Does the Model Read the Room?
Probing and Patching Partner Expertise in Collaborative LLM Dialogue

Mika Okamoto[1] and Gabriele Sarti[2]

[1]Georgia Institute of Technology
[2]Northeastern University
mokamoto7@gatech.edu

Abstract

Does a language model's internal representation of its conversational partner's expertise actually drive how it responds? We investigate this for multi-turn collaborative dialogue—a setting where partner expertise must be actively inferred across turns, not read from a prompt. Using ExpertCollab, a controlled corpus of dialogues between LLM-played personas at four expertise levels, we show that partner expertise is linearly decodable from early residual-stream layers via linear probing, peaking at layer 8 before degrading to near-chance well before the network's midpoint. We then ask whether those early-layer representations are causally active. Counterfactual activation patching—replacing residual-stream activations at a candidate layer with those from a different expertise condition—reveals a sharp dissociation: injecting the expertise difference at probe-peak layers leaves a fixed late-layer readout essentially unchanged, while the same difference injected past the network's midpoint propagates nearly completely, an 11× separation confirmed to be content-specific rather than a generic perturbation artifact. This gap reaches the model's output: late-layer patches shift roughly one in three top predicted tokens; probe-peak patches leave the output essentially unchanged, with two independent diagnostics converging on the same transition point. Probing alone would mislocalize the intervention target by over half the network depth—a structural constraint on any attempt to steer partner-conditioned behavior. All results are for a single model family on a synthetic corpus; generalizability across architectures and naturalistic dialogue remains an open question.

1 Introduction

Skilled communicators adapt to their audience: vocabulary, depth, and framing shift based on what the listener knows and what expertise they bring [Clark and Murphy, 1982, Bell, 1984, Clark, 1996]. Whether LLMs do something analogous internally—and where in the network any such partner representation operates—has practical stakes for audience adaptation, sycophancy, and calibrated expert disagreement [Sharma et al., 2024, Wang et al., 2026, Vennemeyer et al., 2026].

One common way to ask where something is encoded in a neural network is *probing*: training a lightweight classifier on the model's internal representations at each layer to test whether a given attribute can be read off there [Alain and Bengio, 2017, Belinkov, 2022]. Studies using this approach show that user attributes are decodable from intermediate layers [Pan et al., 2026, Choi et al., 2025, Zhu et al., 2024], creating a natural but often wrong inference: that the layer where the probe peaks is where the attribute is *computationally relevant*. A probe tells you where information is readable; it says nothing about whether the network actually uses that representation downstream. The causal complement is *activation patching*: replacing activations at a candidate layer with those from a different condition and measuring whether the change propagates to a later position [Vig et al., 2020, Meng et al., 2022, Wang et al., 2023].[1] A growing body of work finds

[1]Throughout, "causally active" refers specifically to within-forward-pass propagation: whether a modified activation at layer L reaches a fixed downstream readout.

that probe peaks and propagation loci frequently diverge—in vision transformers [Huang and Chang, 2025], sparse autoencoder features [Arad et al., 2025], and user-belief representations [Choi et al., 2025]. We ask whether the same holds for a relational, multi-turn attribute that must be inferred across a conversation rather than read from a single prompt.

