WHY DO LANGUAGE MODELS ANSWER INCORRECTLY DESPITE SUFFICIENT EVIDENCE? A MECHANISTIC STUDY OF RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) can still produce incorrect answers even when sufficient and correct evidence is available. This work studies failures of evidence utilization in the absence of explicit knowledge conflicts. We compare responses to the same question under two conditions: one containing only supporting evidence and the other containing the full context while retaining the same evidence. We combine layer-wise answer scoring with linear probes to track changes in internal representations, and further use bidirectional residual stream patching to test whether the identified states causally affect model responses. Experiments on four open-weight instruction-tuned models across three question-answering benchmarks show that pre-generation states in the middle-to-late layers distinguish answer regression from answer rescue induced by context expansion. When the question and input evidence remain unchanged, patching in the corresponding state from the same question under a correct-answer condition repairs a subset of incorrect responses, whereas patching in the reverse direction degrades a subset of correct responses. After determining and freezing the probe parameters and intervention locations on HotpotQA, we reproduce these effects on MS MARCO and Natural Questions. Across the two types of transfer samples, the repair rates range from 25.4% to 63.7% and from 44.2% to 81.8%, respectively. Equal-norm random-source controls further show that the intervention effect on the overall residual stream depends on the source state, while the effects associated with the attention and multilayer perceptron (MLP) modules vary across models. These findings provide causal intervention evidence for understanding why models sometimes fail to convert correct evidence into correct answers, and provide a basis for studying internal-state-based correction.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) uses external documents to help language models answer questions (Lewis et al., 2020; Asai et al., 2024; Jeong et al., 2024; Shi et al., 2024b). This paradigm has become an important foundation for knowledge-intensive question answering (QA). However, retrieving correct evidence does not guarantee generating a correct answer, as irrelevant information, hard distractor passages, and context position can all reduce answer accuracy (Shi et al., 2023; Amiraz et al., 2025; Zhang et al., 2024b). Sufficient Context distinguishes whether the context provides enough information to support answering a question from whether the model actually produces the correct answer (Joren et al., 2025). This work further focuses on the underlying mechanism: why does a model still fail to generate the correct answer when the context already contains sufficient and correct evidence?

A common explanation is that parametric knowledge competes with external context. Studies on counterfactual contexts, knowledge conflicts, and component-level interventions examine how models choose between internal knowledge and external evidence (Xu et al., 2024; Monea et al., 2024; Jin et al., 2024; Zhao et al., 2025a; Goyal et al., 2025; Li et al., 2025). In contrast, this work focuses on settings without explicit knowledge conflict: why does a model still fail to effectively use the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

provided evidence when no clear parametric answer is observed? We therefore examine whether supporting facts are encoded, integrated with the question, and represented in pre-generation states that affect the answer.

We compare the model's responses to the same question under two evidence conditions: one contains only the supporting evidence, while the other adds additional text while retaining the same supporting evidence, forming the full context. Prior studies show that additional context can either harm or improve answer accuracy (Cuconasu et al., 2024; Wu et al., 2025a). Based on whether the model answers correctly under the two conditions, we divide model behavior into four groups: Regression, Rescue, Persistent Failure, and Robust Control. Among them, Regression and Rescue provide a contrast for analyzing why context expansion induces answer changes in opposite directions.

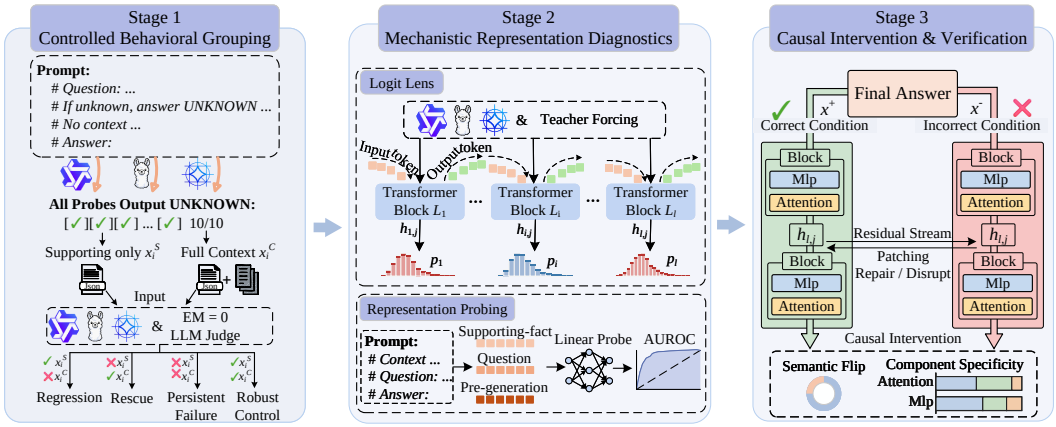


Figure 1: Overview of the experimental framework. Stage 1 selects questions for which ten closed-book probes yield no concrete answer, then compares responses to supporting evidence alone and to the full context containing the same evidence. This yields four behavioral groups: Regression, Rescue, Persistent Failure, and Robust Control. Stage 2 uses teacher-forced layerwise answer scoring and linear probes at three locations—the supporting-fact ends, question end, and pre-generation position—to examine answer preference and response transitions. Stage 3 patches residual-stream states in both directions between correct and incorrect runs of the same question, evaluates changes in freely generated answers, and compares interventions on attention and MLP outputs.

