commitment alters 27.3% of final answers. The dominant failure begins with an end-of-turn (EOT) token committed before any answer content: over 4,000 LLaDA decodings, confidence-first decoding does so in 55.4 to 80.6% of questions, and in a random diagnostic sample 78.9% of wrong answers begin this way. In 54.7% of EOT-first failures, answer-supporting predictions are already present in the canvas.

In Figure 1, asked who plays Peter Rabbit with a passage naming James Corden, LLaDA-1.5 (Zhu et al., 2026) already predicts `James Corden plays the`, yet confidence-first decoding commits EOT at position 30 and later assigns all four passage actors to the role. From the same state, committing an end-of-sequence (EOS) token at position 21, with no answer token supplied, leads the unchanged decoder to `James Corden`. Token-position interventions show that the identity of the committed boundary token governs this effect: where the baseline first commits EOT and EndFirst commits EOS, EOS yields 38.7% exact match (EM) at either tested position, and EOT yields 0% at both (Section 5.3).

These observations point to a minimal intervention: change the competition among candidate commitments exactly where an early decision can first redirect the trajectory, and leave answer content to the model. **EndFirst** runs the base decoder unchanged until $t^\star$, the first step whose proposed write is not EOS, lowers every non-EOS logit once by a uniform offset, and returns control to the original decoder with no extra forward pass; the model's own distribution selects the committed token and position. Proposition 1 shows that $t^\star$ is the earliest step at which any uniform EOS reweighting, applied once or at every step, can alter confidence-first decoding.

We evaluate EndFirst on four retrieval-augmented QA benchmarks and across multiple masked-diffusion decoders. On standard RAG with LLaDA, EndFirst improves EM by 12.2 to 26.6 points on four QA benchmarks. At 32 forward passes it exceeds a 64-pass retrieval-guided decoder on Natural Questions, HotpotQA, and TriviaQA; on top of that decoder it improves EM in all eight LLaDA model-task settings, by 5.3 points on average, and the gains carry across decoding schedules and canvas lengths. Our contributions are threefold:

- We identify **first-commitment lock-in**, a failure mode in which a single early irreversible write redirects subsequent predictions; shared-state branching shows that changing only this commitment alters 27.3% of final answers (Section 3).
- We characterize the mechanism of this failure, showing that early EOT commitments dominate the observed errors, that boundary-token identity drives the effect independently of write position, and that uniform EOS reweighting first changes the trajectory at $t^\star$ (Proposition 1).
- We introduce **EndFirst**, a training-free, single-step intervention that adds no forward passes, outperforms decoding alternatives using up to twice the forward passes, and composes with retrieval guidance (Section 5).

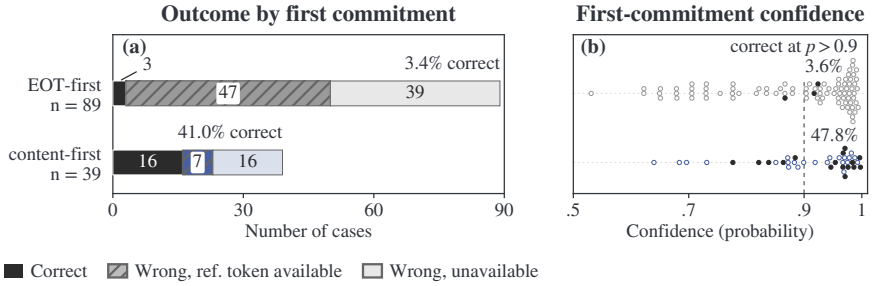


Figure 2: Confidence-first decoding commits EOT before content in most diagnostic decodings, and only 3.4% of these end correct, even at high confidence. (a) Outcomes: correct (dark), failures whose uncommitted predictions overlap the reference (hatched; Appendix D), other failures (light). (b) First-commitment confidence; filled points are correct, hollow wrong.

2 From Predictions to Commitments

Predictions and commitments. Masked diffusion decoding separates prediction from commitment. Let $x^{(t)}$ denote the partially filled canvas at step t and $\mathcal{M}^{(t)}$ its open positions. A forward pass on $x^{(t)}$ produces $p_\theta(\cdot \mid x^{(t)})_i$ for every $i \in \mathcal{M}^{(t)}$, omitting question and passages from the notation. The top-1 tokens form the provisional *predicted canvas*: after another position is filled, the remaining predictions change. In Figure 1 it reads `James Corden plays the`.

The decoder turns one of these provisional predictions into an irreversible decision. A commitment $a = (i, v)$ writes token v at position i and removes i from $\mathcal{M}^{(t)}$. Confidence-first decoding selects $\arg\max_{i \in \mathcal{M}^{(t)}, v} p_\theta(v \mid x^{(t)})_i$, one write per step, never revisiting a write. Written tokens form the *committed canvas*, context for all later passes. A prediction can change; a commitment cannot.

The first non-padding commitment. LLaDA (Nie et al., 2025) uses two terminal tokens: end-of-turn (EOT) ends the assistant turn and EOS pads the canvas after it; every other token is *content*. Confidence-first decoding begins by writing EOS padding at the final canvas position in 95.3 to 99.6% of decodings per group (Appendix B); we exclude this deterministic prefix. We define $t^\star$ as the first step at which the decoder proposes a non-EOS action, and call the action executed there the *first commitment*. Under confidence-first decoding it is EOT or content; EndFirst (Section 4) can replace it with EOS. Decoding runs the full step budget, and the answer is the whole canvas with special tokens removed, never truncated at a terminal token (Appendix A). Because terminal tokens are removed only at scoring time and do not stop decoding, their downstream effect arises through the context they create for later predictions.

Task and evaluation. We study retrieval-augmented short-answer QA. Given a question and its top three retrieved passages, the model fills a 32-position masked canvas and is instructed to answer in at most ten words (Appendix A). Unless stated otherwise, decoding is deterministic and commits one position per forward pass. We use exact match (EM) as the primary metric because the task requires a single direct answer (Appendix G). Our interventions change only which prediction is committed.

3 First-Commitment Lock-In

Prevalence of early EOT. Early EOT commitments are common. Across 4,000 trajectories from the two LLaDA models on Natural Questions and HotpotQA (1,000 questions each), the decoder commits EOT before any content token in 55.4 to 80.6% of examples. The first EOT concentrates near the end of the canvas, at position 30 in 59.8 to 81.3% of cases per group. Despite the ten-word instruction, these trajectories produce 20.4 to 21.1 words on average.

Confidence and outcome. In the random 128-item diagnostic sample (Appendix D, Figure 2a), 69.5% of trajectories are EOT-first and 30.5% content-first; 3.4% of EOT-first trajectories end correct, compared with 41.0% of content-first ones, and 78.9% of wrong answers are EOT-first. The gap persists at high confidence: 60.9% of first commitments have probability above 0.9, yet only

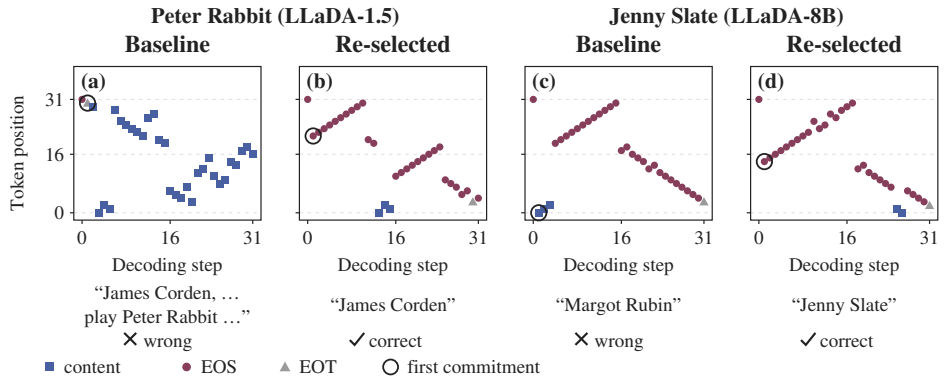


Figure 3: From the same state and budget, EOS in place of the first commitment changes the final answer (Peter Rabbit, left, excerpted; Jenny Slate, right). Blue squares: content; wine circles: EOS; gray triangles: EOT; black rings: first commitment.

16.7% of these trajectories end correct (3.6% EOT-first, 47.8% content-first; Figure 2b). Confidence scores an action under the current state, and committing the action changes that state.

Answer evidence before failure. For 49.5% of baseline failures, at least one uncommitted position at $t^\star$ has a top prediction that overlaps a reference answer, rising to 54.7% among EOT-first failures, against 16.4% under mismatched references (Appendix D). Many failures thus occur while item-specific answer evidence is present in the predicted canvas.

Causal effect of the first commitment. For each diagnostic trajectory, we save the complete decoding state at $t^\star$, replace only the action executed there, and resume the original decoder. Paired trajectories share the question, passages, model state, preceding commitments, and 32 deterministic forward passes; a repeated pass on an unchanged canvas reproduces the same distribution, so any difference in the answer is attributable to the changed commitment. Replaying all 128 diagnostic items with the action selected by EndFirst (Section 4) changes 27.3% of final answers. Conditional on the first action changing, 49.3% of answers differ; among EOT-first trajectories, 36.0% (Appendix F).

Divergent trajectories. In the Peter Rabbit case (Figure 1), LLaDA-1.5 commits EOT at position 30 with probability 0.904 while its predicted canvas already reads `James Corden plays the`, and later assigns all four passage actors to the role; writing EOS at position 21 yields `James Corden` with no answer token supplied. In the Jenny Slate case (Figure 3; Appendix Figure 9), committing the fragment `Marg` turns the predicted `Slate` into `ot`, while writing EOS first keeps `Slate`, yielding `Jenny Slate`.

We refer to this phenomenon as *first-commitment lock-in*: an early irreversible action alters subsequent predictions so that a trajectory that can reach the correct answer converges to an incorrect one. The most common instance begins with an EOT commitment before answer content, and the identity of the committed token determines the effect (Section 5.3, Figure 5a). Changing only a boundary commitment, without supplying any answer token, is sufficient to redirect later content generation. These trajectories isolate commitment selection as a failure source distinct from answer prediction: the model supports a useful continuation, and the decoder makes a different prediction irreversible first.

