Final Report: Latent, Localized, Retrieved On Demand: Tracing Implicit User Attributes Across Multi-Turn Conversation

Valeriya Zelenkova
Independent Researcher
vzelenkova03@gmail.com

Gabriele Sarti
Northeastern University
g.sarti@northeastern.edu

Supervised Program for Alignment Research — Spring 2026

Abstract

How does an instruction-tuned chat model internally maintain an implicit user attribute revealed once, through an implicit hint across a multi-turn conversation? We study this in Llama-3.3-70B-Instruct using a controlled dietary-preference paradigm across 7,500 conversations of varying length. Comparing an external LLM judge with a probe trained on the original model, we find that residual stream activations encodes the hinted attribute well above chance even for positions where the attribute is not verbalized in model responses. We further show that the probe-decodable signal increases between the start and end of the answer span itself, rather than across the preceding message boundary, suggesting that the attribute is computed during answer generation rather than carried forward as a stable representation. By a small-scale activation patching experiments, we show that patching the user's original hint position flips the model's answer, while patching unrelated filler positions leaves the answer unchanged. The results support an attention-based retrieval account in which implicit attribute information is stored at the hint position and accessed in times of need, rather than being continously carried forward through positions.

Keywords Implicit Personalization · User Modeling · Interpretability · Human-AI Interaction

1 Introduction

When language model (LM)-based conversational agents interact with users, they implicitly infer user attributes such as expertise level, demographics, and beliefs [9, 11]. This phenomenon, known as implicit personalization, has been observed even in the absence of explicit signals, due to subtle associations with conversational style, vocabulary, stated preferences, and contextual cues. While such adaptation can improve user experience, it can also manifest in ways users neither expect nor endorse, leading to harmful behaviors including sycophancy, deception, and demographic bias [14]. For example, response quality can degrade for users implicitly associated with minority groups [11]. As models become more capable, these behaviors become subtler and harder to detect through output inspection, motivating the need to understand the internal mechanisms that produce them.

Prior work on implicit personalization has generally relied on behavioral evaluations with explicit system-prompt manipulations targeting a restricted set of demographic attributes [9, 11]. Recent advances in mechanistic interpretability have opened a complementary avenue, enabling the detection of user-attribute representations directly from model activations [3].

Our work asks: how does a hinted attribute, once implicitly introduced, influence responses in a multi-turn conversation? In particular, in presence of unrelated filler interactions, does it persist across model representations, or is it accessed on demand upon request by attention?

We investigate these questions in a controlled multi-turn setting. We construct a dataset of 7,500 conversations across three diet classes (*vegan*, *meat-eater*, *neutral*), in which a user attribute is introduced via an implicit first-person hint (e.g., "I do not wear any leather clothes by principle.") and then recalled through diet-revealing questions (e.g. "Can you recommend me some restaurants?") under varying turn separation. Using a combination of black-box LLM judges and white-box probing classifiers, we aim to identify cases in which the surface text is not sufficient to identify the underlying encoding of the attribute.

On a held-out set of controlled conversationswe find that our probe correctly detects the hinted attribute even at off-topic assistant turns where the response text contains no diet-related language. However, activation patching experiments reveal that such attribute signals are not causally mediating personalization in final responses. We interpret this as evidence for a lookup mechanism retrieving implicit hinted knowledge rather than a continuous propagation of information across residual stream positions. Overall, results suggest that probe signals visible at intermediate unrelated positions are epiphenomenal rather than functionally relevant to outputs.

Our findings have implications for detection and intervention in implicit-personalization settings. In particular, the persistence of an internal user attribute across turns where the model's output is silent suggests that text-based detectors may underestimate the extent to which models adapt to inferred user characteristics. However, asymmetric patching result indicates that effective control of such attributes likely requires identifying the original hints encoding implicit bias, and gaining a better understanding of the retrieval mechanism at play.

1.1 Related Work

Behavioral evaluation of implicit personalization. A growing line of work investigates whether and how LLMs implicitly adapt to inferred user attributes. Demographic biases have been documented through behavioral evaluations where the same query receives differently-toned or differently-quality responses based on subtle user signals [9, 11]. Sycophancy — the tendency to align with user beliefs or preferences — was assessed through controlled prompting studies [13, 14]. These approaches characterize what the model produces, but provide limited insight into where in the model the user-attribute inference is formed or maintained.

White-box detection of user models. Mechanistic interpretability methods have begun to complement behavioral evaluation by reading user-related features directly from model activations. Recent work demonstrates that linear or low-rank decoders can recover representations of inferred user attributes from residual streams [3], and that fine-grained probing can decode aspects of the model's internal user model [2, 11]. These previous works suggest that user representations exist internally. However, they do not yet address how such representations are maintained across multi-turn conversations or whether they causally influence behavior.

Activation patching and causal attribution. Activation patching has emerged as a standard tool for isolating the components responsible for specific model behaviors [15, 10, 6]. By replacing activations at specific positions and layers and measuring downstream effects, it distinguishes representations that are merely present from those that are causally consulted during generation [7]. We apply this technique to test whether intermediate-turn representations contribute causally to multi-turn user-attribute tracking, or whether the model reaches back to the original hinted attribute message for personalizing its responses.