Figure 1 outlines our study of four open-weight instruction-tuned models. We combine layerwise vocabulary projection (Logit Lens), linear probes, and bidirectional residual-stream patching to relate internal answer information to candidate scores and free generation, then compare attention and multilayer perceptron (MLP) outputs.

The main contributions of this work are as follows:

1. We establish a same-question, support-only/full-context contrast that separates Regression, Rescue, Persistent Failure, and Robust Control without observed explicit knowledge conflict.

2. We develop a validation pipeline from representation analysis to generation intervention. By combining layerwise vocabulary projection, linear probes, and bidirectional state patching, we connect internal information readout, changes in answer scores, and free-generation outcomes, and examine their effects through component-level interventions and control analyses.

3. We identify a late-stage answer control phenomenon that reproduces across datasets, together with cross-model differences in component roles. Across four models and three benchmark datasets of different types, probes based on late-network states before answer generation predict Regression and Rescue, while patching these states affects model responses. However, the relative roles of attention and MLP modules vary across models.

2 Related Work

2.1 Evidence Utilization Failures and Knowledge Conflicts

The benefits of retrieval depend on both the external information and how effectively the model uses it (Lewis et al., 2020; Mallen et al., 2023). Self-RAG uses reflection tokens to regulate retrieval and generation (Asai et al., 2024), while Adaptive-RAG selects retrieval strategies based on question complexity (Jeong et al., 2024). REPLUG optimizes the retriever to adapt to a fixed language model (Shi et al., 2024b), whereas RAFT trains models to distinguish useful evidence from distractors (Zhang et al., 2024a). Irrelevant information, hard distractor passages, and positional bias can reduce answer accuracy (Shi et al., 2023; Amiraz et al., 2025; Zhang et al., 2024b), while additional text can also improve responses (Cuconasu et al., 2024; Wu et al., 2025a). For evaluation, RAGTruth annotates generated content that lacks source support or contradicts the source (Niu et al., 2024), and RAGChecker separately diagnoses the retrieval and generation components (Ru et al., 2024). Joren et al. (2025) use context sufficiency to distinguish whether the available information is sufficient and analyze model behavior in answering and refusing to answer.

Research on knowledge conflicts examines how models choose between parametric knowledge and external evidence (Xu et al., 2024), including studies of counterfactual contexts (Monea et al., 2024), attention-head interventions, and the effects of instruction tuning (Jin et al., 2024; Goyal et al., 2025). Astute RAG integrates internal knowledge with external evidence (Wang et al., 2025), while FaithfulRAG addresses factual conflicts (Zhang et al., 2025). Context-Aware Decoding contrasts predictive distributions with and without context (Shi et al., 2024a), and CoCoA further adjusts decoding dynamically (Khandelwal et al., 2025). Choi et al. (2025) modulates reliance on different knowledge sources through conflict-aware soft prompts. Matched-condition protocols distinguish evidence availability from a model's ability to use available evidence (Xia, 2026).

2.2 Internal Mechanisms of External Evidence Utilization

Prior studies analyze representations related to context relevance assessment and supporting-evidence copying (Farahani & Johansson, 2024), as well as the relationships among context utilization, contributions from parametric knowledge, and hallucination (Sun et al., 2025). Research on retrieval heads identifies mechanisms for information extraction from long contexts (Wu et al., 2025b), compares differences between training on synthetic and real data (Zhao et al., 2025b), and constructs unfaithful samples through intervention to improve training and decoding (Huang et al., 2025). Component-level analyses further reveal a division of labor between knowledge coordination and contextual retrieval (Zhao et al., 2025a; Kahardipraja et al., 2025), while also showing that the same attention head can process both parametric and contextual information (Li et al., 2025).

Evidence attribution methods trace how evidence influences answers through context ablation (Cohen-Wang et al., 2024), context-sensitive tokens and gradients (Qi et al., 2024), as well as question-answer circuits, response-distribution attribution, and knowledge-utilization probing (Basu et al., 2025; Li et al., 2026; Wadhwa et al., 2026). Other studies examine retrieval-induced changes in layerwise representations and generation (Yeh & Li, 2026), and use sparse autoencoder interventions to mitigate contextual interference and improve generation faithfulness (Giannetti et al., 2026; Xiong et al., 2026).

2.3 From Representation Analysis to Causal Validation

Ghandeharioun et al. (2024) provides a framework for inspecting hidden representations, while Chuang et al. (2024) adjusts decoding based on prediction differences across layers. Correctness probes reveal discrepancies between internal signals and model outputs, as well as limitations in cross-task generalization (Orgad et al., 2025). Studies on multiple-choice questions further distinguish the formation of answer information from the prediction of option symbols (Wiegreffe et al., 2025). These findings motivate joint analyses of internal representations and generation behavior.

Studies on factual recall, knowledge editing, causal mediation, and circuit analysis provide a foundation for state interventions (Yu et al., 2023; Meng et al., 2022; Vig et al., 2020; Wang et al., 2022; Geva et al., 2023). Linear representations provide a basis for both probing and intervention (Park et al., 2024). Inference-Time Intervention identifies truthfulness-related directions in attention-head

activations, while Representation Engineering studies how population-level representations can be monitored and controlled (Li et al., 2023; Zou et al., 2025). Related work also shows that activation directions can modulate model behavior (Rimsky et al., 2024; Todd et al., 2024). Our intervention instead transfers a state between two evidence conditions for the same question, allowing us to test the effect of a question-specific state difference in both directions. At the same time, localization and editing effects may not align (Hase et al., 2023), and conclusions from interventions depend on the evaluation metric, perturbation method, and input format (Zhang & Nanda, 2024; Opielka et al., 2026).

