Final Report: Probing and Steering Temporal Representations in LLMs for AI Oversight

Justin Shenk
shenk.justin@gmail.com
Independent Researcher
Jyotin Goel, Tejas Dahiya, Michal Mraz, Vansh Gupta, Ipshita Bandyopadhyay,
Ronen Roy, Harleen Kaur Bagga, Avigya Paudel, Archit Sharma,
Tim Obiso, Adrian Molofsky

Supervised Program for Alignment Research — Spring 2026

Abstract

We investigate how large language models internally represent and process temporal information, an open question with direct implications for AI safety — particularly for understanding planning capability, deception detection, and oversight robustness. Across 11 complementary sub-projects, we apply linear probes, causal interventions (activation steering), and behavioral evaluations to models ranging from 1B to 30B parameters. Our findings show that temporal representations are genuinely encoded in LLM activations and can be causally steered, but that careful baselines are essential: in code generation, what appears to be planning signal is fully explained by surface features. We also find safety-relevant phenomena including context fatigue (entropy collapse over long contexts), sycophantic surrender under social pressure, and a tentative link between short-term temporal steering and more harmful outputs. These results motivate both methodological caution in probing studies and further investigation of temporal representations as a lever for AI safety monitoring.

Keywords mechanistic interpretability · temporal reasoning · linear probes · activation steering · AI safety

1 Introduction

Understanding how LLMs represent time internally is relevant to several AI safety concerns. Models that encode temporal horizons may develop planning capabilities, and the ability to detect and steer these representations could inform oversight mechanisms [4, 7]. If a model distinguishes short-term from long-term reasoning, this distinction may also relate to deceptive alignment [3] — where a model behaves differently based on perceived temporal context such as training versus deployment [6].

This project brings together 11 researchers investigating complementary facets of temporal representation in LLMs. Sub-projects span probing for temporal structure in activations, causal steering of temporal horizons, evaluating lookahead planning in code generation, measuring context fatigue over long sequences, detecting sycophantic surrender under social pressure, tracking uncertainty encoding across reasoning steps, and analyzing activation stability under repetition.

Our shared codebase[1] and common methodological toolkit — linear probes on residual stream activations, activation steering via probe-derived directions, and behavioral evaluations — enable cross-model and cross-task comparisons that no single sub-project could achieve alone.

[1] https://github.com/justinshenk/temporal-awareness

Success for this project means producing publishable findings on how temporal information is encoded and used in LLMs, developing reusable methodology for probing studies (including rigorous baseline protocols), and identifying safety-relevant properties of temporal representations that could inform runtime monitoring. Figure 1 illustrates the research structure across the three layers of analysis.

