Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Query Cannot See the Question: A Short Convolution's Reach Decides Which Part of a Query Conditions Retrieval, and What Falls Outside Becomes a Confident Wrong Answer

Maximiliano Rodrigo Speranza

mrsperanza@frba.utn.edu.ar

Independent Researcher https://orcid.org/0009-0005-0413-8554

Research Article

Keywords: state space models, linear attention, associative recall, abstention, selective prediction, mechanistic interpretability, retrieval, hallucination, pre-registration, negative results

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10934426/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The Query Cannot See the Question

A Short Convolution's Reach Decides Which Part of a Query Conditions Retrieval, and What Falls Outside Becomes a Confident Wrong Answer

Maximiliano Speranza
Independent Researcher, Buenos Aires, Argentina
ORCID 0009-0005-0413-8554
mrsperanza@frba.utn.edu.ar

September 2026

Abstract

Linear-attention and state-space models form their query with a short causal depthwise convolution. Kernel size 4 is the de facto default, so the query at a given layer is a function of the current token and three predecessors. We show that this reach is not a detail of local smoothing but a hard limit on *which part of a question is allowed to condition retrieval*, and that what falls outside does not degrade gracefully: it disappears exactly, and the model answers confidently from the part it can still see.

We measure the effect in a small language model with a co-trained persistent archive queried across sequences. The sensitivity of retrieval to a query token, measured as the total-variation movement of the read distribution when that token is replaced, is **0.000000** as soon as the token passes the convolution's reach, and the cut-off moves with the kernel: with kernel 3 (reach 2) the step falls between distances 2 and 3, and with kernel 5 (reach 4) between 4 and 5. Sixty cells out of sixty, two architectures, three seeds each, two query components, five distances.

The behavioural consequence is a specific and common failure. In our task, questions have the form *what is the <**relation**> of <**entity**>?*, with the entity at distance 1 from the read position and the relation at distance 3. With kernel 3, the relation is outside the window in 100% of queries, so the model retrieves on the entity alone. When asked about a relation that was never stated for an entity that *was*, it answers from the wrong stored fact instead of abstaining: correct abstention is 0.5850 to 0.7349 across three seeds. Widening the kernel to 5, which costs 1,024 parameters out of 865,651, lifts it to **0.9931** to **1.0000** with no overlap between conditions, and overall accuracy to 0.988–0.993 against a trivial floor of 0.4065.

We then check the mechanism outside our own model. In `mamba-130m-hf`, 129M parameters, changing one token five to eight positions before the read moves the layer output at every distance and leaves the convolution output at **exactly zero**, 80 cells out of 80. The state sees the whole sequence; the query that reads it does not. Incidentally, the oldest convolution tap in that checkpoint is exactly zero in all 24 layers, so its effective reach is 2 rather than the nominal 3; we report the measurement and not a cause.

Depth changes the law without repealing it, and we measure by how much. In a recurrent model the combination of two query components happens wherever both are available, so the distance that decides is the one between them. Across fifteen question forms, six distances and two checkpoints —`mamba-130m` with 24 layers and `mamba-370m` with 48— the cut in layer 0 is exact and falls at the measured reach in both, while from layer 1 recurrence restores the signal *attenuated*, at 1.077 and 1.028 per token of distance ($r = 0.978$ in both, curves correlated at 0.955). Three times the parameters and twice the depth leave the rate unchanged, which makes the attenuation a property of the architecture. A control holding the distance fixed and varying the filler separates distance from lexical content.

Finally we test whether access predicts behaviour, and report a negative. Fine-tuning `mamba-130m` on the same task with the discriminating component outside the window costs

+0.167, +0.181 and +0.141 of correct abstention at step 100 — three seeds out of three above our pre-registered threshold — and **nothing at all by step** 400. Our small model, four blocks deep, never recovers; the 24-layer model recovers quickly. **Where there are no layers left to pay with, the window sets a ceiling; where there are, it sets a toll.** The strong claim therefore holds for memory consulted from an early layer, and not as a behavioural prediction for a deep model trained to convergence.

Finally we ask how often this configuration arises when nobody is arranging it. Across four question-answering corpora — SQuAD, Natural Questions, TriviaQA and HotpotQA, 33,585 questions — between 90% and 100% of real questions have their discriminating parts further apart than the measured reach, on both definitions that can be computed without annotation, and it worsens with question difficulty, reaching **1.0000** on multi-hop questions. The configuration our synthetic task was built to study is the ordinary case, not a corner case.