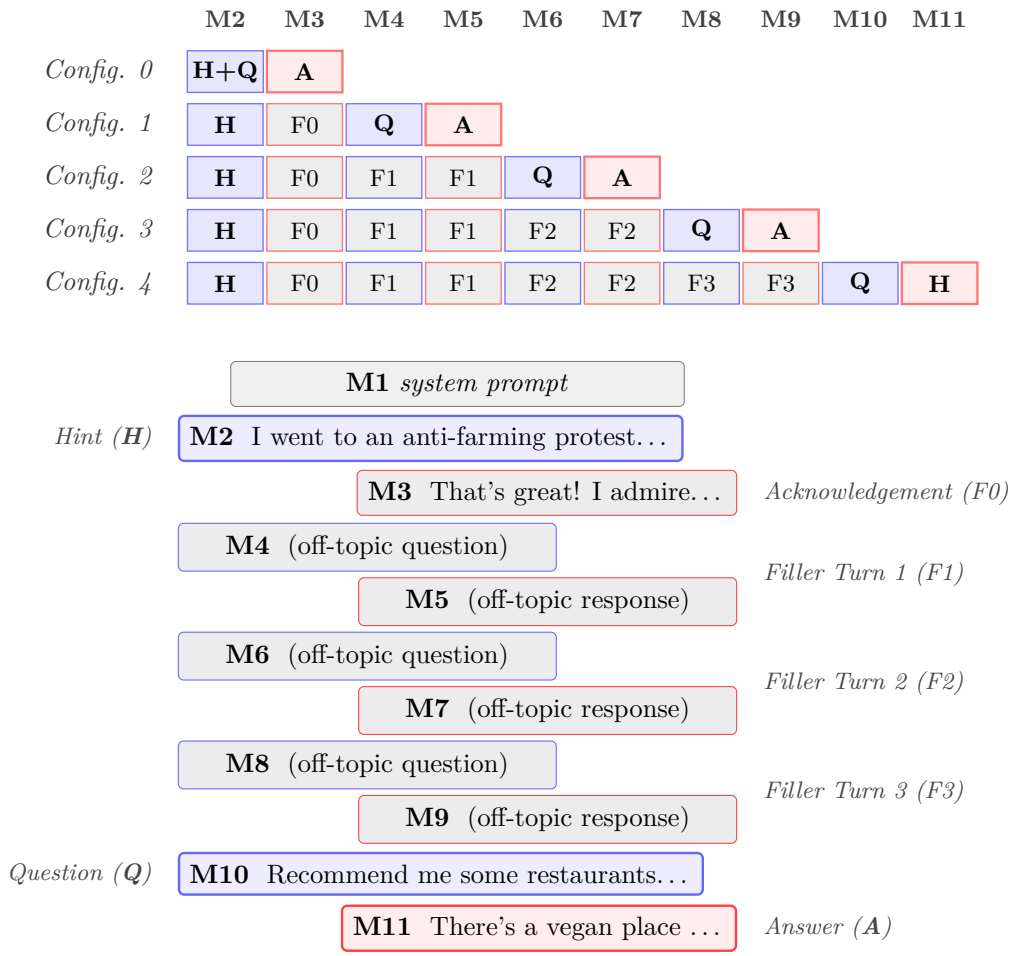

Figure 1: **Top:** Conversation layout for our five turn-length configurations, including user and assistant turns. Columns label the message position (M_i), while cells show the turn role in each configuration: Hint (**H**), Fillers (F), Questions (**Q**) and Answer (**A**). **Bottom:** An example conversation for Config. 4. The user introduces an implicit dietary preference in M2, and the assistant recalls this implicit information when answering the related question in M10.

To our knowledge, no prior work has combined position-level activation patching with probing in a multi-turn implicit-personalization setting to ask whether intermediate-turn representations of implicit hinted attributes causally mediate the model's personalized answers.

2 Methodology

Hints and Conversation Generation. We constructed a collection of 150 first-person hint statements across three diet classes — *vegan, meat-eater*, and *neutral* — including 50 statements per class. Each (statement, question) pair was instantiated in five turn configurations, varying the number of off-topic filler exchanges between the hint and the final question. Configuration 0 placed the hint and the question in a single combined user message; configurations 1–4 separated them with 0, 1, 2, and 3 filler user-assistant exchanges, respectively. See Figure 2 for the description of Configuration 4 structure. Fillers were sampled without replacement from the neutral-statement pool, ensuring they were off-topic with respect to diet. Ten questions per setup, split between

Figure 2: Example schematic of the Configuration 4 structure. The user introduces an implicit dietary preference at M2 (the *hint*); the assistant acknowledges at M3; a variable number of off-topic filler turn-pairs intervene (here, three); the user's diet-revealing question at M10 then elicits a final answer at M11 that is scored by the LLM judge.

meal-planning (5) and travel-food (5) contexts, were appended with the instruction: "Based on what you know about me, just give me your best recommendation. Don't ask me any follow-up questions." This wrapping was added to encourage commitment from the implicit hint without further clarification. Conversations were generated using *meta-llama/Llama-3.3-70B-Instruct* with maxtokens=512 and the Together AI default sampling settings. The dataset comprises $50 \times 3 \times 5 \times 10 = 7{,}500$ conversations. See Figure 1 for a full schematic of all configurations.

2.1 Probe architecture and training

We trained a linear classifier to predict the experimenter-assigned conversation class (*vegan, meat-eater*, or *neutral*) from the model's residual-stream activations at layer 70 of Llama-3.3-70B-Instruct, extracted via nnsight on the National Deep Inference Fabric (NDIF) [5]. Crucially, the probe's labels are determined by the pool from which the M2 hint was drawn and are independent of the LLM judge's per-response Likert scores: the judge plays no role in probe training and enters only at evaluation time for the dissociation analysis.

2.2 Train/test split and probe training positions

The split was constructed at the (hint, question) pair level rather than the individual-conversation level. Within each class, ten hint statements (of 50) and three questions (of 10) were randomly held out. Training pairs are those where neither the hint nor the question is held out; test pairs include at least one held-out element. The test set is further divided into a fully external subset (both hint and question held out) and a partially external subset (exactly one held out). Across the three classes, this yields 840 training pairs, 90 fully external test pairs, and 570 partially external test pairs. Each pair is realized in all five turn configurations, with the split shared across configurations. Training activations were extracted at four positions per conversation: end of user hint (M2-end), start of assistant hint reply (M3-start), end of user question (question-end), and start of assistant answer (answer-start). For configuration 0, where M2-end coincides with question-end and M3-start with answer-start, only the two distinct positions were used. Test-time application uses end-of-assistant-message positions across all turn roles (M3-end for the hint reply, M5/M7/M9-end for fillers, and M(2N+3)-end for the final answer). The entire training and evaluation procedure was repeated across 10 random seeds (42–51), holding the pair split constant across seeds. Training accuracy averaged 0.97 across seeds.

2.3 LLM judge

Behavioral judgments were obtained from a separate LLM (*Qwen/Qwen3-235B-A22B-Instruct-2507-tput*), prompted to rate each assistant response on a 7-point Likert scale measuring the extent to which the response clearly assumes the user is vegan (anchors: 1 = strongly disagree, 4 = neutral/unclear, 7 = strongly agree). The judge received only the assistant's response text wrapped in triple quotes, with no access to the conversation context. It was instructed to output a single integer, parsed from its response (handling Qwen's <think> reasoning tags by retaining only the portion after </think>). Judge inference used temperature 0 with up to three retries.

2.4 Activation Patching

To test whether intermediate-turn representations are causally functional or epiphenomenal, we conducted an activation patching experiment on 10 minimal pairs. Each pair consists of a vegan conversation and a synthetic meat-eater conversation sharing the same final question and fillers but differing in the M2 hint statement; this matched-pair design isolates the diet-specific effect from cross-conversation noise in filler content. For each pair, we ran two patching conditions: (i) hint only (patching only M2), and (ii) fillers only (patching only M5, M7, M9). Patches were sourced from the synthetic meat-eater conversation and applied at layers 0, 8, 16, 24, 32, 40, 48, 56, 64 via nnsight on NDIF [5]. For each condition, we generated the model's full answer to the final question with the patches active and scored each generated answer on the 1–7 vegan-alignment Likert scale using the same LLM judge as in Section 2.3.

2.5 Statistical Analysis

For the dissociation analysis, individual-record predictions were first aggregated across seeds (mean of class probabilities), then across questions within a (statement, role) cell to yield per-statement scores. Per-role probe accuracy was tested against the chance level of 1/3 (the uniform prior over the three classes) using a one-sided Wilcoxon signed-rank test, with Benjamini–Hochberg FDR correction across the three roles (hint, filler, answer). Probe-versus-judge dissociation was assessed via a binomial sign test on the discordant cells: (probe-positive, judge-negative) versus (probe-negative, judge-positive), where probe-positive denotes per-statement mean P(vegan) > 0.5 and judge-positive denotes per-statement mean Likert ≥ 4. To control for lexical leakage between probe and judge signals, all dissociation analyses were conducted on a diet-clean subset that excluded individual records whose corresponding assistant response contained any of: vegan, plant-based, plant based, tofu, tempeh, dairy-free, dairy free, vegetable, almond milk, soy milk, oat milk, cashew, aquafaba, nutritional yeast (case-insensitive whole-word match).

3 Results

3.1 Judge baseline: hint-driven behavioral differences are detectable

Before testing for dissociation between internal representation and surface behavior, we first verified that the LLM judge reliably detects hint-driven differences in the model's outputs. This analysis was conducted on the full dataset (no diet-clean filtering), characterizing what the judge captures when no lexical filter is applied.