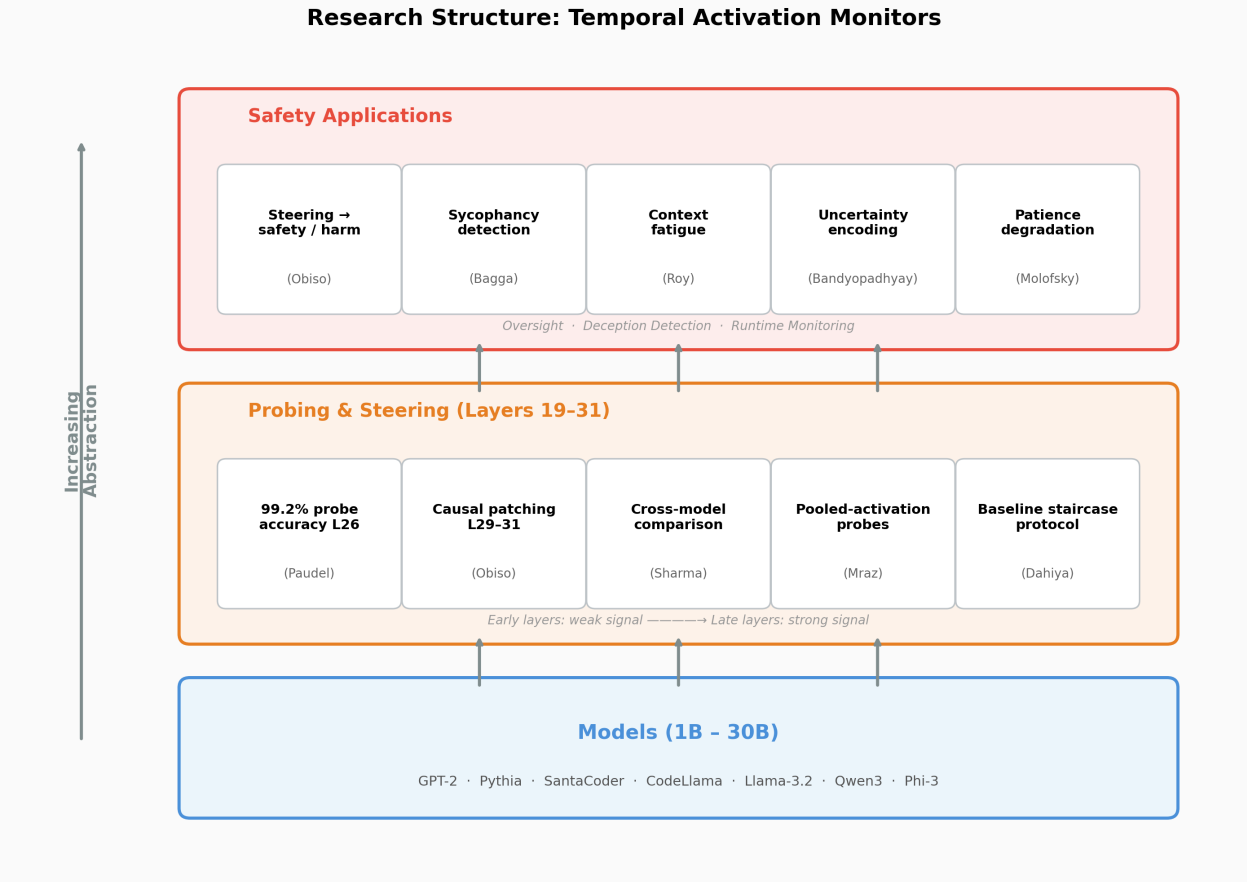

Figure 1: Research structure of the Temporal Activation Monitors project. The bottom layer shows the models under study (1B–30B parameters). The middle layer maps probing and steering contributions across fellows. The top layer shows safety-relevant applications that build on the probing results.

2 Methodology

2.1 Common Framework

All sub-projects share a core methodology: train linear probes on transformer residual stream activations to detect temporal features [1], then validate with causal interventions. We extract activations at each layer from frozen pretrained models and train logistic regression or linear classifiers to predict temporal properties (e.g., short-term vs. long-term horizon, temporal order, uncertainty level). Where probes detect signal, we perform activation steering by adding scaled probe directions to residual streams during generation [4] and measuring behavioral change.

2.2 Models

We evaluate models across four architecture families and a range of scales: GPT-2 (124M–1.5B), Pythia (410M–2.8B), Qwen3 (4B, 8B, 30B-A3B MoE), Llama-3.2 (1B, 3B), Phi-3-mini-4k, CodeLlama-7B, and SantaCoder. This breadth allows us to test whether findings generalize across architectures and whether scale thresholds exist for specific capabilities.

2.3 Datasets and Evaluation

Each sub-project constructs task-specific evaluation sets, with contrastive pair datasets following the Contrastive Activation Addition (CAA) framework [7]. Notable datasets include: 1,000 validated implicit temporal pairs across 40 categories and 8 domains, with no explicit temporal keywords (Paudel); code generation samples with 500 examples across 5 return types for the planning study (Dahiya); sycophancy benchmarks with social pressure conditions (Bagga); and repetition-degradation benchmarks across 4 datasets with 10 repetition counts (Molofsky). We use bootstrap resampling (typically 1,000 iterations) for statistical significance testing and PCA for dimensionality reduction where applicable.

2.4 Methodological Innovations

Three innovations emerged from the sub-projects. First, a **baseline staircase protocol** (Dahiya): chance → bag-of-words → name-only → name+params → probe, which progressively tests whether probe accuracy reflects genuine internal computation or surface features. Second, **pooled-activation probes** (Mraz): aggregating across token positions rather than using only the last token, yielding improved probe performance. Third, **implicit temporal pairs** (Paudel): evaluation data that tests temporal reasoning without temporal keywords, preventing probes from relying on lexical shortcuts.