We give the diagnosis as a recipe that requires no training: change one token of the query and see whether retrieval moves. If it does not, that token is invisible to the search, and any confidence the model expresses about it is unearned — though, as our negative shows, invisible to one layer's query is not the same as unusable by the model.

1 The problem

A model that keeps a persistent memory has to do two things that are usually measured together and are in fact separate. It has to *find* the right entry, and it has to *know* when the entry it is looking for is not there. The second one is what makes the difference between a memory and a confabulator, and it is the harder of the two.

We found that in our own system the second failure had a purely mechanical cause, upstream of anything the model learns. The query used to search the archive is formed by a short causal convolution over the hidden states. In a question of the form *what is the* `<art>` `<noun>` *of* `<entity>`*?*, the entity sits one token before the read position and the noun that names the relation sits three tokens before it. A kernel of size 3 reaches two tokens back. The relation is one token outside the window, deterministically, in every single query.

The model was not ignoring the relation. It could not see it.

Contributions. We state them as claims, because that is what a reader has to be able to check against the evidence.

1. **The reach of the query-forming convolution is a hard cut, not a soft decay.** Retrieval sensitivity to a query token is 0.000000 as soon as the token passes the reach, and the step moves with the kernel. Sixty cells out of sixty (Result 1), and the ablation of the single tap that reaches the relation reproduces it (Result 3).

2. **Widening the window by one tap on each side buys abstention.** Correct abstention on the hard case goes from 0.5850–0.7349 to 0.9931–1.0000 with no overlap across three seeds, for 1,024 parameters out of 865,651 (Result 2).

3. **The dissociation holds in a model that is not ours.** In `mamba-130m`, the layer output moves at every distance while the convolution output stays at exactly zero, 80 cells out of 80. The state sees the whole sequence; the query that reads it does not (Result 4).

4. **In a deep model the window attenuates rather than blocks, at a rate that does not depend on model size.** 1.077 and 1.028 per token of distance in 24 and 48 layer checkpoints, $r = 0.978$ in both, with a control that separates distance from lexical content (Result 5).

5. **Access does not imply behaviour, and we report that as a negative.** A component that is provably invisible to one layer's query still gets learned, only more slowly. The effect is +0.16 at step 100 across three seeds and gone by step 400 (Result 6).

6. **The configuration is the common case, not a corner case.** Across four QA corpora and 33,585 questions, 90–100% of real questions have their discriminating parts beyond the measured reach, rising to 1.0000 on multi-hop questions (Result 7).

7. **A diagnosis that costs nothing.** Replace one token of a query and check whether retrieval moves. It requires no training, no gradient and no labels, and it applies to any model that consults a memory through a locally formed query.

What we got wrong, and report anyway. A control that could not have come out any other way, and a headline statistic that did not survive sweeping its threshold. Both are in Result 5, with the test that catches each class of mistake.

2 Related work

Short convolutions are ubiquitous and their size is inherited, not measured. Mamba-2 (Dao and Gu, 2024) uses a causal depthwise convolution of size 4 by default. In the reference implementation of Qwen3-Next, which uses Gated DeltaNet (Yang et al., 2025) for its linear-attention layers, the configuration exposes `linear_conv_kernel_dim`, documented as "kernel size of the convolution used in linear attention layers" and **defaulting to 4** (Hugging Face, 2026). Ablations in this line report that removing the convolution hurts, which establishes that local context matters; none reports what is lost when the window is present but too narrow for the query at hand. A reach of three tokens is therefore not an exotic configuration, it is the setting in deployed models.

The closest theoretical result. Li et al. (2024) study N-gram associative recall and prove, by construction, that a causal filter of length N suffices, matching the filter to the n-gram. That is the same family of idea and we cite it as the precursor. Three things separate it from what we report. Their query is a *contiguous* n-gram, the last N tokens with nothing irrelevant in between, whereas what governs our case is the *distance* of a component with filler between the two parts that matter. Their Theorem 1 is an existence result and they run no experiment in which the filter is too short. And they report the opposite in practice, stating that "while K/Q convolution helps in theoretical constructions for N-gram AR, in real experiments, they don't provide noticeable performance benefits". We measure a large effect in a regime they do not test, namely a persistent *cross-sequence* memory with a compositional query.