3 METHOD

3.1 PROBLEM SETUP AND BEHAVIORAL GROUPING

For each sample i, let q_i denote the question, y_i the gold answer, and C_i the full context. The supporting evidence S_i consists of the supporting facts in C_i. For each question, we sample ten closed-book responses without external evidence using the same prompt and `temperature=1`. We retain only samples for which all ten responses are `UNKNOWN`, reducing the likelihood that an observable closed-book answer competes with the supplied evidence. Details of evidence construction, prompting, and decoding configurations are provided in Appendices A and C.

For each retained sample, we construct a support-only input $x_i^S = (S_i, q_i)$ and a full-context input $x_i^C = (C_i, q_i)$. We normalize answers and compute Exact Match (EM) following the official HotpotQA evaluation procedure. For responses with EM=0, Qwen3.8-27B judges whether they are semantically equivalent to the gold answer (Qwen Team, 2026). We denote correctness under the two conditions as $z_i^S, z_i^C \in 0, 1$, where 1 indicates a correct answer and 0 an incorrect answer. This yields four behavioral groups: Regression $(1, 0)$, Rescue $(0, 1)$, Persistent Failure $(0, 0)$, and Robust Control $(1, 1)$. Subsequent representation analyses and interventions use a high-confidence subset. Responses under both conditions must satisfy the corresponding criterion: a correct response must have EM=1, while an incorrect response must have EM=0 and be judged semantically nonequivalent to the gold answer.

3.2 LAYERWISE ANSWER SCORING AND REPRESENTATION PROBING

For a decoder-only model with L Transformer blocks, let $h_{\ell,j}$ denote the residual state at token position j, with $\ell = 0$ before the first block and $\ell = 1, \ldots, L$ after each block.

Layerwise Answer Scoring. At each residual position, we apply the model's final normalization layer and output head to obtain a vocabulary distribution p_ℓ. Let g and c be the fixed correct and incorrect answer candidates for a sample. For Regression and Rescue, they are the answers generated under the two conditions. For Persistent Failure, g is the first official gold answer in recorded order, converted to the predefined natural answer form, and c is the incorrect full-context response. We score the same candidate pair under both contexts; for Robust Control, we score only the correct answer.

We use teacher forcing, feeding the preceding tokens of a candidate answer to compute the probability of the next token. Let $S_\ell(a \mid x)$ denote the mean token log-probability of candidate answer a. We define the Whole-Answer Margin as $M_\ell(x; g, c) = S_\ell(g \mid x) - S_\ell(c \mid x)$. To compare answer preferences under an identical prefix, we further define the First-Divergence Margin as

$$F_\ell(x; g, c) = \log p_\ell(g_d \mid x, g_{<d}) - \log p_\ell(c_d \mid x, g_{<d}),$$

where d is the first token position at which the two answers differ, with the preceding token prefixes satisfying $g_{<d} = c_{<d}$. If one answer is a complete prefix of the other, the model's end-of-sequence token represents the termination of the shorter answer. For both margins, a positive value indicates a preference for the correct candidate, whereas a negative value indicates a preference for the incorrect candidate. We report these natural-log probability margins and their changes in nats, where one nat is one unit on the natural-log scale. Implementation details are provided in Appendix B.1.

For samples with an incorrect candidate, we compute the changes in answer preference induced by context expansion:

$$\Delta_{i,\ell}^M = M_\ell(x_i^C; g_i, c_i) - M_\ell(x_i^S; g_i, c_i),\ \Delta_{i,\ell}^F = F_\ell(x_i^C; g_i, c_i) - F_\ell(x_i^S; g_i, c_i).$$

A positive value indicates that the full context shifts the answer preference toward the correct candidate, whereas a negative value indicates a shift toward the incorrect candidate.

Representation Probing. At each residual position, we extract the supporting-fact end state (the mean representation of the final tokens of the supporting facts), the question-end state (the representation of the final question token), and the state immediately before answer generation (the pre-generation state, also called the Answer-Decision State). We train L_2-regularized logistic regression models on standardized representations to predict answer correctness under each of the two context conditions, yielding six correctness probes in total. For Regression and Rescue samples, we additionally compute the representation difference at the same layer and token position: $d_{i,\ell,p} = h^C_{i,\ell,p} - h^S_{i,\ell,p}$, where p denotes one of the three token positions above, and $h^S_{i,\ell,p}$ and $h^C_{i,\ell,p}$ are the corresponding representations. After standardizing the representation differences, we train three transition-direction probes to distinguish Rescue from Regression. All nine probes are trained independently at each layer.

For all probes, we select the regularization strength on the HotpotQA training set using five-fold stratified cross-validation, with Area Under the Receiver Operating Characteristic Curve (AUROC) as the evaluation metric. We then train the final probes on the full training set, fix the standardization parameters and probe parameters, and evaluate them on the independent HotpotQA validation set and two target datasets. For HotpotQA, we additionally construct a metadata-only baseline using question type, difficulty, text length, and the number of supporting facts to test whether the representations provide predictive information beyond these attributes. Training details are provided in Appendix B.2.

3.3 Bidirectional State Patching and Free-Generation Validation

Bidirectional State Patching. To test the effects of representations on model responses, we perform bidirectional residual stream patching on Regression and Rescue samples. The candidate residual positions consist of the positions with the highest cross-validation AUROC on the training set for each of the nine probes, the early-layer position $\lceil L/4 \rceil$, and the final position L.