4 EndFirst: Re-Selecting the First Commitment

From diagnosis to intervention. This distinction motivates intervening directly on the commitment decision. An effective intervention should change which prediction becomes irreversible without specifying answer content. EOS is a content-neutral alternative already scored by the same forward pass: it carries no answer word, the model assigns it probability e_i at every open position i, and in both cases of Section 3 the recovering write is EOS where e_i is high. EndFirst modifies only the competition between EOS and non-EOS actions and lets confidence-first decoding select the resulting commitment.

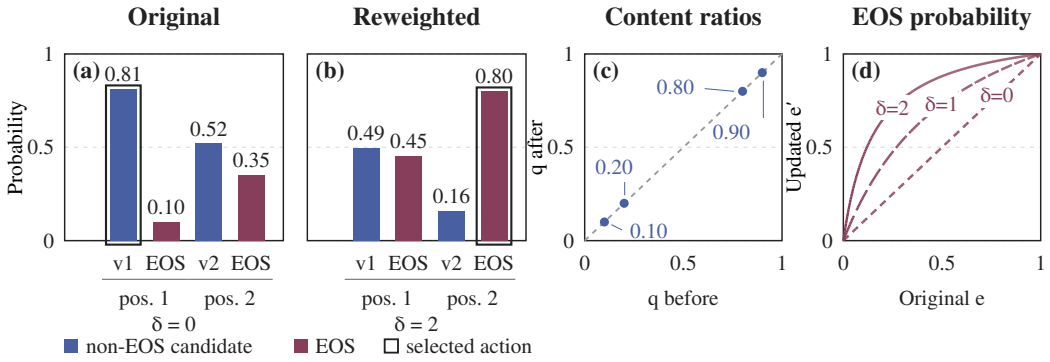


Figure 4: EndFirst changes cross-position competition while preserving within-position non-EOS ratios. (a, b) Illustrative distributions before and after $\delta = 2$, with selected actions outlined. (c) Unchanged conditional non-EOS probabilities $q = p\,(1-e)^{-1}$. (d) The EOS transformation in Equation 1 (Section 4). EndFirst re-selects once and resumes ordinary decoding at the same forward-pass budget.

Reweighting the commitment distribution. With e the EOS probability at an open position and p that of any non-EOS candidate there (content or EOT), subtracting a constant $\delta \geq 0$ from every non-EOS logit gives

$$p' = \frac{p}{1 + (\exp(\delta) - 1)\, e}, \qquad e' = \frac{\exp(\delta)\, e}{1 + (\exp(\delta) - 1)\, e}. \tag{1}$$

All non-EOS candidates at a position share one denominator, so the transform preserves their ordering, including EOT against content, and promotes no answer token over another. It raises the EOS odds at every position by the same factor $\exp(\delta)$, penalizing non-EOS writes most where EOS mass is high. Since e' increases strictly with e, a re-selected EOS lands where the model's own EOS probability is highest, up to ties (Appendix K). It thus changes the boundary competition and prescribes neither answer content nor a write position.

The earliest effective intervention point. The next result shows that uniform EOS reweighting before $t^\star$ is redundant.

Proposition 1 (EOS-prefix invariance) *Consider deterministic confidence-first decoding with one write per step, on the model's or a guided distribution, and at any step apply Equation 1 with the same $\delta \geq 0$ at every open position. If the original action at that step writes EOS, the reweighted decoder takes the same action. Hence the bias, even if applied at every step, leaves the decoder's EOS-only prefix unchanged and can first alter the trajectory at $t^\star$, the first proposed non-EOS write.*

Reweighting lifts no non-EOS probability above the winning EOS and preserves the EOS order across positions, so an EOS action is unchanged; induction gives the prefix claim (Appendix C). Before $t^\star$ the reweighting leaves every executed action unchanged; at $t^\star$ the base decoder first prefers a non-EOS action, a decision the reweighting can change. EndFirst therefore intervenes once, at $t^\star$, the step where an every-step bias also first acts (Section 5.3).

EndFirst. EndFirst runs the unmodified decoder until $t^\star$, the first step whose top proposed write is not EOS (Section 2). At that step only, it applies Equation 1 with

$$\delta = \frac{c\,|\mathcal{M}^{(t^\star)}|}{L}, \tag{2}$$

where $L = 32$ is the canvas length and c sets the strength, so δ is largest when the first non-EOS proposal comes early. Confidence-first selection on the reweighted distributions picks the write, and all later steps use the original decoder (Figure 1b). On ARAM it reweights the guided logits; Appendix H covers block and threshold decoding. Because EndFirst reuses the distributions already computed at $t^\star$, it requires no additional forward pass.

Table 1: EndFirst consistently improves both standard and retrieval-guided decoding. EM and token F1 (%) on 1,000 questions per task; EndFirst uses $c = 4$ with standard RAG and $c = 2$ with ARAM; small numbers give the gain over the same base decoder. Bold: best; italics: second best per column within each model.

			NQ		HotpotQA		TriviaQA		T-REx	
	Method	Passes	EM	F1	EM	F1	EM	F1	EM	F1
LLaDA-8B	Standard RAG	32	19.6	33.7	6.5	16.2	44.9	56.0	16.6	22.5
	+ EndFirst	32	*36.5* +16.9	*49.2* +15.5	*22.4* +15.9	*31.1* +14.9	*71.5* +26.6	*77.5* +21.5	28.8 +12.2	33.2 +10.7
	ARAM	64	33.8	46.3	17.8	26.5	69.4	75.4	*31.3*	*35.0*
	+ EndFirst	64	**38.5** +4.7	**50.4** +4.1	**24.7** +6.9	**32.6** +6.1	**75.1** +5.7	**80.1** +4.7	**33.4** +2.1	**36.6** +1.6
LLaDA-1.5	Standard RAG	32	20.5	34.6	6.6	16.3	45.5	56.8	16.2	22.5
	+ EndFirst	32	*35.0* +14.5	*48.4* +13.8	*22.2* +15.6	*31.3* +15.0	*71.2* +25.7	*77.7* +20.9	29.0 +12.8	33.8 +11.3
	ARAM	64	32.4	45.1	16.4	25.7	68.5	75.3	*29.6*	*34.2*
	+ EndFirst	64	**38.4** +6.0	**50.3** +5.2	**24.3** +7.9	**32.9** +7.2	**75.1** +6.6	**80.6** +5.3	**32.2** +2.6	**36.1** +1.9
Dream	Standard RAG	32	38.2	52.7	24.4	33.2	75.1	81.8	29.3	33.9
	+ EndFirst	32	39.4 +1.2	*53.7* +1.0	**26.3** +1.9	*34.7* +1.5	75.5 +0.4	82.0 +0.2	29.4 +0.1	33.9 0.0
	ARAM	64	*40.2*	*53.7*	25.8	*34.7*	**76.6**	**82.4**	*32.9*	*37.0*
	+ EndFirst	64	**40.6** +0.4	**54.1** +0.4	*26.2* +0.4	**35.2** +0.5	*76.5* −0.1	*82.3* −0.1	**33.1** +0.2	**37.2** +0.2

Table 2: EndFirst outperforms decoding alternatives at equal or lower inference cost. EM (%) on 1,000 questions per task; small numbers give the gain over the best 32 or 33-pass baseline in that column; the dashed line separates these methods from ARAM at 64 passes. Bold: best; italics: second best per column.

		LLaDA-8B				LLaDA-1.5			
Method	Passes	NQ	HotpotQA	TriviaQA	T-REx	NQ	HotpotQA	TriviaQA	T-REx
Standard RAG	32	19.6	6.5	44.9	16.6	20.5	6.6	45.5	16.2
Frontier-16	32	21.2	9.9	50.1	16.5	23.6	11.2	54.3	18.4
UNCODE	32	21.2	10.5	51.3	12.8	24.0	12.4	55.5	15.3
SPREAD	33	22.8	12.4	54.5	17.3	24.8	13.0	58.0	18.0
EndFirst (c=4)	32	*36.5* +13.7	*22.4* +10.0	*71.5* +17.0	28.8 +11.5	*35.0* +10.2	*22.2* +9.2	*71.2* +13.2	29.0 +10.6
EndFirst (c=8)	32	**39.8** +17.0	**26.8** +14.4	**75.3** +20.8	*30.0* +12.7	**38.5** +13.7	**25.9** +12.9	**74.8** +16.8	*29.3* +10.9
ARAM	64	33.8	17.8	69.4	**31.3**	32.4	16.4	68.5	**29.6**

5 Experiments

Setting. We study short-answer retrieval-augmented QA on Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and T-REx (Elsahar et al., 2018), with 1,000 questions per task, each with its top three retrieved passages, decoded on a 32-position canvas at temperature 0. We compare three decoders: *Standard RAG* runs confidence-first decoding for 32 passes; ARAM adds retrieval-guided logit adjustment in 64 passes (Kim & Ye, 2026); and SPREAD adds query-guided unmasking in 33 passes (Yu et al., 2026a). EndFirst runs on standard RAG and ARAM at $c = 2$ unless marked otherwise. We report exact match (EM), our headline metric, and F1 (Rajpurkar et al., 2016) over all outputs, including empty and unresolved ones; a lenient containment metric would also credit restating the question or hedging (Appendix G). Details, answer normalization, and paired intervals for every comparison are in Appendices A and E to H.

Intervention strength. We use $c = 2$ as the default. For standard RAG we also report $c = 4$, selected on a disjoint 64-question calibration set, and study stronger doses in Section 5.4.

