

# Where Does Conversational Repair Happen? Meaning Revision Inside Language Models

Vedanth Nanesha

National Institute of Technology Karnataka, Surathkal  
vedanthnanesha04@gmail.com

## Abstract

Conversational repair, the process by which interlocutors detect and correct misunderstandings, is central to human language use but rarely studied in the internal representations of large language models (LLMs). We ask: when a model commits to an incorrect interpretation of an ambiguous utterance and the user corrects it, where and how does the corrected interpretation emerge across layers? We construct a controlled diagnostic dataset of 291 three-turn dialogues, pairing each repair run (an ambiguous user turn, a model commitment to a default reading, and a user correction) with a surface-matched control that performs no repair, isolating the correction’s effect. Across six instruction-tuned models from three families, we find that (i) repair revises meaning in a narrow band of mid-to-late layers, where activation patching causally shifts output to the corrected reading; (ii) when repair fails behaviourally, the corrected reading is causally absent rather than weak; (iii) attention, not the feed-forward sublayer, carries the revision; and (iv) revision is stronger when the misreading is attributed to the assistant rather than the user, indicating more than generic in-context override. These results characterize conversational repair as a localisable, model-dependent computation and provide a reusable diagnostic for meaning revision in interaction.

## 1 Introduction

Language understanding in interaction is not a single-shot mapping from an utterance to a meaning. It is a dynamic process in which interpretations are proposed, found wanting, and revised. In conversation analysis this process is known as repair (Schegloff et al., 1977). A speaker produces an ambiguous turn, a recipient displays an understanding of it, and, when that understanding is incorrect, the original speaker initiates a correction. This last configuration, in which the recipient of the

trouble source corrects an understanding that has already been displayed, is termed third-position repair (Schegloff, 1992). Repair is one of the principal mechanisms by which interlocutors maintain common ground and negotiate meaning rather than merely transmit it (Clark, 1996), making it central to interactive language understanding.

Large language models are now widely deployed as conversational agents, and they are routinely subjected to exactly this dynamic: a user asks something, the model answers under one reading, and the user pushes back. Existing evaluation is largely behavioural, leaving open what happens inside the model when an interpretation is revised: where in the network does the old interpretation give way to the new one, and is that revision the same process that produces the right answer for an unambiguous question?

Taking a mechanistic interpretability approach, we build a controlled diagnostic dataset of three-turn dialogues, sourcing the ambiguity from existing, externally validated resources and authoring the repair; every item is structured as a minimal pair whose repair run corrects the model’s committed reading and whose surface-matched control does not, so their contrast isolates the correction (Figure 1).

Using this dataset we run behavioural and mechanistic analyses on six instruction-tuned models from three families, organised around six research questions. **RQ1** (localisation): across layers, where does the corrected interpretation emerge once the model has read the repair turn? **RQ2** (causality): is that location causally responsible for the output, or only correlated with it? **RQ3** (failure): when repair fails behaviourally, is the internal revision merely weak, or is it absent? **RQ4** (mechanism): which components carry the revision, and how is it organised across attention heads and model families? **RQ5** (speaker attribution): does the model revise a misreading attributed to the assistant role

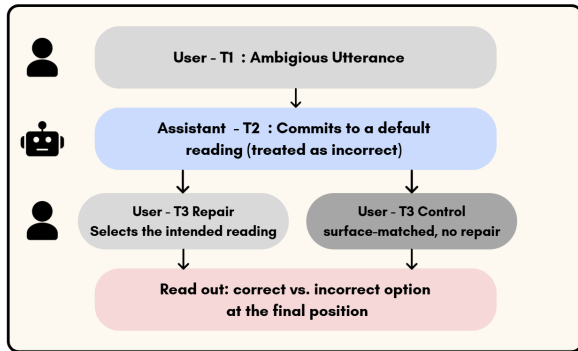


Figure 1: Two-pass minimal-pair design, shown as a chat. The two runs share the ambiguous user turn (T1) and the model’s committed reading (T2) and differ only in the third user turn: a repair that selects the intended reading versus a surface-matched control that performs no repair. We read the correct-minus-incorrect option signal at the final position, with the logit lens across layers and with activation patching of the late-layer residual from the repair run into the control run.

differently from the same misreading attributed to the user? **RQ6** (selectivity): does the model revise only when the correction is warranted, or does it defer to the user regardless of the truth?

Our contributions are as follows. (1) We introduce a controlled, minimal-pair diagnostic dataset for conversational repair, covering five ambiguity types and four repair formats, together with a normalised revision margin for reading out meaning revision. (2) Using this diagnostic we give a mechanistic account of repair across the six instruction-tuned models: the corrected reading is localised to the later layers, is causally responsible for the output there, is absent rather than merely weak when repair fails behaviourally, and is carried by attention whose head-level organisation is model-dependent. (3) Through a speaker-attribution manipulation that holds the committed reading fixed while changing its speaker, we find that revision tends to be stronger for assistant-role commitments than for user-role assertions of the same content, which suggests that the attributed speaker of a commitment, and not only its content, shapes how it is revised. (4) Using an adversarial factual control, we separate genuine repair from mere deference and find that models resist a false correction far more than a valid one, though this resistance varies widely across models.

## 2 Related Work

**Repair and ambiguity in dialogue.** Conversation analysis provides the theoretical vocabulary of repair (Schegloff et al., 1977; Schegloff, 1992). In NLP, ambiguity has been characterised through entailment (Liu et al., 2023) and through open-domain question answering with multiple valid readings (Min et al., 2020), and clarification has been studied in conversational QA (Guo et al., 2021). These resources treat ambiguity and clarification behaviourally. We reuse their ambiguity material but add a controlled repair structure and analyse internal representations.

**Interpretability methods.** The logit lens and its trained refinement, the tuned lens (Belrose et al., 2023), decode intermediate representations into vocabulary space and reveal that predictions are refined across layers. Causal tracing and activation patching localise computations to specific components (Meng et al., 2022; Wang et al., 2023). We adopt both as tools in our work.

**In-context override and knowledge conflict.** The closest prior work studies competition between a model’s parametric knowledge and conflicting in-context information. Such competition is resolved in late layers and is controllable through attention (Ortu et al., 2024; Jin et al., 2024), and grounded and ungrounded responses have distinct causal signatures (Monea et al., 2024). Repair is mechanistically a pragmatic instance of this competition: the model must override a default interpretation displayed in the assistant role. We make this connection explicit, and we show that the head-level organisation of repair differs from the few-dominant-heads pattern reported for factual counterfactuals.