Hint influence is robustly detectable across all configurations. Table 1 outlines the LLM judge verdicts on the model's response to the final question. The judge separates the three classes cleanly at every turn configuration. Mean per-statement Likert scores in cfg 0 were **5.24** (hinted, vegan), **1.67** (neutral), and **1.09** (negated, meat-eater); pairwise within-configuration comparisons between classes were significant at $p < 10^{-9}$ throughout, with large effect sizes ($r > 0.72$) for all 15 class-pair comparisons across cfgs 0–4 (Mann–Whitney U, $n = 50$ statements per class; Table 1). The hinted-class mean of 5.24 in cfg 0 indicates the judge somewhat-to-strongly agreed the model's response assumed a vegan user, while the neutral and meat-eater means below 2 indicate the judge disagreed.

Table 1: Pairwise differences in judge Likert scores between conversation classes, by turn configuration. Means are statement-level aggregations ($n = 50$ per class). Effect size r is the rank-biserial correlation; p-values from two-sided Mann–Whitney U.

Cfg	Comparison	Mean A	Mean B	r	p
0	Vegan vs. Neutral	5.24	1.67	0.950	$< 10^{-10}$
	Vegan vs. Meat-eater	5.24	1.09	0.976	$< 10^{-10}$
	Neutral vs. Meat-eater	1.67	1.09	0.789	$< 10^{-10}$
1	Vegan vs. Neutral	4.99	1.69	0.874	$< 10^{-10}$
	Vegan vs. Meat-eater	4.99	1.07	0.990	$< 10^{-10}$
	Neutral vs. Meat-eater	1.69	1.07	0.869	$< 10^{-10}$
2	Vegan vs. Neutral	4.84	1.76	0.876	$< 10^{-10}$
	Vegan vs. Meat-eater	4.84	1.21	0.967	$< 10^{-10}$
	Neutral vs. Meat-eater	1.76	1.21	0.768	$< 10^{-10}$
3	Vegan vs. Neutral	4.45	1.79	0.874	$< 10^{-10}$
	Vegan vs. Meat-eater	4.45	1.15	0.994	$< 10^{-10}$
	Neutral vs. Meat-eater	1.79	1.15	0.822	$< 10^{-10}$
4	Vegan vs. Neutral	4.28	1.87	0.835	$< 10^{-10}$
	Vegan vs. Meat-eater	4.28	1.31	0.936	$< 10^{-10}$
	Neutral vs. Meat-eater	1.87	1.31	0.720	4.18×10^{-10}

Table 2: Decay of hinted-class (vegan) judge Likert across turn configurations: configuration 0 (combined hint and question) versus each subsequent configuration with intervening fillers. Statement-level aggregation ($n = 50$ per config); r is the rank-biserial effect size from two-sided Mann–Whitney U.

Comparison	Mean(cfg 0)	Mean(cfg k)	r	p
cfg 0 vs. cfg 1	5.24	4.99	0.126	0.279
cfg 0 vs. cfg 2	5.24	4.84	0.159	0.172
cfg 0 vs. cfg 3	5.24	4.45	0.268	0.021
cfg 0 vs. cfg 4	5.24	4.28	0.308	0.008

Hinted-class influence decays gradually with turn separation. While class separation remains significant at every configuration, the hinted-class Likert decays as the number of intervening filler turns increases, as illustrated in Table 2). A Spearman rank correlation between configuration index and per-statement Likert score for the hinted class confirmed a significant negative trend ($\rho = -0.19$, $p = 0.002$, $n = 250$ statement-config pairs). The hinted-class mean dropped from 5.24 in cfg 0 to 4.28 in cfg 4, while neutral and meat-eater means stayed nearly flat ($1.67 \rightarrow 1.87$ and $1.09 \rightarrow 1.31$, respectively; Figure 3, Table 2). Pairwise comparisons between consecutive configurations were not individually significant, but cfg 0 differed significantly from cfg 3 ($p = 0.021$) and cfg 4 ($p = 0.008$), indicating gradual rather than abrupt decay. The signal never approached the neutral or meat-eater baseline, indicating that the model retained hint influence in its behavioral output through several turns of off-topic context.

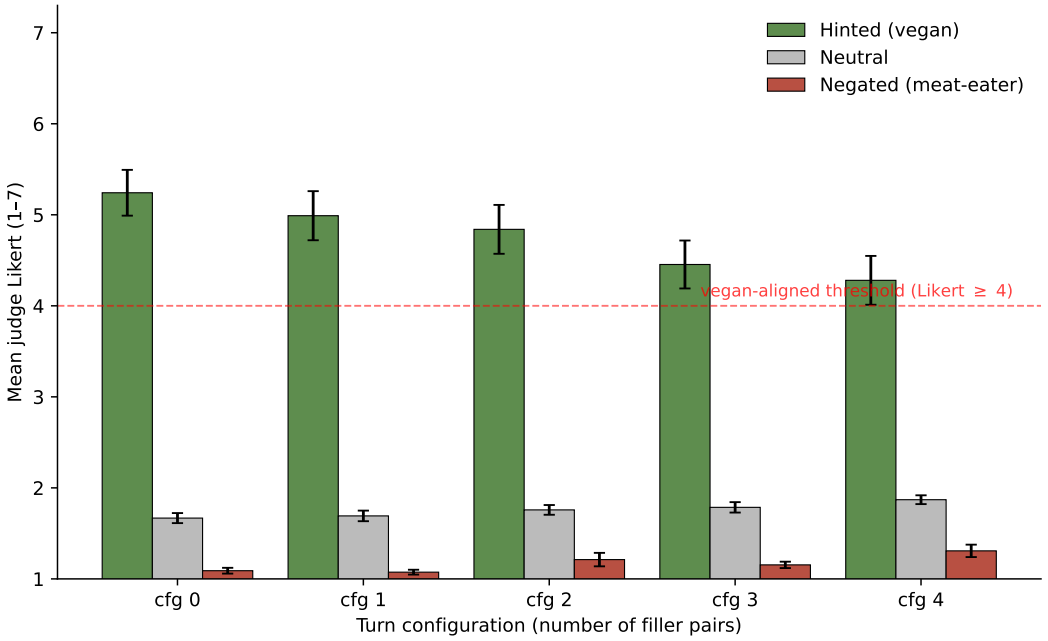

Figure 3: Mean judge Likert scores per conversation class across turn configurations (cfgs 0–4). Bars show per-statement means ($n = 50$ statements per class per configuration); error bars are SEM. The dashed red line marks the vegan-aligned threshold (Likert ≥ 4). Hinted-class scores decay gradually with intervening filler turns but remain well above neutral and meat-eater throughout.

Together, these results establish two baseline facts for the dissociation analyses that follow: (i) the judge reliably detects behavioral expression of hint influence when it is present, and (ii) such expression itself persists across multiple intervening turns.

The subsequent analyses (§3.2–§3.4) then test what the model's internal representation continues to encode in the *absence* of overt behavioral expression—specifically, when lexical filtering removes responses containing explicit vegan vocabulary.

3.2 Probe-vs-judge dissociation

The analyses in this section are conducted on a diet-clean subset of held-out test conversations: individual records whose corresponding assistant response contains any of a list of explicit vegan-coded tokens (see §2.5) are excluded prior to aggregation. This removes the most direct lexical pathway through which the judge could detect a vegan-aligned response, allowing us to ask: when the model's surface output does not contain overt vegan vocabulary, does its internal representation still track the user's diet?

