Does the Silent Workspace Anticipate the Thought? Trace-Aligned Jacobian-Lens Analysis of a Reasoning Model

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

The Jacobian lens transports a residual-stream vector into the final-layer basis with the average input–output Jacobian, $\text{lens}_l(h) = \text{unembed}(J_l h)$, reading the tokens a model is *silently disposed to say later*; the resulting "J-space" has been proposed as an emergent workspace. Reasoning ("thinking") models make this claim newly testable: they *externalize* their reasoning as an explicit `<think>` trace, which provides a ground-truth future to check the silent readout against. On Qwen3-8B with a pretrained Jacobian lens, evaluated on a held-out, reasoning-typed probe suite, we read the silent workspace at positions *inside* the model's own thinking trace and ask whether it anticipates the tokens the model goes on to write. It does: the silent readout lands in the realized near-future thinking window with recall 0.88 versus a frequency-controlled floor of 0.15, and the effect is uniform across four reasoning types (recall, multi-hop, analogy, arithmetic). A horizon sweep with the immediate next token *excluded* shows the readout genuinely reaches beyond next-token prediction (excluded recall rising from 0.01 to 0.41 as the window widens). We are candid about the limits: on this clean excluded-next metric the Jacobian lens holds only a small edge over an ordinary logit-lens readout (peak gap +0.025), and two planned axes—a thinking-versus-non-thinking ablation and causal steering—were not run. We also report a methodological finding: on a chat/reasoning model that answers in prose, a first-token answer-fidelity metric does not transfer, which is precisely why reading *inside* the trace is the right probe.

1 Introduction

A model that will emit a token does not decide to at the last instant; the disposition is present, silently, in earlier layers. The logit lens [6] reads such dispositions by projecting a hidden state through the unembedding; the tuned lens [2] sharpens that readout. The *Jacobian lens* [1] transports a residual vector through the average input–output Jacobian $J_l = \mathbb{E}[\partial h_{\text{final}}/\partial h_l]$ before decoding, $\text{lens}_l(h) = \text{unembed}(J_l h)$, and frames the set of these silent readable dispositions as a low-dimensional workspace (J-space) in which a model holds a thought before writing it down.

The workspace framing invites an obvious question that base models cannot answer directly: *does the silent workspace actually predict what the model will say?* A reasoning model answers it. Such a model, prompted to think, writes an explicit `<think>` trace before its answer, and that trace is a ground-truth "future": the tokens the model *did* go on to produce. We can therefore read

*Corresponding author: deep@lab.cloud

the silent workspace at a position partway through the trace and check it against the model's own realized reasoning. On Qwen3-8B [8], a hybrid-thinking model, with the pretrained Jacobian lens released for it [5], we make this measurement on a held-out probe suite. Contributions:

- a **trace-alignment** test showing the silent readout anticipates the model's realized near-future reasoning tokens, far above a frequency-controlled floor (§4);
- a **horizon sweep** with the immediate next token excluded, isolating how far ahead the readout reaches, and an honest account of the *small* margin by which the Jacobian lens beats an ordinary logit-lens readout (§5);
- a **methodological finding**—first-token answer fidelity does not transfer to a chat/reasoning model that answers in prose (§3)—which motivates reading inside the trace.

2 The Jacobian lens and setup

For a residual vector h at layer l, $\text{lens}_l(h) = \text{unembed}(J_l h)$ with $J_l = \mathbb{E}[\partial h_{\text{final}}/\partial h_l]$ estimated over a corpus. We use the pretrained neuronpedia lens for Qwen3-8B-Instruct [5], fit on `Salesforce/wikitext` [4]; the model is frozen and the lens is *loaded*, not fit.

Probes and splits. 200 hand-built probes, 50 in each of four reasoning cells (single-fact *recall*, *multi-hop*, *analogy*, *arithmetic*), group-aware split 145 train / 55 test; the lens-fitting corpus is disjoint from all probes.