## 3 The Repair Diagnostic Dataset

**Design principles.** The dataset is built to satisfy three requirements. First, the ambiguity must be genuine and visible to the model, rather than dependent on private knowledge. We therefore source ambiguity from existing datasets rather than inventing it. Second, every item must be a minimal pair, so that a repair run and a control run differ only in the third turn. Third, the model’s default reading must be the incorrect one, so that there is a misunderstanding for the repair to overturn; we enforce this with a per-model behavioural filter (Section 4).

| INT (intent), contrastive repair |                                                                                                         |
|----------------------------------|---------------------------------------------------------------------------------------------------------|
| Turn 1                           | Who is the longest-serving manager in the Premier League?                                               |
| Turn 2                           | By total time in charge, that is Sir Alex Ferguson, who managed Manchester United for over two decades. |
| Turn 3 (re-pair)                 | Not by total years. I mean measured by the number of games managed.                                     |
| Turn 3 (control)                 | By total years in charge he is far ahead of anyone else in the history of the league.                   |
| Probe                            | So who is it?                                                                                           |
| Readings                         | correct: Arsène Wenger; incorrect: Alex Ferguson                                                        |

Table 1: A worked dataset item. The repair and control runs share Turns 1 and 2 and differ only in Turn 3. The default reading (most total time) is incorrect; the repair selects the intended reading (most games). Further examples are provided in Appendix A.

**Item structure.** Each item is a three-turn dialogue. Turn 1 is an ambiguous user utterance. Turn 2 is an assistant turn that commits to a plausible but incorrect default reading and makes that reading observable. Turn 3, in the repair condition, is a user correction that selects the intended reading; in the control condition it is a length-matched, non-corrective filler that does not introduce the correct reading or its answer. A probe question, identical across conditions, reads out the model’s interpretation through a constrained two-choice question whose options are the correct and incorrect readings, presented in a randomised order to control position bias. Table 1 gives a worked example, and further examples spanning the other ambiguity types are provided in Appendix A.

**Taxonomies.** We annotate two independent dimensions. The ambiguity type is one of referential (REF), intent or sense (INT), discourse and coreference (COREF), pragmatic and scope (PRAG), or missing information (MISS). The repair format is one of explicit, contrastive, elliptical, or clarification, and a separate boolean flag marks meta-repair turns that explicitly comment on the misunderstanding.

**Sourcing.** Referential, intent, and missing-information items are drawn from AmbigQA (Min et al., 2020), which builds on Natural Questions (Kwiatkowski et al., 2019); its multiple disambiguated readings supply the two interpretations and their answers. Pragmatic and scope items are drawn from the ambiguous sentences of AMBIENT (Liu et al., 2023). Coreference items are drawn

|       | by ambiguity type | by source    |
|-------|-------------------|--------------|
| REF   | 69                | AmbigQA: 148 |
| INT   | 40                | MultiWOZ: 60 |
| MISS  | 99                | AMBIENT: 42  |
| COREF | 42                | WSC: 41      |
| PRAG  | 41                |              |
| Total | 291               |              |

Table 2: Composition of the repair diagnostic dataset. Each item is a minimal pair (a repair run and a control run).

from the Winograd Schema Challenge (Levesque et al., 2012), whose two candidate antecedents form the two readings. To broaden the dataset beyond a question-answering register, we additionally include task-oriented items in the style of MultiWOZ (Budzianowski et al., 2018), in which a user makes a booking-style request, the assistant commits to a default value for an under-specified slot, and the user repairs that value; here the two readings are the two slot values. Only the repair scaffolding (Turn 2, Turn 3, the control turn, and the probe) is authored; the ambiguity and the gold readings are sourced. We manually reviewed the authored scaffolding for each item to check that the repair selected the intended reading, the control did not leak the answer, and the probe remained identical across conditions. Table 2 summarises the composition, which spans two registers: open-domain question answering and task-oriented dialogue.<sup>1</sup>

## 4 Experimental Setup

**Models.** We study seven instruction-tuned models: Qwen2.5 at 0.5B, 3B, 7B, 14B, and 32B (Yang et al., 2024), Phi-3.5-mini (Abdin et al., 2024), and Llama-3.1-8B (Grattafiori et al., 2024), the six larger of which pass the behavioural gate and carry the mechanistic analyses. The Qwen series gives a five-point scaling sweep within a single family. All experiments use greedy decoding and a fixed random seed for the option ordering, so that runs are deterministic and comparable.

**Read-out.** Behaviourally, we record whether the model selects the correct or the incorrect reading on the two-choice probe. For the internal analyses we use the residual stream at the final token position, where the model is about to answer. We report the logit difference between the correct and incorrect

<sup>1</sup>Data and code: <https://github.com/vedanthnanesha/conversationalrepair>

| Model        | rep. | pers. | ctrl. | effect |
|--------------|------|-------|-------|--------|
| Qwen2.5-0.5B | 0.48 | 0.52  | 0.47  | 0.01   |
| Qwen2.5-3B   | 0.69 | 0.31  | 0.12  | 0.57   |
| Qwen2.5-7B   | 0.99 | 0.01  | 0.01  | 0.99   |
| Qwen2.5-14B  | 0.92 | 0.08  | 0.24  | 0.68   |
| Qwen2.5-32B  | 0.96 | 0.04  | 0.21  | 0.75   |
| Phi-3.5-mini | 0.73 | 0.27  | 0.12  | 0.61   |
| Llama-3.1-8B | 0.94 | 0.06  | 0.49  | 0.45   |

Table 3: Behavioural repair gate ( $n = 291$ ). rep.: repair success; pers.: persistence of the incorrect reading; ctrl.: control baseline; effect: repair minus control.

option letters, averaged over both option orderings, with bootstrap confidence intervals over items.

**Per-model minimal-pair filter.** A valid minimal pair requires that, without repair, the model holds the incorrect reading. We therefore keep, for each model, only items whose control run yields the incorrect reading and whose repair run yields the correct reading. This filtered subset is the basis for all mechanistic analyses, and it differs across models because their default readings differ. Sample sizes therefore vary across analyses, and the failure analyses use the complementary set of items where repair does not succeed.

## 5 Experiments and Results

### 5.1 Behavioural repair and scaling

Before looking inside the models, we first measure whether they incorporate the repair at all. Table 3 reports, for each model, the repair success rate, the persistence of the incorrect reading, the control baseline, and the resulting effect. Repair competence scales with model size within the Qwen family, from no measurable effect at 0.5B to near-perfect repair from 7B onward. The 0.5B model fails the gate and is excluded from the mechanistic analyses. The larger models (14B, 32B) and Llama show higher control baselines, meaning that for some items their default reading is already the intended one; the per-model filter removes those items, leaving 124 to 287 valid minimal pairs per model.