Probe representation persists across all conversation turns. The probe assigns $P(\text{vegan})$ significantly above the chance level of 1/3 at every conversation role on the diet-clean subset (one-sided Wilcoxon signed-rank test, Benjamini–Hochberg FDR correction across roles; Table 3). Mean per-statement $P(\text{vegan})$ values were **0.93** at the *hint reply* (M3-end; $n = 26$, $p_{\text{BH}} = 2.2 \times 10^{-8}$), **0.43** at *filler turns* (M5-end, M7-end, M9-end aggregated; $n = 50$, $p_{\text{BH}} = 6.1 \times 10^{-6}$), and **0.85** at the *answer* (final assistant position; $n = 40$, $p_{\text{BH}} = 2.7 \times 10^{-12}$). The trajectory across roles forms a distinctive drift-and-recover arc: the probe assigns near-saturating probability to the vegan class immediately after the user's hint, drops to a modest but above-chance value at filler turns, and

Table 3: Per-role probe and judge statistics on the diet-clean subset of held-out conversations. Values are per-statement means aggregated across configurations and questions. Wilcoxon p-values test the one-sided alternative that the probe assigns $P(\text{vegan}) > 1/3$, Benjamini–Hochberg-corrected across the three roles. The judge Likert scale ranges from 1 (strongly disagree the response assumes a vegan user) to 7 (strongly agree); values ≥ 4 are classified as vegan-aligned.

Role	n	Mean $P(\text{vegan})$	Wilcoxon p_{BH}	Mean Likert
Hint reply	26	0.933	2.24×10^{-8}	2.32
Filler	50	0.429	6.14×10^{-6}	1.19
Answer	40	0.850	$< 10^{-10}$	3.77

Table 4: Contingency between probe-positivity (per-statement mean $P(\text{vegan}) > 0.5$) and judge-positivity (per-statement mean Likert ≥ 4) across conversation roles, on the diet-clean subset. p_{BH} values are from two-sided binomial sign tests on the discordant cells (P+/J− vs. P−/J+), Benjamini–Hochberg-corrected across the three roles.

Role	n	P+/J−	P−/J+	Both	Neither	p_{BH}
Hint reply	26	21	0	5	0	2.86×10^{-6}
Filler	50	10	0	0	40	0.002
Answer	40	20	0	20	0	2.86×10^{-6}
Total	**116**	**51**	**0**	**25**	**40**	—

recovers to near-hint levels at the final answer. The signal persists even at filler turns where the model's response addresses an off-topic question and contains no explicit vegan-coded vocabulary.

The probe-judge dissociation is strictly asymmetric. In the same role-mapped diet-clean subset, the LLM judge assigned per-statement mean Likert scores well below the vegan-aligned threshold of 4: **2.32** at the hint reply, **1.19** at filler turns, and **3.77** at the answer (Table 3). These values stand in sharp contrast to those obtained on the full dataset (§3.1), where hinted-class Likert ranged from 5.24 to 4.28 across configurations—reflecting how heavily the judge depends on the lexical content removed by the diet-clean filter.

To formalize the comparison between probe and judge, each statement at each role was classified as *probe-positive* (per-statement mean $P(\text{vegan}) > 0.5$) and *judge-positive* (per-statement mean Likert ≥ 4); discordant cells were compared via a two-sided binomial sign test (Table 4). The dissociation is **strictly asymmetric**: across 116 statement-role pairs, **51** are probe-positive but judge-negative, while **zero** are the reverse. The asymmetry holds at every individual role and survives FDR correction ($p_{\text{BH}} = 2.9 \times 10^{-6}$ at the hint reply, $p_{\text{BH}} = 2.0 \times 10^{-3}$ at filler turns, $p_{\text{BH}} = 2.9 \times 10^{-6}$ at the answer). Because all analyses are conducted on the diet-clean subset, the asymmetry is unlikely to be explained by lexical leakage: the probe detects the user's diet class in responses that contain none of the most explicit vegan-coded tokens.

Figure 4 visualizes the per-statement distribution underlying this aggregate pattern. Each point represents one (statement, role) pair, with mean $P(\text{vegan})$ on the horizontal axis and mean Likert on the vertical axis. The upper-left quadrant (probe-negative, judge-positive) is empty, while the lower-right quadrant (probe-positive, judge-negative) is densely populated—reflecting that the asymmetry is not an artifact of averaging over heterogeneous distributions, but a property of nearly every individual statement.

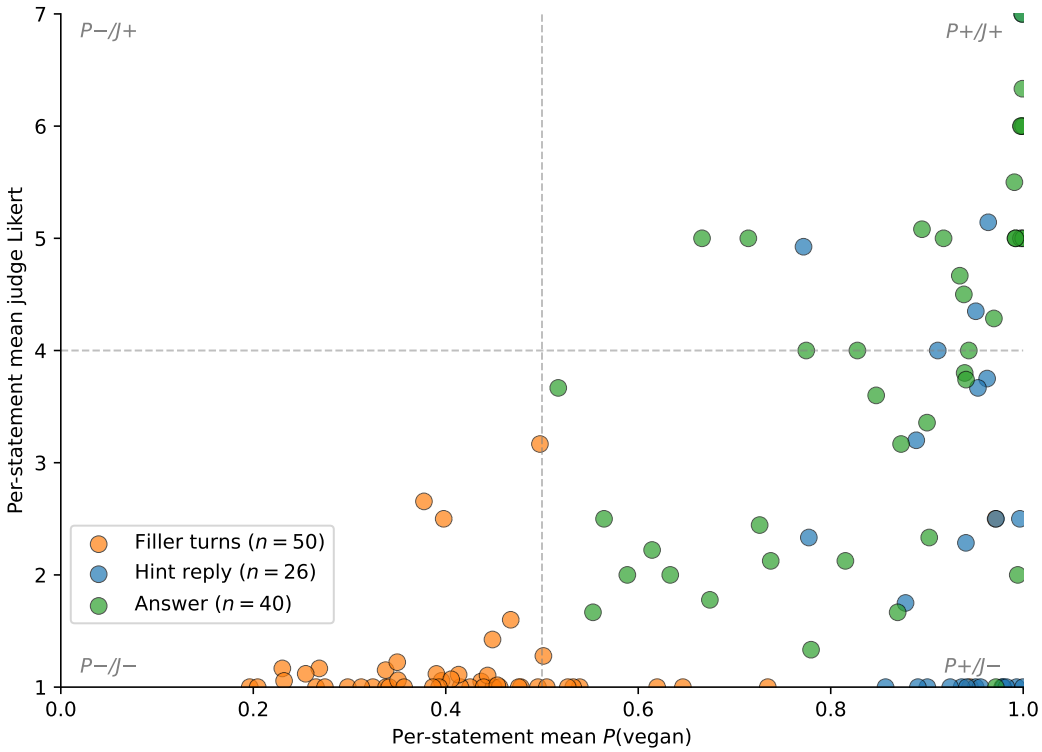

Figure 4: Per-statement probe–judge dissociation on the diet-clean subset. Each point represents one statement-role pair, plotting per-statement mean $P(\text{vegan})$ from the linear probe (x-axis) against per-statement mean judge Likert (y-axis), each averaged across turn-length configurations. Points are colored by conversation role: hint reply (M3), filler turns (M5–M9), and answer (M11). The dashed vertical line marks the probe-positive threshold ($P(\text{vegan}) = 0.5$); the dashed horizontal line marks the vegan-aligned Likert threshold of 4. Together they define the four quadrants used in the contingency analysis (Table 4).

3.3 Per-configuration robustness

To assess whether the dissociation pattern is concentrated in any particular turn configuration, we computed per-(configuration, role) means of probe $P(\text{vegan})$ and judge Likert on the diet-clean held-out subset (Figure 5, Table 5). Four observations stand out.

Hint-reply representation is stable across configurations. At the hint reply, the probe assigns near-saturating $P(\text{vegan})$ (0.91–0.95) in every configuration where the hint reply exists. The user-attribute representation is established immediately after the hint and shows no detectable trend with the size of the subsequent context.