3 Results

3.1 Temporal Representations Are Real and Steerable

Linear probes across Qwen3-4B, Llama-3.2-3B, and Phi-3-mini confirm that temporal and process-state information is genuinely encoded in LLM activations [8]. Paudel achieves 99.2% probe accuracy at layer 26 of Qwen3-4B for classifying short-term vs. long-term thinking on implicit pairs (no temporal keywords), with shuffled-label controls at chance. Steering works best at layers 19–22 (not the probe's best layer), with layer 22 at $\alpha = 50$ yielding approximately 4× more probability on long-term completions. Interactive results are available online.[2]

Obiso finds that temporal computation is causally localized to the last three layers (29–31) of Qwen3-4B via causal patching [5], and that probe accuracy updates token-by-token during autoregressive generation — an "internal clock" that tracks temporal context in real time. The linear component of temporal representations transfers across domains, while the nonlinear component is domain-specific, supporting linear probes as the appropriate tool for universal temporal monitoring.

Sharma confirms temporal representations exist across all three models tested but vary in robustness: Qwen3 shows the most distributed and robust signal, Llama exhibits strong but less clean generalization, and Phi displays weaker, more localized behavior.

[2]Paudel: https://avi161.github.io/temp-steering-results/; Paudel slides: https://docs.google.com/presentation/d/10pftLZsDfvfAfPHRN3k_f5jSffTAO_5SYrJj2QAULkI/