### 5.2 Where the corrected reading appears

We next locate, layer by layer, where the corrected reading becomes decodable. Using the logit lens, we read the correct-minus-incorrect logit difference at every layer for both the repair and the control run, and average over the filtered items. Figure 2 shows the result for Qwen2.5-7B. Through the early and

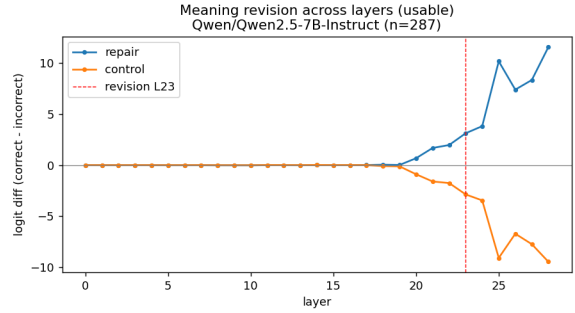


Figure 2: Layer-wise logit difference (correct minus incorrect) for the repair and control runs, Qwen2.5-7B. The curves coincide and favour the incorrect reading until the late layers, then separate sharply.

middle layers the two curves are indistinguishable and both favour the incorrect reading, even though the model has already read the repair turn. The curves separate sharply in the late layers, with the repair run swinging to the correct reading while the control run remains on the incorrect one. We define the revision layer as the first layer at which the repair curve crosses zero and the separation reaches a quarter of its maximum.

Table 4 reports the revision layer and the final-layer separation, with confidence intervals, across models. The revision occurs at a relative depth of 56 to 89 percent. Llama revises earliest, and within the Qwen family the revision sits at a stable relative depth (77 to 89 percent) as the model is scaled from 3B to 32B. Because the raw separation is measured in each model’s own logit units and is therefore not comparable across models, we also report a scale-free read-out: the normalised revision margin  $M_{\text{rep}}$ , defined as  $p(\text{correct}) - p(\text{incorrect})$  under a two-way softmax over the two reading tokens at the final layer, which lies in  $[-1, 1]$ . In the original assistant-role condition,  $M_{\text{rep}}$  on the repair run is 0.84 to 0.97 across all six models, whereas on the matched control run the corresponding margin is strongly negative ( $-0.48$  to  $-0.97$ ); the normalised revision effect (repair minus control margin) is thus 1.4 to 1.9 in every model, a consistency that the raw logit separation obscures. A per-type breakdown for Qwen2.5-7B places the revision layer at the same layer for all five ambiguity types, with comparable final separations, indicating that the late-layer revision is not specific to any one kind of ambiguity. The revision layer is likewise identical across the four repair formats, although elliptical repairs, which supply only a corrected fragment, produce a weaker final separation than

| Model        | rev.   | sep. [95% CI]     | $M_{\text{rep}}$ | peak |
|--------------|--------|-------------------|------------------|------|
| Llama-3.1-8B | L18/32 | 10.9 [9.8, 12.0]  | 0.90             | L30  |
| Phi-3.5-mini | L22/32 | 37.9 [35.7, 40.1] | 0.84             | L28  |
| Qwen2.5-3B   | L32/36 | 24.1 [22.9, 25.1] | 0.86             | L34  |
| Qwen2.5-7B   | L23/28 | 21.0 [20.3, 21.8] | 0.97             | L26  |
| Qwen2.5-14B  | L37/48 | 41.4 [39.6, 43.2] | 0.92             | L46  |
| Qwen2.5-32B  | L53/64 | 32.3 [31.1, 33.6] | 0.97             | L62  |

Table 4: Localisation of meaning revision. rev.: probing revision layer; sep.: final-layer separation between repair and control in raw logits;  $M_{\text{rep}}$ : normalised final-layer revision margin for the repair run,  $p(\text{correct}) - p(\text{incorrect})$  in  $[-1, 1]$ ; peak: causal peak layer from patching (Section 5.3).

explicit or contrastive repairs; this ordering holds in every model we examine, and is detailed in Appendix E. The same early-overlap, late-separation pattern holds for the remaining models, as shown in Appendix B, Figure 8. The effect is also about the correction rather than its wording, since rewording each repair turn into several meaning-preserving paraphrases leaves Qwen2.5-7B’s repair success at 0.87, against 1.00 for the original phrasing.

### 5.3 Where the corrected reading is caused

Decodability is correlational, and the logit lens is moreover an approximate probe: reading intermediate states directly through the unembedding can under- or over-state what a layer represents, which is precisely what motivated trained refinements such as the tuned lens (Belrose et al., 2023). We therefore treat the probing curves as evidence of where the corrected reading becomes *decodable*, and establish causation with a lens-independent method, activation patching. For each item we run the repair pass and cache its last-token residual at every layer. We then run the control pass and overwrite a single layer’s last-token residual with the cached repair value, and we record whether the control output flips to the correct reading. Patching the final token position avoids the length mismatch between the repair and control turns. Figure 3 shows that for Qwen2.5-7B patching has no effect through the early layers, then rises sharply to a near-complete flip rate in the late layers. The causal peak layers in Table 4 coincide with the probing revision layers, so the place where the corrected reading is decodable is also where the corrected answer decision becomes causally available at the final position. Across all six models the causal flip rate at the peak is 0.90 or above. Appendix B, Figure 9, shows the per-layer flip rate for the re-

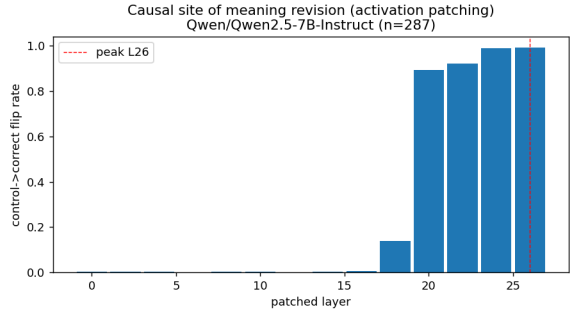


Figure 3: Causal flip rate as a function of the patched layer, Qwen2.5-7B. Patching the late-layer residual of the repair run into the control run flips the output to the corrected reading.

maining models, where the same sharp late-layer onset is present.