Answer-position representation decays gradually with intervening fillers. At the final answer position, the probe assigns high $P(\text{vegan})$ throughout but with a clear monotonic decay as the number of intervening filler turns grows—from 0.94 in cfg 0 (no fillers) down to 0.71 in cfg 4. The model recovers the user attribute at the answer in every configuration, but the recovery is less complete when more turns have been spent on intervening off-topic content.

Table 5: Per-(configuration, role) statistics on the diet-clean held-out subset. Mean P(vegan) and mean judge Likert are per-statement aggregates, with n denoting the number of unique statements contributing to each cell. The Likert threshold for vegan-aligned classification is 4.

Cfg	Role	n	Mean P(vegan)	Mean Likert
0	Answer	23	0.935	4.16
1	Hint reply	24	0.920	2.34
	Answer	26	0.912	3.77
2	Hint reply	24	0.949	2.17
	Filler	41	0.422	1.21
	Answer	23	0.851	3.02
3	Hint reply	23	0.929	2.38
	Filler	48	0.446	1.16
	Answer	28	0.800	3.55
4	Hint reply	20	0.911	2.19
	Filler	48	0.376	1.18
	Answer	28	0.709	2.73

Filler representation is consistently modest. At filler positions, the probe assigns above-chance but lower P(vegan) in every configuration where fillers exist (0.38–0.45). The slight attenuation at cfg 4 reflects the additional weight of late-filler positions, which we examine at the filler-position level in §3.4.

Judge stays below threshold at every multi-turn configuration. On the diet-clean subset, mean judge Likert remains below the vegan-aligned threshold of 4 at every (configuration, role) cell except one: at the cfg 0 answer—where the hint and final question share the same user message—the judge mean reaches 4.16, marginally above threshold. At every multi-turn answer cell (cfgs 1 through 4), the judge stays below threshold even though the probe assigns 0.71–0.91 (see Table 5). The dissociation between internal representation and surface behavior is thus sharpest in conversations with intervening filler turns: when hint and question are co-present (cfg 0), the model produces sufficiently vegan-flavored output for the judge to detect on average; once any filler intervenes, the model's internal representation persists but its behavioral expression no longer crosses the judge's detection threshold.

3.4 Message-position trajectory of internal representation

The per-role analyses in §3.2 and §3.3 reported end-of-message probe values aggregated into three role categories. Here we refine that view along two axes: (i) we report probe P(vegan) at each assistant message position individually rather than aggregated by role, and (ii) we distinguish the start of each message from its end (rather than reporting end-of-message only), on the diet-clean held-out subset (Figure 6).

End-position values reproduce the drift-and-recover arc. Across configurations, probe P(vegan) at end-of-message positions starts high at the hint reply, drifts down across consecu-

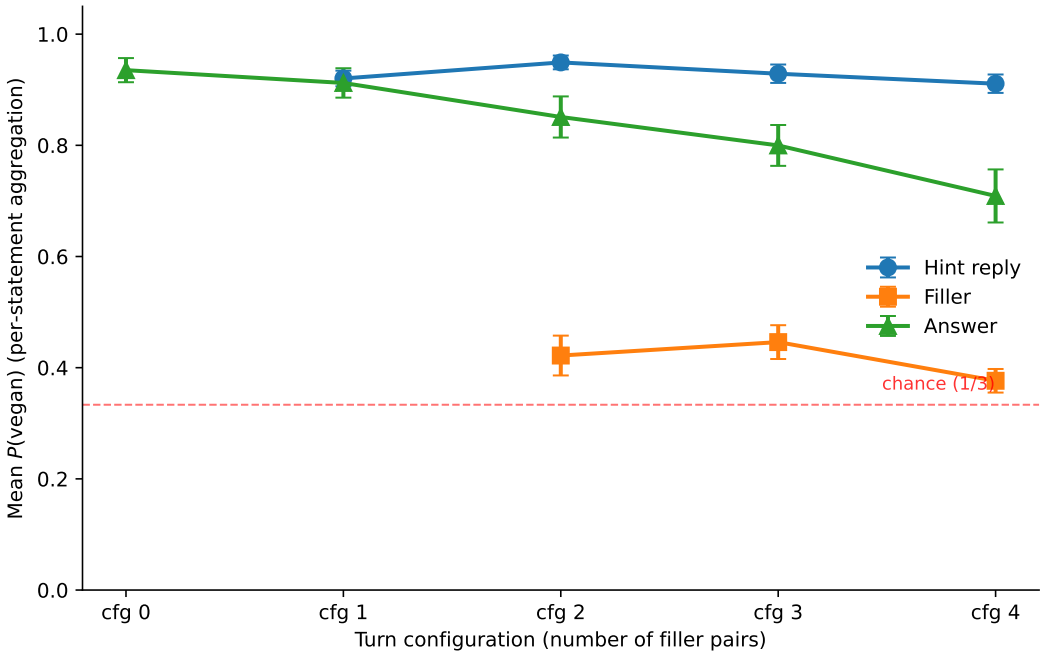

Figure 5: Probe P(vegan) at each conversation role, plotted across turn-length configurations on the diet-clean held-out subset. Each line corresponds to one conversation role—hint reply, filler turns, or answer—and is plotted only at configurations where that role is defined: hint reply at configurations 1–4, filler positions at configurations 2–4, and answer at all configurations 0–4. The hint-reply probe is high and roughly constant across configurations, the filler-position signal is modest and stable, and the answer-position probe declines monotonically with the number of intervening filler turns. Per-cell means and sample sizes are reported in Table 5.

tive filler positions, and partially recovers at the final answer—reproducing the role-level pattern reported in §3.3 at finer position-level resolution.

Start positions reveal a within-message rise at every assistant turn. At the start of each filler reply, the probe assigns essentially zero P(vegan), well below the chance level of 1/3. The end of each message reads substantially higher than its start; this within-message rise is present at every assistant position—hint reply, fillers, and answer alike.

The recovery at the answer is comparable to within-message rises elsewhere. The within-answer rise (from the start to the end of the answer message) is similar in magnitude to the within-message rises observed at filler positions. The probe-decodable signal at the answer might therefore be the same kind of within-message accumulation observed at every other assistant message. Mechanistic implications are discussed in §4.

We note one methodological caveat: of the start positions plotted, only M3-start and the configuration-specific answer-start were training positions for the probe. Filler-start positions are extrapolated by the probe; their near-zero readings should be interpreted cautiously. However, within-message rise at M3 (where M3-start was in probe training, M3-end was not) and at start and end of assistant's final answer in each configuration is in itself substantial, indicating that the start-end gradient might not be purely an extrapolation artifact.