For each sample, let x^+ and x^- denote the input conditions producing correct and incorrect answers. At each tested location, the supporting-fact ends, question-end, or pre-generation position—Repair replaces the residual state in x^- with the corresponding state from x^+; Damage reverses the transfer. Supporting-fact ends are matched in recorded fact order, while the other states are aligned by semantic role. Only residual-stream vectors at the target token indices are replaced; the receiving input, token indices, and positional embeddings remain unchanged. Even when the contexts differ in length, subsequent computation, including position-dependent attention, uses the receiving sequence's positions. We measure intervention effects using F_L, the First-Divergence Margin at the final residual position:

$$E^{\text{repair}} = F_L^{\text{repair}} - F_L(x^-), \qquad E^{\text{damage}} = F_L(x^+) - F_L^{\text{damage}}.$$

F_L^{repair} and F_L^{damage} denote the margins after intervention. A positive repair effect indicates that Repair increases the relative score of the correct candidate, while a positive damage effect indicates that Damage decreases this score. We additionally divide the raw effect by the margin difference between the two unpatched conditions to measure the magnitude of the change relative to the original gap. The definition is provided in Appendix B.3.

On the HotpotQA training set, for each candidate residual position, we compute the four mean effects of pre-generation interventions across the two behavioral groups and the two intervention directions, and then take their equally weighted average. We select the position with the highest score as the fixed main intervention position for subsequent independent validation.

Free-Generation Validation. At the main intervention position, we apply the same bidirectional patching and use autoregressive generation to test whether the complete answer changes. We include three controls: final-question-token patching at the same layer, early-layer pre-generation patching, and identity write-back, which respectively test the effects of token position, network depth, and the intervention operation itself. Intervention outputs are evaluated using the same EM and semantic-judge procedure.

Before patching, we regenerate an unperturbed answer under the same question and context condition. We consider baseline correctness reproduced when this answer receives the same correct-or-incorrect label as the originally recorded answer; the answer strings need not match.

For behavioral group b and intervention direction k, we define the semantic flip rate as $R_{b,k} = \frac{N_{b,k}^{\text{flip}}}{N_{b,k}^{\text{valid}}}$, where the denominator is the number of valid samples and the numerator is the number of samples exhibiting the expected flip: from incorrect to correct for Repair and from correct to incorrect for Damage. Generation configurations and details on sample integrity are provided in Appendix D.

3.4 Component Intervention and Source Specificity

Within the network block corresponding to the fixed main intervention position, we separately replace the overall residual output, attention-module output, and MLP-module output at the pre-generation position, and compare their effects on answer preference. For each output type, we perform both Repair and Damage under three conditions: same-question paired replacement, equal-norm random control, and identity write-back.

Same-question paired replacement uses the corresponding state from the other context condition of the same question. For the equal-norm random control, we use within-group pairing with a fixed random seed to assign another question from the same behavioral group as the state source for each sample. Let u_i denote the state to be intervened on, δ_i^P the change induced by paired replacement, and δ_j^R the state difference from source question j under the same network block, component, and intervention direction. We construct $\tilde{u}_i^R = u_i + \delta_j^R \frac{\|\delta_i^P\|_2}{\|\delta_j^R\|_2}$. This control preserves the change direction from another question while matching its magnitude to that of the same-question paired replacement, allowing us to test whether the intervention effect depends on the direction of the state change for the current question. We measure Directional Specificity as $\text{Spec} = E^P - E^R$. A positive value indicates that same-question paired replacement more strongly drives the expected score change, whereas a negative value indicates a larger effect from the random control. Finally, using the overall residual intervention as a reference, we compute $E^{\text{Attn}} - E^{\text{MLP}}$ for the same sample and intervention direction to compare the paired intervention effects of attention module and MLP module.

4 Experiments

4.1 Data, Models, and Evaluation Setup

Datasets. We select three question-answering benchmarks to examine whether the findings generalize across different question and evidence settings. HotpotQA covers multi-hop encyclopedic question answering (Yang et al., 2018). MS MARCO QA focuses on web question answering for real-world search queries (Bajaj et al., 2018). Natural Questions (NQ) constructs question-answering tasks from real Google queries and Wikipedia pages (Kwiatkowski et al., 2019). Prompts and decoding settings are provided in Appendix A; context construction, evaluation rules, and retained sample counts are detailed in Appendix C.