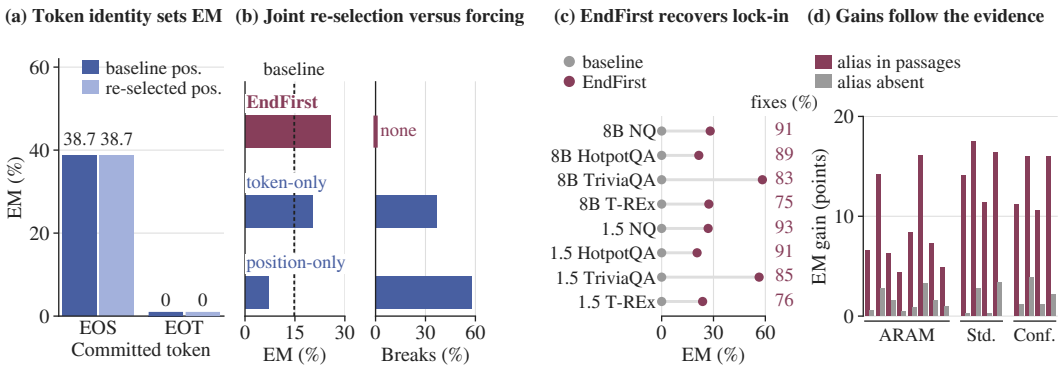


Figure 5: Mechanistic controls for the first commitment. The end-of-turn (EOT) token closes the assistant turn; EOS pads the canvas after it. (a) On the 31 items where EndFirst replaces a baseline EOT with EOS, committing EOS gives 38.7% EM at either position, and EOT gives 0% at either. (b) On all 128 diagnostic items, EndFirst changes 27.3% of answers and breaks none; token-only and position-only change 52.3% and 54.7% yet reach only 20.3% and 7.0% EM (dashed: baseline, 14.8%). (c) On questions whose first two baseline commitments are terminal (6.6 to 37.5%; Table 9), baseline (grey) EM is 0% and EndFirst (wine) 20.5 to 58.3%; right: share of all EndFirst fixes. (d) EndFirst's EM gain with and without a reference alias in the passages, reaching 4.4 to 17.5 points with one, all 16 settings of Table 10 (Std.: standard RAG; Conf.: confirmation).

5.1 Low Forward-Pass Cost and Composition

EndFirst beats ARAM with half the forward passes. At $c = 4$, EndFirst improves standard RAG by 12.2 to 26.6 EM points across the eight LLaDA settings, and at 32 passes it exceeds ARAM at 64 on NQ, HotpotQA, and TriviaQA for both LLaDA models, by 2.1 to 5.8 EM points (Tables 1 and 2; Appendix G). A single localized intervention outperforms retrieval-guided decoding with half as many model evaluations.

EndFirst and retrieval guidance compose. At ARAM's 64 passes and the same default $c = 2$, ARAM + EndFirst improves EM in all eight LLaDA settings, by 5.3 points on average (Table 1). On T-REx, where standard RAG's output contains a reference alias, EndFirst at $c = 4$ leads ARAM by 3.9 and 6.4 points, and ARAM + EndFirst is the best arm, keeping 94.9% and 93.9% of the items ARAM answers beyond $c = 4$ (Appendix G). The two methods act on different stages of generation: ARAM reshapes token distributions using retrieved evidence, and EndFirst changes which supported prediction becomes irreversible context.

The gains hold on held-out questions. On 256 held-out questions per group, EndFirst adds 6.3 to 7.8 points at $c = 2$ and 16.8 to 18.4 at $c = 4$; at $c = 4$ it loses none of standard RAG's correct answers (Appendix Table 6; Appendix G).

5.2 Precise Answer Formation

The gain lies in how the answer is formed. EndFirst converts diffuse or overextended answers into precise commitments. With $c = 4$, standard-RAG outputs shrink from 11.5 to 15.2 normalized words to 3.6 to 4.5 words on NQ and HotpotQA, and in 65.5 to 93.8% of EndFirst's fixes per setting the baseline had already mentioned the reference answer inside a longer response (Appendix G). EndFirst turns answer-supporting generation into an exact final answer. Shortening alone does not reproduce the gain: with the same deterministic prefix rule truncating both outputs, EndFirst leads by 1.9 to 8.2 EM points in all 16 LLaDA settings of Appendix G. Truncated to 6.50 words, standard RAG with LLaDA-8B on NQ still trails the 6.60-word answers of EndFirst at $c = 2$ by 8.2 points (Appendix Table 8). No shorter canvas reaches it (Appendix Table 11): the 16-position canvas matches EndFirst's length within 1.6 words at only 2.3 to 3.2 EM. These comparisons establish that output shortening does not account for the improvement. The gain is largest when the retrieved passages contain the answer, reaching 4.4 to 17.5 EM points in all 16 settings (Figure 5d; Appendix Table 10): EndFirst improves how available evidence becomes a final commitment.

Table 3: EndFirst improves every base decoder. EM (%), c=2, 32 passes; †EndFirst on standard RAG; more in Appendix Table 13.

		NQ			HotpotQA		
Model	Base decoder	Base	EndFirst	Gain	Base	EndFirst	Gain
8B	Standard RAG	19.6	**29.3**	+9.7	6.5	**14.6**	+8.1
	Block	19.4	**25.1**	+5.7	9.5	**13.5**	+4.0
	Threshold	19.7	**28.4**	+8.7	6.5	**13.0**	+6.5
	UNCODE†	21.2	**29.3**	+8.1	10.5	**14.6**	+4.1
1.5	Standard RAG	20.5	**28.4**	+7.9	6.6	**14.7**	+8.1
	Block	22.0	**27.0**	+5.0	11.1	**14.9**	+3.8
	Threshold	20.4	**27.3**	+6.9	6.5	**13.2**	+6.7
	UNCODE†	24.0	**28.4**	+4.4	12.4	**14.7**	+2.3

Table 4: A stronger dose improves every setting. EM (%), standard RAG; Fixes and Losses: questions only c=8 or only c=4 answers exactly.

			EndFirst				
Model	Task	RAG	c=4	c=8	Gain	Fixes	Losses
8B	NQ	19.6	*36.5*	**39.8**	+3.3	34	1
	HotpotQA	6.5	*22.4*	**26.8**	+4.4	44	0
	TriviaQA	44.9	*71.5*	**75.3**	+3.8	40	2
	T-REx	16.6	*28.8*	**30.0**	+1.2	13	1
1.5	NQ	20.5	*35.0*	**38.5**	+3.5	35	0
	HotpotQA	6.6	*22.2*	**25.9**	+3.7	40	3
	TriviaQA	45.5	*71.2*	**74.8**	+3.6	36	0
	T-REx	16.2	*29.0*	**29.3**	+0.3	7	4

5.3 Why the First Commitment Matters

Token identity versus write position. Where the baseline first commits EOT and EndFirst replaces it with EOS ($n = 31$), we cross the written token with its position (Figure 5a; Appendix Table 15). EOS makes 38.7% of these answers exact at either position and EOT none at either (exact McNemar $p = 0.00049$ per pair). EOS at the baseline's late position yields 1.74-word answers, whereas moving EOT earlier only shortens answers from 16.35 to 11.74 words. The committed EOS changes the state from which content is predicted, moving the first content commitment from step 2 to a median of step 27 (Appendix Figure 8b).

Re-selection versus forced perturbation. On all 128 diagnostic items, forcing EndFirst's first token at the baseline's position (token-only) or the baseline's first token at EndFirst's position (position-only) changes about twice as many answers as EndFirst but reaches lower EM (Figure 5b), and breaks 36.8% and 57.9% of the baseline's correct answers, while EndFirst raises EM from 14.8% to 25.8% and breaks none (Appendix F). The baseline answers none of the 31 EOT-to-EOS items, so these losses fall elsewhere: forcing either component alone breaks answers, while EndFirst lets the model choose token and position together among its own supported actions.

Where corrections occur. Questions on which the baseline spends its first two commitments on terminal tokens, a stratum defined from the baseline trajectory alone, are 6.6 to 37.5% of each setting yet hold 75.0 to 93.4% of all corrections by ARAM + EndFirst at $c = 2$ across the eight LLaDA settings (Figure 5c; Appendix Table 9). The baseline answers none of them; EndFirst answers 20.5 to 58.3% and breaks none.

Every-step bias. A fixed $\lambda = 0.25$ subtracted from the non-EOS logits at every step reaches 14 to 26% of EndFirst's gain. Tuned per group to match EndFirst's mean answer length, without EM or F1, the bias lands on EndFirst's operating point: the same answer on 78.0 to 82.2% of questions and EM within 1.3 points in three groups and 2.3 in the fourth (Appendix Tables 7 and 6). The tuned every-step bias converges to the same operating point because its first effective action occurs at $t^\star$, exactly where EndFirst intervenes (Proposition 1): it matches EndFirst's first action on 75.0 to 90.2% of confirmation questions, after 7 to 10 calibration runs per group (Appendix E). EndFirst locates the intervention point directly from the decoding structure.

5.4 Dose, Direct Baselines, and Generality

Dose strength. A stronger dose improves accuracy in every setting (Table 4). Moving from $c = 4$ to $c = 8$ adds 0.3 to 4.4 EM points, fixing 249 additional questions and losing 11 across the eight model and task combinations, and $c = 8$ beats ARAM by 5.9 to 9.5 points on NQ, HotpotQA, and TriviaQA for both LLaDA models; the full sweep from $c = 0$ to $c = 8$ rises monotonically in every group (Appendix Table 11).

Direct baselines at equal budget. At the same 32-pass budget and the default $c = 2$, EndFirst beats UNCODE (Huang et al., 2026a), which recalibrates the unmasking priority at every step, in all eight LLaDA settings, by 2.3 to 13.4 EM points (Table 3; Appendix Table 13). Frontier-1 and Frontier-16, which write only within k positions of the leftmost mask as in the width-16 gate of Yeom et al. (2026), trail EndFirst by at least 3.5 points (Table 2; Appendix Table 11).

Generality across models and decoders. The gain transfers beyond the default decoder. At $c = 2$, EndFirst improves LLaDA-MoE on NQ and HotpotQA, improves both dense LLaDA models under block and confidence-threshold decoding (Table 3), and gains 9.6 to 17.8 points on TriviaQA and T-REx under standard decoding (Appendix Table 13). On a 64-position canvas EndFirst gains 3.4 to 5.8 EM points (Appendix Table 14), at most 1.6% run-time overhead (Appendix H). Dream exhibits a different commitment pattern: its first content token appears only after 21.4 and 24.5 terminal commitments (Appendix B), so EndFirst triggers near the end of decoding (mean step 23.8 and 26.9 of 32), when few positions remain open, and leaves ARAM's accuracy intact. On the LLaDA models, whose first commitment binds early, it adds 5.3 points on average.

6 Related Work

Commitment order and boundary preferences. For masked diffusion language models (Lou et al., 2024; Zhu et al., 2026), concurrent studies tie commitment order to reasoning and uncertainty (Yeom et al., 2026; Yu et al., 2026b; Lee et al., 2026) and control it with a leftmost-mask gate (Yeom et al., 2026) or deferred boundary commitments (Shu et al., 2026). UNCODE recalibrates the unmasking priority at every step to delay the boundary-first and trivial-token writes of confidence-based samplers (Huang et al., 2026a). These methods treat early boundary writes as one class. Our analysis separates an early boundary's presence from its identity: under shared-state interventions, an early EOS recovers answers and an early EOT locks them in.

