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.

Where Abstention Lives: A Pre-Registered Four-Way Comparison of Abstention Interfaces in a Small Language Model with Versioned Memory

Maximiliano Rodrigo Speranza

maximiliano.speranza@gmail.com

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

Research Article

Keywords: selective prediction, abstention, hallucination, language model memory, versioned knowledge, pre-registration, negative results

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10839567/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.

Where Abstention Lives: A Pre-Registered Four-Way Comparison of Abstention Interfaces in a Small Language Model with Versioned Memory

Maximiliano Speranza
Independent Researcher, Buenos Aires, Argentina
ORCID: 0009-0005-0413-8554
maximiliano.speranza@gmail.com

August 2026

Abstract

Teaching a language model to say "I don't know" has at least three published implementations, and they are never compared against each other. The uncertainty can be a token in the vocabulary (`[IDK]`), a separate binary head (SelectiveNet), or an entry in the memory the model reads from (pointer sentinel, the no-answer score of SQuAD 2.0). Each was introduced in a different architecture, on a different task, against a different baseline. Nobody has asked the obvious question: *holding the model, the data, the seeds and the training budget fixed, does it matter where the abstention decision lives?*

We answer it in a 863,730-parameter language model trained from scratch, with a co-trained persistent archive and versioned facts, where an answer is a single token and the metric is therefore exact. Four interfaces are compared pairwise across 27 training units, with every prediction hash-frozen before the corresponding data existed.

It matters, and the ordering is stable. A separate binary head—**129 parameters,** 0.015% **of the model**—reduces the false-abstention rate by a factor of 2 to 2.8× against the vocabulary token at matched budget in 3/3 seeds, and moves the training frontier at which abstention becomes learnable down by ≈ 0.10 of accuracy margin: with a head, a model that recovers 10 points worse can still be taught to keep quiet. Renormalizing the token's embedding vector—the explanation the `[IDK]` paper offers for its own failure in small models—**does not help**, leaving a gap of +0.1465 in false abstention against the head. The mechanism is not the norm of the vector; it is that two decisions of different natures compete for one softmax.

The memory slot fails, and fails informatively. Its attention mass converges to 0.4074, 0.4046 and 0.4020 against an empirical base rate of absent questions of 0.4048: it learned the prior, not membership. A 400-point threshold sweep rescues nothing (Youden's J between +0.038 and +0.078). Read together with an archive matching score that sits at chance (AUC 0.4984), this says something sharper than "the slot did not work": **in a co-trained memory read by softmax, absence has no representation, and adding a place for it to live is not sufficient to create one.**

We state explicitly what these results do *not* license. The claim supported here is *"separating the head makes abstention learnable in this regime"*, not *"the model knows when it does not know"*: the residual ceiling is one of calibration, not capacity (AUC 0.777–0.998), and seven pre-registered attempts to obtain the decision without labels all landed in AUC 0.50–0.67.

Keywords: selective prediction; abstention; hallucination; language model memory; versioned knowledge; pre-registration; negative results

1 The question nobody asked

A model that answers everything is a model that answers wrong when it should have said nothing. The field's responses to this fall into three architectural families, each with a canonical citation:

- **In the vocabulary.** Add a token that means "I don't know" and let the ordinary cross-entropy push probability mass onto it. This is the `[IDK]` token (Cohen et al., 2024), and it inherits the training machinery for free.
- **In a separate head.** Add an output whose only job is to decide whether to answer. This is selective prediction: SelectiveNet (Geifman & El-Yaniv, 2019) trains prediction, selection and auxiliary heads end-to-end; Deep Gamblers is the same move with a different loss.
- **In the memory.** If the model reads from a retrievable store, give the store an entry that means "nothing here". This is the pointer sentinel (Merity et al., 2016) and the no-answer score of SQuAD 2.0 (Rajpurkar et al., 2018).

These are not three versions of one experiment. They live in different architectures, address different tasks, and are each compared against a baseline without abstention rather than against each other. The comparison that would tell a practitioner *which one to build* has not been run.