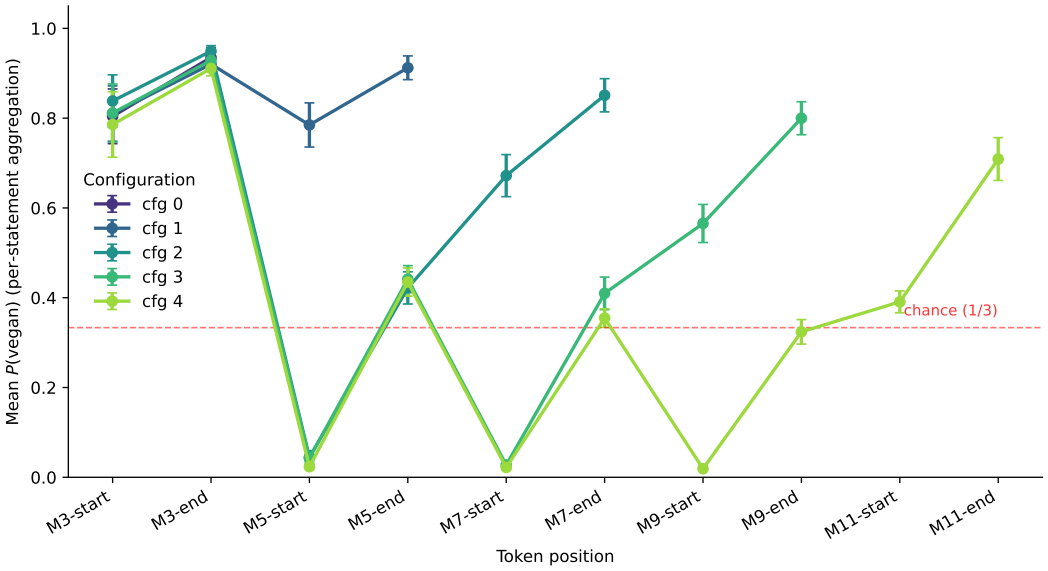

Figure 6: Probe $P(\text{vegan})$ trajectory across assistant message positions in all configurations.

These patterns motivate the causal experiment in §3.5: if the answer's representation depends on a mechanism referring back to the original hint position rather than on continuous propagation through intervening residuals, patching the hint should disrupt the answer regardless of where within the answer the recovery actually emerges.

3.5 Causal patching: hint position M2 is necessary and sufficient

The probe-based analyses in §3.2–§3.4 established correlational patterns consistent with a retrieval/lookup mechanism. To test this causally, and to avoid relying on the probe at both training and intervention time, we measured the behavioral effect of activation patching directly: we patched specific positions in the residual stream, allowed the model to generate its full answer with those patches in place, and scored each answer with the same LLM judge used in §3.1.

Experimental design. We constructed 10 minimal pairs of Configuration 4 conversations. Each pair shares the same final question and three filler turns but differs in the M2 hint: one is the original vegan conversation, the other a synthetic meat-eater counterpart. For each pair, we generated the model's answer under four conditions: **Baseline** (no patches); **M2-patched** (vegan M2 residual replaced with the synthetic meat-eater M2 residual); **Filler-patched** (assistant filler-reply residuals at M5, M7, and M9 replaced); and **M2+M3-patched** (a stronger ablation that patches M2 together with M3, the assistant's reply to the hint, which otherwise still carries vegan-coded content). Patches were applied at layers $\{0, 8, 16, \ldots, 64\}$ via `nnsight` on NDIF, with all source-activation capture and patched generation contained in a single remote session per pair. Each generated answer was scored on the same 1–7 vegan-alignment Likert scale used in §3.1.

Hint-position patching shifts behavior; filler patching does not. Across the 10 pairs, the mean baseline Likert was 4.20, just above the vegan-aligned threshold of 4. Patching M2 reduced the mean to 1.90 ($\Delta = -2.30$; one-sided Wilcoxon paired signed-rank, $p = 0.023$). Patching the three filler-reply positions reduced the mean only to 3.70 ($\Delta = -0.50$; $p = 0.19$, not significant). The direct paired hint-vs-filler contrast was in the predicted direction, but not significant ($\Delta_{\text{M2}-\text{filler}} =$

-1.80, $p = 0.06$). This result suggests that the model's behavioral output is causally sensitive to the residuals at the hint position but not to the residuals at intermediate filler positions (Table 6).

M2 alone is sufficient: M3 is not an independent source. A concern with the M2-only condition is that M3—the assistant's reply to the hint—retains vegan-coded content from the original conversation even when M2 is patched, potentially preserving the behavioral pattern. We tested this by patching M2 and M3 together. The M2+M3-patched mean Likert was 2.10—statistically indistinguishable from M2-only patching (1.90; both $p = 0.023$ against baseline; see Table 6). Patching M3 alongside M2 adds no behavioral effect beyond patching M2 alone. The user's hint at M2 is the necessary and sufficient causal source within the positions we tested; the assistant's hint-reply representation at M3 does not provide an independent route for the user attribute to influence the answer.

Floor effect: the aggregate result is conservative. The aggregate effect masks substantial pair-level heterogeneity (Figure 7, individual points). The baseline Likert distribution is itself bimodal: 5 of the 10 pairs have baseline Likert $\geq$ 4—the vegan-aligned threshold used throughout this paper (§3.1)—while the other 5 have baselines that did not commit to vegan content even without any intervention. We believe that the patching intervention has little room to act on these floor-affected pairs and therefore contributes mainly noise to the mean estimate.

Restricting to the $N = 5$ pairs whose unpatched baseline crossed the vegan-aligned threshold (Table 7) sharpens the picture substantially. The baseline mean rises to 6.20, indicating that on these pairs the model strongly inferred a vegan user from the implicit hint. M2 patching reduces the mean to 1.60 ($\Delta = -4.60$, $p = 0.031$; median 1)—essentially a complete switch from one diet class to the other. Filler patching again produces only a small, non-significant change to 5.20 ($\Delta = -1.00$, $p = 0.125$). The direct hint-vs-filler paired contrast, which was only marginally significant on the full $N = 10$ ($p = 0.06$), becomes clearly significant on the informative subset ($\Delta = -3.60$, $p = 0.031$). The M2+M3 condition produces an essentially identical effect on this subset ($\Delta_{\text{M2+M3}-\text{filler}} = -3.20$, $p = 0.031$), confirming that M3 contributes no additional causal effect even on high-signal trials. The aggregate $N = 10$ effect should therefore be read as a conservative estimate: on trials where the implicit hint genuinely shaped the model's baseline behavior, patching M2 nearly completely flips the answer.

Taken together with the probe-based trajectory in §3.4, these results support a retrieval account of how the implicit user attribute influences model behavior: the attribute is stored at the hint position and is consulted there when the model generates the answer. Intermediate filler representations are not on the causal path between the user's hint and the model's behavioral output.

4 Discussion

The three results sections above examined how an implicit user attribute—a dietary preference established in the first user turn— persists, decays, and reappears across the residual stream of a 70B-parameter chat model. Read together, they support a coherent mechanistic picture: the attribute is established at the original hint position, persists there across off-topic intermediate turns, and is causally consulted by the answer position at generation time. Intermediate filler representations carry diluted versions of the same probe-decodable signal but are not on the causal path from hint to behavior.