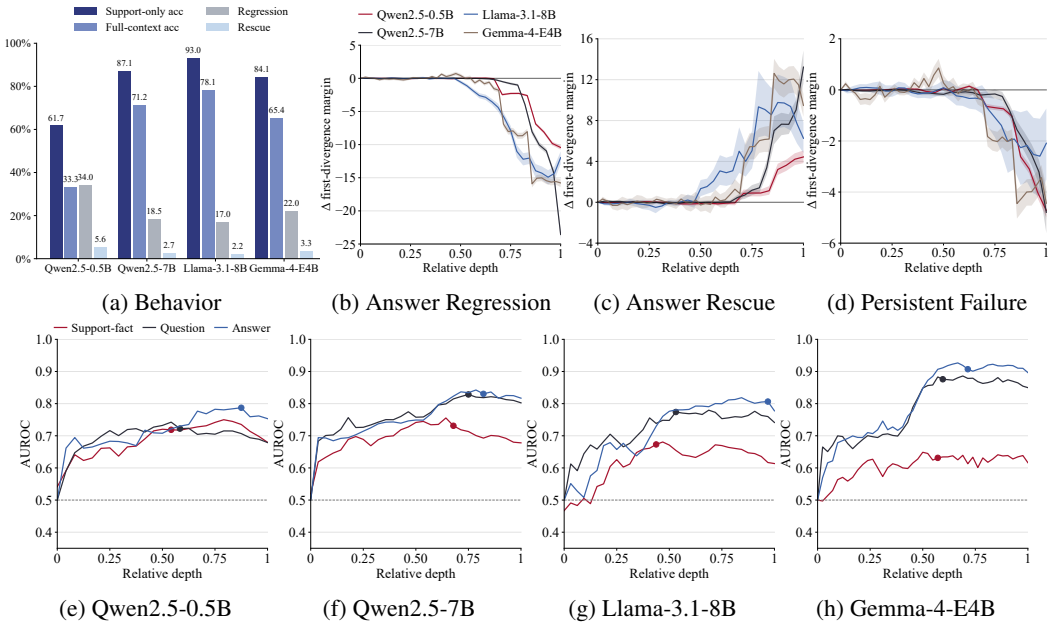
Models. We conduct experiments with four open-weight instruction models: Qwen2.5-0.5B-Instruct, Qwen2.5-7B-Instruct (Qwen et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Gemma-4-E4B-Instruct (Team et al., 2026). Hereafter, we omit "-Instruct" from model names in figures and tables; all experiments use the instruction-tuned variants listed above. This selection enables both scale comparisons within the same model family and architectural comparisons across different families.

Evaluation. The HotpotQA training set is used to fit the probes and select intervention positions, while the validation set is used for independent evaluation. We evaluate the frozen probes and intervention locations on MS MARCO train, dev, and test, and on NQ train and dev. Reporting the official dev and test splits separately checks whether the patterns observed on the target training splits persist on other splits; exception handling is described in Appendix D. We perform 10,000 bootstrap resamples at the independent-sample level to estimate 95% confidence intervals (CIs), and

apply Holm correction within predefined test families. Statistical significance is determined using the corrected criterion ($p < 0.05$). Details are provided in Appendix B.4.

4.2 Divergence in Behavioral Changes and Answer Representations

Obs. 1. Correct evidence can still lead to incorrect answers. When only supporting evidence is provided, the error rates of the four models still range from 7.0% to 38.3%. After switching to the full context, accuracy further decreases by 14.8% to 28.4%, while both Regression and Rescue occur across all models (Figure 2a). These results show that evidence sufficiency and context organization jointly shape the conditions under which answer failures arise. Semantic accuracy and high-confidence behavioral-group counts are reported in Table 4 of Appendix D and Table 3 of Appendix C, respectively.



(a) Behavior (b) Answer Regression (c) Answer Rescue (d) Persistent Failure

(e) Qwen2.5-0.5B (f) Qwen2.5-7B (g) Llama-3.1-8B (h) Gemma-4-E4B

Figure 2: Results on the HotpotQA dev set. (a) reports the semantic accuracy under the two evidence conditions, and the Regression and Rescue rates. (b–d) show the layerwise changes in answer preference for the Regression, Rescue, and Persistent Failure groups; margin changes are measured in nats. Shaded regions indicate 95% CIs. (e–h) show the AUROC of frozen transition-direction probes for the four models. Support-fact, Question, and Answer denote the supporting-fact end, question-end, and pre-generation states, respectively. Circles mark probe evaluation positions selected on the HotpotQA training set; the dashed line at 0.5 indicates chance-level AUROC.

Obs. 2. Context-induced differences in answer scores expand in the middle and late layers. For the Regression and Rescue groups, changes in the First-Divergence Margin are close to zero in the early layers and generally expand in the negative and positive directions, respectively, in the middle and late layers (Figure 2b, Figure 2c). The Persistent Failure group shows a negative change in the late layers (Figure 2d), indicating that additional context can reduce the relative score of the correct candidate even when the correctness label remains unchanged. Thus, layerwise scoring reveals changes in answer preference that are not captured by the final behavioral labels. Whole-Answer Margins show the same overall direction. Their trajectories and final-layer CIs are reported in Figure 3 and Table 7 of Appendix E.

Obs. 3. Regression and Rescue are more distinguishable at the question-end and pre-generation states. At the layers selected for each probe on the training set, the validation-set AUROC at the pre-generation position is higher than that at the supporting-fact final-token position for all four models. The layerwise differences are particularly pronounced for Gemma and Llama (Figure 2(e–h)). These positional differences indicate that information distinguishing Regression from Rescue exhibits different degrees of decodability across token positions as computa-

tion proceeds, providing a basis for subsequent tests of the intervention effects at the question-end and pre-generation states. Complete probe results and the metadata-only baseline are provided in Appendix F.

4.3 From Answer-Score Changes to Free-Generation Flips

Obs. 4. Patching pre-generation states in the middle and late layers can change answer preferences and free-generation responses. For the Regression group, patching the corresponding state from the correct-answer condition of the same question increases the First-Divergence Margin by an average ranging from 7.28 to 14.93 nats across the four models, while reverse patching decreases the margin by an average ranging from 5.75 to 14.14 nats. The Rescue group shows the same direction of effects.

Free-generation interventions at the same positions further change answer correctness. Across the four models, the repair rates range from 54.3% to 79.1% for the Regression group and from 63.1% to 80.8% for the Rescue group. Reverse Damage changes correct answers to incorrect ones at rates ranging from 36.6% to 70.2% and from 51.5% to 84.0%, respectively. After Holm correction, all 16 scoring effects across the two behavioral groups are significantly positive. For each of the four models, all 12 comparisons with the early-layer, question-end, and identity write-back controls show significantly higher flip rates for the main intervention after Holm correction. Scoring effects and their 95% CIs are reported in Table 11 of Appendix G. Free-generation flip rates, valid denominators, and control results are reported in Tables 12 and 13 of Appendix H.