There is a reason it is hard to run in a large model: to attribute a difference to the interface, everything else has to be held fixed—the same pretrained weights, the same data order, the same optimizer state, the same budget, several seeds—and in a model where a single training run is expensive, that grid is not affordable. We run it in a model small enough that the grid is affordable and the internals are inspectable, and we are explicit throughout about what that buys and what it costs (§7).

Contributions.

1. The four-way paired comparison itself, with matched seeds and matched budget: 27 training units, all verdicts emitted by scripts frozen before the data existed.
2. A *measured* refutation of the explanation that Cohen et al. (2024) give for its own failure in small models. They attribute it to the initialization of the `[IDK]` embedding; we implement the most generous version of that hypothesis and it fails exactly as the unmodified token does.
3. The location of the training frontier at which abstention becomes learnable, and how far the architecture moves it (≈ 0.10 of accuracy margin).
4. A negative on the memory slot with an identified mechanism—convergence to the prior—and the accompanying observation that the archive's matching score does not encode presence at all.
5. An explicit statement of the ceiling: what is left is calibration, not capacity, and seven label-free routes to the cut have been closed.

2 Related work

Selective prediction. The reject option is old (Chow, 1970); its modern deep form is SelectiveNet (Geifman & El-Yaniv, 2019), which separates "do I answer?" from "what do I answer?" into distinct heads trained jointly. **The separation itself is not new, and this paper does not claim it.** What SelectiveNet is not: a language model, with a vocabulary softmax, or with memory.

Abstention inside a language model's vocabulary. Cohen et al. (2024) add an `[IDK]` token and shift probability mass toward it within the same cross-entropy. Three properties of that work matter here. It never compares against a separate binary head. It does not name the competition for probability mass, though it mitigates it in passing with a hyperparameter Π that caps at 0.5 the mass displaceable to `[IDK]` so the gold token "remains competitive". And it reports that the method *fails* in very small models (pythia-70m, pythia-160m), attributing this to "numerical precision issues with the initialization of `[IDK]`". Our model lives in exactly that regime, which makes the attribution testable (§5.2).

Abstention inside a retrieval interface. The pointer sentinel (Merity et al., 2016) adds a sentinel that competes with the copy distribution; SQuAD 2.0 (Rajpurkar et al., 2018) scores no-answer against the spans; energy-based OOD detection (Liu et al., 2020) thresholds the score. **The null slot is not an invention of this work**; what is new is running it as the third arm of a controlled contrast inside a model whose memory is co-trained rather than external.

Why memory makes this harder, not easier. Joren et al. (2024) report that retrieval augmentation *worsens* abstention: added context raises the model's confidence and produces more hallucination rather than more "I don't know". A model with an archive plus abstention is working against that effect, not benefiting from it.

Versioned facts. Wang et al. (2025) document proactive interference when a key is updated several times in context, and the recency shortcut that follows. They propose no architectural fix, evaluate only large pretrained models, and—the gap that matters for us—never ask for the *superseded* value. A bench that asks only for the current value cannot separate "prefers the latest" from "uses the order", because the shortcut answers both.

3 The bench

Model. A language model trained *from scratch*, with no pretrained component: 863,730 parameters, 3.5 MB, a closed vocabulary of 242 human-readable tokens. The architecture is a delta rule recurrence plus a **co-trained persistent archive** with a learned order stamp in the key, read by softmax and injected at block 0. The archive persists across sessions while the recurrent state is reset between them, so it is the only possible bridge from one session to another.

Task. Facts are stated in one session, corrected in another—including elliptical corrections of the form "no, it's beto" that do not name the entity—and queried in a third:

```
session 1  USER  the director of norte is ana
session 2  USER  no , the director of norte is beto
session 7  USER  who directs norte ?
           MODEL beto
```

The answer is a **single token**, which makes the metric exact and deterministic: no LLM judge, no interpretive parser. The model is asked for the current value, for the superseded value (`anterior`), and for facts that are not in the archive at all, in which case the correct behaviour is abstention. During abstention training, 40% of questions have no answer ($p_{\text{nose}} = 0.4$).