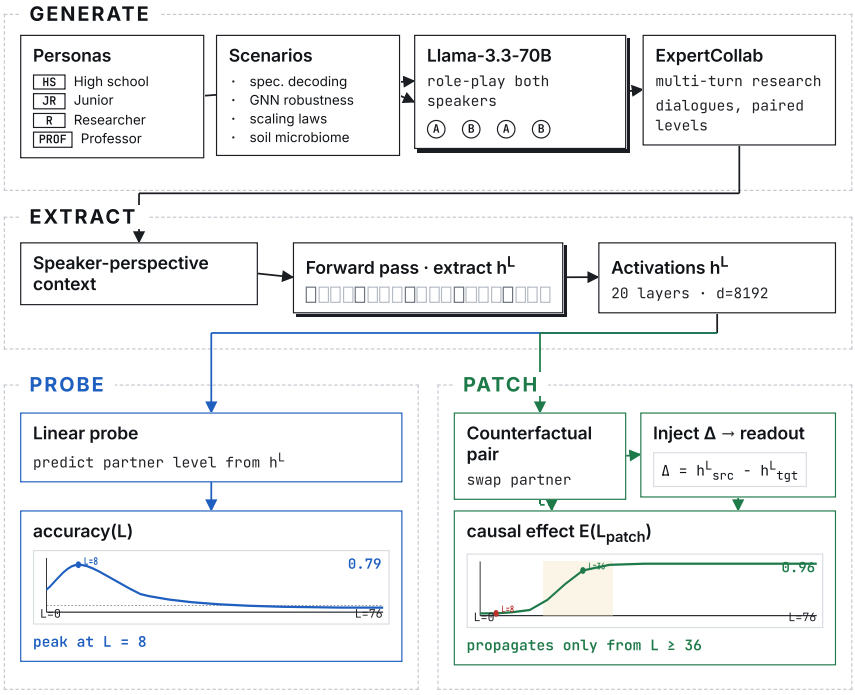

Figure 1: `Llama-3.3-70B` role-plays both speakers in each ExpertCollab dialogue (**Generate**); residual-stream activations are extracted at the participant-perspective position every four layers (**Extract**); linear probes test where partner expertise is linearly readable (**Probe**, blue); counterfactual activation patching tests whether early-layer representations propagate to a fixed late-layer readout (**Patch**, green).

Our analysis yields three findings: partner expertise shows an early probe peak that degrades before the network's midpoint, contrasting with statically-specified attributes that persist throughout; the probe-peak layer is causally inert while mid-to-late layers propagate nearly completely; and this propagation boundary coincides with the generation boundary, confirmed by two independent diagnostics converging on the same transition point. Full counterfactual generation—running the model to completion under modified activations across multiple turns—remains the primary open extension.

Contributions. (1) A concrete demonstration that, for a relational multi-turn attribute, probing mislocalizes the intervention target by over half the network—extending the decodability≠causality finding from vision transformers [Huang and Chang, 2025] and sparse autoencoder features [Arad et al., 2025] to conversational partner modeling. (2) A *participant-perspective* activation extraction methodology, together with ExpertCollab, a controlled corpus of multi-turn research-planning dialogues between LLM-played personas at four expertise levels. (3) A counterfactual patching design with norm-matched random-delta controls and behavioral linkage measurements connecting the propagation finding to the model's output distribution. All results are for a single model family on a synthetic corpus; generalizability across architectures and naturalistic dialogue remains an open question.

2 Related Work

Implicit personalization. LLMs systematically adjust outputs based on inferred user attributes [Jin et al., 2024]. A prominent downstream failure is sycophancy [Sharma et al., 2024], with mechanistic accounts linking it to user-opinion representations overriding learned knowledge in deeper layers [Wang et al., 2026] and distinct sycophantic behaviors mapping to separable internal subspaces [Vennemeyer et al., 2026]; Papadatos and Freedman [2024] use linear sycophancy probes as a fine-tuning signal. LLMs can represent beliefs and mental states of conversational partners in their activations [Zhu et al., 2024], but this literature focuses on vignette-style tasks rather than free-form multi-turn generation. We extend this to a less stylized setting: tracking a collaborator's competence level during generation, where the attribute must be inferred across turns. A growing literature applies probes specifically to multi-turn dialogue [Jaipersaud et al., 2025], with Ichmoukhamedov and Martens [2025] cautioning that probes trained on short conversational formats may not generalize to longer ones.

Causal localization and the decodability/propagation gap. Transformer language models maintain a running representation vector—the *residual stream*—to which each layer adds its computation via a skip connection, rather than replacing the previous representation outright [Elhage et al., 2021]. What a probe reads at layer L therefore reflects everything written into this stream up to that point, but that information need not persist forward if subsequent layers overwrite or ignore it. Activation patching [Vig et al., 2020, Geiger et al., 2022, Meng et al., 2022, Wang et al., 2023] tests this directly by injecting a modified activation at some layer and measuring whether the change survives to a downstream position. Prior work finds that probe peaks and causal loci frequently diverge: Huang and Chang [2025] demonstrate this in vision transformers; Arad et al. [2025] find it for sparse autoencoder features; Choi et al. [2025] show that user-attribute representations are reliably decoded from intermediate layers but that actually steering behavior requires interventions across multiple layers. Our experiment extends this framework to a relational, multi-turn conversational attribute.

Layered organization of transformer LMs and linear probing. Transformer language models organize linguistic information hierarchically by depth [Tenney et al., 2019, Jawahar et al., 2019, Geva et al., 2021]: a probe peaking in early layers most naturally exploits surface regularities—hedging patterns, vocabulary formality, syntactic complexity. Linear probes [Alain and Bengio, 2017, Belinkov, 2022] test whether a property is linearly encoded at a given layer; Hewitt and Liang [2019] introduce the selectivity criterion to distinguish probe expressivity from representation content. We restrict to linear probes for tractability in the high-dimensional, low-sample regime ($p = 8192 \gg n$) and for compatibility with the downstream causal step.