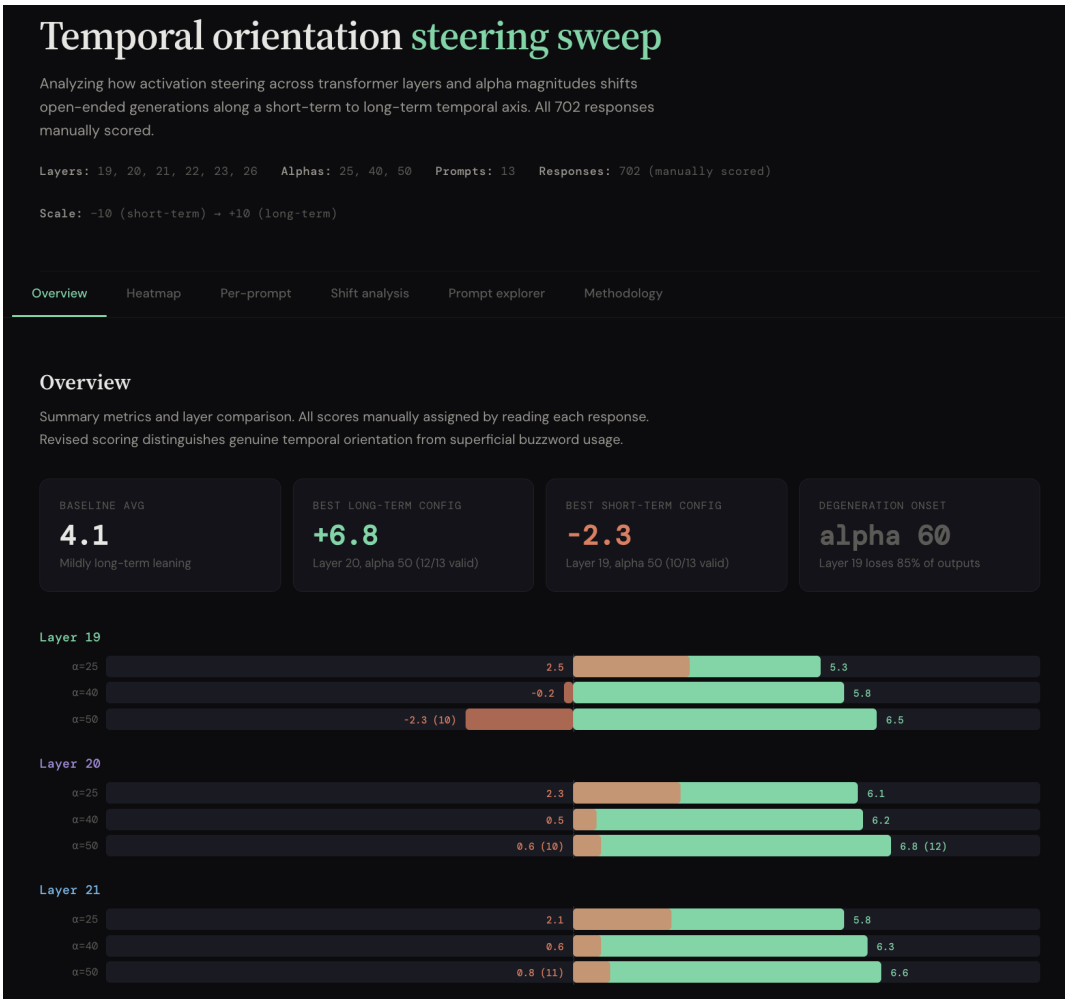

Figure 2: Temporal orientation steering sweep (Paudel). Layer-by-layer and alpha-by-alpha results from 702 manually scored responses, showing best long-term config (+6.8, layer 20, α=50) and best short-term config (−2.3, layer 18, α=50).

3.2 Baselines Matter: Deflating Apparent Planning Signal

Dahiya's investigation of lookahead planning in code generation provides a cautionary result. Across all 11 models tested (GPT-2 family, Pythia 410M–2.8B, SantaCoder, CodeLlama-7B, Llama-3.2-1B), initial probing suggested 80–91% accuracy at predicting return types — apparently strong evidence of planning. However, the baseline staircase reveals this is fully explained by surface features: a name+params baseline (a classifier trained on function name and parameter names alone) achieves 87–95% accuracy, consistently beating the probe ($p \approx 1.0$ across every model and layer). Misleading-name experiments confirm models follow parameter names 56% of the time and function names 22–34%. Generation-time probing shows probe accuracy actually *decays* during generation, and acrostic behavioral tests (4–20%, near chance) confirm these models cannot perform structural planning. This frames as a methodology contribution: a "baseline staircase" protocol for deflating spurious probing results.

shift temporal preference, even in general open-ended generation tasks.

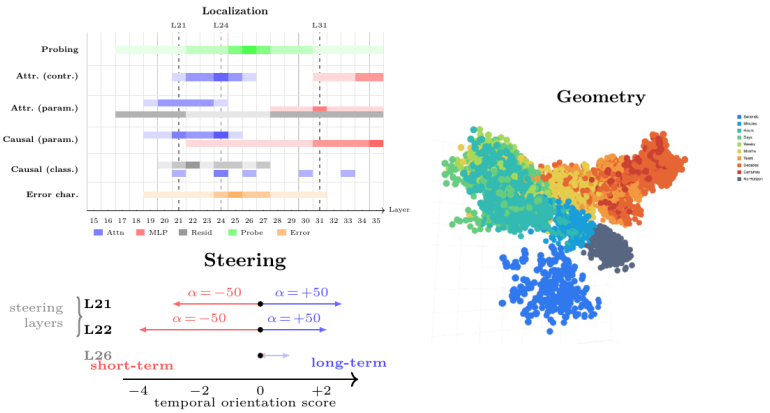

Figure 1: (Top-left) Six rows summarize where the temporal-preference signal sits across the network. Five localization methods (probing, attributional contrastive, attributional parametric, causal parametric, causal classification) converge on a subgraph in layers 17–35; darker shading indicates stronger signal. The bottom row (*Error char.*) shows that an unrelated meta-cognitive variable, cumulative reasoning reliability, is also decodable above 95% across L19–L31 within the same subgraph, indicating the localized region encodes more than just temporal preference (Section 5.1; Appendix R). (Top-right) Time horizon geometry within the identified subgraph (Section 5.2). (Bottom) CAA steering shifts temporal preference bidirectionally at layers 19–22 but weakly at the best probing layer (L26), illustrating the probing–steering dissociation (Section 5.3).

We focus on `Qwen3-4B-Instruct-2507`: its non-thinking-only operation keeps computation inside a fixed prompt template (enabling token-aligned attribution and patching), and its latent preferences are stable under prompt perturbations, trading breadth across model families for depth in tracing a concept from localization through geometry to intervention. Our methodology integrates four localization pipelines plus dedicated behavioral and steering instruments that differ in prompting paradigm, localization technique, scale, and resolution. Because these pipelines approach localization from fundamentally different angles, their independent convergence on the same subgraph components strongly suggests that our findings reflect the genuine model structure.