Together, the score and generation effects identify an intervenable answer-selection stage in the middle and late layers; the controls further locate the effect to the patched layer and token position.

4.4 Source Specificity of States and Component Differences

Obs. 5. Overall residual patching exhibits source specificity. On the HotpotQA validation set, the same-question paired source produces a larger directional effect on the First-Divergence Margin than the equal-norm random source under all 16 conditions formed by four models, two behavioral groups, and two intervention directions, Repair and Damage. The mean differences range from 2.51 to 14.87 nat, and all remain significant after Holm correction. With change magnitudes matched, the larger same-question effect indicates source specificity (Table 14 of Appendix I).

Obs. 6. Intervention effects of attention and MLP differ across models. At the selected intervention positions, the paired intervention effects of attention are higher than those of MLP by values ranging from 1.14 to 3.14 nat for Qwen2.5-0.5B and from 1.30 to 1.47 nat for Llama. For Gemma, the MLP effects are higher than the attention effects by values ranging from 1.04 to 2.14 nat. All of these differences remain significant after Holm correction. For Qwen2.5-7B, the attention advantage mainly appears in the Regression group, with differences ranging from 2.40 to 3.41 nat, while the component differences in the Rescue group do not reach significance (Table 15 of Appendix I). The source controls further show that, for Llama, the paired MLP effects are lower than those from the random source by values ranging from 0.46 to 1.17 nat, whereas the paired effects of both the overall residual and attention are higher than those from the random source (Table 14).

4.5 Cross-Dataset Validation under Frozen Settings

We directly apply the probe parameters and intervention positions determined on the HotpotQA training set to the two target datasets, MS MARCO and NQ. Per-model results are provided in Appendices J and K.

Obs. 7. Transition-direction information in pre-generation states transfers across datasets. On MS MARCO and NQ, the AUROC of the frozen probes at the pre-generation position ranges from 0.6288 to 0.9767. Across the four models, all 20 tests are significantly above the random level after Holm correction, and all have higher AUROC than the corresponding probes at the supporting-fact final-token position. For the latter, only 4 tests are significantly above the random level after correction (Appendices J.1 and K.1). This positional difference remains consistent across different question-answering types and data splits, indicating that pre-generation states contain transferable information for distinguishing Regression from Rescue.

Obs. 8. Bidirectional patching at frozen positions changes answer scores and generation outcomes across datasets. On MS MARCO and NQ, all 80 directional effects on the First-Divergence Margin across the four models are significantly positive after Holm correction (Appendices J.2 and K.2). In free generation, both behavioral groups exhibit repairs from incorrect to correct answers and reverse damage from correct to incorrect answers. Repair rates range from 25.4% to 81.8%, and damage rates range from 12.7% to 81.8% (Tables 20 and 29). Among the 240 paired comparisons between the main intervention and the early-layer, question-end, and identity write-back controls, 236 show significantly higher flip rates for the main intervention after Holm correction (Table 6); the corresponding MS MARCO and NQ free-generation results are reported in Appendices J.3 and K.3. The affected model, split, and behavioral-group cells are marked in Table 1.

Component-level controls further show that all 80 paired-source effect differences for the overall residual state on the target splits are significantly positive, while the relative effects of attention and MLP vary across models and behavioral groups (Appendices J.4 and K.4). Table 1 summarizes free-generation repair and damage rates across all six evaluation splits.

Table 1: Free-generation semantic flip rates (%) across dataset. Each cell reports the Repair / Damage rate. † indicates that, compared with the three controls (early-layer, question-end, and identity write-back), at least one comparison does not pass the Holm-corrected significance test; all other values pass all three tests ($p < 0.05$). CIs are provided in Tables 12, 20, and 29; HotpotQA valid sample sizes are in Table 12.

Model	Behavior	HotpotQA	MS MARCO			NQ	
		dev	train	dev	test	train	dev
Qwen2.5-0.5B	Regression	54.3 / 36.6	26.0 / 21.4	25.4 / 20.5	29.4 / 24.9	47.9 / 29.1	51.0 / 31.9
	Rescue	63.1 / 51.5	39.3 / 27.5	30.6 / 30.5	33.3 / 29.3	58.0 / 45.1	70.0 / 47.8
Qwen2.5-7B	Regression	64.2 / 44.1	28.4 / 18.4	30.1 / 19.2	29.9 / 20.4	44.2 / 25.1	44.7 / 18.5
	Rescue	66.0 / 61.7	56.7 / 53.7	62.0 / 55.6	60.7 / 60.4	54.6 / 43.1	63.6 / 47.2
Llama-3.1-8B	Regression	79.1 / 70.2	38.0 / 41.0	36.3 / 37.7	37.5 / 39.9	62.4 / 62.7	59.6 / 55.4
	Rescue	80.8 / 84.0	63.7 / 63.1	56.2 / 52.1	56.5 / 56.5	76.8 / 66.9	81.8 / 81.8
Gemma-4-E4B	Regression	71.7 / 50.4	26.0 / 15.3	25.6 / 15.2	25.8 / 12.7	52.8 / 28.3	54.8 / 18.6
	Rescue	65.2 / 60.9	52.0 / 44.6	59.1 / 52.4	56.0 / 44.0†	72.6 / 53.3	80.0 / 30.0†

5 Conclusion

This work asks why a model can answer incorrectly despite receiving sufficient and correct evidence, even when no explicit competing parametric answer is observed. Across four models and three question-answering benchmarks, answer-related information becomes more distinguishable near generation, and bidirectional patching shows that pre-generation states can affect both candidate scores and complete answers. Transferring the state from the correct-answer condition of the same question repairs some incorrect responses; reversing the transfer damages some correct responses without changing the receiving input. These effects persist when probes and intervention positions selected on HotpotQA are applied to MS MARCO and NQ, while the relative effects of attention and MLP outputs vary across models. Together, the results locate an intervenable part of evidence utilization in the answer-selection process and suggest that improving RAG reliability requires attention to how models use retrieved evidence after it enters the context.