3 Methods

We now describe the corpus construction, activation extraction protocol, and experimental design that instantiate this framework.

3.1 ExpertCollab

ExpertCollab comprises synthetic research-planning dialogues between two LLM-played agents at four expertise levels—*high school student* (HS), *junior researcher* (JR), *researcher* (R), and *professor* (PROF)—generated by `Llama-3.3-70B-Instruct-Turbo` via Together AI. Each agent receives a persona prompt specifying engagement style, characteristic blind spots, and interaction tendencies appropriate to its expertise level; agents communicate in ways that reflect their level rather than announcing it. Synthetic dialogue permits controlled construction of expertise counterfactuals impossible to arrange naturalistically; LM-played personas reproduce individual-level behavioral patterns with notable fidelity [Park et al., 2026]. The single generator controls for style and topic-distribution confounds; we discuss the resulting same-model confound in section 5.

Personas specify no explicit credentials or job titles—expertise must be inferred from how an agent communicates, not from stated affiliations. All 28 conversations were used without modification. The corpus

covers seven distinct expertise-level pairings (drawn from same-level and cross-level combinations of the four levels) across four research scenarios (speculative decoding, GNN robustness, scaling laws, soil microbiome), yielding $n = 28$ base conversations of 12 turns each with randomized first speaker.

3.2 Participant-Perspective Probing

When analyzing a conversation, a key design choice is whose perspective to take. Rather than feeding the model the full transcript as a passive reader, we extract activations at the position the model occupies when it is actually about to generate its next turn—what we call the *participant perspective*. This matters because the same exchange looks different from each speaker's viewpoint: from speaker s's perspective, the partner's prior turns are the input (`user`) and s's own prior turns are its previous outputs (`assistant`). Formatting the context this way with `add_generation_prompt=True` ensures the extracted representation reflects what the model knows at the moment of generation—its working state as it decides what to say next, not a static reading of the transcript.

Residual-stream activations are extracted at the final token position of every fourth layer of `Llama-3.3-70B-Instruct` (layers 0–76, 20 checkpoints total, $d = 8192$) via NDIF/`nnsight`. Both speakers' activations are pooled into role-agnostic targets ($n = 56$ total speaker-turn instances). L2-regularized logistic regression probes [Alain and Bengio, 2017] ($C = 0.01$, stratified 5-fold cross-validation; C selected to prevent degenerate solutions in the high-dimensional low-sample regime; we did not tune C against layer profiles) are trained per (layer, turn) cell on two targets: `partner_level` (4-class: the partner's expertise level) and `own_level` (4-class: the generating speaker's own level). Given the small sample size ($n \approx 9$–11 per class per fold), per-cell CV accuracy estimates carry substantial uncertainty; results should be interpreted in terms of the overall layer profile rather than point estimates at individual cells.

Surface-feature baseline. To test how much of the neural probe signal reflects simple lexical regularities rather than deeper processing, we also train a logistic regression classifier on five surface-linguistic statistics computed from each speaker's conversational history at extraction time: hedging rate (fraction of hedge-word tokens), mean sentence length, question rate, type-token ratio, and vocabulary complexity (mean inverse log-frequency of content words). The baseline uses the same `partner_level` target and cross-validation protocol ($C = 0.01$, stratified 5-fold).

3.3 Counterfactual Activation Patching

Experiment design. Probing tells us where partner expertise is *readable* in the network. The patching experiment asks whether those representations are *causally active*—whether modifying them actually changes what the network computes downstream. Concretely: if we replace the residual-stream vector at layer L during a forward pass with one taken from a different expertise condition, does that substitution survive to a fixed late-layer readout at layer 76? If so, the representation at layer L is on the causal path; if not, it is effectively overwritten or bypassed before the output.

Formally, for a pair of dialogues differing only in one participant's expertise level, we define the *expertise delta* $\Delta = \mathbf{h}^L(\text{source}) - \mathbf{h}^L(\text{target})$, where $\mathbf{h}^L$ is the residual-stream vector at layer L. We add Δ to the target's stream at layer L and run the forward pass to completion, then check whether the late-layer readout shifted toward the source.

Counterfactual pairs. Each dialogue has two participant roles. We hold one participant's persona constant (participant B) while varying the other's expertise level across all four levels (participant A), creating a set of conversations that differ only in A's expertise. Pairing these across all ordered expertise-level combinations yields the counterfactual pairs used for analysis. The primary analysis uses $n = 48$ cross-level pairs in the *partner-encoding* ablation: we patch participant B's residual stream (whose context changes as their partner's expertise varies, so Δ captures how B encodes its partner) and read out the `partner_level` probe at layer 76.