Figure 3: Temporal preference concepts (Paudel). *Top-left:* localization — probe accuracy across layers. *Top-right:* geometry of temporal-preference activations colored by horizon, from seconds to centuries. *Bottom:* steering effect across layers 19–22.

3.3 Safety-Relevant Phenomena

Sycophancy. Bagga finds that social pressure causes answer flip rates of 10–62.5%, but true sycophantic surrender occurs in only 1.5–15.5% of cases. Output confidence is an unreliable safety monitor: 23.4% of high-confidence correct answers still surrendered under pressure, including some at confidence 1.0. Hidden-state signal for predicting surrender emerges only in the final layer quartile, but the critical probe remains underpowered (11 positive examples; ∼1,000 needed).

Context fatigue. Roy demonstrates that entropy collapse is a real problem as context fills, with models growing more confident without corresponding accuracy gains. This is partially counteracted by in-context learning, where models sometimes improve at a task pattern over the context window.

Uncertainty encoding. Bandyopadhyay finds that uncertainty is strongly encoded at the first reasoning step but weakens after subsequent reasoning, is distributed across layers, and only weakly predicts future states. Across datasets, uncertainty encoding varies significantly and consistently degrades from step 1 to step 2.

Steering and safety. Obiso observes that steering toward long-term thinking monotonically increases output safety, while short-term steering produces more violent completions — a finding that, if confirmed, would link temporal horizon representations directly to safety-relevant behavior.

Table 1. Code return-type prediction across 11 models (507 examples, 5 types). The N+P position baseline matches or exceeds the best probe in every model. Three additional models (Qwen 1.5B/7B/14B) replicated independently with the same null result.

MODEL	PROBE	N+P	GAP
GPT-2 SMALL (117M)	80.3	84.6	−4.3
GPT-2 MEDIUM (345M)	78.7	87.6	−8.9
GPT-2 XL (1.5B)	80.7	87.4	−6.7
PYTHIA-410M	80.9	89.2	−8.3
PYTHIA-1B	82.3	90.1	−7.9
PYTHIA-1.4B	83.2	90.0	−6.7
PYTHIA-2.8B	83.2	91.1	−7.9
SANTACODER (1.1B)	91.1	93.1	−2.0
CODELLAMA-7B	88.6	94.5	−5.9
LLAMA-3.2-1B	86.6	93.7	−7.1
LLAMA-3.2-1B-INST	83.6	94.3	−10.6

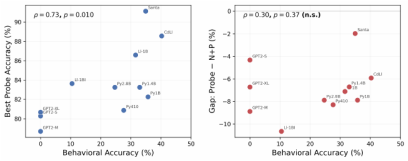

Figure 1. Left: behavioral accuracy predicts probe accuracy ($\rho = 0.73$, $p = 0.010$). Right: we do not find a significant relationship

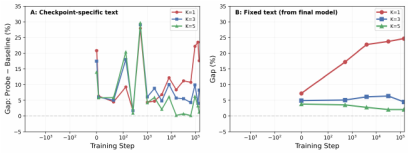

Figure 2. Future Lens gap at $K=1, 3, 5$ for Pythia-2.8B. (A) Checkpoint-specific text (17 checkpoints): volatile at early steps but steep decay by step 4,000. (B) Fixed text from the final model (5 checkpoints): confirms narrowing on identical evaluation data. $K=1-K=5$ spread grows from 3.4 pp at initialization to 22.7 pp at convergence.

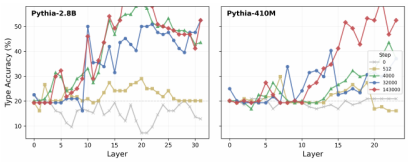

Figure 3. Logit lens type accuracy (zero-shot) at each layer across five training checkpoints. Pythia-2.8B (left) and Pythia-410M (right) both develop from chance (20%, dashed) to 52%. No

Figure 4: Baseline staircase result (Dahiya). The name+params baseline matches or exceeds the residual-stream probe across all 11 models tested, deflating the apparent planning signal.

Additionally, smaller models (4B) show stronger deception detection than larger ones (9B), suggesting internal "honesty" may degrade with scale.