**Causal controls.** To verify that the flip reflects the repair signal rather than an arbitrary perturbation, we compare four patches at the peak layer for Qwen2.5-7B (Appendix B, Table 11). A random vector with matched norm flips the output only at chance (0.51), which rules out a generic perturbation effect. Patching a different item’s repair residual, or the mean repair residual, both matched on the target option, also flips the output. This indicates that by the final token the residual encodes the answer decision rather than item-specific semantic content; the item-specific repair information is integrated earlier, consistent with the late-layer separation that probing reveals. The pattern holds across all six models: the random control stays at chance (0.48 to 0.57) while the own-repair patch reaches 0.93 to 1.00 (Appendix B, Figure 11).

### 5.4 Failed repair lacks the internal signature

The smaller models provide enough repair failures to connect behaviour to mechanism: Phi-3.5-mini (79 items) and Qwen2.5-3B (85). In both, the failure probing curve never crosses zero, so the internal state remains committed to the incorrect reading. Causally, patching failures at the peak layer flips only 0.21 of Phi control runs ([0.15, 0.27]) and 0.15 of Qwen2.5-3B runs ([0.11, 0.21]), against around 0.90 for successful items. Failed repair is therefore not a weak version of successful repair; across both families the corrected reading is largely absent from the representation (Appendix B, Figure 10).

Final-layer probabilities make the overwrite contrast explicit: the incorrect reading has probability 0.01 to 0.07 on successful repairs, 0.60 to 0.85 on

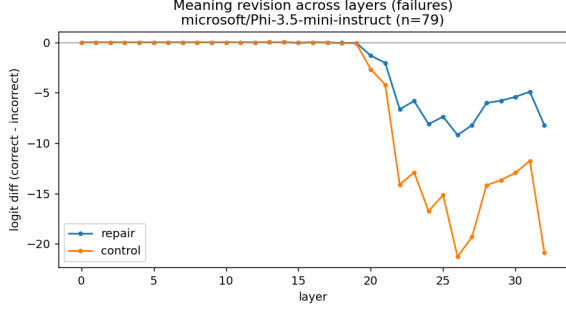


Figure 4: Layer-wise logit difference for behavioural repair-failure items, Phi-3.5-mini. The repair curve never crosses zero, indicating that the corrected reading never overtakes the committed interpretation.

failures, and 0.74 to 0.99 in controls. Persistence is thus the signature of repair failure rather than a general property of repair. Figure 4 shows the failure curves, and read as a classifier the revision-layer margin separates eventual successes from failures with AUROC 0.97 for Phi-3.5-mini and 0.98 for Qwen2.5-3B, confirming that the repair outcome is largely fixed by the revision layer rather than later (Appendix F).

### 5.5 Attention, not the feed-forward sublayer

The sublayer- and head-level analyses in this section and the next use a finer per-sublayer and per-head causal sweep. We run these on three models chosen to span the model families and scales in our study (Qwen2.5-7B, Phi-3.5-mini, and Llama-3.1-8B), which is sufficient to establish the qualitative cross-family contrast we report. We decompose the revision into the attention and feed-forward contributions of each layer by patching each sublayer output separately. In all three models the attention output carries the repair signal more strongly than the feed-forward output (Qwen: 0.18 versus 0.13; Llama: 0.56 versus 0.28; Phi: 0.09 versus 0.02). Figure 5 shows this for Qwen2.5-7B, where the attention curve lies above the feed-forward curve through the revision window; the same ordering holds for Phi and Llama (Appendix C). This is consistent with prior work attributing in-context override to attention (Ortu et al., 2024; Jin et al., 2024).

### 5.6 Distributed versus concentrated heads

We finally ask which attention heads carry the revision. We patch the contribution of each individual head, in a layer window around the revision layer, from the repair run into the control run. No single

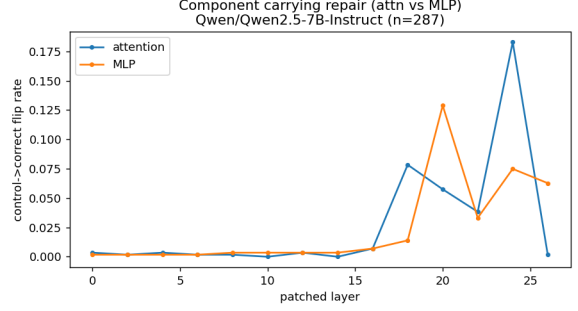


Figure 5: Attention versus feed-forward component patching for Qwen2.5-7B. Patching only the attention output of a layer recovers more of the corrected reading than patching only the feed-forward output, identifying attention as the dominant carrier.

head is sufficient on its own, with the strongest single head flipping at most a few percent of items for Qwen and Phi. To test whether a set of heads is sufficient, we patch the top-ranked heads together and sweep their number. The result reveals a clear cross-model difference. For Qwen2.5-7B the cumulative flip rate climbs from 0.05 for one head to 0.88 for sixteen heads, so repair is distributed across a head population. For Phi-3.5-mini the cumulative effect is weaker and saturates near 0.45. For Llama-3.1-8B, by contrast, a single head already flips 0.47 of items and four heads reach 0.88, so repair is concentrated in a few heads. The carrier of repair is therefore always attention, but its head-level organisation ranges from concentrated to fully distributed across model families. Figure 6 shows the cumulative head-patching curves for all three models, and the head-level causal maps are given in Appendix D (Figure 14 for Qwen2.5-7B, Figure 13 for Phi and Llama).

**Repair-reading heads.** Having located the causal heads, we ask what they compute. From the decision position, we measure the attention each head pays to the user’s repair turn. The top causal heads attend to the repair turn substantially more than other heads in the same layer window, by a factor of 1.5 in Qwen2.5-7B, 1.9 in Llama-3.1-8B, and 2.3 in Phi-3.5-mini. The heads that causally carry repair are therefore the heads that read the user’s correction, which connects the causal location to a concrete function and helps explain why repair is realised in attention rather than in the feed-forward sublayer.

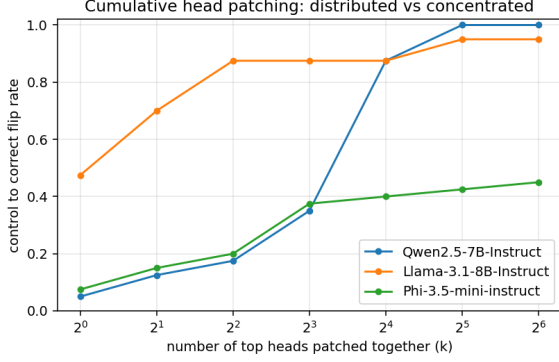


Figure 6: Cumulative head patching. The flip rate as more top-ranked heads are patched together reveals a concentrated organisation for Llama (a single head suffices) and a distributed organisation for Qwen and Phi (many heads are required).