Readout and normalization. We fix the readout at $L_{\text{probe}} = 76$, downstream of every candidate patch layer $L_{\text{patch}} \in \{0, 4, \ldots, 76\}$, so any nonzero effect E requires propagation through the intervening com-

putation. To measure propagation, we compare the probe's output on the patched forward pass against two baselines—what the probe would output if the full source activation were present (p_{from}) versus the unpatched target (p_{to}):

$$E(L_{\text{patch}}, T) = \frac{p_{\text{patched}} - p_{\text{to}}}{p_{\text{from}} - p_{\text{to}}}.$$

$E \approx 1$ means the patched stream fully recovered the source's late-layer representation; $E \approx 0$ means the injected delta did not reach the readout. Cells where $|p_{\text{from}} - p_{\text{to}}| < 0.05$ are masked to exclude pairs where the probe cannot distinguish the two conditions (20.1% of cells). All reported intervals are 95% cluster bootstrap CIs resampling pairs ($B = 1000$).

Controls. Setting $L_{\text{patch}} = L_{\text{probe}}$ (trivial recovery) verifies the measurement is working: the probe reads the patched activation directly, so E should be ≈ 1 regardless of content. Replacing Δ with a norm-matched random vector—same magnitude, random direction—tests whether any early-layer null result reflects content-specific failure to propagate or simply generic decay of perturbations over network depth.

Behavioral metrics. To connect the propagation finding to the model's actual output, we record two additional quantities on the same patched and unpatched forward passes across all $n = 48$ pairs: (1) *token overlap*—the fraction of top-5 predicted tokens at the generation position shared between the patched and unpatched run, and (2) *representation shift* $(1 - \cos\theta)$—cosine dissimilarity between the patched and unpatched residual-stream activations at layer 76. Unlike the probe-based E, these metrics require no baseline normalization and are computed for every pair at all candidate patch layers.

4 Results

4.1 Partner Expertise Is Decodable Early but Does Not Persist

The probing results (fig. 2) reveal a structural asymmetry between two qualitatively different encoding tasks. `own_level`—the speaker's own expertise, specified in the system prompt—has mean CV accuracy exceeding 0.60 at every layer from 0 through 76. The network's skip connections carry this information forward throughout the entire model without degradation, establishing that the architecture is capable of maintaining an attribute across depth when it is statically provided.

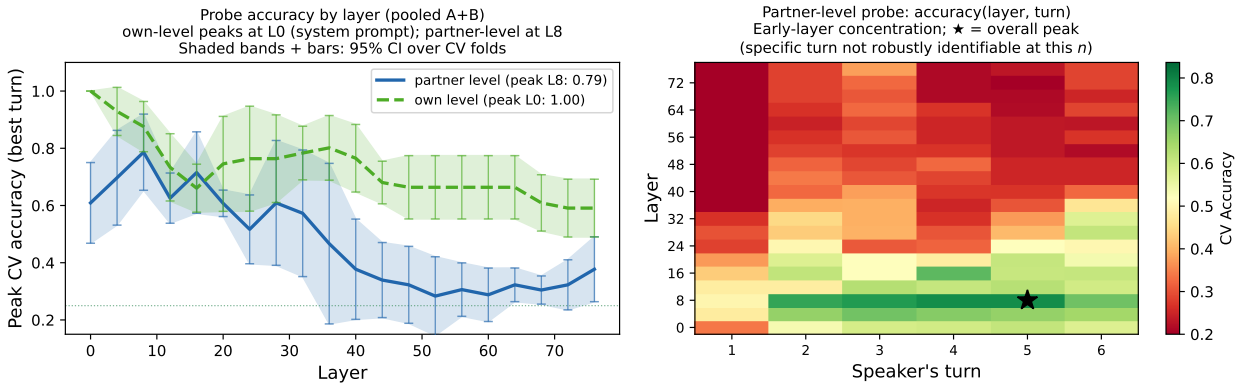

Figure 2: Pooled participant-perspective probe accuracy ($n = 56$). **Left:** Peak CV accuracy by layer; partner-level peaks at layer 8 (acc = 0.786) and declines to near-chance by layer 40; own-level stays above 0.60 throughout. **Right:** Partner-level accuracy heatmap (layer × turn); early-layer concentration is robust, but the starred maximum (layer 8, turn 5) is not reliably identifiable at this sample size.