Trace-alignment metric. For each probe we build the prompt with thinking enabled and greedily generate a bounded `<think>` trace. At sampled positions t along the trace we read the lens at the last position of (prompt + generated tokens up to t) and take its top-1 token. We score whether that top-1 lands in the model's realized future window $\text{gen}[t{+}1 : t{+}1{+}H]$ (*future-recall*), and, as a clean "reaches beyond the next token" measure, in $\text{gen}[t{+}2 : t{+}1{+}H]$ (*next-token-excluded recall*). The best readout layer is chosen on a held-out half of probes and reported on the other half. A **shuffled-trace floor** pairs each top-1 with a different sample's future window, controlling for token frequency. We compare the J-lens against the logit lens throughout.

3 Reading position: the last prompt token is not the answer

On a reasoning model the natural base-model readout—the lens at the last *prompt* position—does not read the answer: its top tokens are `<think>` and "ponder," i.e. the model's silent disposition there is *to begin thinking*. Reading at the answer onset (the first token after `</think>`) also fails as a fidelity probe: only 42/200 probes close `</think>` within a 256-token budget, and a chat/reasoning model answers in prose ("*The* capital of France is Paris"), so the first post-`</think>` token is a sentence preamble, not the gold entity—both the J-lens and logit lens score 0.0 first-token Hit@1. This is a metric-transfer failure, not a lens failure, and it motivates reading *inside* the trace, where the "next token" is a genuine content token.

4 Trace alignment: the silent workspace anticipates the reasoning

Reading the lens at 604 positions inside the held-out traces, the silent J-lens readout lands in the model's realized near-future thinking window (horizon $H{=}24$) with recall **0.879**, versus a

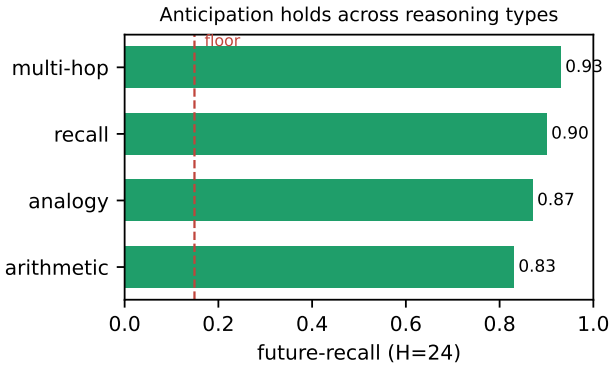

Figure 1: Trace-alignment future-recall (H=24) by reasoning type. The dashed line is the frequency-controlled shuffled floor; anticipation holds across all four cells.

frequency-controlled shuffled floor of **0.131** (a lift of 0.75). The logit lens is a close companion at 0.869. So the workspace genuinely anticipates the model's own reasoning, validated against the trace as ground truth. The effect is **uniform across reasoning types** (Fig. 1): future-recall is 0.90 (recall), 0.93 (multi-hop), 0.87 (analogy), 0.83 (arithmetic) type.

5 How far ahead? A horizon sweep

To ask *how far* the readout reaches, we sweep the horizon H and, crucially, also report recall with the immediate next token **excluded** (Fig. 2). Future-recall rises and saturates near 0.88 by $H{\approx}24$ for both lenses. The excluded-next recall rises from ≈ 0.01 at $H{=}2$ to **0.41** at $H{=}48$ for the J-lens (logit 0.40), confirming the readout points at tokens that recur *later* in the reasoning, not merely the next one. The honest caveat is the *margin*: on this clean excluded-next metric the Jacobian lens sits above the logit lens at every horizon ≥ 8 but by only $+$**0.025** at its peak ($H{=}16$–32). The two readouts are largely comparable at reaching ahead; the Jacobian's linear future-transport adds little over the identity readout inside a coherent reasoning trace. We state the headline as *the workspace anticipates the reasoning*, not *the Jacobian lens is much better than the logit lens.*

6 Limitations

(i) The Jacobian lens's edge over the logit lens is *small* on the excluded-next metric; our positive claim is about the workspace, not about the lens beating the logit baseline. (ii) We did *not* run a thinking-versus-non-thinking ablation, so we cannot say the workspace is sharpened by thinking. (iii) We did *not* run causal steering, so alignment here is correlational, not causal. (iv) The generation budget (≤ 384 tokens) closes `</think>` on only about half the probes; the trace-alignment metric does not depend on closing, but a larger budget would broaden the analysis. (v) We study one model (Qwen3-8B); a horizon-sweep with a next-token-excluded floor and larger models are natural extensions.