### 5.7 Whose commitment is it? Assistant-role versus user-role

Every analysis so far treats the incorrect reading in Turn 2 as an assistant-role commitment. This is what makes the phenomenon repair rather than ordinary in-context override: the model is conditioned on a displayed understanding attributed to the assistant role. We test whether that attribution matters mechanistically. We construct a second version of every item in which Turn 2 is word-for-word identical but is attributed to the *user* rather than the assistant, so that the model revises the same committed reading when it is attributed to the user. The only change between the two versions is the role tag on Turn 2, so any difference is attributable to speaker role alone. We call the original setting **assistant** and the new one **user**.

We read the normalised revision margin  $M_{\text{rep}}$  (Section 5.2) on the repair run in both settings and define the *assistant-role advantage*  $A = M_{\text{rep}}(\text{assistant}) - M_{\text{rep}}(\text{user})$ . A positive  $A$  means the corrected reading is established more strongly when the misreading is attributed to the assistant role rather than to the user. Table 5 reports  $A$  for all six models. It is positive in every model, and its bootstrap confidence interval excludes zero in four of the six; the two largest Qwen models, whose repair margin is already near ceiling, show the same direction without reaching significance. Behaviourally, the assistant-role condition succeeds at least as often as the user-role condition in every model. Figure 7 shows that the assistant/user difference is absent through the early and middle layers and emerges only in the late layers, coinciding with

| Model        | $M_{\text{rep}} \text{ assistant}$ | $M_{\text{rep}} \text{ user}$ | $A$ [95% CI]          |
|--------------|------------------------------------|-------------------------------|-----------------------|
| Qwen2.5-3B   | 0.86                               | 0.82                          | 0.043 [0.005, 0.084]  |
| Qwen2.5-7B   | 0.96                               | 0.92                          | 0.040 [0.015, 0.067]  |
| Qwen2.5-14B  | 0.93                               | 0.91                          | 0.019 [-0.007, 0.046] |
| Qwen2.5-32B  | 0.97                               | 0.95                          | 0.016 [-0.005, 0.043] |
| Phi-3.5-mini | 0.84                               | 0.73                          | 0.110 [0.053, 0.170]  |
| Llama-3.1-8B | 0.91                               | 0.82                          | 0.087 [0.056, 0.125]  |

Table 5: Assistant-role advantage.  $M_{\text{rep}}$  is the normalised final-layer revision margin on the repair run;  $A = M_{\text{rep}}(\text{assistant}) - M_{\text{rep}}(\text{user})$  is positive when repairing an assistant-role commitment yields a stronger corrected reading than repairing a user-role assertion of the same content.  $A > 0$  in all six models and its 95% confidence interval excludes zero in four.

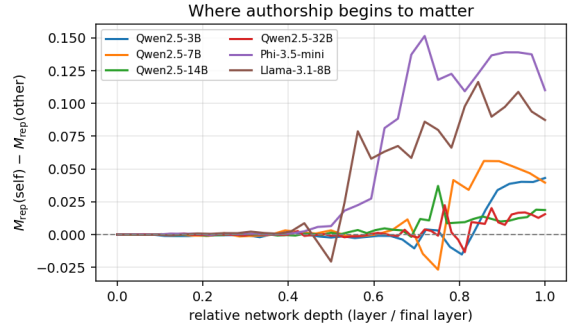


Figure 7: Per-layer assistant/user difference in  $M_{\text{rep}}$  against relative network depth. The difference is near zero through the early and middle layers and fans out positive in the late layers, coinciding with the revision window.

the revision window of Section 5.2.

A plausible reading is that a user-role assertion behaves like a competing knowledge source, of the kind studied in factual knowledge conflict, and so lingers, whereas a self-attributed displayed understanding is more cleanly overwritten by the repair pathway. The manipulation is deliberately minimal, and the *user* framing places a stated-then-corrected belief in the user’s turns, a slightly different discourse configuration; we therefore read the effect as evidence that speaker attribution matters, rather than as an exact measure of its size.

### 5.8 Repair versus sycophancy

A natural worry is that what we have called repair is just sycophancy, with the model following any user pushback whether or not it is warranted. Our ambiguity items cannot settle this, because their two readings are both legitimate and neither is false. We therefore build a separate adversarial control from 300 factual questions in MMLU (Hendrycks et al., 2021), each with one correct answer and a plausible

| Model        | held | caving | repair |
|--------------|------|--------|--------|
| Qwen2.5-3B   | 253  | 0.31   | 0.69   |
| Qwen2.5-7B   | 240  | 0.33   | 0.99   |
| Qwen2.5-14B  | 250  | 0.19   | 0.92   |
| Qwen2.5-32B  | 248  | 0.11   | 0.96   |
| Phi-3.5-mini | 261  | 0.58   | 0.73   |
| Llama-3.1-8B | 205  | 0.21   | 0.94   |

Table 6: Repair versus sycophancy. held: items (of 300) whose correct answer the model holds under a neutral turn; caving: fraction of those on which the model abandons the correct answer when the user falsely insists on a distractor; repair: genuine repair success from Table 3. No model caves as readily as it accepts a genuine correction, but resistance to a false one varies widely across models (caving 0.11 to 0.58).

distractor. The assistant states the correct answer, the user then insists, wrongly, on the distractor, and the same two-choice probe asks which option is factually correct. For each model we keep only the questions whose answer it holds under a neutral follow-up, so the test is restricted to facts the model actually knows; the sycophancy rate is the fraction of those on which it abandons the correct answer under an adversarial correction.

No model abandons a known fact as readily as it accepts a genuine correction, so repair is not merely deference to the most recent user turn. Resistance is far from complete, however, and varies across models, with caving ranging from 0.11 (Qwen2.5-32B) to 0.58 (Phi-3.5-mini) and falling with scale within the Qwen family. The caving and repair rates come from different datasets and probes, so we read their contrast qualitatively rather than as a precise ratio.

## 6 Discussion

Across the six models the same picture recurs. The correction is taken up only in the later layers, and the layers where the corrected reading becomes decodable are also the layers where it is causally responsible for the output; when repair fails, that revision is absent rather than merely weak; and in every case it is attention, not the feed-forward sub-layer, that carries it. This places repair within the family of in-context override computations, alongside factual knowledge conflict. At the same time, the head-level analysis distinguishes repair from the few-dominant-heads pattern reported for factual counterfactuals: in two of our three models repair is distributed across a head population rather than concentrated. The degree of distribution is a prop-

erty of the model family, not of the phenomenon, which suggests that conclusions about head specialisation drawn from a single model should be treated with caution.