`partner_level` behaves differently in every respect. It must be inferred from the partner's utterances rather than read from a prompt entry, and the probe profile reflects this: accuracy builds across turns as the model accumulates evidence, reaching its maximum at layer 8 (acc = 0.786, turn 5 in the best-turn

analysis), then degrading monotonically to near-chance by layer 40. We note that the specific turn of peak accuracy should be interpreted cautiously given the small n; with roughly 9–11 training instances per fold, per-turn estimates are noisy and the turn-5 optimum is not robustly identifiable. The consistent finding is the early-layer peak and subsequent degradation, not the specific turn coordinate. The peak is later, lower, and impermanent.

The profile—early peak, monotone decline, absent at late layers—is consistent with an early-layer encoding built primarily on surface-linguistic features (hedging density, vocabulary formality, syntactic complexity) that dominate early transformer representations [Tenney et al., 2019, Geva et al., 2021]. The surface-feature baseline tests this directly: a logistic regression on five lexical statistics achieves 0.32 accuracy on the same 4-class target—above chance (0.25), but 0.45 points below the neural probe's 0.786. The early-layer encoding captures structure that simple surface statistics miss, making the question of whether it propagates forward non-obvious. The patching experiment answers it.

4.2 Early-Layer Deltas Do Not Propagate; Mid-to-Late Layers Do

Probing established that partner expertise is linearly readable from early-layer representations; the question is whether those representations are causally active. Protocol calibration is verified: with $L_{\text{patch}} = L_{\text{probe}} = 76$, per-turn E values are $\{1.09, 1.00, 1.00, 1.00, 0.98, 1.00\}$ (mean 1.01), confirming the measurement is working correctly. The main layer profile (fig. 3, left) gives a clear answer: at the probe-peak layers (0 and 4), $E = 0.009$ and $E = 0.012$—the injected delta leaves the late-layer readout essentially unchanged, meaning the network's intervening computation overwrites or ignores the modification entirely. The early band (layers 0–12) yields mean $E = 0.085$ [0.050, 0.123]. E begins rising at layer 8 and transitions sharply through layers 20–36, reaching 0.496 at layer 20, 0.623 at layer 28, and 0.858 at layer 36 ($E = 0.858$ [0.801, 0.911]). By layer 36 propagation is nearly complete, and layers 40 through 72 all exceed 0.90. The early-layer representation the probe can read is not on the causal path to the output.

The random-delta control (fig. 3, right) confirms the early-layer null is content-specific, not a generic property of perturbation decay. Replacing the expertise delta with a norm-matched random vector produces confidence interval widths 10–15× larger than the real experiment with no monotone structure. It is the expertise-specific direction of the delta, not its magnitude, that determines whether it propagates.

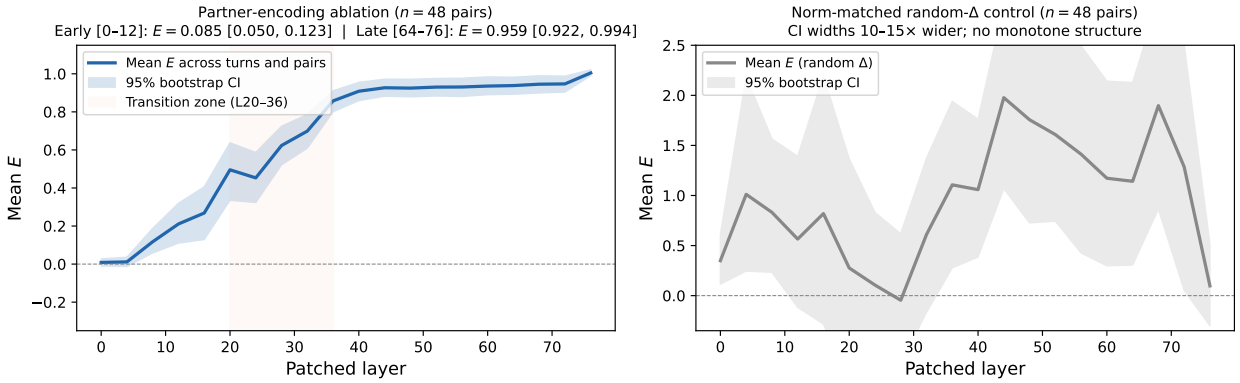

Figure 3: Layer profile of normalized causal effect E ($n = 48$ pairs). **Left:** E rises monotonically from near zero at early layers to near one at late layers (11.3× separation); shaded region marks the transition zone (layers 20–36). **Right:** Norm-matched random-Δ control yields no monotone structure and CI widths 10–15× larger, ruling out generic perturbation decay.