Metrics and gate. `vigente` is accuracy on the current value; `nose` is the abstention rate on questions without an answer; `falsa_abst` is the abstention rate on questions that *do* have one. The pre-registered gate, fixed before any of the campaigns reported here and never adjusted:

Gate: `nose` ≥ 0.50 **and** `falsa_abst` ≤ 0.10

Both halves are required together, because a model that answers `NOSE` to everything scores `nose` = 1.000 and must be visible as the failure it is.

The shortcut, and the margin over it. With $p_{\text{nose}} = 0.4$, the policy *never abstain* scores 0.5906 on this bench. We therefore report the **margin over the shortcut**: the model's `vigente` at the end of the base phase minus 0.5906. This variable turns out to govern whether abstention is learnable at all (§5.4).

Protocol. Every campaign has a pre-registration frozen with its SHA before implementation, a deviations file, and per-unit reporting. **Results are never reported as a mean alone**: bimodality across seeds is a measured property of this bench, and an average would hide exactly what is of interest.

Use of a language model. All experiment code, training campaigns and report generation in this work were carried out with the assistance of a large language model. The study design, the pre-registrations and their frozen hashes, the choice of controls and the verification of every reported number are the author's own. The model is not an author and bears no accountability for the work.

4 The four interfaces

All four start from the *same* base checkpoint and re-initialize Adam, so the abstention phase is matched:

condition	where abstention lives	published counterpart
`token`	an entry in the value softmax	`[IDK]` (Cohen et al., 2024)
`escala`	idem, with the vector renormalized	the `[IDK]` paper's own explanation
`cabeza`	a separate binary output (+129 params)	SelectiveNet (Geifman & El-Yaniv, 2019)
`slot`	an entry in the memory (+256 params)	pointer sentinel (Merity et al., 2016)

`escala` deserves a word, because it is the control that could have saved the entire architecture. A weight inspection showed `NOSE` with an input-embedding norm of 0.367 against 1.011 and 1.028 for value tokens: it was competing in the same softmax with a vector three times shorter. `escala` renormalizes it to the mean norm of the value vectors on both input and output sides. **If the problem were the norm, `escala` would fix it and no new architecture would be needed.**

A declared asymmetry: `slot` introduces 256 new parameters and `cabeza` 129. The difference is 127 parameters out of 863,730 (0.015%), and it is declared here rather than discovered later—though as it happens, `slot` loses, so the objection "it won because it had more parameters" never arises.

5 Results

5.1 Vocabulary versus head

Twenty-one units (7 configurations × 3 conditions) at 14,000 steps. On the five *hard* units—those where the base model has a low margin over the shortcut:

condition	passes the gate
`token`	**0 of 5**
`escala`	**0 of 5**
`cabeza`	**4 of 5**

The single failure, at the lowest margin of the series, misses on `falsa_abst` = 0.1189 against the 0.10 required. `nose` **was never the problem** in any condition (0.54–0.93): all of them detect absence. What fails is always `falsa_abst`—they keep quiet too often.

A sanity check pre-registered as a non-inferiority test *failed upward*: `cabeza` was required to keep `vigente` within ±0.05 of `token` and exceeded it in 3 of 7 units (+0.0633, +0.0668, +0.0928), always by recovering *better*. We record the failure as such rather than adjusting the threshold after seeing the number.

5.2 The norm is not the mechanism

`escala` averages 0.2394 false abstention across the five hard units against 0.0929 for `cabeza`: a gap of **+0.1465**, almost three times the 0.05 the pre-registration asked for.

The negative is clean rather than an implementation failure. The renormalization survives training (it enters at 1.000 by construction and closes at 0.899 input / 1.131 output); the norm stayed where it was put and it still did not help. And a weight probe explains *why* the premise was wrong in the first place: the 0.367 that motivated the condition is from the **input** matrix, while the softmax that decides the answer uses the **output** matrix—where `NOSE` was already 1.77–2.05× *longer* than an average value token in the five hard units, and 0.825× in the easy ones. `escala` therefore *raised* the norm in the easy units and *lowered* it in the hard ones: the same intervention doing opposite things depending on the unit.