Finally, the speaker-attribution manipulation points to a difference between repair and generic in-context override. Holding the committed reading fixed, the model revises an assistant-role misreading somewhat more completely than the same misreading asserted by the user. Viewed through the lens of conversational grounding, even a modest asymmetry of this kind is natural, since revising one’s displayed understanding is how speakers restore common ground after a misunderstanding (Clark, 1996). The models revise readily when a correction is warranted but are far more reluctant to abandon a known fact when the user insists on a falsehood, though this reluctance varies across models (Section 5.8).

## 7 Conclusion

We presented a controlled diagnostic dataset and a mechanistic study of conversational repair in language models, adapting interpretability tools to a previously unstudied linguistic phenomenon and contributing a reusable resource, a localisation of repair-induced meaning revision that is consistent across model families, and a mechanistic account of when repair fails. We hope the dataset and this framing are useful to others studying how interactive systems build, lose, and restore shared understanding. We see this as a first step toward how language models handle meaning in interaction, a largely unexplored complement to the study of factual recall and reasoning. Several questions follow naturally. How is repair realised in casual, multi-party, and multilingual dialogue, where trouble and its resolution are messier than in our controlled pairs? Do models maintain an internal representation of their own commitments, distinct from propositions they merely entertain, and can it be read or steered? When a model yields to an unwarranted correction, is it overriding a representation of the truth or simply failing to form one? We offer the diagnostic and this framing as a starting point for moving from models that produce language toward models we can understand, and correct, in dialogue.

## Limitations

Our ambiguity material is drawn from two registers, open-domain question answering and task-

oriented dialogue, but roughly half of the items remain question-answering in style. The late-layer revision we observe may therefore partly reflect how these models perform answer selection, and confirming the pattern on casual and multi-party dialogue is important future work. Our read-out is a constrained two-choice probe; free-form repair in naturalistic settings may behave differently. The head-level analysis uses an item subset and a layer window for tractability, and reports causal flip rates rather than a full circuit. Finally, although we scale to 32B parameters within one family, all models studied are instruction-tuned and below 100B parameters, and larger or differently trained models may organise repair differently. For reproducibility, all experiments use open-weight instruction-tuned models and publicly available source datasets, and every read-out requires only forward passes with no additional training.

## Ethics Statement

The dataset is built from publicly available research datasets and contains no personal or sensitive data. The work is analytical and poses no foreseeable risk of harm. A clearer understanding of when models do and do not revise their interpretation after correction may support more transparent and trustworthy interactive systems.

## References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *arXiv preprint arXiv:2404.14219*.
- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the Tuned Lens](#). *arXiv preprint arXiv:2303.08112*.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Herbert H. Clark. 1996. *Using Language*, 1 edition. Cambridge University Press, Cambridge.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. [Abg-CoQA: Clarifying ambiguity in conversational question answering](#). In *Proceedings of the 3rd Conference on Automated Knowledge Base Construction (AKBC 2021)*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. [Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural Questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The Winograd schema challenge](#). In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012*, pages 552–561, Rome, Italy. AAAI Press.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. [We’re afraid language models aren’t modeling ambiguity](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. 2024. [A glitch in the matrix? locating and detecting language model grounding with fakepedia](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6828–6844, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. [Competition of mechanisms: Tracing how language models handle facts and counterfactuals](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Emanuel A. Schegloff. 1992. [Repair after next turn: The last structurally provided defense of intersubjectivity in conversation](#). *American Journal of Sociology*, 97(5):1295–1345.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. [The preference for self-correction in the organization of repair in conversation](#). *Language*, 53(2):361–382.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: A circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

## A Additional dataset examples

We provide four further dataset items, chosen to illustrate the remaining ambiguity types, repair formats, and the task-oriented register. As in the main text, the repair run and the control run share Turns 1 and 2 and differ only in Turn 3, and the probe is a constrained two-choice question over the two readings.

| Task-oriented (MISS), explicit repair |                                                                                    |
|---------------------------------------|------------------------------------------------------------------------------------|
| Turn 1                                | Can you book a table for dinner at La Bella Vita tonight?                          |
| Turn 2                                | Sure, I can book a table for one person at La Bella Vita tonight. Is that correct? |
| Turn 3 (repair)                       | Actually, I meant for three people.                                                |
| Turn 3 (control)                      | That sounds perfect, when would you like to arrive?                                |
| Probe                                 | How many people will be dining?                                                    |
| Readings                              | correct: three; incorrect: one                                                     |

Table 7: A task-oriented item in the style of a booking dialogue. The assistant commits to a default value for an under-specified party-size slot, and the repair turn corrects it.

| COREF (coreference), contrastive repair |                                                                                                                        |
|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| Turn 1                                  | The new hospital is right next to the interstate, which makes it easy for patients to get there.                       |
| Turn 2                                  | That is a real benefit; being right off the highway means patients can reach the hospital quickly when they need care. |
| Turn 3 (repair)                         | I did not mean getting to the hospital. I meant it makes it easy for patients to get to the interstate from it.        |
| Turn 3 (control)                        | That kind of easy access really does make a genuine difference for people during an emergency.                         |
| Probe                                   | So what is it that is easy for patients to get to?                                                                     |
| Readings                                | correct: the interstate; incorrect: the hospital                                                                       |

Table 8: A coreference item. The pro-form “there” is ambiguous between two antecedents introduced in Turn 1.

## B Cross-model localisation of revision

The results below support the claims in Sections 5.2 and 5.3 that the localisation of meaning revision holds for most models we study, not only the headline model.

Figure 8 reproduces the probing analysis for three additional models. In each panel the vertical axis is the correct-minus-incorrect logit difference read from the residual stream by the logit lens, and

| PRAG (scope), contrastive repair |                                                                                                            |
|----------------------------------|------------------------------------------------------------------------------------------------------------|
| Turn 1                           | I asked the participant whether they were a US citizen or a Green Card holder.                             |
| Turn 2                           | So you were checking which of the two statuses applies to them, citizen or Green Card.                     |
| Turn 3 (re-pair)                 | Not which one. I was asking whether they held either status at all, since they might have neither.         |
| Turn 3 (control)                 | It is a fairly standard eligibility question that we ask nearly everyone at intake.                        |
| Probe                            | So what was I actually asking the participant?                                                             |
| Readings                         | correct: whether they held either status, possibly neither; incorrect: which of the two statuses they hold |

Table 9: A pragmatic scope item. The disjunction admits an exclusive reading (which one) and an inclusive reading (either, possibly neither).