The heatmap (fig. 4) shows that this depth-dependence holds regardless of where in the conversation the patch is applied: every column is red at the bottom (early layers, near-zero propagation) and green at the top (late layers, near-complete propagation), with the same transition band running horizontally across all six speaker turns. The causal structure is a property of network depth, not conversational position.

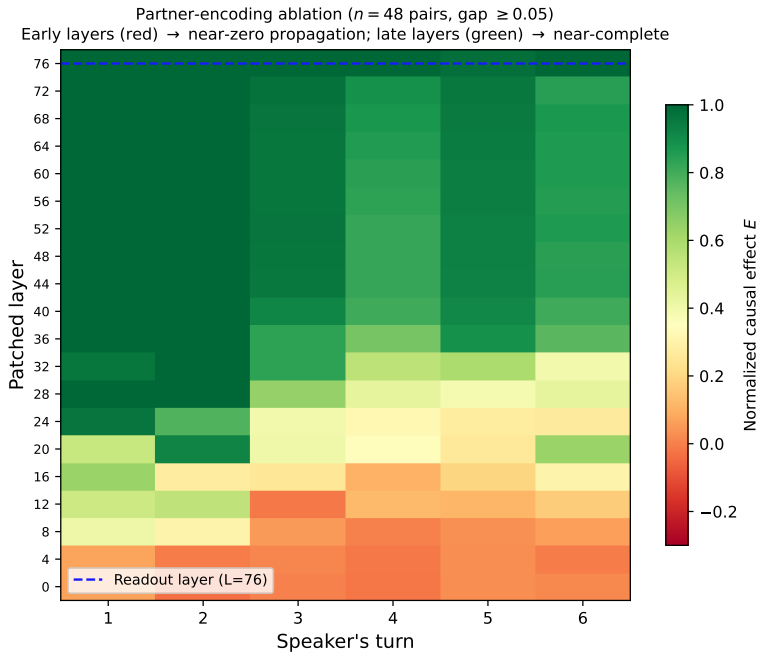

Figure 4: $E(L_{\text{patch}}, T)$ heatmap ($n = 48$ pairs, 20.1% of cells masked). The early-layer null and late-layer recovery are uniform across all turns, indicating a property of network depth rather than conversational position.

Directionality and domain stratifications are consistent with the main result. Patching toward a higher-expertise source ($n = 24$) and patching toward a lower-expertise source ($n = 24$) yield $E = 0.653$ [0.597, 0.713] and $E = 0.660$ [0.614, 0.704] respectively—a 0.007-point difference with fully overlapping confidence intervals, indicating that recovery is direction-symmetric. Domain recovery varies modestly: `gnn_robustness` $E = 0.691$, `scaling_laws` $E = 0.668$, `speculative_decoding` $E = 0.659$, `soil_microbiome` $E = 0.598$; the soil scenario, the only domain outside machine learning, shows the lowest recovery, consistent with expertise-discriminating signals being less sharply differentiated there.

A complementary ablation patching the other participant's stream and reading out the own-level probe showed qualitatively similar early/late dissociation, but the near-perfect own-level probe accuracy at all layers produces small normalization denominators and unstable E values. We treat this as exploratory.

The probe finds partner expertise at layer 8; patching at layer 8 moves the late-layer readout by 0.118. The patch needs to reach layer 36 before it moves the readout by more than 0.85. The probe peak and the propagation zone are in different parts of the network, and the gap is content-specific.

4.3 The Propagation Gap Reaches the Model's Output

Whether this internal propagation structure reaches the model's output is what the behavioral metrics answer (fig. 5). Token overlap declines gradually from the earliest patch layers—0.97 at layers 0–4, falling to 0.85 by layer 8 and settling near 0.67 by layer 32. Even low-magnitude perturbations at shallow layers produce a small output-side footprint, consistent with the network's additive (residual) structure: any injected delta contributes to the final-layer computation and can displace marginal predicted tokens. But the effect is small and graded, not the sharp transition we would expect if early layers were causally relevant.

Representation shift ($1 - \cos\theta$) tells the sharper story. It is essentially flat (≤ 0.01) at layers 0–4, then rises sharply beginning at layer 8, reaching 0.40 by layer 32 where it saturates. This onset matches the transition in E—and it does so without using a trained probe: cosine dissimilarity at layer 76 requires no classifier, yet identifies the same boundary as the probe-based causal effect. Two independent diagnostics, designed around entirely different readouts, converge on the same layer-8 transition, which strongly implies this is a structural property of the computation rather than an artifact of either readout method.