> **This is a direct, measured disagreement with Cohen et al. (2024).** They attribute the failure of `[IDK]` in small models to the initialization of its embedding. In this bench the initialization does explain where the asymmetry *comes from*—`NOSE` is never trained during the base phase and stays where it was put—but it does not explain the *failure*, because correcting it explicitly changes nothing. What remains is that two decisions share one softmax.

5.3 Budget: what the gate was measuring, and what survives

The result above was measured with every unit at 14,000 steps, and a pre-registered follow-up asked whether `token` simply arrives later. It partly does. Extended to 20,000 steps (within the original learning-rate horizon, so the schedule is untouched), four of the five failing units cross the gate; false abstention falls in 5/5 and Spearman(step, `falsa_abst`) is negative in 5/5.

We therefore retract the framing, in the pre-registration's own words: "`cabeza` passes the gate where the others fail" is not true once the others get 6,000 more steps. What survives is the paired comparison at matched budget:

`falsa_abst` at 20,000	`cabeza`	`token`	ratio
seed 0	**0.0633**	0.1296	2.0×
seed 1	**0.0421**	0.0713	1.7×
seed 2	**0.0166**	0.0458	2.8×

> **Budget explains crossing the gate, not the advantage.** A gate is a threshold, and a threshold converts a continuous difference into a yes/no that depends on where it is placed. What does not depend on the threshold is that the difference is still there, with the same sign, in 3/3 seeds.

There was no trade: `vigente` rose in all five units over the same extension, up to +0.32. The model did not buy abstention by answering worse.

5.4 The frontier, and how far the architecture moves it

Whether abstention is learnable at all is governed by the margin over the shortcut, and this matters practically: **if a model must be trained to near-perfection before it can be taught to say "I don't know", the method is useless for anything real**, because no useful model reaches 1.0 accuracy in its domain.

Thirteen units per condition, spanning the margin range by cutting base runs *by `vigente` value* rather than by step number, then training 2,000 abstention steps from each cut:

	cut falls between	Spearman(margin, `falsa_abst`)
`token`	+0.2358 and +0.2826	−0.8033
`cabeza`	+0.1489 and +0.1672	−0.8886
the head moves the frontier ≈ 0.10 of margin downward		

Thirteen units ordered by margin with **no inversions**, and the transition is a step rather than a slope: between two points separated by 0.047 of margin, `falsa_abst` falls from 0.2237 to 0.0855, a factor of 2.6×. In the band from +0.1672 to +0.2358, `token` fails 4/4 and `cabeza` passes 4/4.

The operational number. With a separate head it suffices to train to a margin of $\approx +0.17$; with the vocabulary token one must reach $\approx +0.26$. **With** 129 **parameters one can teach a model that recovers** 10 **points worse to keep quiet.**

A confound, and the cell that resolved it. Margin and degree of training are entangled in the sampled range: all high-margin points came from truncated runs and all low-margin points from complete ones. The missing cell is a model trained to completion that is nonetheless stuck at a low margin. One seed failed to produce cuts and delivered exactly that cell—an easy-task run, trained to completion, margin +0.2124—and a pre-registration froze two mutually exclusive predictions before it was measured. `token` failed (0.1885) and `cabeza` passed (0.0902): **the margin is not a proxy for task difficulty**. With a caveat we state rather than bury: the binary verdict for `cabeza` rests on the final tick. What is robust is the paired contrast, 7/7 valid points in the same direction (sign test, $p = 0.0078$; means 0.2718 vs. 0.1430).

5.5 The memory slot converges to the prior

Two vectors are added to the archive, $k_\varnothing$ and $v_\varnothing$, competing as one more column of the read. Following the design, the null is *never* required to win the read—the model demonstrably answers correctly with the right entry in second place—only to carry relative mass. Three seeds, 26,000 steps, paired against `cabeza` controls.

	seed 0	seed 1	seed 2
accuracy (`cabeza` → `slot`)	0.9705 → 0.8991	0.7769 → 0.7859	0.8351 → 0.8999
`nose` (`cabeza` → `slot`)	0.9119 → **0.0000**	0.5416 → **0.0000**	0.7298 → **0.0000**

The model trains (the blocking sanity criterion passes 3/3) and then never abstains, in any seed. A clean zero is the signature of a broken instrument, and here it was one the first time—an evaluation script tested `==` where it needed `in` and could not emit abstention at all. The fix *raises* accuracy in all three units (0.7572 → 0.8991, 0.4307 → 0.7859, 0.7744 → 0.8999), i.e. it favours the slot, and `nose` still reads zero. A regression check reproduced every previously published unit bit for bit.