Why the field could not see it. The canonical benchmark for this capability, MQAR (Arora et al., 2024), has a *single-token* query: the key is looked up and the value returned. By construction such a query cannot have one part inside the window and another part outside. Zoology attributes 82% of the recall gap to the convolution applying a fixed, input-independent filter, which is a statement about *adaptivity*. Ours is about *reach*, and the two are independent: our kernel-5 model has exactly the same input-independence and the failure disappears.

From the attention side. Multi-token attention (Golovneva et al., 2025) convolves over attention weights so that attention can be conditioned on several tokens at once, motivated by the observation that a single token is not enough to locate a match. It moves in the same direction and confirms ours by contrast: it never asks which part of the query goes blind, and does not connect the question to abstention.

3 Setup

The substrate is a 3.5 MB language model, 865,651 parameters, $d = 128$, four blocks, trained from scratch on a closed synthetic language of 242 tokens with a co-trained persistent archive. Facts are stated in one sequence, revised in another and queried in a third, **with the recurrent state reset between sequences**, so retrieval cannot be solved by carrying activations forward.

Queries are read at the final token of the question. The read is injected in block 0 **before the mixer**, which matters for the mechanism: at that point the hidden state is still the token embedding, with no recurrent mixing, so the convolution's window is literally everything the query can see. This also explains why the effect below can be exactly zero rather than merely small.

Four answer categories are scored separately. `current` and `previous` are questions with an answer in the archive. `nose_ent` asks about an entity never mentioned, the easy absence. `nose_rel` asks about a relation never stated *for an entity that was*, which is the case that looks like a real hallucination, because the retrieval finds a genuine neighbouring fact.

4 Result 1: the window is a step function, and it is exactly zero outside

We measure sensitivity by replacing one component of the question with a different one of the same type and recording the total-variation distance between the read distributions. To vary distance without touching the question, the read position is advanced r places over the padding that already follows the question mark, which is exactly equivalent to appending r fillers.

model	kernel (reach)	d=1	d=2	d=3	d=4	d=5	d=6
`v3` (3 seeds)	3 (2)	0.71–0.74	0.15–0.25	**0.000000**	**0.000000**	**0.000000**	**0.000000**
`kq3` (3 seeds)	5 (4)	0.36–0.46	0.10–0.15	0.05–0.35	0.07–0.33	**0.000000**	**0.000000**

Sixty cells out of sixty match the prediction made before running, namely that sensitivity is positive for $d \leq$ reach and zero beyond it. One cell initially read 0.0106 where a zero belonged; the cause was that the "previous value" phrasing is two tokens longer, so advancing the read position clipped against the tensor bound, and clipping *shortens* the distance being measured. Those samples are now discarded rather than clipped and the cell reads 0.000000.

5 Result 2: widening the window buys abstention

unit	kernel	`current`	`nose_ent`	`nose_rel`	false abstention
`kq3_s0`	5	1.0000	0.9538	**0.9931**	0.0000
`kq3_s1`	5	0.9964	0.9726	**1.0000**	0.0030
`kq3_s2`	5	1.0000	0.9356	**1.0000**	0.0000
`v3_s0`	3	1.0000	0.9851	0.6090	0.0000
`v3_s1`	3	1.0000	1.0000	0.5850	0.0000
`v3_s2`	3	0.9927	0.9414	0.7349	0.0064

There is no overlap between conditions on `nose_rel`. Overall accuracy, counting a correct abstention as correct, is 0.988–0.993 for kernel 5 against a trivial floor of 0.4065, the score of a model that abstains on everything.

The cost is real and we state it. `nose_ent`, the easy case, drops slightly, from 0.941–1.000 to 0.936–0.973. A reading that only reports a pooled abstention number hides this trade.

It is not capacity. The kernel-5 model adds 1,024 parameters, 0.12%, and the kernel-3 control already retrieves the right entry with probability 1.0000. What changes is access, and Result 1 measures the access directly.

6 Result 3: ablating the tap that reaches the relation

Turning off one convolution tap at a time in the trained kernel-5 model, without retraining, damages `current` by 0.483, 0.582 and 0.530 for the tap that reaches the relation, against 0.372, 0.013 and 0.195 for the tap that reaches a function word and 0.218, 0.093 and 0.157 for the tap that reaches the article. The relation tap has the *smallest* learned magnitude of the five, so per unit of weight it costs about twice as much as either control.