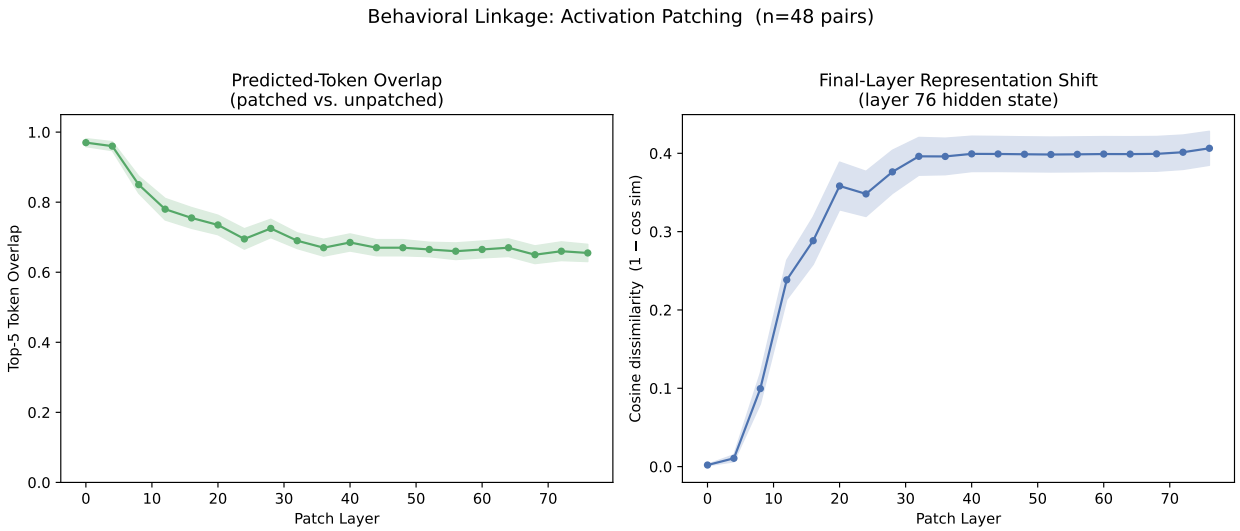

Figure 5: Behavioral effects of activation patching ($n = 48$ pairs). **Left:** Top-5 token overlap declines gradually from 0.97 at layers 0–4 to ≈0.67 by layer 32. **Right:** Cosine dissimilarity ($1 - \cos\theta$) at layer 76 is flat ($\leq$0.01) at layers 0–4 then rises sharply from layer 8, matching the onset in E without requiring a trained probe.

Together the metrics close the chain. Probing told us where partner expertise is readable (layer 8). Patching told us where it propagates (layer 36+). The behavioral measurements confirm that the propagation boundary is also the generation boundary: patches at probe-peak layers leave the model's output nearly untouched, while the same delta injected past the network's midpoint shifts roughly one in three of the model's top predicted tokens. The gap between readable and causally active extends all the way to what the model says.

5 Discussion and Future Work

Practical implications. The central finding constrains where interventions on partner-conditioned behavior would need to be applied. If a practitioner wanted to modify how a model adapts to its conversational partner—whether to reduce sycophancy, calibrate explanation depth, or steer expert disagreement—probing would direct them to the layers where the partner-expertise signal is strongest. The patching result shows that intervening there would have negligible downstream effect; the intervention would need to target the latter half of the network, where the representation actually propagates. The mislocalization spans over half the network—a concrete instance of the broader principle that decodability does not imply causal relevance [Huang and Chang, 2025, Arad et al., 2025].

What the early-layer signal is. The surface-feature baseline rules out the simplest explanation for the early-layer peak: the 0.45-point gap between lexical accuracy (0.32) and neural probe accuracy (0.786) shows the early-layer encoding captures structure that surface statistics do not. The profile is consistent with richer representations—discourse dynamics, reciprocal question–answer structure, topic-authority negotiation—that aggregate measures miss. This reframes the main finding: the gap between probe peak and propagation zone is not a story about the probe finding noise. The model forms a rich early encoding of partner expertise that nonetheless fails to route forward. Whether this representation is overwritten by subsequent processing, written into a non-propagating subspace, or simply fails to connect to the channels that carry information to later layers remains an open mechanistic question. Cross-model replication and nonlinear readouts can begin to constrain these possibilities.