The mechanism is visible in the attention mass itself:

mass on the null slot	seed 0	seed 1	seed 2
questions *with* an answer	0.4074	0.4046	0.4020
questions *without* an answer	0.4086	0.4065	0.4037
AUC (without vs. with)	0.5190	0.5313	0.5182
empirical base rate of unanswerable questions: 829/2048 = **0.4048**			

> **The slot learned the prior.** 0.4074/0.4046/0.4020 against a base rate of 0.4048 is the exact optimum of the binary cross-entropy for a predictor with no usable signal. Because the prior sits below any sensible decision threshold, "it never abstains" is not a badly chosen cutoff but the *necessary consequence* of having learned the base rate instead of membership.

We tested the cutoff objection anyway, sweeping 400 thresholds over the mass. No threshold passes the gate in any seed; the best attainable point abstains on $\approx$ 80% of unanswerable questions and on $\approx$ 75% of answerable ones (Youden's J between +0.038 and +0.078, i.e. noise). The two populations do not separate, so no cut can separate them.

Two observations worth recording. First, a declared risk that did *not* materialize: the slot absorbs $\approx$ 40% of the read mass on every query and accuracy holds at 0.90/0.79/0.90, with the superseded-value accuracy not degrading (and rising +0.111 in one seed). The read tolerates losing four tenths of its mass to a constant vector. Second, $v_\varnothing$ started at exactly zero and ended with norm 1.34–1.71: the model did not use it as a pure gate but learned a value for it, injecting a fixed vector into every answer—a learned read bias, not a signal of absence.

5.6 The result that frames the slot's failure

The slot's negative is not an isolated disappointment. Independently and earlier, the archive's own matching score—the maximum match of the query against the stored keys—separates "the answer is present" from "the answer is absent" at AUC 0.4984 and 0.5022: chance to two decimals. The top-2 margin and the logsumexp give the same. Three controls that could have failed did not: logits reconstructed from the extracted score match the model's bit for bit, the score varies, and the archive is used. Re-measured at the position of *maximum focus*—where the read concentrates up to 65% of its mass on a single entry—it is still chance (0.5007, 0.5077).

> **The model selects an entry forcefully, and the force of the selection does not encode whether what it was looking for is there. It always finds something, and it finds it with the same conviction when there is nothing to find.**

This is the expected consequence of the mechanism: a softmax read sums to one, so the best entry wins with comparable mass whether or not a good candidate exists, and nothing in training ever pressured the magnitude of the match to mean "present". **An external retriever gets the empty set for free; moving the index inside the network destroys it**—and, as §5.5 shows, adding a place for absence to live does not restore it.

6 What we do not claim

Three restraints, stated because the results invite overreach.

This is calibration, not knowledge. The supported claim is *"separating the head makes abstention learnable in this regime"*. The residual ceiling is one of calibration: the AUC of the head's logit is 0.777–0.998, so the information needed to decide is present—what is misplaced is the cutoff. On the single unit that fails the gate, a threshold chosen on one half of the data and judged on the other passes it. The sentence *"the model knows when it does not know"* is not supported by anything here.

Everything usable passes through labels. Seven pre-registered attempts to obtain the cut without supervision—a Gaussian mixture on the logit, an uncalibrated dispersion threshold, an internal disagreement monitor under masking, a key-tie detector, and others—all landed in AUC 0.50–0.67, against 0.777–0.998 for the supervised head. Two of them failed for a reason worth reporting, because it constrains what a future detector must measure. In this bench the model **never fabricates content**: out-of-archive answers occur at rate 0.0000 across all four difficulty levels; every wrong answer is a real archived value attached to the wrong entity. So when a question has no answer, the model does not guess into the void—it anchors on *another real entry*, and its answer is as stable under masking as a correct one. The disagreement monitor confirmed the model reads the archive (98–99% of correct answers change when the supporting entries are masked) and still could not discriminate, because **disagreement measures whether the answer came from the archive, not whether it came from the *correct* entry**. Any future label-free detector has to separate those two things.