A pre-registered criterion phrased on the abstention metric failed, and we report why: ablating the tap removes the relation from *every* query, so the model abstains more, and a metric that rewards abstention cannot fall. The damage appears in accuracy and in false abstention instead.

7 Result 4: the dissociation holds in a real model

The results above are measured in a small model, so we checked the mechanism in one that is not ours: `state-spaces/mamba-130m-hf`, 129,135,360 parameters, on CPU, with no training. In Mamba the short convolution is applied to x *before* B, C and Δ are computed, and C_t is the analogue of the read query.

We change a single token at distance d from a read position and measure, in layer 0, the movement of the convolution output and of the layer output. Ten read positions, two texts.

	d=1	d=2	d=3	d=5	d=7	d=8
`conv1d` output	0.8–3.7	0.5–1.6	**0.0**	**0.0**	**0.0**	**0.0**
layer output, median	0.164	0.043	0.022	0.018	0.040	0.012
range over 10 cells	0.09–0.96	0.02–0.06	0.01–0.05	0.01–0.04	0.01–0.10	0.01–0.02

The layer output moves at *every* distance, 80 cells out of 80, the smallest movement over all of them being 6.3×10^{-3}: the state does see the whole sequence. The convolution output is exactly zero beyond its reach. **That is the dissociation.** The model sees the context; the query that reads its state does not.

An incidental finding, reported with its uncertainty. We had predicted movement at $d = 3$, since the kernel is 4. It is exactly zero. The reason is that **the oldest tap of the convolution is exactly zero in all** 24 **layers**, which we confirmed both from the weights and by intervention: changing one token moves the convolution output at exactly three consecutive positions, not four. So the effective window in this checkpoint is 3, not 4, and the reach is 2. We do not know the cause. A weight that is exactly zero across 24×1536 entries does not come from a gradient, so a conversion artefact is plausible, but we did not verify that and we do not claim it. The fact is reproducible in three lines. Its practical consequence, if it generalises, is that the nominal kernel is an upper bound on the reach and should be measured rather than assumed.

8 Result 5: in a deep model the window attenuates rather than blocks, and the rate is the same in two model sizes

Result 4 measured access in layer 0 of a real model. The obvious objection is that a deep recurrent model has many layers in which to move information forward, so a hard zero in one layer need not matter. We measured that directly, and the answer is more useful than either extreme.

What is measured, and why here the distance is between the two *parts*. In our small model the query is formed at the last token and reads an external archive, so the distance that decides is the one from each component to the read position. A recurrent language model has no external archive: the memory is the state, it is updated at *every* position, and each token gets its own turn to condition retrieval as it enters. Answering requires *combining* the entity and the relation, and that combination can happen at any position where both are available. The distance that decides is therefore the one separating the relation from the entity, and we measure the movement of the `conv1d` output *at the entity's position* when the relation token is replaced.

Setup. Fifteen question forms spanning six relation-to-entity distances, eight random contexts of four facts each, all layers, in `state-spaces/mamba-130m-hf` (24 layers) and `state-spaces/mamba-370m-hf` (48 layers). The measured reach of the convolution is 2 in both. Attenuation is reported against the mean of the d=2 forms.

	d=2	d=3	d=4	d=5	d=6	d=7
layer 0 (both models)	moves	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
attenuation at layer 1, `130m`	1.10	2.31	3.06	4.64	4.65	6.92
attenuation at layer 1, `370m`	1.04	3.21	3.17	4.47	5.72	6.47

- **In layer 0 the cut is exact and falls at the measured reach**, in both models: d=2 moves, $d \geq 3$ is exactly zero. This is Result 1 reproduced in a model that is not ours and at a different depth.
- **From layer 1 on it is no longer zero.** Recurrence does carry the relation forward. But the signal arrives attenuated, and the attenuation grows with distance: a linear fit gives 1.077 per token ($r = 0.9784$) in `130m` and 1.028 per token ($r = 0.9777$) in `370m`. The two curves correlate at 0.9549.
- **Three times the parameters and twice the depth do not change the rate.** That makes the attenuation a property of the architecture rather than of a particular checkpoint.

The control that adjudicates, and one that did not. If attenuation depended on which words sit in the gap rather than on how many, the effect would be lexical and not architectural. We therefore held the distance fixed and varied the filler: at d=5, four different fillers (*of the person named*, *of that other person*, *in the file for*, *of our dear friend*) spread by a factor of 1.98, against a range of 1.10 to 6.92 across distances. Distance decides; the filler does not.