| MISS (missing constraint), elliptical meta-repair |                                                                               |
|---------------------------------------------------|-------------------------------------------------------------------------------|
| Turn 1                                            | What is the tallest ride at Six Flags Over Texas?                             |
| Turn 2                                            | The tallest is Titan, a steel hypercoaster with a drop of more than 250 feet. |
| Turn 3 (re-pair)                                  | No, you have misunderstood me. The tallest wooden one.                        |
| Turn 3 (control)                                  | That one looks genuinely terrifying. I have heard the first drop is unreal.   |
| Probe                                             | So which ride do I mean?                                                      |
| Readings                                          | correct: The New Texas Giant; incorrect: Titan                                |

Table 10: A missing-constraint item with an elliptical, meta-repair Turn 3 that explicitly flags the misunderstanding.

the two curves are the repair run and the control run. The reading is the same across models: the curves are indistinguishable and favour the incorrect interpretation through the early and middle layers, then separate in the late layers as the repair run alone moves to the correct interpretation. The layer at which the separation begins differs across models, which is the variation in revision depth reported in Table 4, but the qualitative shape is invariant.

Figure 9 reproduces the causal patching analysis. The vertical axis is the fraction of control runs whose output flips to the correct reading when the repair run’s residual at a given layer is spliced in. For the successful items of Phi-3.5-mini and Llama-3.1-8B the flip rate remains at the unpatched baseline in the early layers and rises sharply to a near-complete flip in the late layers, mirroring the probing onset and confirming that the late layers are causally responsible. The third panel restricts the analysis to Phi-3.5-mini’s behavioural repair-failure items, where the flip rate stays low at every

| Patch at peak layer             | flip rate |
|---------------------------------|-----------|
| Own repair residual             | 1.00      |
| Different item (option-matched) | 1.00      |
| Mean repair residual (matched)  | 1.00      |
| Random vector (matched norm)    | 0.51      |

Table 11: Causal controls for patching, Qwen2.5-7B at L26 ( $n = 150$ ). A random vector flips the output only at chance, ruling out a generic perturbation effect.

layer; this is the causal counterpart of the failure result in Section 5.4.

The same failure signature appears in a second family. Figure 10 repeats both analyses for the 85 repair-failure items of Qwen2.5-3B: the probing curve never crosses into the correct region, and patching the failure items flips at most 0.15 of control runs at any layer, against 0.94 for its successful items. Failure as the absence of internal revision therefore holds for Qwen2.5-3B as well as Phi-3.5-mini.

Figure 11 reports the causal-control comparison of Section 5.3 across the six models. The repair patch flips almost every control run, while a random vector with matched norm never exceeds chance, which confirms that the effect is specific to the repair signal rather than to a generic perturbation at the peak layer.

## C Attention versus feed-forward

Additional component-patching results support the claim in Section 5.5 that attention, rather than the feed-forward sublayer, carries the repair signal, and show that the ordering holds beyond the main-paper model.

Figure 12 shows, for Phi-3.5-mini and Llama-3.1-8B, the flip rate obtained when only the attention output of a layer is patched and when only the feed-forward output is patched. In both models the attention curve lies above the feed-forward curve at its peak, matching Figure 5 for Qwen2.5-7B. The absolute size of the effect varies across models, with Llama showing the strongest single-component effects, but the qualitative ordering is consistent.

## D Per-model head maps

The per-model head maps below support the claim in Section 5.6 that the head-level organisation of repair differs across models.

Figure 13 shows the attention-head causal maps for Phi-3.5-mini and Llama-3.1-8B. Each cell is

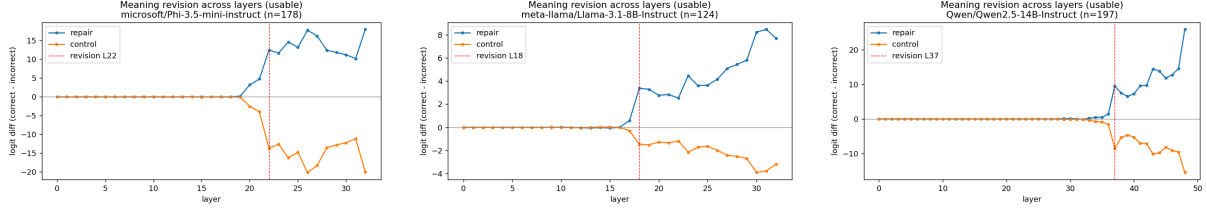


Figure 8: Layer-wise probing curves for Phi-3.5-mini (left), Llama-3.1-8B (centre), and Qwen2.5-14B (right). In every model the repair and control curves coincide in the early and middle layers and separate in the late layers.

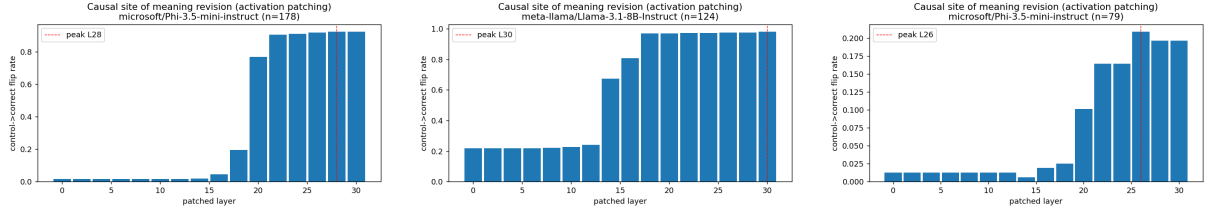


Figure 9: Causal patching flip rate per layer for Phi-3.5-mini (left) and Llama-3.1-8B (centre), and for the Phi-3.5-mini repair-failure items (right). Successful items reach a near-complete flip in the late layers; the failure items remain low at every layer.

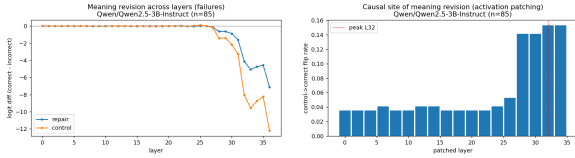


Figure 10: Repair-failure items of Qwen2.5-3B. Left: the probing curve never crosses into the correct region. Right: causal patching flips at most 0.15 of control runs at any layer. Both mirror the Phi-3.5-mini failure result.

the flip rate obtained by patching a single head at a single layer, with brighter cells indicating a larger causal effect. The Llama map contains a small number of bright cells, the visual signature of the concentrated organisation in which a few heads carry most of the effect. The Phi map is more uniform and dim, consistent with a distributed organisation in which no single head is sufficient. Together with the cumulative curves of Figure 6, these maps show that the same phenomenon is implemented by a few heads in one model and by a broad population in another.