Decoding control for retrieval-augmented generation. Retrieval aids knowledge-intensive tasks and can also distract (Izacard et al., 2022; 2023; Shi et al., 2023). In masked diffusion, ARAM interpolates retrieval-conditioned and query-only logits at every step, much as classifier-free guidance, contrastive decoding, and DoLa reshape each step's distribution (Kim & Ye, 2026; Ho & Salimans, 2022; Li et al., 2023; Chuang et al., 2024); SPREAD unmasks query-relevant positions first, SARDI refreshes retrieval from provisional tokens, and DLLM-Searcher overlaps search with reasoning (Yu et al., 2026a; Jünger et al., 2026; Zhao et al., 2026). These methods shape the evidence and distributions the decoder draws on; which supported prediction first becomes irreversible context is a separate decision, which EndFirst sets, composing with ARAM.

Termination and length control. Rainbow Padding and VoidPadding retrain padding targets against EOS overflow, DAEDAL adjusts canvas length during denoising, Prophet commits all remaining positions once the answer converges, and SureLock, Fast-dLLM, and SDTT cut denoising computation (Kim et al., 2026; Liu et al., 2026; Li et al., 2026a;b; Oba et al., 2026; Wu et al., 2026; Deschenaux & Gulcehre, 2025). In autoregressive decoding, EOS biases and length penalties decide when generation stops (Wu et al., 2016; Murray & Chiang, 2018; Stahlberg & Byrne, 2019). In a masked canvas a committed EOS is also context for every later prediction, and any uniform EOS bias first acts at $t^\star$ (Proposition 1). EndFirst uses this role of EOS.

Revision and remasking. Non-autoregressive translation, Mask-Predict, learned remasking, and draft-and-verify decoding revise or revoke earlier writes after they condition later ones (Gu et al., 2018; Lee et al., 2018; Ghazvininejad et al., 2019; Huang et al., 2026b; Hong et al., 2026), and autoregressive exposure-bias studies trace errors to conditioning on the model's own outputs (He et al., 2021; Arora et al., 2022). The causal effect of one irreversible write under a fixed model state remains open; this work isolates it and selects the first write before it becomes context.

7 Conclusion

Masked diffusion generation separates provisional prediction from irreversible commitment, making commitment selection part of the generation process itself. By branching from identical intermediate states, we show that changing a single early commitment redirects 27.3% of final answers, and that the identity of an early boundary commitment determines whether answer-supporting predictions develop into an exact response. EndFirst acts on this decision once, at the earliest point where uniform EOS reweighting can alter the trajectory. Without additional forward passes, it substantially improves retrieval-augmented QA, outperforms decoding alternatives of equal and higher cost, and composes with retrieval guidance. Prediction quality and commitment quality are distinct dimensions of masked diffusion decoding.

AI USE STATEMENT

In this work, we used generative AI tools to refine hypotheses, provide feedback on the research methodology and experiment design, implement experiment and analysis code, support the qualitative review of changed answers and of baseline outputs, and interpret results. The semantic categories of the 14 exact-match gains in the diagnostic sample come from an AI-assisted review (Appendix E). Additionally, we used generative AI tools to create scientific figures, identify and check related literature, format references, draft and edit the paper text, and check submission requirements. We have reviewed all AI-assisted work, and reported numbers were checked against saved model outputs and logged scores. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A lists the model checkpoints and revisions, tokenizers, decoding settings, the literal prompt, and answer extraction. Appendix B gives the settings and forward-pass budgets of every compared method, Appendix D the diagnostic sample and support measure, and Appendix E the calibration and confirmation splits, answer normalization, and the seeded paired-bootstrap protocol behind all reported intervals.

REFERENCES

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.58. URL https://aclanthology.org/2022.findings-acl.58/.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *The Thirteenth International Conference on Learning Representations*, pp. 5007–5045, 2025. URL https://openreview.net/forum?id=uZ5K4HeNwd.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1544/.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel decoding of conditional masked language models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pp. 6112–6121, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633/.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. Scaling diffusion language models via adaptation from autoregressive models. In *The Thirteenth International Conference on Learning Representations*, pp. 5046–5073, 2025. URL https://openreview.net/forum?id=j1tSLYKwg8.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *ICLR*, 2018. URL https://openreview.net/forum?id=B1l8BtlCb.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5087–5102, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.415. URL https://aclanthology.org/2021.emnlp-main.415/.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Feng Hong, Geng Yu, Yushi Ye, Haicheng Huang, Huangjie Zheng, Ya Zhang, Yanfeng Wang, and Jiangchao Yao. Wide-in, narrow-out: Revokable decoding for efficient and effective DLLMs. In *The Fourteenth International Conference on Learning Representations*, pp. 40132–40150, 2026. URL https://openreview.net/forum?id=XtLQHlNLxy.

Pengcheng Huang, Tianming Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Tong Xiao, Zulong Chen, and Maosong Sun. Empirical analysis of decoding biases in masked diffusion models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6853–6876, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.311. URL https://aclanthology.org/2026.acl-long.311/. arXiv:2508.13021v3.

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't settle too early: Self-reflective remasking for diffusion language models. In *The Fourteenth International Conference on Learning Representations*, pp. 56998–57026, 2026b. URL https://openreview.net/forum?id=BsZeTuB5fD.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023. URL https://www.jmlr.org/papers/v24/23-0037.html.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Paul Jünger, Justin Lovelace, Linxi Zhao, Dongyoung Go, and Kilian Q Weinberger. Self-augmenting retrieval for diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uiE6BIx2vC.

BumJun Kim, Dongjae Jeon, Dueun Kim, Wonje Jeung, and Albert No. Rainbow padding: Mitigating early termination in instruction-tuned diffusion LLMs. In *The Fourteenth International Conference on Learning Representations*, pp. 18977–19001, 2026. URL https://openreview.net/forum?id=cznTlh7Msz.

Jaemin Kim and Jong Chul Ye. Adaptive guidance for retrieval-augmented masked diffusion models, 2026. URL https://arxiv.org/abs/2603.17677.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1149. URL https://aclanthology.org/D18-1149/.

Jihoon Lee, Hoyeon Moon, Kevin Zhai, Arun Kumar Chithanar, Anit Kumar Sahu, Soummya Kar, Chul Lee, Souradip Chakraborty, and Amrit Singh Bedi. TEST-TIME SCALING IN DIFFUSION LLMS VIA HIDDEN SEMI-AUTOREGRESSIVE EXPERTS. In *The Fourteenth International Conference on Learning Representations*, pp. 46880–46899, 2026. URL https://openreview.net/forum?id=L5y7in91vd.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Beyond fixed: Training-free variable-length denoising for diffusion large language models. In *The Fourteenth International Conference on Learning Representations*, pp. 91715–91731, 2026a. URL https://openreview.net/forum?id=Ic2A2gCseC.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. Diffusion language model knows the answer before it decodes. In *The Fourteenth International Conference on Learning Representations*, pp. 135235–135253, 2026b. URL https://openreview.net/forum?id=g88nt4ieTG.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Chunyu Liu, Zhengyang Fan, Kaisen Yang, and Alex Lamb. Voidpadding: Let [void] handle padding in masked diffusion language models so that [eos] can focus on semantic termination, 2026. URL https://arxiv.org/abs/2606.17999.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://aclanthology.org/W18-6322/.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 50608–50646. Curran Associates, Inc., 2025. doi: 10.52202/085713-1689. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/48b383b24230e0e6e649d9c98dae4d8c-Paper-Conference.pdf.

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. Stopping computation for converged tokens in masked diffusion-LM decoding. In *The Fourteenth International Conference on Learning Representations*, pp. 143558–143580, 2026. URL https://openreview.net/forum?id=PzhNnMepgl.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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, pp. 103131–103167, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Yingte Shu, Yuchuan Tian, Chao Xu, Yunhe Wang, and Hanting Chen. Deferred commitment decoding for diffusion language models, 2026. URL https://arxiv.org/abs/2601.02076.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331/.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *The Fourteenth International Conference on Learning Representations*, pp. 57027–57051, 2026. URL https://openreview.net/forum?id=3Z3Is6hnOT.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa,

Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. URL https://arxiv.org/abs/1609.08144.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

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

Jewon Yeom, Jaewon Sok, Seonghyeon Park, Jeongjae Park, Hwiyeong Lee, and Taesup Kim. Answer first, reason later: When commitment order costs accuracy in diffusion language models, 2026. URL https://arxiv.org/abs/2608.05687.

Chuanyue Yu, Jiahui Wang, Yuhan Li, Heng Chang, Ge Lan, Qingyun Sun, Jia Li, Jianxin Li, and Ziwei Zhang. Unlocking the potentials of retrieval-augmented generation for diffusion language models, 2026a. URL https://arxiv.org/abs/2601.11342.

Longxuan Yu, Yu Fu, Shaorong Zhang, Hui Liu, Mukund Varma T, Greg Ver Steeg, and Yue Dong. Thinking out of order: When output order stops reflecting reasoning order in diffusion language models, 2026b. URL https://arxiv.org/abs/2601.22035.

Jiahao Zhao, Shaoxuan Xu, Zhongxiang Sun, Fengqi Zhu, Jingyang Ou, Yuling Shi, Chongxuan Li, Xiao Zhang, and Jun Xu. Dllm-searcher: Adapting diffusion language model for efficient search agents. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2498–2508. ACM, July 2026. doi: 10.1145/3805712.3809643. URL http://dx.doi.org/10.1145/3805712.3809643.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.524. URL https://aclanthology.org/2026.acl-long.524/.

A Experimental Details

Checkpoints and decoding. LLaDA-8B and Dream are Instruct checkpoints; the LLaDA-1.5 checkpoint is based on LLaDA-8B-Instruct. Each model is loaded with its matching tokenizer; Table 5 lists the Hugging Face checkpoints. Every main-table model runs NQ, HotpotQA, TriviaQA, and T-REx. The 256-question confirmation and the short-canvas controls use only the two LLaDA models on NQ and HotpotQA; the diagnostic trajectories use those same four model-task groups. Other control budgets are in Appendix B.