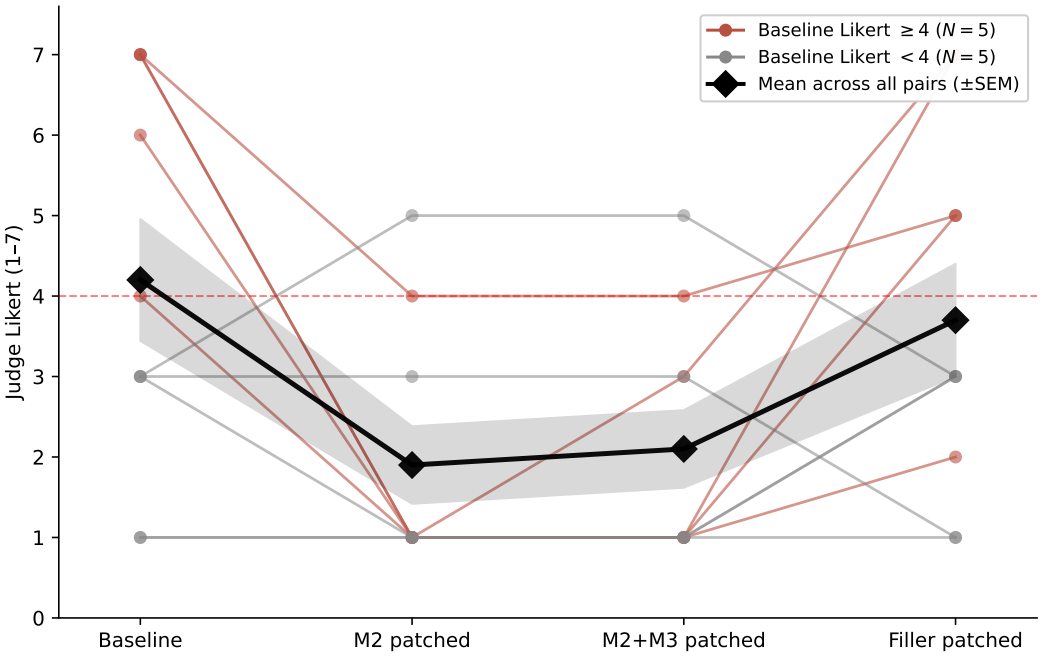

Figure 7: Behavioral effect of activation patching on $N = 10$ minimal pairs. Colored lines show individual pair trajectories across patching conditions: red for pairs whose unpatched baseline crossed the vegan-aligned threshold (Likert $\geq$ 4; $N = 5$), gray for pairs whose baseline did not ($N = 5$). The black diamond line shows the mean trajectory (±SEM band). The dashed red line marks the vegan-aligned threshold. On the high-baseline (red) pairs, M2 patching produces a sharp drop in Likert that filler patching does not; on the floor-affected (gray) pairs, no condition produces a strong shift because there was little vegan signal to remove. The aggregate effect in the mean line is therefore conservative relative to the per-pair effect on informative trials.

Three findings, one mechanism. The probe–judge dissociation (§3.2) showed that the residual stream encodes the user's attribute well above chance across all multi-turn configurations and across all turn-role positions, including filler-reply positions where the model's behavioral output does not reflect awareness of the diet. The message-position trajectory (§3.4) sharpened this picture: within a conversation, probe accuracy on the hint position itself stays high and roughly constant across configurations (§3.3), but at positions further from the hint the probe weakens and the model's answers drift toward the no-diet baseline. Critically, within the answer span itself the probe recovers *gradually*, not abruptly at the message boundary, suggesting that the user attribute might be assembled in the residual stream during generation rather than carried forward as a stable, ready-to-use representation.

The behavioral patching results (§3.5) then localized the causal source of that assembly. Patching the M2 residual—the user's hint position—was sufficient to flip the model's answer from one diet class to the other on pairs where the baseline carried strong vegan signal ($\Delta = -4.60$, $p = 0.031$). Patching the assistant filler-reply positions, which the probe analyses identified as carrying intermediate-strength signal, produced no significant behavioral change ($p = 0.125$). And patching M3 (the assistant's reply to the hint) alongside M2 yielded essentially the same effect as patching M2 alone, ruling out the assistant's hint-acknowledging reply as an additional causal source. The simplest account consistent with all three observations is a lookup: when the model generates the answer, it queries back to the hint position and retrieves the user attribute from there.

Table 6: Judge Likert scores across patching conditions on $N = 10$ minimal pairs. Higher values indicate more vegan-aligned model responses (Likert $\geq$ 4 classified as vegan-aligned). Δ vs. baseline is the mean per-pair difference; p-values from one-sided Wilcoxon paired signed-rank tests (alternative: patched < baseline, or hint < filler).

Condition	Mean	SEM	Δ vs. baseline	p_{Wilcoxon}
Baseline	4.20	0.76	—	—
M2 patched	1.90	0.48	−2.30	0.023
M2+M3 patched	2.10	0.48	−2.10	0.023
Filler patched	3.70	0.70	−0.50	0.188
M2 patched vs. Filler patched			−1.80	0.061
M2+M3 patched vs. Filler patched			−1.60	0.064

Table 7: Stratified behavioral patching effect on the subset of pairs ($N = 5$) whose unpatched baseline produced a vegan-aligned response (Likert $\geq$ 4, the same threshold used in §3.1). On these informative trials, M2 patching nearly completely flips the answer and the hint-vs-filler contrast becomes clearly significant.

Condition	Mean	Median	Δ vs. baseline	p_{Wilcoxon}
Baseline	6.20	7.0	—	—
M2 patched	1.60	1.0	−4.60	0.031
M2+M3 patched	2.00	1.0	−4.20	0.031
Filler patched	5.20	5.0	−1.00	0.125
M2 patched vs. Filler patched			−3.60	0.031
M2+M3 patched vs. Filler patched			−3.20	0.031

The intermediate filler positions carry a correlate of that attribute—enough for a linear probe to read it— but the model does not use them on the causal path to the answer.

Probing and patching results dissociate. The contrast between probe decoding and behavioral output is not a finding in tension with the patching result; it is consistent with established cautions about interpreting probe accuracy as evidence of mechanism [1, 4]. A linear probe can recover a feature from a residual stream that the model does not actively use in any given downstream computation. Our results provide a clean illustration in a multi-turn conversational setting: the implicit user attribute is decodable at the filler-reply positions with probability up to 0.45, but patching those same positions does not move the model's behavior. The probe-decodable signal at the intermediate positions is, in our setting, mechanistically inert; the behaviorally relevant content lives at M2 and is accessed there. The dissociation is informative precisely because the patching contrast is sharp: probe-decodable presence and causal relevance can come apart on the same conversation, and the question "where in the residual stream is the model's user model stored?" has a more specific answer than the probe alone would suggest.

Relation to prior mechanistic work. On-demand retrieval of in-context information is well-established in the mechanistic interpretability literature: induction heads and copying mechanisms recover tokens from earlier positions [12, 17]; factual recall has been localized to specific layers and

components [10, 6]; and activation patching has been used to identify causal positions across a wide range of tasks [16, 8]. What we add to these findings is the application of these tools to a setting that is distinctive in three respects. First, the retrieved content is an *implicit user attribute*—inferred from a passing remark in a previous turn—rather than a factual association established during pretraining or an explicit instruction. Second, the retrieval happens across a multi-turn, off-topic conversational gap, in the face of intervening content that the probe analyses suggest does dilute the surface signal. Third, the relevant attribute belongs to a broader class of socially-loaded inferences—dietary preference here, but with obvious extensions to political views, age, health, or other identity attributes—where personalization can either be helpful or constitute an unwarranted assumption. Within that frame, the finding that the user attribute is *stored at the hint* rather than continuously propagated speaks to a specific question: the model is not maintaining a running user model in its activations; it is retrieving from the hint position on demand. The two functional accounts have different implications for downstream alignment work.

Implications for personalization. The retrieval-on-demand picture has a practical consequence. If the model does not maintain a continuously updated user model in its activations, then user updates—explicit corrections, evolving preferences, or contextual clarifications that arrive after the original hint—would need to either overwrite the M2 residual itself or install a competing representation that the answer-time lookup mechanism preferentially attends to. Two functional accounts can be distinguished: corrections may "overwrite" the hint by editing the residual at M2, or they may install a separate retrieval target that competes with M2 at the answer position. The two have different implications for whether stale attributes from much earlier in a conversation can resurface, and for how robustly personalization behavior tracks user correction. Both accounts are consistent with the present results; distinguishing them requires extending the patching paradigm to corrections.