ETHICS STATEMENT

This study uses publicly available question-answering datasets and open-weight models. It does not involve human participants, animal experiments, or environmentally sensitive materials. We identified no significant ethical risks associated with the conduct of this study and verified the accuracy of the reported results.

REPRODUCIBILITY STATEMENT

We specify the behavioral grouping, probe training, intervention-location selection, and bidirectional patching procedures in the main text. The appendices provide the prompts and decoding settings, dataset construction and sample counts, mathematical definitions, evaluation criteria, control interventions, statistical procedures, and computational resources. We will release the experimental code and processed sample manifests upon acceptance to support independent reproduction.

AI USE STATEMENT

We used large language models to check the manuscript's formatting and refine its language. We carefully reviewed and verified all LLM-generated content for accuracy and completeness. We take responsibility for the final content of this work.

REFERENCES

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in RAG. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.892. URL https://aclanthology.org/2025.acl-long.892/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

Samyadeep Basu, Vlad I Morariu, Ryan A. Rossi, Nanxuan Zhao, Zichao Wang, Soheil Feizi, and Varun Manjunatha. On mechanistic circuits for extractive question-answering. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VvSWiNIuPL.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. Conflict-aware soft prompting for retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26981–26995, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1371. URL https://aclanthology.org/2025.emnlp-main.1371/.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. Contextcite: Attributing model generation to context. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95764–95807. Curran Associates, Inc., 2024. doi: 10.52202/079017-3035. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/adbea136219b64db96a9941e4249a857-Paper-Conference.pdf.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 719–729, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657834. URL https://doi.org/10.1145/3626772.3657834.

Mehrdad Farahani and Richard Johansson. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16966–16977, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.943. URL https://aclanthology.org/2024.emnlp-main.943/.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Christian Giannetti, Giovanni Trappolini, Nicola Tonellotto, Fabrizio Silvestri, and Pietro Lio. The mechanics of interference: Defusing distractors in RAG via sparse autoencoder interventions. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11999–12007, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.583. URL https://aclanthology.org/2026.findings-acl.583/.

Sachin Goyal, Christina Baek, Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 68067–68099, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/aa27ac7aca4e462da1439b43ceebc04c-Paper-Conference.pdf.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,

Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan

Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668. Curran Associates, Inc., 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3927bbdcf0e8d1fa8aa23c26f358a281-Paper-Conference.pdf.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, and Bing Qin. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16896–16913, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.826. URL https://aclanthology.org/2025.acl-long.826/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 20310–20334, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/33dffa2e3d2ab74a783d1a8c292f66d9-Paper-Conference.pdf.

Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. The atlas of in-context learning: How attention heads shape in-context retrieval augmentation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 118164–118208. Curran Associates, Inc., 2025. doi: 10.52202/085713-3940. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ab49f5e61eadbe979373c133232a96a0-Paper-Conference.pdf.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. CoCoA: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6835–6855, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.348. URL https://aclanthology.org/2025.emnlp-main.348/.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34074–34104. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25c.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Ruizhe Li, CHEN CHEN, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. Attributing response to context: A jensen–shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 146805–146837, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ed67dff7cb96e7e86c4d91c0d5db49bb-Paper-Conference.pdf.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/

068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with fakepedia. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6828–6844, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.369. URL https://aclanthology.org/2024.acl-long.369/.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality $\neq$ invariance: Function and concept vectors in llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 83393–83417, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/86b3697c4eb7792c951831636bfdacd5-Paper-Conference.pdf.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a712d461e57201efe35d429a6f1731c1-Paper-Conference.pdf.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.347. URL https://aclanthology.org/2024.emnlp-main.347/.

Qwen, 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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21999–22027. Curran Associates, Inc., 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/27245589131d17368cccdfa990cbf16e-Paper-Datasets_and_Benchmarks_Track.pdf.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 50250–50279, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7daf60e805e596c3bd1e843e72ea5560-Paper-Conference.pdf.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux,

Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ae163cb8788970e53b4fd9578141139-Paper-Conference.pdf.

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

Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazedeh. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM Companion '26, pp. 132–139, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723582. doi: 10.1145/3779211.3793209. URL https://doi.org/10.1145/3779211.3793209.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models.

In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

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, 2022. URL https://arxiv.org/abs/2211.00593.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how lms answer multiple choice questions. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 78293–78318, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c248154176c08147e82c0b30961604f7-Paper-Conference.pdf.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5019–5039, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.250. URL https://aclanthology.org/2025.acl-long.250/.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9b77f07301b1ef1fe810aae96c12cb7b-Paper-Conference.pdf.