We report a control that failed for an instructive reason. We first tried holding d=2 and lengthening the question *after* the entity. It returned values identical to the base condition to the last decimal in all 24 layers, and it had to: nothing that follows a position can affect that position. It was causality wearing the costume of a control, and it is worth stating the test that catches this class of mistake before the data arrive — *if the hypothesis were false, could this control come out differently?*

A headline we discarded. Reading off the layer at which attenuation first drops below 1.5 gives $0, 9, 9, 13, 19, 21$ for $d = 2 \ldots 7$, about four layers per token of distance, with $r = 0.971$. Sweeping the threshold from 1.2 to 2.5 moves that slope from 3.97 to 1.54 and breaks the ordering, so the quantity is an artefact of the threshold and we do not report it. The layer-1 attenuation depends on no threshold, and that is what the paragraph above states.

What this changes about the claim. The strong form of the law — exactly zero outside the reach — holds *per layer*, and holds in the layer where a read is injected. In a deep model it becomes a law of attenuation: the window does not decide what the model can eventually combine, it decides what that combination costs. This is the quantitative version of the limitation we stated rather than measured in the first draft, and it sharpens where the strong claim applies: memory consulted from an early layer, where there are no layers left to pay with.

9 Result 6: the access law does not become a behavioural failure once the model has layers to pay with

Results 1–5 measure *access*. This one asks whether access predicts *behaviour*, and the answer is a clean negative that we think is the most useful part of the paper, because it bounds the claim rather than extending it.

Setup, pre-registered before running. We fine-tune `state-spaces/mamba-130m-hf` on the same task, four facts in context, the answer a single token, and evaluate on the form the model was trained on. Three conditions, three seeds each, 800 steps, 512 examples per evaluation: `near` trains and evaluates on d=2, where relation and entity share a window in layer 0; `far` on d=5, where the movement in layer 0 is exactly zero; `far+near` mixes both. The blocking criterion was `current` ≥ 0.90 and the primary one was `near` $-$ `far` ≥ 0.10 in ≥ 2 of 3 seeds.

The result, as a curve, because the endpoint alone would hide it.

step	`near`	`far`	`near` − `far`, per seed		
100	0.9850	0.8221	**+0.167**	**+0.181**	**+0.141**
200	0.9881	0.9209	+0.095	+0.106	+0.000
300	0.9881	0.9489	+0.095	+0.011	+0.012
400	0.9881	0.9886	−0.012	+0.011	+0.000
800	0.9485	0.9921	−0.131	+0.000	+0.000

At step 100 the effect is present, large, and unanimous: three seeds out of three above the pre-registered threshold, with a mean of +0.163. By step 400 it is gone. **The window decides how much the model has to spend to learn the discrimination, not whether it can learn it.**

How the pre-registered criteria read, and we read them as written. At the 800-step horizon all three `far` seeds finish at ≥ 0.95, so the ceiling guard we declared in advance fires and the primary criterion is **not evaluable**: there is no headroom left for `near` to be better. The velocity criterion, an area under the curve declared post-hoc after a pilot but before this campaign, gives +0.022, +0.040, +0.021 — three positive signs out of three, and all well below the 0.10 it asked for. It **does not pass**. Mixing forms bought nothing here (+0.013, +0.009, −0.003), and the no-harm criterion passes (false abstention ≤ 0.0063).

The area under the whole curve was the wrong statistic for a transient effect: it averages eight points of which only the first two carry signal, so it dilutes +0.16 into +0.03. We report that as a lesson rather than as a rescue — the criterion stands as written and does not pass.

Why this is coherent with Result 5 rather than a refutation of it. Result 5 measured that the window's cut is exact in layer 0 and that recurrence restores the signal over the following layers, attenuated but present. A model with 24 layers therefore has room to pay the toll, and what we observe is exactly the price being paid: a slower start, then parity. The contrast with our small model is the point, and it is a controlled one, since the two vehicles differ in depth:

	depth	effect of keeping the discriminating part outside the window
small model, 26,000 steps	4 blocks	**permanent**: 0.544, 0.678, 0.506 against 0.993, 0.980, 1.000
`mamba-130m`, 800 steps	24 layers	**transient**: +0.16 at step 100, zero from step 400

Where there are no layers left to pay with, the window sets a ceiling. Where there are, it sets a toll. The strong claim of this paper therefore holds for memory consulted from an early layer — retrieval-augmented reads injected before the mixer, shallow adapters, single-layer memory heads — and does not hold, as a behavioural prediction, for a deep model trained to convergence on the task.