Table 5: Main-table checkpoints and decoding. All three use a 32-token canvas, 32 commitment steps, one block of length 32, zero-temperature greedy top-1 tokens, and no remasking. Standard RAG and ARAM choose positions by confidence; SPREAD uses query relevance. Their forward counts are 32/64/33, respectively. Dream reads shifted predictor logits; LLaDA reads same-position logits.

Model	Hugging Face checkpoint	Tokenizer / mask / EOS
LLaDA-8B	`GSAI-ML/LLaDA-8B-Instruct`	matching tokenizer; 126336 / 126081
LLaDA-1.5	`GSAI-ML/LLaDA-1.5`	matching tokenizer; 126336 / 126081
Dream	`Dream-org/Dream-v0-Instruct-7B`	matching tokenizer; 151666 / 151643

Experiments run on NVIDIA A800-SXM4-80GB GPUs; model weights are in bfloat16 and decision logits are cast to float32.

Questions and retrieved context. The evaluation uses development sets with pre-retrieved passages: NQ (2,837 questions), HotpotQA (5,600), TriviaQA (5,359), and T-REx (5,000); the TriviaQA and T-REx sets are those used by ARAM. From each task we sample 1,000 questions uniformly without replacement under a fixed random seed; sampled questions have unique IDs, nonempty questions, three passages, and answer aliases. Retrieval is fixed: every method uses the same model, prompt, tokenizer, and passages, and receives the first three retrieved passages in their released order, separated by blank lines in the prompt, with no further truncation.

The task-shared held-out NQ and HotpotQA pools exclude normalized questions used in any earlier run, reject within-pool duplicates, require a nonempty answer and at least three passages, then shuffle eligible questions with a fixed random seed. The first 64 per task are calibration and the next 256 are confirmation; both LLaDA models use the same task pool. The trajectory analyses use a random diagnostic sample of 128 model-question items, 32 per LLaDA model-task group (Appendix D). Appendix E gives the calibration and scoring protocol.

Comparison implementations. Standard RAG is confidence-first decoding: one retrieval-conditioned model call per step, choose each open position's top-1 token, then commit the position with the highest softmax confidence. It uses 32 calls, one 32-position block, no sampling and no remasking. We reimplement SPREAD from its paper: it encodes the raw question alone without special tokens in one additional forward call, averages the last normalized-layer hidden states, and selects each open canvas position by sigmoid of cosine similarity to that question vector. The token at that position is its greedy top-1 prediction; 32 canvas calls plus the question call give 33. We reimplement ARAM's guidance between retrieval-conditioned and prior distributions: each step evaluates the same canvas with the three-passage prompt and a prior prompt whose sole context is `No relevant context available.`. For prior and context distributions p_0, p_c with log probabilities ℓ_0, ℓ_c, it uses

$$s = \max\left(0, \sum_v (p_c - p_0)(\ell_c - \ell_0)\right), \qquad H = -\sum_v p_c \ell_c,$$
$$w = \tanh\left(\frac{\beta s}{H + 10^{-6}}\right), \qquad \ell = \ell_0 + w(\ell_c - \ell_0).$$

Its β is 0.1 for both LLaDA checkpoints and 0.5 for Dream, and its 32 steps use 64 calls. No sampling temperature, beam, or remasking is added. The EndFirst settings are in Appendix E.

Literal user prompt and answer extraction. The following is the exact user-message content constructed for every method, including its line breaks. {context} is the first three passages in their released order, each prefixed Passage 1:, Passage 2:, or Passage 3: and joined by two newlines. {query} is the raw question.

```
Use the following passages as context if they are helpful,
but provide a SHORT and DIRECT answer.

RULES:
- First, check the context passages for the answer
- If the context contains the answer, use it
- If not, use your general knowledge
- Answer must be 1-10 words maximum
- Do NOT include any explanations or extra text
- Do NOT repeat or summarize the passages

Context:
Passage 1:
{passage_1}

Passage 2:
{passage_2}

Passage 3:
{passage_3}

Question: {query}

Short Answer:
```

This text is passed as a single user message through each model's released chat template, with the generation prompt appended; Dream's template also adds its default system message.

The answer is the full left-to-right canvas decoded with special tokens removed, including interior ones. Decoding does *not* truncate at the first EOS or EOT, and generation continues to the full step budget. All reported scores, including those of canvases with an interior EOS, use the scoring rule in Appendix E.

B Settings and Budgets

The main task is retrieval-conditioned short-answer QA with the question and top three retrieved passages, a 32-position masked canvas, deterministic confidence-first decoding, and one write per forward pass. Standard RAG uses 32 passes; SPREAD uses 33; ARAM and EndFirst on ARAM use 64. Short-canvas controls use 4, 8, or 16 positions and the same number of passes. The separate fixed-0.25 every-step EOS-bias and frontier controls use 32 passes. Confirmation uses a 32-position canvas and 32 passes per arm; EndFirst adds no forward pass. On the 4,000 confidence-first decodings of Section 3, the step-0 write is EOS padding at the last canvas position in 95.3 to 99.6% of decodings per model-task group, and a content token is the very first write, ahead of the EOS padding, in only 0.1% of decodings; on average, 5.6 to 8.0 terminal tokens precede the first content token. The short-canvas controls keep a median first-content step of 2. On the 16-position canvas, the step-0 write is a terminal token at the last position in 99.9 to 100% of questions per group, and the first terminal token lies at position 14 or later in 93.5 to 95.4%; the model fills the remaining positions with one compressed sentence, so no output is empty. On the confirmation sets, logged mean run time per question differs between EndFirst and standard RAG by -0.004 to $+0.001$ seconds. With standard RAG, Dream commits on average 21.4 and 24.5 end tokens before its first content token on NQ and HotpotQA; the Dream trigger statistics come from a separate logged run of 64 questions per task with the ARAM decoder at 64 passes; there EndFirst triggers at mean step 23.8 and 26.9 and changes the output on 2 and 1 questions.

C Proof of Proposition 1

Let the original action at the step write EOS at position j, so $e_j \geq p_i(v)$ and $e_j \geq e_i$ for every open i and non-EOS v. By Equation 1, $p'_i(v) \leq p_i(v) \leq e_j \leq e'_j$, and $e'_j \geq e'_i$ because e' increases strictly with e. Any tie after reweighting was a tie before it, so the same tie-break selects the same action. The canvas after an unchanged action is unchanged, and induction over the steps before $t^\star$ gives the prefix claim. □

D Diagnostic Sample and Support Measure

The diagnostic sample contains 32 randomly selected model-question items in each of four LLaDA model-task groups, for 128 in total. NQ1908 (Jenny Slate) is a selected illustration outside the random sample. At the baseline's first proposed non-EOS action, support excludes the selected action position and every already committed position. For each remaining masked position, decode its top-1 token while skipping special tokens; lowercase it, remove ASCII punctuation and English articles, and collapse whitespace. A token supports a reference if its normalized text has at least two alphanumeric characters and is a substring of any normalized reference answer. This lexical measure yields 49.5% $(54/109)$ among baseline failures, with 54.7% $(47/86)$ for EOT-first and 30.4% $(7/23)$ for content-first failures. Two controls show that this overlap is specific to the item. Scoring each failure against the references of every other diagnostic item gives 16.4% on average. A stricter whole-word variant, in which a predicted token must equal a reference word other than a function word, gives 31.2% of failures (36.0% EOT-first, 13.0% content-first) against 0.8% with other items' references.

E Confirmation Protocol and Semantic Review

For each task, the two models share 64 calibration and 256 confirmation questions, disjoint by task-qualified ID, source index, and normalized query; the confirmation questions are also disjoint from the diagnostic sample. A fixed random seed shuffles eligible questions after duplicate and prior-input exclusion. Each arm uses the same prompt, top three retrieved passages, 32-position canvas, and 32 deterministic passes. EndFirst applies $2(32 - t^\star)/32$ once at the baseline's first proposed non-EOS write. The every-step EOS bias subtracts a constant from non-EOS logits at every step; per group the constant is selected on calibration questions to minimize the difference in mean normalized output length from the schedule, without EM or F1, and then frozen. Length matching is enforced on calibration questions only.

For scoring, we lowercase the decoded answer, delete ASCII punctuation and English articles, and collapse whitespace. EM is exact equality to a normalized reference alias, and F1 uses multiset token overlap; both take the maximum over all aliases, and empty or unresolved outputs remain in the denominator. The 95% intervals are paired percentile bootstrap intervals over $10{,}000$ resamples of questions with a fixed seed. Table 7 applies the same protocol to the $1{,}000$-question sets. Table 6(a) gives full per-group EM, F1, and mean words for baseline, EndFirst, and every-step EOS bias; panel (b) gives EM differences with paired intervals and discordant counts. Each group has $n = 256$. Figure 6 shows EndFirst's per-group gains with their paired intervals, and Table 6 gives the full statistics.

An AI-assisted qualitative review of the 14 exact-match gains in the random diagnostic sample finds that in 78.6% (11) EndFirst commits cleanly to the entity the model had already found; two select a reading of an ambiguous question, and one repairs a role assignment.

Calibration and paired comparison. The selected every-step coefficients are 1.5, 1.75, 1.5625, and 1.875 in the table's row order. The initial grid is $\{0, 0.25, 0.5, 1, 2, 4\}$; bracket refinement minimizes absolute mean-word error, breaking ties toward smaller coefficients, with a 0.2-word stopping tolerance and at most six refinements. The selected runs used one to four refinements, so each group has 7 to 10 calibration runs. EndFirst and the every-step EOS bias have the same pre-action canvas at the EndFirst trigger on all 256 questions per group. Their trigger actions agree on 75.4%, 87.5%, 75.0%, and 90.2% of the 256 items. On confirmation, every-step-minus-EndFirst EM intervals contain zero in all four groups.

Table 6: Full confirmation statistics. Panel (a) gives EM and F1 (%) and mean normalized words per arm; panel (b) gives EM differences in points with paired 95% intervals. Discordant: questions only EndFirst / only every-step answers correctly.