**Attention knockout.** As a further probe of how the repair is read, we block the decision position from attending to the repair-text span while leaving the probe visible, and measure whether the output reverts to the incorrect reading. The effect is negligible: repair-run accuracy falls by at most 0.02 across the six models, and a control knockout of the assistant’s Turn-2 span has no effect. The cor-

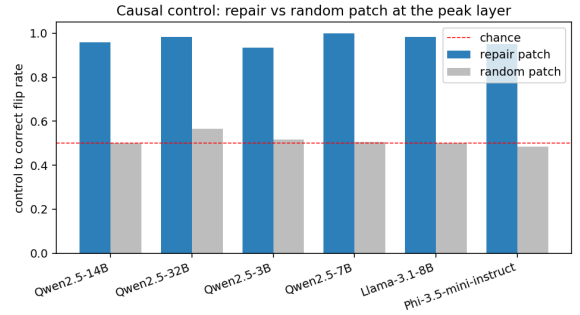


Figure 11: Causal control across models: the own-repair patch flips the control output near 1.0, while a random patch with matched norm stays at chance (dashed line).

rection is therefore routed to the decision position through intermediate positions rather than through a single direct attention edge. This is consistent with the distributed head organisation: no single attention path is necessary, even though the causal heads, in aggregate, attend preferentially to the repair turn.

## E Revision by repair format

The breakdown below supports the observation in Section 5.2 that the locus of revision is stable across repair formats while its strength varies. Table 12 breaks the Qwen2.5-7B probing analysis down by the four repair formats. The revision layer is identical (L23) for every format, but the final separation is markedly smaller for elliptical repairs, which supply only a corrected fragment, than for explicit,

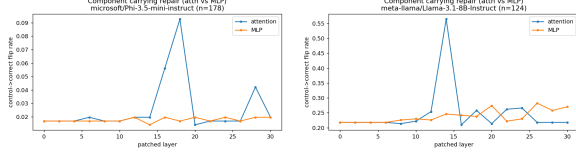


Figure 12: Attention versus feed-forward component patching for Phi-3.5-mini (left) and Llama-3.1-8B (right). Attention is the dominant causal carrier in both, as in Qwen2.5-7B (Figure 5).

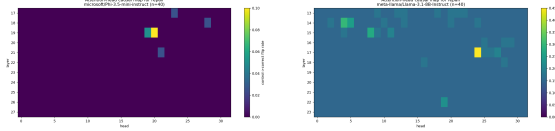


Figure 13: Attention-head causal maps for Phi-3.5-mini (left) and Llama-3.1-8B (right).

contrastive, or clarification repairs. This ordering, with elliptical repairs weakest, is consistent across all six models: the elliptical separation is the smallest of the four formats in every case, while the revision layer is format-invariant within each model. Meta-repair turns, which explicitly flag the misunderstanding, are too few in the filtered set ( $n = 3$ ) to support a separate estimate.

| Repair format | n   | revision | final sep. |
|---------------|-----|----------|------------|
| Explicit      | 58  | L23      | 22.8       |
| Contrastive   | 112 | L23      | 21.9       |
| Clarification | 72  | L23      | 21.2       |
| Elliptical    | 45  | L23      | 16.3       |

Table 12: Probing revision by repair format, Qwen2.5-7B. The revision layer is identical across formats; elliptical repairs yield the weakest separation.

Figure 15 shows the same comparison across all six models, with the final separation normalised within each model. Explicit, contrastive, and clarification repairs reach comparable strength, while elliptical repairs are the weakest in every model.

## F Separating successful and failed sep. by the revision margin

Here we expand the AUROC analysis of Section 5.4. For each item, we read the repair-run margin  $M_{\text{rep}}$  at each layer and ask how well it separates behavioural successes from failures. We run this on the full early-warning set for the two failure-rich models: Phi-3.5-mini (80 failures) and Qwen2.5-3B (90). Figure 16 shows AUROC near chance through the early and middle layers, then

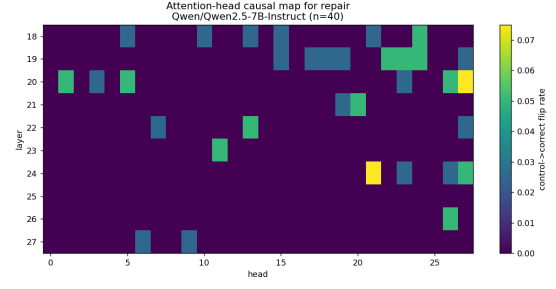


Figure 14: Attention-head causal map for Qwen2.5-7B in the layer window around the revision layer. No single head dominates; the causal effect is spread across many late-layer heads.

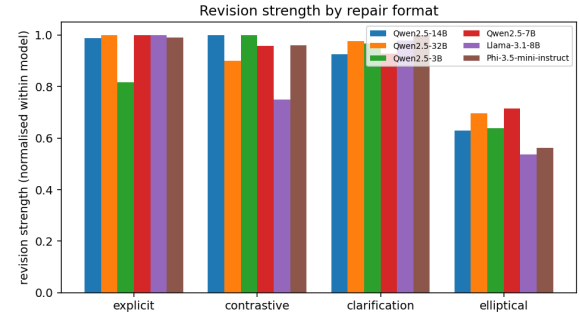


Figure 15: Revision strength by repair format, normalised within each model. Elliptical repairs produce the weakest revision in every model.

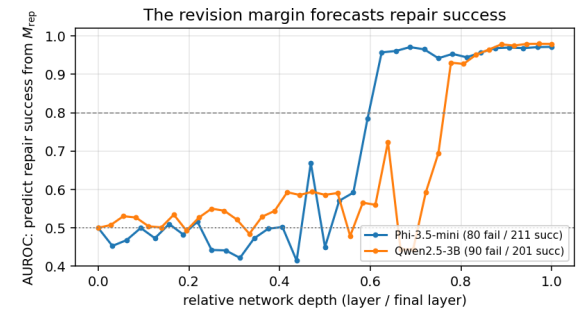


Figure 16: AUROC for separating behavioural repair successes from failures using the revision margin  $M_{\text{rep}}$  at each layer, against relative network depth. The margin is at chance early and rises to near one at the revision window.

near one at the revision window; final AUROC is 0.97 and 0.98, and the margin first exceeds 0.8 at 62 percent relative depth for Phi-3.5-mini and 78 percent for Qwen2.5-3B.