What we are not entitled to conclude. That the effect is negligible in general. Our task has four facts and saturates, our horizon is 800 steps, and the distance we could probe tops out at d=7 because the attenuation is measured in a fixed template. A larger distance, a harder task, or a shallower model could all move this. What the negative does establish is that *measuring access is not enough*: a component that is provably invisible to one layer's query can still govern the answer, and any claim that a model "cannot use" such a component has to be tested behaviourally rather than inferred from the convolution.

10 Result 7: in real questions, the discriminating parts are almost always outside the window

Everything above was measured on a closed synthetic language where the generator fixes the distance. The obvious question is what that distance looks like when nobody is fixing it. We measured it.

What is measured. The distance X between the parts of a question that have to be combined, in tokens of the Mamba BPE. That distance has to be *defined*, and the definition decides the number, so we report three and believe only what survives all three:

- X_1, the span from the first to the last content word;
- X_2, from the last content word to the end of the question;
- X_3, from the interrogative to the last capitalised token, a proxy for the entity.

Content words are everything outside a closed list of function words. Four corpora, chosen to differ in who wrote the questions and why.

corpus	who wrote the questions	$P(X_1 > 2)$	$P(X_3 > 2)$
SQuAD	annotators looking at a paragraph	0.9625	0.9045
Natural Questions	**real Google searches**	**0.9978**	—
TriviaQA	trivia written by humans	0.9943	0.9104
HotpotQA	multi-hop, two inferential steps	**1.0000**	0.9624

n for X_1 is 10,570, 3,610, 12,000 and 7,405 questions respectively. For X_3 it is 6,610, —, 8,326 and 5,269, that is 62% to 71% of each corpus, because X_3 is defined only when the question contains an interrogative followed by a later capitalised token; in Natural Questions, which is lowercased throughout, it is never defined. The threshold is the reach we *measured* in Mamba, which is 2. At the nominal reach of 3 the figures for X_1 are 0.9144, 0.9900, 0.9837 and 0.9997.

Three things this says. It replicates. Between 90% and 100% of real questions have their discriminating parts further apart than the reach of the query-forming convolution, across four corpora written for different purposes. Natural Questions carries the most weight here, because those are queries people actually typed into a search engine rather than items produced for a benchmark.

It gets worse with difficulty. $0.9625 \to 0.9943 \to 0.9978 \to 1.0000$, from SQuAD to HotpotQA. Of 7,405 multi-hop questions, not one has its content parts within two tokens. The regime where one would most want the model to reason carefully is the regime where its query sees least.

X_2 confirms the mechanism rather than contradicting it. The last content word sits 0.18 tokens from the end in Natural Questions and 1.25–1.34 in the others, so the final piece is almost always *inside* the window and everything else is outside. That is the geometry of our synthetic task — entity inside, relation outside — appearing in language nobody wrote for this experiment.

The definition bounds the number from above, and we say by how much. X_1 assumes the two parts that must be combined are the outermost content words, which makes it an upper bound. The opposite assumption, that they are the two *last* content words, is a lower bound, and it gives 0.186, 0.211, 0.295 and 0.264. That version has to be computed over content *words* and not over BPE pieces: in three of the four corpora the last two content *tokens* belong to the same word 90% to 94% of the time, so a sub-token version of the same quantity measures the width of a single word and bounds nothing. Which figure is the right one depends on which two parts actually discriminate, and our corpora do not annotate that. X_3, the interrogative-to-entity distance, is the proxy closest to the quantity we mean, and it sits at 0.90 to 0.96. We report the bracket rather than the most favourable end of it.

A closed form is not available, and it matters which way it fails. It is tempting to fit a normal and read off $1 - \Phi\left(\frac{\text{reach}-\mu}{\sigma}\right)$. None of the twelve distributions is normal: skewness runs from $+0.93$ to $+3.71$ and excess kurtosis up to $+20.95$, all with a heavy right tail, as a distance must, since it is bounded below and unbounded above. The right tail is precisely the region $X >$ reach that the quantity is about, so a normal fitted to the centre would understate exactly the part that decides. The empirical distribution needs no assumption: the figures above are counts.