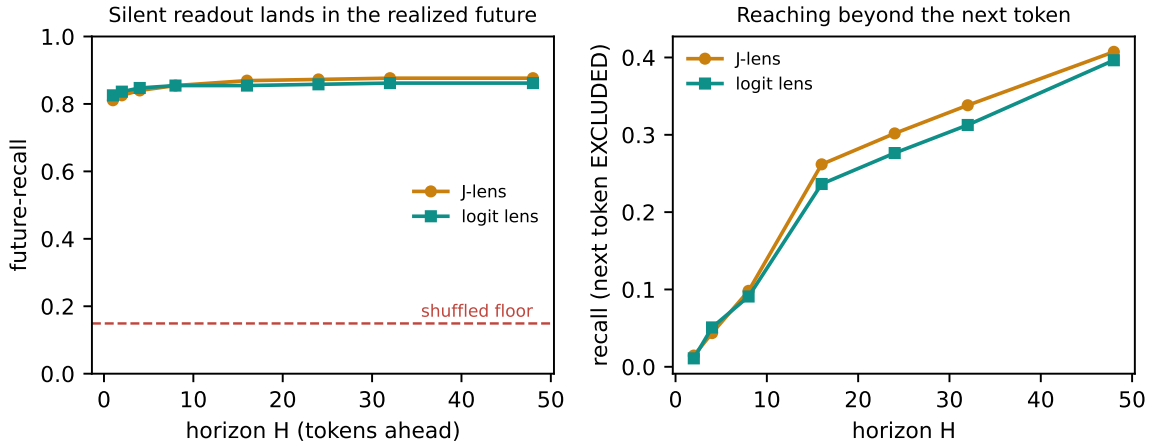

Figure 2: Horizon sweep. Left: future-recall vs. horizon for both lenses, far above the shuffled floor. Right: recall with the immediate next token *excluded*—both lenses reach beyond the next token; the J-lens leads by a small margin.

7 Related work

The logit lens [6] and tuned lens [2] are our readout baselines; the Jacobian lens and its workspace framing [1] are the method we probe. Anticipating tokens a model will emit several steps later from a single hidden state was studied by Future Lens [7]; we validate the silent readout against a reasoning model's realized `<think>` trace rather than its final output. Reading a model's disposition against its own realized generation also connects to work on self-consistency of chain-of-thought and to activation-level intervention [3, 9]; our contribution is to use a reasoning model's explicit trace as ground truth for the *silent* readout, and to measure how far ahead of the pen that readout reaches.

8 Conclusion

On an open reasoning model, the silent J-space workspace demonstrably anticipates the model's own near-future reasoning tokens, uniformly across reasoning types and well above a frequency floor, validated against the model's realized `<think>` trace—a test only a thinking model affords. The honest counterweight is that a plain logit-lens readout does nearly as well, so the workspace's predictive power is not specific to the Jacobian transport. We hope the trace-as-ground-truth methodology, and the next-token-excluded horizon measure, are useful to others studying what reasoning models are about to say.

Availability. All artifacts (the analysis code, a pointer to the pretrained neuronpedia lens, and the 200-probe suite with splits) are available from the authors on request. Total compute ≈ 1.5 GPU-hours.

References

[1] Anthropic. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread. https://transformer-circuits.pub/2026/workspace/. Code: https://

github.com/anthropics/jacobian-lens, 2026. URL https://transformer-circuits.pub/2026/workspace/.

[2] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

[3] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023. URL https://arxiv.org/abs/2306.03341.

[4] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

[5] Neuronpedia. Neuronpedia: Pretrained jacobian lenses for the Qwen3 family. Hugging Face. https://huggingface.co/neuronpedia/jacobian-lens, 2026. URL https://huggingface.co/neuronpedia/jacobian-lens.

[6] nostalgebraist. Interpreting GPT: The logit lens. LessWrong. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

[7] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023. URL https://arxiv.org/abs/2311.04897.

[8] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

[9] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.