Scope and generalizability. All results reported here are for a single model (`Llama-3.3-70B-Instruct`) on ExpertCollab, a synthetic corpus generated by the same model. Three limitations follow directly.

First, the specific layer boundaries (probe peak at layer 8, propagation onset at layer 36) are properties of this architecture and should not be taken as universal constants; they may shift substantially for models of different depth, width, or training regime. Second, because ExpertCollab is generated by the same model being probed, the expertise representations we study may reflect self-recognition of the model's own generation patterns rather than genuine inference about an external interlocutor—a form of circularity that naturalistic corpora or cross-model generation would mitigate. Third, the corpus is small ($n = 28$ base conversations, $n = 56$ speaker-instances for probing, $n = 48$ pairs for patching) and the reported peak CV accuracies carry uncertainty that the current analyses do not fully quantify. We treat these results as establishing a proof of concept and characterizing one instance of the probe-peak/propagation-zone dissociation for a relational dialogue attribute, rather than as claims about LLMs in general.

Open questions. Cross-model replication is the highest priority: all results are for a single model family, and the single-generator confound in ExpertCollab means the probed model also generated the corpus. Running the same pipeline on a second model family with ExpertCollab held fixed would test architecture-invariance and dissolve this confound. More directly, the identified propagation zone is a natural target for steering experiments: activation steering vectors [Turner et al., 2024, Arditi et al., 2024] applied at layers 36+ would test whether the partner-expertise representation can be externally modified, connecting this internal finding to the sycophancy and audience-adaptation behaviors that motivate the problem. Nonlinear readouts and own-level patching remain secondary open questions.

6 Conclusion

We asked whether the layer where a probe finds a conversational partner's expertise is the layer from which that information propagates forward. It is not. Partner expertise is decodable at layer 8 but causally inert there: injecting the corresponding activation difference at probe-peak layers leaves the late-layer readout unchanged, while injecting the same difference at layers 36+ shifts it by over 85%. The gap is content-specific, direction-symmetric, and stable across turns and domains. It also reaches the model's output: the same transition that marks where deltas begin propagating internally marks where the model begins generating differently—with late-layer patches shifting roughly one in three top predicted tokens, and two independent diagnostics converging on the same layer-8 onset. For anyone seeking to steer how a model adapts to its conversational partner, the probe would point to the wrong part of the network by over half its depth.

Acknowledgements

This work was carried out as part of the SPAR Spring 2026 cohort under the mentorship of Gabriele Sarti. We thank the National Deep Inference Fabric (NDIF) for compute access used in all activation extraction and patching experiments. We are grateful to Afitab Iyigun, Jacopo Zacchigna, Hengxu Li, Vinit Patel, and Valeriya Zelenkova for helpful discussions throughout the project.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=ryF7rTqgl. arXiv:1610.01644.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10241–10259, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48 (1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Allan Bell. Language style as audience design. *Language in Society*, 13(2):145–204, 1984. doi: 10.1017/S004740450001037X.

Dami Choi, Vincent Huang, Sarah Schwettmann, and Jacob Steinhardt. Scalably extracting latent representations of users. https://transluce.org/user-modeling, November 2025.

Herbert H. Clark. *Using Language*. "Using" Linguistic Books. Cambridge University Press, 1996.

Herbert H. Clark and Gregory L. Murphy. Audience design in meaning and reference. In Jean-François Le Ny and Walter Kintsch, editors, *Language and Comprehension*, volume 9 of *Advances in Psychology*, pages 287–299. North-Holland, 1982. doi: https://doi.org/10.1016/S0166-4115(09)60059-5. URL https://www.sciencedirect.com/science/article/pii/S0166411509600595.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting vits. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. URL https://openreview.net/forum?id=RyC58MT4aT.

Timour Ichmoukhamedov and David Martens. Exploring the generalization of llm truth directions on conversational formats, 2025. URL https://arxiv.org/abs/2505.09807.

Brandon Jaipersaud, David Krueger, and Ekdeep Singh Lubana. How do llms persuade? linear probes can uncover persuasion dynamics in multi-turn conversations, 2025. URL https://arxiv.org/abs/2508.05625.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, 2024.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=niUroX9EOd.

Henry Papadatos and Rachel Freedman. Linear probe penalties reduce LLM sycophancy. In *Workshop on Socially Responsible Language Modelling Research*, 2024. URL https://openreview.net/forum?id=6N2yES22rG.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. Llm agents grounded in self-reports enable general-purpose simulation of individuals, 2026. URL https://arxiv.org/abs/2411.10109.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in LLMs, 2026. URL https://openreview.net/forum?id=d24zTCznJu.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.