Hallucination here is misattribution, not invention. The preceding paragraph has a consequence for detection in general: in a model with an archive, the failure mode is a true value placed on the wrong entity. That is *harder* to catch than invention, because any verification of the form "does this datum exist?" passes it. We flag the scope: with a closed vocabulary, free fabrication is excluded by construction, so what is measured is that output stays confined to archived content.

7 Limitations

- **Scale.** 863,730 parameters and a synthetic 242-token language. Nothing here transfers automatically: Cohen et al. (2024) works in large models and *fails* at pythia-70m/160m, which is precisely the asymmetry that makes small-model results non-predictive of large ones. The direction of the transfer question is open in both directions.
- **The language is the interface of the test, not the object of study.** A single-token answer is what makes the metric exact without an LLM judge. No claim is made about open natural language.
- **Seeds and bimodality.** Bimodality across seeds is a measured property of this bench. We report per unit and never as a mean alone; with two units in a cell, difficulty and non-convergence cannot be separated. Two conditions have one seed per cell.
- **The abstention-phase budget was chosen by the author**, and §5.3 shows it mattered: one framing of the main result did not survive extending it. The paired comparison at matched budget did.
- **One implementation detail contradicts its own design**: the slot's decision threshold was inherited as mass > 0.5, which is stricter than the design's own stipulation of relative mass. The threshold sweep shows this does not explain the result, but it should be fixed if the condition is revisited.

- **Directed literature search, not a systematic review.** A work using different vocabulary for the same idea—"rejection head", "versioned memory", "fact supersession"—may have been missed.

8 Conclusion

Where the abstention decision lives changes whether it can be learned at all, and the ordering is stable across seeds, budgets and margin regimes. A separate binary head costing 0.015% of the parameters halves to nearly thirds the false-abstention rate against a vocabulary token at matched budget, and lowers by ≈ 0.10 of accuracy margin the point at which a model becomes teachable to keep quiet. The explanation is not the norm of the token's vector—the most generous version of that hypothesis was implemented and failed—but the fact that two decisions of different natures were competing for one softmax.

The memory slot, which on paper is the most elegant of the three because it makes "nowhere" an expressible answer in the space where the model already operates, converges to the base rate and never abstains. Together with an archive matching score pinned at chance, this locates a structural property of co-trained memory: **absence has no representation there, and providing a place for it is not enough to create one**. An external retriever can return the empty set; an internal index read by softmax cannot, and the difference is not repaired by adding a null entry.

What remains open is the part that matters most and that none of these results reach: every route to the decision that does not pass through labels has failed, and it has failed for a specific, measured reason—the model, when it does not know, anchors on another real memory, which from the outside is indistinguishable from knowing.

Data and code availability

All pre-registrations with their SHA hashes, deviation files, analysis scripts, per-seed raw results and machine-generated reports are archived in the project repository. Every verdict reported here is emitted by a script frozen before the corresponding data existed.

Declarations

Funding. None. **Competing interests.** The author declares no competing interests. **Author contributions.** Sole author; responsible for design, implementation, analysis and writing. **Ethics.** Not applicable; no human or animal subjects, all data synthetic. **Author note.** This work was carried out without institutional sponsorship, supervision or funding, and is reported in a personal capacity.

References

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [IDK] token. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2412.06676.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. *International Conference on Machine Learning (ICML)*, 2019. arXiv:1901.09192.

Hailey Joren et al. Sufficient context: A new lens on retrieval augmented generation systems. arXiv:2411.06037, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv:1609.07843, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. *Association for Computational Linguistics (ACL)*, 2018.

Chupei Wang et al. Unable to forget: Proactive interference reveals working memory limits in LLMs beyond context length. arXiv:2506.08184, 2025.

Supplementary Files

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

- tripodees.pdf