(a)		Standard			EndFirst			Every-step		
Model	Task	EM	F1	Words	EM	F1	Words	EM	F1	Words
8B	NQ	14.84	29.84	12.80	22.66	37.41	7.98	23.44	37.85	8.30
8B	HotpotQA	4.30	13.30	15.98	11.72	21.21	10.27	11.72	20.84	10.17
1.5	NQ	14.06	29.17	12.86	21.48	36.49	8.18	22.66	36.59	8.25
1.5	HotpotQA	4.69	13.77	15.98	10.94	20.06	10.32	11.72	20.93	10.27

(b) Model	Task	λ	EndFirst −Standard	Every-step −EndFirst	Discordant
8B	NQ	1.5	7.81 [4.69, 11.33]	0.78 [−2.73, 4.30]	9/11
8B	HotpotQA	1.75	7.42 [4.30, 10.94]	0.00 [−1.95, 2.34]	4/4
1.5	NQ	1.5625	7.42 [4.30, 10.55]	1.17 [−1.95, 4.69]	8/11
1.5	HotpotQA	1.875	6.25 [3.52, 9.38]	0.78 [−0.78, 2.34]	1/3

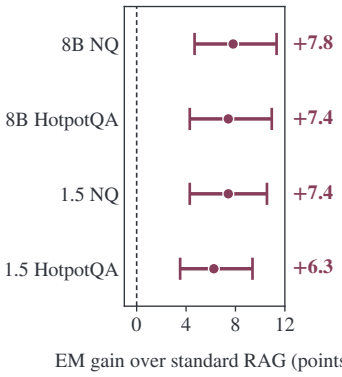


Figure 6: Held-out confirmation (256 questions per group): EM gain of EndFirst over standard RAG with paired 95% intervals. Right column: gain in points.

Control lengths and budgets. The 16-position canvas has mean normalized lengths 8.13/8.01 words on NQ and 7.91/7.85 on HotpotQA for LLaDA-8B/1.5; EndFirst gives 6.60/6.61 and 8.91/9.06. Short canvases change both canvas length and the number of forward passes. The 1,000-question control arms are reported in EM.

F Natural-Candidate Branches

Five selected cases have 28 total continuations: 18 natural-candidate branches plus a baseline and EndFirst trajectory for each case. Candidates are ranked using the original model distribution; reference labels enter only at scoring. The cases establish multiple available continuations at shared states. On all 128 random diagnostic items, replaying with EndFirst's first write changes 27.3% [19.5, 35.2] of final answers: 49.3% [38.0, 60.6] of the items whose executed action changes, and 36.0% [25.8, 46.1] of EOT-first items. The token-only and position-only arms change 52.3% and 54.7% of answers, but reach only 20.3% and 7.0% EM against EndFirst's 25.8%. As a separate check on all 128 diagnostic items, the token-only and position-only arms (EndFirst's first token at the baseline's position, and the baseline's first token at EndFirst's position) cause 7 and 11 exact-match losses relative to the baseline, versus none for EndFirst.

Table 7: 1,000-question comparison with the every-step EOS bias tuned to EndFirst's answer length. Panel (a) gives EM and F1 (%) and mean normalized words per arm; panel (b) gives EM differences in points with paired 95% intervals (protocol as above). Discordant: questions only EndFirst / only every-step answers correctly.

(a)		Standard			EndFirst			Every-step		
Model	Task	EM	F1	Words	EM	F1	Words	EM	F1	Words
8B	NQ	19.6	33.7	11.7	29.3	43.0	6.6	30.6	43.8	6.5
8B	HotpotQA	6.5	16.2	15.2	14.6	23.6	8.9	14.4	23.3	8.8
1.5	NQ	20.5	34.6	11.5	28.4	42.7	6.6	30.7	44.3	6.6
1.5	HotpotQA	6.6	16.3	15.0	14.7	23.9	9.1	15.2	24.2	8.6

(b)			EndFirst	Every-step		Same
Model	Task	λ	−Standard	−EndFirst	Discordant	answer (%)
8B	NQ	1.5	9.7 [7.9, 11.6]	1.3 [−0.1, 2.7]	19/32	80.8
8B	HotpotQA	1.75	8.1 [6.4, 9.9]	−0.2 [−1.6, 1.2]	27/25	78.0
1.5	NQ	1.5625	7.9 [6.3, 9.6]	2.3 [1.0, 3.7]	12/35	82.2
1.5	HotpotQA	1.875	8.1 [6.4, 9.9]	0.5 [−0.8, 1.8]	18/23	81.1

G Answer Precision, Evidence, and the Constant

All numbers in this appendix use the scoring and paired-bootstrap protocol of Appendix E. The 16 LLaDA settings are the eight of Table 1 (ARAM, 64 passes, 1,000 questions), standard RAG on NQ and HotpotQA (32 passes, 1,000 questions; Table 11), and the four confirmation groups (256 questions).

Why exact match. We examine the 2,120 LLaDA baseline outputs that contain a normalized reference alias but fail EM: all such outputs on the standard-RAG and confirmation sets, and those among the changed answers of Table 1. In 13.0% of them the matched alias also appears verbatim in the question, so restating the question satisfies containment, and in 3.2% the match is not at a word boundary. To estimate how often such an output asserts a competing answer (an alternative value, a disjunction, or a list with non-reference items offered as the answer; entities in other roles do not count), fixed lexical rules flag candidate outputs, and an AI-assisted manual audit labels 30 random outputs from each of three strata (rule-flagged, entity-flagged, and unflagged). The stratified estimate is 17.4% [9.8, 25.0], and in 9.3% (±5.6) the reference is not the first candidate stated. Containment counts all of these as correct; EM credits an answer only when the model commits to it. Among EndFirst's fixes, the alias its answer matches already appears word for word in the baseline output in 65.5 to 93.8% per LLaDA setting, usually in mid-sentence: EndFirst turns an answer the model had found into a clean commitment.

Post-hoc truncation. Truncating the baseline at its first terminal token changes no answer in any setting, because the decoded baseline already ends there. The rule-based truncation keeps the first line of the answer and cuts it at the first sentence end (`.`, `!`, or `?` followed by whitespace or the end of the string) or at the first comma followed by whitespace. The oracle takes, per question, the best of the untruncated baseline, the token-truncated baseline, the rule-truncated baseline, and both truncations combined; it uses the reference and serves only as an upper bound on any truncation of the baseline. Table 8 shows that EndFirst exceeds the truncated baseline with a paired interval above zero in 14 of the 16 settings and the oracle in 13. When the same rule truncates both outputs, EndFirst leads in all 16 settings, by 1.9 to 8.2 points, with every interval above zero. With standard RAG, truncation brings the baseline to EndFirst's length or below, e.g., 6.50 against 6.60 words for LLaDA-8B on NQ and, on confirmation, 6.69 against 7.98, and EndFirst still leads by 8.2 and 6.2 points.

Two terminal commitments first. Table 9 isolates the questions on which the baseline spends its first two commitments on terminal tokens (EOS or EOT, in either order) and commits its first content token third; membership uses the baseline trajectory only. These questions are 6.6 to 37.5% of each

Table 8: EndFirst against the truncated baseline. EM (%); differences are EndFirst minus the truncated baseline or minus the per-question truncation oracle, in EM points with paired 95% intervals. Both truncated: EndFirst minus the baseline, with the same rule truncating both outputs.

			EM (%)			EndFirst minus		Both
Decoder	Model	Task	Base	Trunc.	EndFirst	Truncated	Oracle	truncated
ARAM	8B	NQ	33.8	33.0	38.5	5.5 [3.7, 7.4]	3.4 [1.8, 5.1]	4.6 [3.2, 6.0]
	8B	HotpotQA	17.8	19.4	24.7	5.3 [3.6, 7.1]	5.0 [3.4, 6.7]	5.2 [3.6, 6.9]
	8B	TriviaQA	69.4	68.7	75.1	6.4 [4.9, 8.0]	5.6 [4.2, 7.1]	5.6 [4.2, 7.1]
	8B	T-REx	31.3	31.6	33.4	1.8 [0.8, 2.9]	1.7 [0.7, 2.8]	1.9 [0.9, 2.9]
	1.5	NQ	32.4	31.3	38.4	7.1 [5.3, 9.0]	4.7 [3.1, 6.4]	6.2 [4.7, 7.7]
	1.5	HotpotQA	16.4	18.1	24.3	6.2 [4.4, 8.0]	5.9 [4.2, 7.7]	6.3 [4.7, 8.0]
	1.5	TriviaQA	68.5	68.3	75.1	6.8 [5.3, 8.4]	6.0 [4.6, 7.5]	6.0 [4.5, 7.5]
	1.5	T-REx	29.6	29.9	32.2	2.3 [1.2, 3.5]	2.2 [1.1, 3.4]	2.6 [1.5, 3.7]
Standard	8B	NQ	19.6	21.1	29.3	8.2 [6.1, 10.3]	7.2 [5.2, 9.2]	8.2 [6.5, 10.0]
	8B	HotpotQA	6.5	10.9	14.6	3.7 [1.6, 5.7]	3.6 [1.5, 5.6]	6.7 [5.0, 8.4]
	1.5	NQ	20.5	22.0	28.4	6.4 [4.5, 8.4]	5.2 [3.4, 7.1]	6.6 [5.1, 8.2]
	1.5	HotpotQA	6.6	11.6	14.7	3.1 [1.1, 5.1]	3.0 [1.0, 5.0]	6.3 [4.7, 7.9]
Confirmation	8B	NQ	14.8	16.4	22.7	6.2 [2.0, 10.5]	5.1 [1.2, 9.0]	7.4 [4.3, 10.5]
	8B	HotpotQA	4.3	10.5	11.7	1.2 [−2.7, 5.1]	1.2 [−2.7, 5.1]	4.7 [2.0, 7.4]
	1.5	NQ	14.1	16.4	21.5	5.1 [1.2, 9.0]	3.9 [0.0, 7.8]	6.6 [3.9, 9.8]
	1.5	HotpotQA	4.7	10.2	10.9	0.8 [−3.1, 4.7]	0.4 [−3.5, 4.3]	4.3 [1.6, 7.0]

setting. The baseline answers none of them correctly; EndFirst answers 20.5 to 58.3% correctly and breaks none. They hold 75.0 to 93.4% of EndFirst's fixes.

Table 9: Questions on which the baseline (ARAM) spends its first two commitments on terminal tokens and commits its first content token third, for the eight LLaDA settings of Table 1. Share: percentage of the 1,000 questions. Fixes: questions fixed here, out of all questions EndFirst fixes in the setting.