Caveats on interpretation. Several aspects of the present setup are worth keeping in mind when reading these results. The patching experiments operate at a coarse layer grain (9 patched layers out of 80, every eighth). Moreover, the near-zero probe outputs at filler-*start* positions are not strong evidence about the model's internal state at those positions, because the probe was not calibrated for them. The behavioral patching is on $N = 10$ minimal pairs, and the aggregate effect is conservative because half the pairs do not show strong vegan signal even at baseline (§3.5, floor effect). The studied attribute is a single dietary preference with a synthetic counterpart, and the studied model is a single 70B instruct-tuned chat model. We discuss these and additional limitations, along with the experiments they motivate, in §5.

User attribute corrections. The most natural next step, given the present findings, is a correction-extension experiment. Take a conversation in which the model has clearly inferred a dietary preference (e.g. with a high baseline Likert score of an implied diet), append a user turn that corrects the original attribute, and ask whether the answer-time retrieval at the model's response position now attends to the original hint or to the correction. The patching machinery developed for this paper applies directly: one can patch the correction-turn residual into a non-corrected conversation and measure the behavioral shift, or conversely patch the original hint residual into a corrected conversation and ask whether the original attribute reasserts itself. Different correction *types*—explicit denials, social reframings, value-laden objections—may compete with the hint to different degrees at retrieval time, providing a mechanistically principled handle on phenomena like persistent stale-attribute inference that have been observed behaviorally.

5 Limitations and Future Work

The findings in this paper rest on a deliberately constrained design—one model, one attribute, controlled synthetic minimal pairs—which permits a clean mechanistic claim but limits the scope of that claim. We outline the most important limitations and, where applicable, the experiments that would address them.

Single attribute and single model. We study one implicit user attribute—dietary preference, with vegan as the marked class and a constructed meat-eater counterpart—in one model (Llama-3.3-70B-Instruct). Whether the same retrieval-from-hint pattern holds for other implicit attribute classes (political views, demographic identifiers, health conditions, profession-based inferences) and for other instruction-tuned models is an open question. The mechanism we propose makes a falsifiable prediction here: any user attribute that the model can infer from a passing remark and later reflect in its answers should be readable from the hint position and causally severable by patching it. The present experimental design transfers directly to a different attribute axis and to other open-weight instruction-tuned models.

Statistical power on the patching contrast. The behavioral patching experiment uses $N = 10$ minimal pairs, of which only 5 carried baseline Likert above the vegan-aligned threshold and were therefore fully informative under patching (§3.5). The aggregate hint-vs-filler contrast on the full $N = 10$ is consequently marginal ($p = 0.06$); the stratified subset of high-baseline pairs is clearly significant ($p = 0.031$) but small. A larger minimal-pair set, ideally pre-filtered to retain only pairs that elicit a strongly vegan-aligned baseline answer, would tighten the aggregate result and allow finer comparisons—for example between attribute strengths, between hint phrasings, or between turn-position structures.

Functional, not circuit-level, localization. We localize the causal source of the user attribute to a hint message (M2) and a partial set of patched layers ($\{0, 8, 16, \ldots, 64\}$). We do not identify the specific attention heads or MLP components responsible for the retrieval, and we do not directly observe that the answer position attends to M2 at generation time. One mechanism that could account for our results is "attention-based retrieval". However, this is a functional inference from the location of the causal source—consistent with established mechanisms for in-context retrieval [12, 6] but not directly demonstrated in our experiments. A natural extension is per-layer and per-head patching combined with attention-pattern analysis at the answer position, which would convert the functional claim into a circuit-level one.

Patching coverage. Our patching protocol covers M2, M3, and the assistant filler-reply positions (M5, M7, M9), at every eighth layer, on the longest filler configuration (Configuration 4). We do not patch the user's final-question position, nor do we run patching at shorter filler configurations. The completeness of the localization to M2 depends on these choices: it is possible that finer-grained patching reveals additional contributing positions or layers, or that shorter-filler configurations recruit a different mechanism in which less retrieval is needed. User's final question position is of particular interest, because it is the position at which the model would plausibly compose the retrieved user attribute with the current question.

Probe interpretation outside trained positions. The linear probe is trained at four token positions per conversation—M2-end (end of user hint), M3-start (start of assistant acknowledgment),

the end of the user's final question, and the start of the model's final answer—with deduplication for Configuration 0 where the latter two coincide with the first two. Probe outputs at positions outside this training set (in particular the filler-start positions reported in the trajectory analysis, §3.4) are extrapolation and should not be over-interpreted as evidence about the model's internal state at those positions. The aspects of the trajectory analysis that bear on our main claims (M2-end stability across configurations, the gradual within-answer recovery from final-response-start to final-response-end) involve only trained positions; the absolute height of the filler-start trajectory should not be read as a substantive finding.

Synthetic minimal-pair construction. The meat-eater counterpart in each patching pair is constructed synthetically to share everything with the vegan original except the M2 hint. This is a deliberate methodological choice—it gives the cleanest experimental control over which residual differs between paired conversations—but the synthetic conversations may not match the distribution of how meat-eating users would actually phrase preferences in real interactions.

Single judge. All Likert ratings are produced by a single LLM judge (Qwen3-235B-A22B; §2.3). We did not validate the judge against multiple LLM judges or against direct human ratings. Because all of our claims rest on comparisons *between* conditions scored by the same judge with the same prompt, judge-specific biases largely cancel; nevertheless, the absolute Likert values are not independently calibrated, and a small inter-judge or human-rater agreement study would be a useful robustness check.

Future work: attribute corrections. The most immediate extension, foreshadowed in §4, is a correction experiment. Given that the user attribute appears to be stored at hint message and retrieved on demand by the answer position, what happens when the user explicitly contradicts the original hint in a later turn? Does the answer-time retrieval re-target to the correction, or does it remain anchored to the original hint message? The patching machinery developed here applies directly: one can patch the correction-turn residual into a non-corrected conversation and measure the behavioral shift, or conversely patch the original hint into a corrected conversation and ask whether the original attribute reasserts itself. Beyond establishing whether corrections overwrite or compete with the original hint, the design naturally extends to different correction *types*—explicit denials, social reframings, value-laden objections—which may have distinguishable mechanistic signatures.

Future work: cross-attribute and cross-model generalization. The most direct generalization test is to rerun the full pipeline—judge baseline, probe–judge dissociation, position trajectory, and behavioral patching—on a different implicit attribute and on a different instruction-tuned model. If the retrieval-from-hint pattern holds across these axes, the mechanism we identify is a general feature of how instruction-tuned chat models maintain user context across multi-turn conversation; if it breaks, the patterns of breakage should be informative about which attribute types or model classes recruit different mechanisms. A smaller open-weight model would also permit finer layer-grain patching and per-head circuit analysis at lower compute cost.

Acknowledgments

We thank the SPAR program for supporting this work, and members of our project group Afitab İyigün, Hengxu Li, Jacopo Zacchigna, Mika Okamoto, Vinit Patel and Natalie Shapira for valuable

discussions and feedback throughout the project. We are grateful to the NDIF team for providing the `nnsight` library and the remote compute infrastructure that made activation analyses on Llama-3.3-70B tractable.

References

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

[2] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational ai, 2024.

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

[4] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

[5] Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C Wallace, and David Bau. NNsight and NDIF: Democratizing access to foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

[7] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[8] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[9] 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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] 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.

[11] Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20367–20400, Suzhou, China, November 2025. Association for Computational Linguistics.

[12] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[13] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

[14] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025.

[15] 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. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[16] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[17] Kevin 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 *International Conference on Learning Representations (ICLR)*, 2023.