3.4 Emerging Directions

Goel is developing an analysis of token trajectories for multihop questions, with an initial outline established. Gupta is investigating phase transitions in temporal horizon representations, aiming to establish how models transition from short-horizon to long-horizon reasoning and identify features that support each. Molofsky has completed Phase 1 behavioral experiments (8 models × 4 datasets × 10 repetition counts) with Phase 2 targeting sparse autoencoders [2], finding that patience degradation is model- and task-specific: Llama-3.1-8B shows clear accuracy degradation on temporal reasoning (TRAM), Qwen3-8B shows a warm-up effect, and Qwen3-30B-A3B (MoE) remains stable.

Fellow Summary

4 Challenges & Open Questions

Scale limitations. The absence of lookahead planning across all models up to 7B parameters raises the question of whether genuine structural planning requires significantly larger models (70B+), where chain-of-thought reasoning capabilities emerge. Testing at this scale would require additional compute resources.

Statistical power. Several sub-projects face sample size constraints. The sycophancy probe (Bagga) has only 11 positive examples for the critical hidden-state prediction task and needs approximately 1,000 to be conclusive. Scaling these experiments is a priority for follow-on work.

Linear vs. nonlinear representations. Obiso's finding that linear probe components transfer across domains while nonlinear components are domain-specific raises interpretive questions: are we capturing the safety-relevant signal with linear probes, or missing critical nonlinear structure?

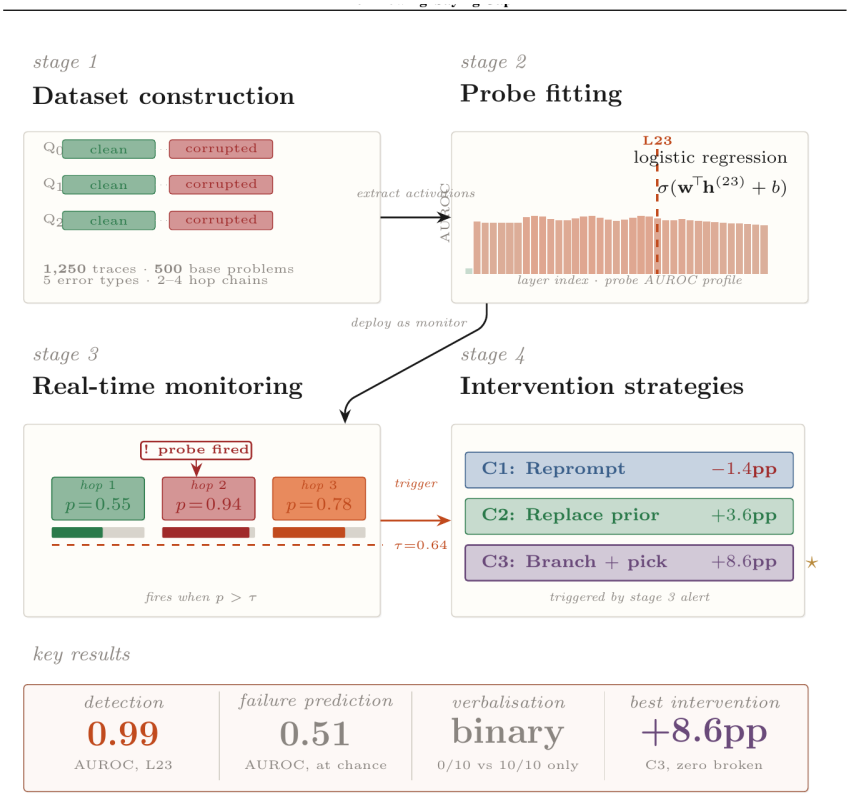

Figure 1. **Methodology overview.** (Stage 1) A contrastive dataset pairs clean and corrupted-prior variants of 500 multi-hop math problems. (Stage 2) A linear probe is fit on residual-stream activations at each layer; layer 23 peaks at AUROC = 0.99. (Stage 3) The fitted probe is deployed as a real-time monitor: it scores each hop and fires when its score crosses threshold $\tau = 0.64$. (Stage 4) Three intervention

Figure 5: Knowing-saying gap methodology overview (Goel). A pipeline that contrasts clean and corrupted-prior variants of ∼500 multi-hop math problems, then fits a probe to residual-stream activations, with the probe firing at $p > 0.64$ as a real-time signal for intervention.