Model	Task	Share (%)	Baseline EM	EndFirst EM	Gain [95% CI]	Fixes
8B	NQ	17.1	0.0	28.1	28.1 [21.6, 35.1]	48 / 53
8B	HotpotQA	31.7	0.0	21.5	21.5 [17.0, 26.2]	68 / 76
8B	TriviaQA	8.4	0.0	58.3	58.3 [47.6, 69.0]	49 / 59
8B	T-REx	6.6	0.0	27.3	27.3 [16.7, 37.9]	18 / 24
1.5	NQ	21.2	0.0	26.9	26.9 [21.2, 33.0]	57 / 61
1.5	HotpotQA	37.5	0.0	20.5	20.5 [16.5, 24.5]	77 / 85
1.5	TriviaQA	10.1	0.0	56.4	56.4 [46.5, 66.3]	57 / 67
1.5	T-REx	9.3	0.0	23.7	23.7 [15.1, 32.3]	22 / 29

Evidence containing the answer. We split questions by whether a normalized reference alias occurs in the three retrieved passages. Every input has exactly three passages. Table 10 shows that EndFirst's gain concentrates on questions whose evidence contains the answer, in every LLaDA setting.

Dose strength c. The value $c = 2$ was fixed before the main experiments and not selected by any sweep; the $|\mathcal{M}|/L$ form was developed on 64 development questions per task for NQ and HotpotQA, which share no question, by ID or normalized query, with the evaluation, calibration, or confirmation sets. Table 11 varies c with standard RAG on the 1,000-question sets. EM rises monotonically with c from 1 to 4 in all four groups, and $c = 4$ exceeds $c = 2$ by 6.6 to 7.8 points, e.g., +7.2 on NQ and +7.8 on HotpotQA for LLaDA-8B. On the 64 calibration questions per group, all 20 runs ($c \in \{1, 2, 4, 8, 16\}$ in four groups) are complete. There, $c = 4$ beats $c = 2$ in all four groups by 6.2 to 9.4 points, every paired interval above zero (23.4 to 31.2 against 15.6 to 21.9 EM), while answers shorten from 7.3 to 7.7 normalized words to 3.0 to 3.9. The default $c = 2$ is the smallest grid dose at which the mean calibration answer in every group falls within the 1 to 10 words the prompt asks for (7.3 to 7.7 words; $c = 1$ leaves HotpotQA at 11.8 and 12.3). The larger doses $c = 8$ and $c = 16$ stay at or above $c = 4$ in every group, by at most 4.7 points, every paired interval touching zero, while answers shrink to 1.7 to 2.3 words. The reported $c = 2$ is therefore a conservative, untuned setting. Against ARAM at 64 passes, standard RAG with EndFirst at $c = 4$ and 32 passes gains +2.7 $[0.9, 4.6]$, +4.6 $[2.6, 6.7]$, and +2.1 $[0.5, 3.7]$ EM on NQ, HotpotQA, and TriviaQA for LLaDA-8B and +2.6 $[0.8, 4.5]$, +5.8 $[3.8, 7.9]$, and +2.7 $[1.0, 4.5]$ for LLaDA-1.5

Table 10: EM gain of EndFirst over its base decoder, split by whether the retrieved passages contain a reference alias. EM points with paired 95% intervals. Share: percentage of questions with the alias in the passages.

Decoder	Model	Task	Share (%)	Alias in passages	Alias not in passages
ARAM	8B	NQ	68.2	6.6 [4.5, 8.7]	0.6 [0.0, 1.6]
	8B	HotpotQA	36.0	14.2 [10.3, 18.3]	2.8 [1.4, 4.2]
	8B	TriviaQA	87.3	6.3 [4.7, 8.0]	1.6 [0.0, 3.9]
	8B	T-REx	40.5	4.4 [2.5, 6.7]	0.5 [−0.3, 1.3]
	1.5	NQ	68.2	8.4 [6.3, 10.6]	0.9 [0.0, 2.2]
	1.5	HotpotQA	36.0	16.1 [12.2, 20.0]	3.3 [1.6, 5.0]
	1.5	TriviaQA	87.3	7.3 [5.6, 9.2]	1.6 [0.0, 3.9]
	1.5	T-REx	40.5	4.9 [2.7, 7.4]	1.0 [0.3, 1.8]
Standard	8B	NQ	68.2	14.1 [11.6, 16.7]	0.3 [0.0, 0.9]
	8B	HotpotQA	36.0	17.5 [13.9, 21.7]	2.8 [1.2, 4.4]
	1.5	NQ	68.2	11.4 [9.1, 13.9]	0.3 [0.0, 0.9]
	1.5	HotpotQA	36.0	16.4 [12.8, 20.3]	3.4 [2.0, 5.0]
Confirmation	8B	NQ	66.4	11.2 [6.5, 15.9]	1.2 [0.0, 3.5]
	8B	HotpotQA	29.3	16.0 [8.0, 24.0]	3.9 [1.1, 6.6]
	1.5	NQ	66.4	10.6 [6.5, 15.3]	1.2 [0.0, 3.5]
	1.5	HotpotQA	29.3	16.0 [8.0, 25.3]	2.2 [0.6, 4.4]

(paired bootstrap over the 1,000 questions, 10,000 resamples); the paired F1 intervals also lie above zero. On the 256 held-out confirmation questions per group, $c = 4$ reaches 21.5 to 32.8 EM, 16.8 to 18.4 points above standard RAG and 10.2 to 10.9 above $c = 2$, every paired interval above zero, and it breaks no answer that either arm gets right. At $c = 4$, EndFirst also lies above the every-step EOS-bias frontier (fixed $\lambda \in \{1, 2, 4\}$, interpolated linearly at EndFirst's mean answer length) on HotpotQA, by 4.3 and 2.7 EM points for LLaDA-8B and LLaDA-1.5.

Table 11: EndFirst beats every alternative at equal passes. EM (%), 1,000 questions per group; bold/italics: best two; small: gain over the best alternative; c=4 selected on 64 calibration questions per group. †Tuned to EndFirst's answer length.

		LLaDA-8B		LLaDA-1.5	
Method	Passes	NQ	HotpotQA	NQ	HotpotQA
Standard RAG	32	19.6	6.5	20.5	6.6
4-position canvas	4	19.0	12.9	18.8	12.5
8-position canvas	8	13.6	7.9	12.5	7.3
16-position canvas	16	2.3	3.2	2.6	2.6
Every-step, λ=0.25	32	22.1	7.7	22.5	7.7
Every-step, tuned†	32	30.6	14.4	30.7	15.2
Frontier-1	32	19.4	8.7	21.6	10.0
Frontier-16	32	21.2	9.9	23.6	11.2
SPREAD	33	22.8	12.4	24.8	13.0
UNCODE	32	21.2	10.5	24.0	12.4
EndFirst (c=1)	32	24.7	10.5	24.6	10.8
EndFirst (c=2)	32	29.3	14.6	28.4	14.7
EndFirst (c=4)	32	*36.5*+5.9	*22.4*+8.0	*35.0*+4.3	*22.2*+7.0
EndFirst (c=8)	32	**39.8**+9.2	**26.8**+12.4	**38.5**+7.8	**25.9**+10.7

Table 12 compares $c = 8$ with $c = 4$ on the 1,000-question evaluation sets: $c = 8$ adds up to 4.4 EM points and improves every setting, fixing 249 questions and losing 11 across the eight settings.

H SCHEDULES, CANVASES, AND FURTHER BENCHMARKS

This appendix gives the protocol behind Table 3 in Section 5.4 and Figure 7. All of its settings use the default dose $c = 2$ and the scoring and paired-bootstrap protocol of Appendix E. LLaDA-MoE is a mixture-of-experts LLaDA; it runs standard RAG on the 32-position canvas at 32 passes, like the two dense LLaDA models. With semi-autoregressive block decoding (four blocks of eight positions, 32 passes, standard RAG), EndFirst's single global trigger fires exactly once on every

Table 12: Stronger dose $c = 8$ against $c = 4$ on the 1,000-question evaluation sets, standard RAG, 32 passes. EM (%); c8−c4 in EM points with paired 95% intervals (10,000 resamples). Fixes / losses: questions only $c = 8$ / only $c = 4$ answers exactly. Words: mean normalized words.

Model	Task	Standard	$c = 4$	$c = 8$	c8−c4 [95% CI]	Fixes / losses	Words $c = 4$	Words $c = 8$
8B	NQ	19.6	36.5	39.8	+3.3 [2.2, 4.5]	34 / 1	3.57	2.08
	HotpotQA	6.5	22.4	26.8	+4.4 [3.2, 5.7]	44 / 0	4.30	2.07
	TriviaQA	44.9	71.5	75.3	+3.8 [2.6, 5.1]	40 / 2	2.44	1.77
	T-REx	16.6	28.8	30.0	+1.2 [0.5, 2.0]	13 / 1	1.79	1.45
1.5	NQ	20.5	35.0	38.5	+3.5 [2.4, 4.7]	35 / 0	3.75	2.43
	HotpotQA	6.6	22.2	25.9	+3.7 [2.5, 5.0]	40 / 3	4.55	2.66
	TriviaQA	45.5	71.2	74.8	+3.6 [2.5, 4.8]	36 / 0	2.53	1.83
	T-REx	16.2	29.0	29.3	+0.3 [−0.3, 1.0]	7 / 4	2.02	1.79

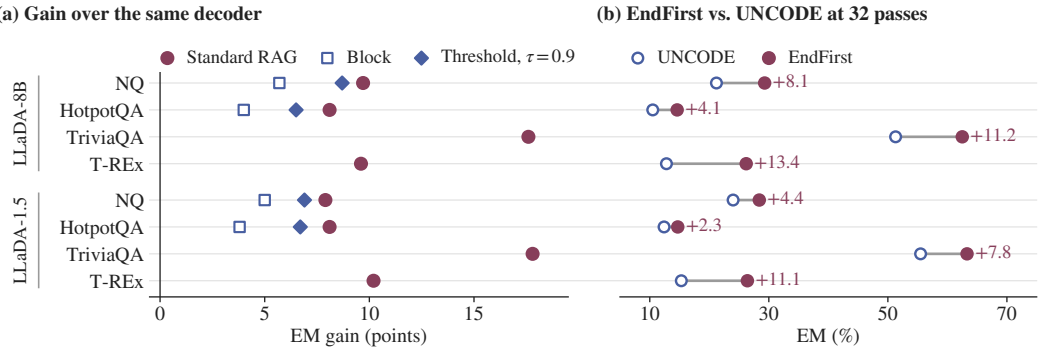


Figure 7: EndFirst ($c = 2$) raises EM under every decoder, model, and task we test. (a) Gain over the same decoder at the same passes, 1,000 questions per point. (b) EM of UNCODE and of standard RAG with EndFirst; labels give the gain in points. Base EM, LLaDA-MoE, and paired intervals: Appendix H.

question. Confidence-threshold parallel decoding commits every open position whose top-1 confidence exceeds $\tau = 0.9$ in the same pass, so the number of passes varies per question; EndFirst applies Equation 2 at the first pass that proposes a non-EOS write, and the threshold rule then selects the writes. On the 64- and 128-position canvases, L in Equation 2 is the canvas length, and the per-question budget scales with it. UNCODE replaces confidence-first selection with a priority recalibrated at every step; we run it at standard RAG's 32-pass budget on the same 1,000 questions per task and compare it with standard RAG plus EndFirst at that budget. Across all logged runs, EndFirst's run-time overhead is at most 1.6%. Table 13 gives every gain of Figure 7 with its paired interval, and Table 14 the 64- and 128-position canvas runs.