What this does not show. These are syntactic proxies, not the true semantic parts of a question, and in Natural Questions X_3 could not be computed at all because search queries are not capitalised. More importantly, this measures the *geometry of the language*, not the behaviour of a model: Result 6 established that being outside the window delays learning rather than preventing it. What it does establish is that the configuration our synthetic task was built to study is not a corner case. It is the ordinary case.

11 The diagnosis, as a recipe

The measurement in Result 1 requires no training and no gradient. For any model that consults a memory with a query built from a local window:

1. Take a query whose parts are separated, so that some component sits at distance d from the read position.
2. Replace that component with a different one of the same type.
3. Compare the retrieval distributions.

If they are identical, that component is invisible to the search, and the model's confidence about it is unearned. With a kernel of 4, the default in deployed linear-attention layers, the reach is three tokens, so any part of a question further back than that does not enter that layer's query.

Code and data. All code, the frozen pre-registrations with their SHA256, and the raw JSON from which every number in this paper is computed are available at `https://github.com/SperanzaMax/delta-archivo`. Results 4 and 5 need no training and no GPU: they download a public checkpoint and probe it, and they reproduce in minutes on a laptop.

12 Limitations

The substrate is a small model on a closed synthetic language, and the distance between the relation and the read position is fixed by the generator, so the effect here is deterministic where in natural text it would be a distribution. The measurement is of *access*, not use: a token inside the window is not necessarily used, and indeed the article tap is inside and contributes little. Finally, the read in our model is injected before the mixer, which is what makes the zero exact, and that is why the strong claim is made for memory consulted from early layers. Result 5 measures what happens when it is not: recurrence does carry the signal forward, but attenuated by a factor of ≈ 1.05 per token of distance, replicated at two depths. What that measurement does *not* say is whether the attenuated signal suffices for the task. It is a measurement of magnitude, not of use, and in the upper layers the attenuation tends to 1 without our knowing whether that means the information was recovered or that the representation came to be dominated by something else.

13 Conclusion

A short convolution decides which part of a question is allowed to condition a memory read, and what falls outside does not arrive weakly. It does not arrive. That is a property of the architecture, it costs nothing to check, and it is invisible from the outside because the model answers with full confidence from the part it can still see.

The rest of the paper is about how far that statement travels. It survives leaving our own model, in two public checkpoints of different depth, where the cut in the first layer is exact and falls at the measured reach rather than the nominal one. It survives changing the model size, since the attenuation past the first layer runs at essentially the same rate in 24 and 48 layers. And it stops, in a way we can state precisely, when the model has depth to spend: a component outside the window is learned anyway, just later.

That last result is the one we would keep if we could keep only one. It converts a claim about access into a claim about cost, and cost is something an architect can trade. Where there are no layers left to pay with, the window sets a ceiling; where there are, it sets a toll. Memory read from an early layer, an adapter, or a single retrieval head sits on the first side of that line, and those are exactly the places where retrieval is being bolted onto language models today.

We think the most useful thing here is not the fix but the measurement, because the measurement is free. Change one token of a query, look at whether the search moves, and you learn whether the model is entitled to the confidence it is about to express. We were surprised by how often the answer was that it is not, and equally surprised, later, that being unable to see something turned out not to mean being unable to learn it.

References

Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. ICML 2024.

Yang, S., Kautz, J., and Hatamizadeh, A. (2025). Gated Delta Networks: Improving Mamba2 with Delta Rule. ICLR 2025. arXiv:2412.06464.

Hugging Face Transformers documentation, `Qwen3NextConfig`, parameter `linear_conv_kernel_dim`, default 4. Accessed 2 September 2026.

Li, M., Zhang, X., Huang, Y., and Oymak, S. (2024). On the Power of Convolution Augmented Transformer. arXiv:2407.05591.

Arora, S., Eyuboglu, S., Timalsina, A., Johnson, I., Poli, M., Zou, J., Rudra, A., and Ré, C. (2024). Zoology: Measuring and Improving Recall in Efficient Language Models. ICLR 2024. arXiv:2312.04927.

Golovneva, O., Wang, T., Weston, J., and Sukhbaatar, S. (2025). Multi-Token Attention. arXiv:2504.00927.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ventanaes.pdf