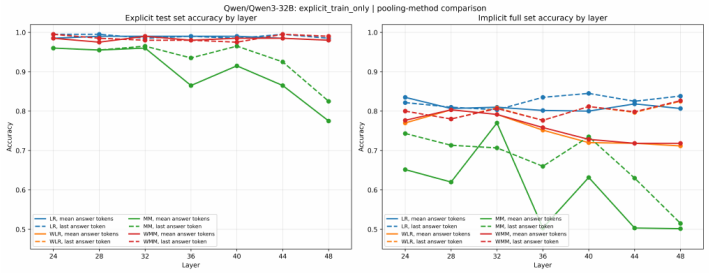

Figure 1: Layerwise explicit-trained probe accuracies from the probe-variation experiment. The left panel reports explicit test set accuracy; the right panel reports full implicit set accuracy.

68 steering remains a strong temporal horizon readout even though LR and whitened variants attain

Figure 6: Layerwise probe accuracies from the probe-variation experiment (Gupta). *Left:* explicit test-set accuracy; *right:* full implicit-set accuracy.

Fellow	Research Question	Key Finding	Target
Jyotin Goel	Token trajectories for multi-hop reasoning	Initial outline on trajectory analysis	Publication
Tejas Dahiya	Lookahead planning in code generation	No planning found; baseline staircase deflates all signal	Publication
Michal Mraz	Improved probe training for temporal features	Pooled probes outperform last-token; causal steering confirmed	Publication
Vansh Gupta	Phase transitions in temporal horizon	Establishing short-to-long horizon transition	Publication
Ipshita Bandyopadhyay	Uncertainty encoding across reasoning steps	Uncertainty strong at step 1, degrades after reasoning	Publication
Ronen Roy	Context fatigue in long contexts	Entropy collapse is real; partially offset by ICL	Publication
Harleen Kaur Bagga	Sycophancy detection via hidden states	Sycophancy in 1.5–15.5% of cases under social pressure; confidence unreliable as monitor	Publication
Avigya Paudel	Temporal steering on implicit pairs	99.2% probe accuracy; 4× steering effect at layer 22	Pub + Demo
Archit Sharma	Cross-model temporal representation comparison	Signal varies: Qwen robust, Phi weak, Llama intermediate	Publication
Tim Obiso	Causal localization of temporal computation	Last 3 layers; linear component transfers cross-domain	Publication
Adrian Molofsky	Patience degradation under repetition	Degradation is model- and task-specific; SAE phase next	Publication

Table 1: Summary of fellow research questions, key findings, and target outputs.

Surface cue confounds. The baseline staircase result (Dahiya) demonstrates that probe accuracy can be entirely explained by surface features. This motivates broader adoption of rigorous baselines across all probing sub-projects and careful evaluation of whether temporal probes in other contexts might face similar confounds.

5 Future Work

Several threads extend beyond the SPAR program and remain open for follow-on work:

1. **Scale underpowered experiments**: expand the sycophancy dataset (Bagga) to ∼1,000 positive examples to properly power the hidden-state prediction probe. Complete Phase 2 (activation probes + SAEs) of the patience-degradation study (Molofsky) across the remaining models, targeting precursor gap detection.

2. **Test generalization**: 1,000 validated implicit temporal pairs across 40 categories (Paudel) remain available as a clean held-out set. Cross-model causal-patching analysis (Sharma) is the natural next step for determining whether temporal representations are functionally equivalent across architectures.

3. **Confirm steering–safety link**: Obiso's observation that short-term steering increases violent completions requires systematic confirmation with larger sample sizes and controlled generation.

4. **New directions**: extend the phase-transition hypothesis for temporal horizon representations (Gupta), complete the token-trajectory analysis for multihop reasoning (Goel), and further characterize the interaction between context fatigue and in-context learning (Roy).

5. **Publication targets**: the baseline staircase methodology (Dahiya) and temporal steering results (Obiso, Paudel) are the most publication-ready outputs. Sub-projects are being prepared for ICML 2026 workshops, with several also targeting COLM 2026; supporting blog posts and web demos cover preliminary and negative findings.

References

[1] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[2] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[3] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[4] 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.

[5] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

[6] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

[7] Nina Rimsky, Alexander Matt Turner, Wes Gurnee, and Neel Nanda. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.