I Token-Position Decomposition

Table 15 reports the token-position controls, including answer counts, commitment timing, and output length. The arms separate the identity of the first committed token from the position where it is written, on the 31 diagnostic questions where the baseline first commits EOT and EndFirst commits EOS at that step. Exact two-sided McNemar tests confirm that the written token carries the effect. At either position, EOS is correct where EOT is not on 12 questions and the reverse never holds ($p = 0.00049$ each). Moving EOS between the two positions changes no exact-match outcome ($p = 1$), and neither does moving EOT ($p = 1$); the two EOS arms give identical outputs on all but one question. On the other 21 EOT-first questions whose first action EndFirst changes, it writes a content token at a new position, and all four arms reach the same EM of 4.8%.

J Additional Trajectory: Jenny Slate

Figure 9 shows the additional Jenny Slate trajectory. It illustrates how replacing the first content commitment changes subsequent predictions on the shared canvas.

Table 13: EndFirst ($c = 2$) across models, decoders, and tasks, and against UNCODE (Figure 7). EM (%) on a 32-position canvas. Each row pairs a base decoder with the same decoder plus EndFirst at the same passes; against UNCODE, EndFirst is standard RAG with EndFirst. Gain: EM points with a paired 95% percentile-bootstrap interval (10,000 resamples). 1,000 questions per row. Better arm in bold.

Base decoder	Model	Task	Base	EndFirst	Gain [95% CI]
Standard RAG	LLaDA-MoE	NQ	20.6	**23.7**	+3.1 [2.1, 4.2]
		HotpotQA	10.9	**13.2**	+2.3 [1.3, 3.4]
	LLaDA-8B	TriviaQA	44.9	**62.5**	+17.6 [15.2, 20.0]
		T-REx	16.6	**26.2**	+9.6 [7.8, 11.5]
	LLaDA-1.5	TriviaQA	45.5	**63.3**	+17.8 [15.5, 20.2]
		T-REx	16.2	**26.4**	+10.2 [8.4, 12.1]
Block	LLaDA-8B	NQ	19.4	**25.1**	+5.7 [4.3, 7.2]
		HotpotQA	9.5	**13.5**	+4.0 [2.5, 5.6]
	LLaDA-1.5	NQ	22.0	**27.0**	+5.0 [3.6, 6.5]
		HotpotQA	11.1	**14.9**	+3.8 [2.4, 5.2]
Threshold, $\tau = 0.9$	LLaDA-8B	NQ	19.7	**28.4**	+8.7 [7.0, 10.5]
		HotpotQA	6.5	**13.0**	+6.5 [4.9, 8.1]
	LLaDA-1.5	NQ	20.4	**27.3**	+6.9 [5.4, 8.5]
		HotpotQA	6.5	**13.2**	+6.7 [5.2, 8.3]
UNCODE	LLaDA-8B	NQ	21.2	**29.3**	+8.1 [6.0, 10.3]
		HotpotQA	10.5	**14.6**	+4.1 [2.3, 5.9]
		TriviaQA	51.3	**62.5**	+11.2 [8.6, 13.7]
		T-REx	12.8	**26.2**	+13.4 [11.2, 15.6]
	LLaDA-1.5	NQ	24.0	**28.4**	+4.4 [2.4, 6.4]
		HotpotQA	12.4	**14.7**	+2.3 [0.5, 4.1]
		TriviaQA	55.5	**63.3**	+7.8 [5.4, 10.2]
		T-REx	15.3	**26.4**	+11.1 [9.0, 13.2]

Table 14: Longer canvases, LLaDA-8B with standard RAG. Columns as in Table 13: Base is the base arm's EM (%), EndFirst the same arm with EndFirst at $c = 2$, Gain the EM difference in points with a paired 95% interval over the n scored questions. The per-question pass budget scales with the canvas length.

Setting	Task	Base	EndFirst	Gain [95% CI]	n
64-position canvas	NQ	34.8	38.2	+3.4 [1.8, 5.0]	500
	HotpotQA	21.6	27.4	+5.8 [3.8, 8.0]	500
128-position canvas	NQ	39.2	39.4	+0.2 [−0.4, 1.0]	500
	HotpotQA	28.0	28.0	0.0 [0.0, 0.0]	500

K Reweighting Illustration

Figure 4 illustrates the three properties of Section 4 on illustrative two-position distributions.

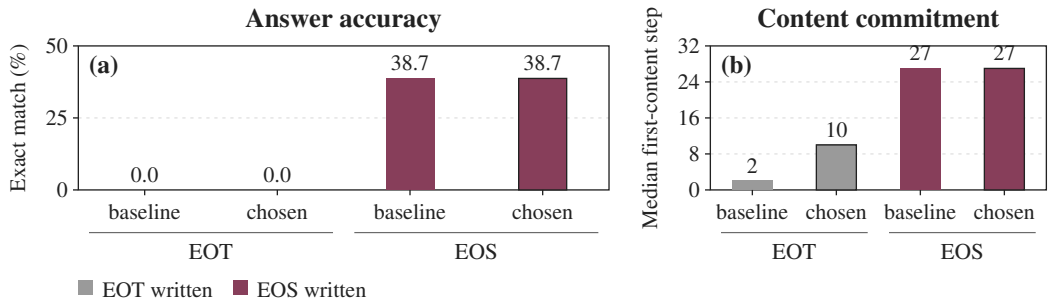


Figure 8: Writing EOS recovers the answer. Where the baseline first commits EOT and EndFirst replaces it with EOS ($n = 31$; two LLaDA models, standard RAG), moving EOT to EndFirst's position leaves exact match at 0.0%, whereas EOS reaches 38.7% at either position. (a) Exact match by token and position. (b) Median first-content commitment step. Baseline/chosen: the baseline's EOT position or EndFirst's. Appendix Table 15 gives lengths and further measures.

Table 15: Token identity and write position, separated. EM is in percent and empty-answer entries are counts, both over the $n = 31$ items. Diagnostic subset where the baseline first commits EOT and EndFirst's re-selection commits EOS at that step ($n = 31$), two LLaDA models, standard RAG. Committing EOS yields the same exact match whether it is written at the baseline's EOT position or at the position EndFirst chose; committing EOT at the chosen position does not. Writing EOS moves the median first content commitment from step 2 to step 27.

Arm	Token @ position	Mean pos.	EM (%)	Empty	First content commit (median)	Words
Baseline	EOT @ baseline	28.2	0.0	0	2	16.35
Position-only	EOT @ chosen	20.2	0.0	0	10	11.74
Token-only	EOS @ baseline	28.2	38.7	0	27	1.74
EndFirst	EOS @ chosen	20.2	38.7	0	27	1.74

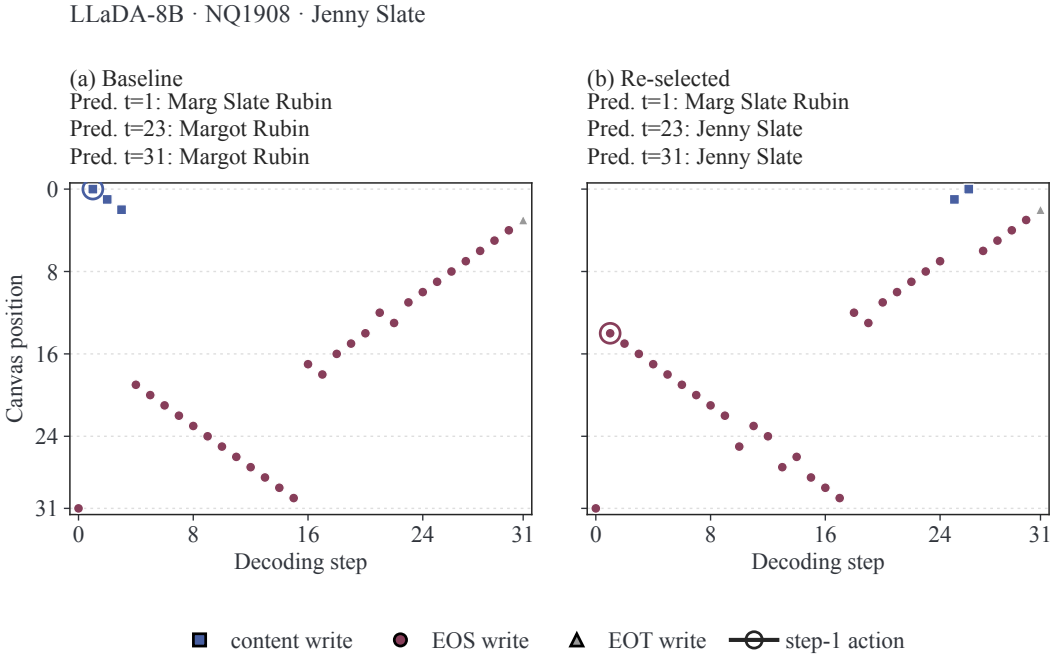


Figure 9: A second logged trajectory shows a change from "Margot Rubin" to "Jenny Slate" (LLaDA-8B, NQ1908; selected real case). Top-1 decoded answer excerpts are shown at selected steps; lower panels plot all 32 writes by step and canvas position. Black rings mark the first commitment. Colors mark token class only.