Haizhou Xia. Diagnosing evidence utilization in long-context and retrieval-augmented language models under matched evidence conditions, 2026. URL https://arxiv.org/abs/2606.06758.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 112549–112577, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b6f6bfbd260fbf2f5acb0a1d6439ca0e-Paper-Conference.pdf.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Samuel Yeh and Sharon Li. How retrieved context shapes internal representations in RAG. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15446–15468, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.758. URL https://aclanthology.org/2026.findings-acl.758/.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024a. URL https://arxiv.org/abs/2403.10131.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 60755–60775. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1943. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ffdbbe354893979367f93e2121e37dd-Paper-Conference.pdf.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 102978–103012. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3443. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/94f7c80260eea8d52a5c0ec7b1637c04-Paper-Conference.pdf.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77885–77910. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/zhao25ad.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A Prompts and Decoding Settings

A.1 Experimental Prompts and Decoding Configuration

Closed-book probing for all four models on the three benchmark datasets is performed through an OpenAI-compatible interface with 10 repeated requests, using `temperature=1` and

`max_tokens=1024`. A sample is retained only if all requests succeed and every returned response is `UNKNOWN`. The closed-book probing prompt is shown in the Prompt 1.

Prompt 1: Closed-Book Prompt

You are being tested for knowledge available from your model parameters.

Answer the user's question using only factual knowledge you already know. Do not use external context. Do not guess.

Output rules:

- If you know the answer with confidence, output only the shortest answer phrase.

- If you do not know the answer with confidence, output exactly "UNKNOWN".

- Do not explain.

- Do not output labels, JSON, markdown, or extra punctuation.

Question answering under both evidence conditions uses the prompt shown in the Prompt 2, with the corresponding context and question inserted into the indicated fields.

Prompt 2: Evidence-Based QA Prompt

Answer the question using only the context supplied. The context contains sufficient information to answer the question.

Output rules:

- Output only the shortest complete answer.

- Do not explain your reasoning.

- Do not output labels, JSON, markdown, or extra commentary.

Context: context

Question: question

Answer:

Prompt 3: Semantic Judge Prompt

You are a strict context-grounded answer evaluator.

Determine whether the Model Answer fully and correctly answers the Question. Use the Gold Answer as the expected answer and the Reference Supporting Facts as authoritative evidence for resolving meaning, identity, aliases, dates, numbers, units, and answer scope.

Judge EQUIVALENT only when:

- The model answer identifies the same required entity, value, date, number, or complete set as the gold answer; or

- Different wording or additional precision is explicitly supported by the supporting facts and does not change the answer required by the question.

- Differences are limited to harmless aliases, abbreviations, formatting, capitalization, articles, punctuation, implied units, or non-contradictory answer framing.

Judge NOT_EQUIVALENT when:

- The model answer is only partially correct.

- It gives the wrong entity, value, date, number, scope, or set.

- It omits any required entity or value.

- It adds an unsupported or contradictory entity, value, or claim.

- It merely mentions a relevant entity without answering the question.

Do not solve the question using outside knowledge. Do not explain your judgment. Output exactly EQUIVALENT or NOT_EQUIVALENT.

Question: question

Gold Answer: gold_answer

Reference Supporting Facts: supporting_facts

Model Answer: model_answer

Context-based question answering is also performed through the OpenAI-compatible interface, with `temperature=1` and `max_tokens=2048`. Layer-wise scoring and state patching use the same prompts described above. Scoring is performed with teacher forcing without answer sampling. Free-form generation uses the same prompt as input and applies greedy decoding from the beginning of the answer; the generation-length limit for each model is provided in Appendix D. State patching is performed once during prompt prefill.

Correctness under the two context conditions and correctness of free-form answers after state patching are both evaluated by first computing normalized EM. For responses with EM=0, Qwen3.8-27B is used to determine whether the response is semantically equivalent to the gold answer. The judge uses `temperature=1` and `max_tokens=1024`. Prompt 3 shows the HotpotQA version. For MS MARCO and NQ, the prompt provides all official gold answers and accepts any one of them; the remaining decision criteria are unchanged.

B Mathematical Details

B.1 Residual Stream and Layer-Wise Output Readout

In the main text, $h_{\ell,j}$ denotes the state at residual location ℓ and input-token position j. For layer-wise readout, this state is passed through the model's final normalization operation Norm and output projection W_U to obtain the next-token probability distribution:

$$p_\ell(v \mid x_{\le j}) = \mathrm{softmax}\left(W_U \,\mathrm{Norm}(h_{\ell,j})\right)_v .$$

Here, v is a vocabulary token, and j denotes the position of the last token currently included in the input. Consider a candidate answer $a = (a_1, \ldots, a_T)$. To compute the probability of its t-th token a_t, the model input contains the question, the evidence, and the preceding $t-1$ tokens of the candidate answer, $a_{<t}$. Under this shared input, we read out the probability assigned to a_t. Teacher forcing is applied token by token, and the log-probabilities are averaged:

$$S_\ell(a \mid x) = \frac{1}{T}\sum_{t=1}^{T} \log p_\ell(a_t \mid x, a_{<t}).$$

For the correct candidate g and incorrect candidate c, the Whole-Answer Margin is $M_\ell(x; g, c) = S_\ell(g \mid x) - S_\ell(c \mid x)$. Using the mean log-probability reduces the direct effect of answer length on the total score. The First-Divergence Margin F_ℓ instead compares the next token after the shared token prefix of the two candidates, so both probabilities are computed under the same input prefix. If the token sequence of one candidate is exactly a prefix of the other, the model's end-of-sequence token is used at the first divergence position to represent the termination of the shorter answer. The same candidate pair is used under both context conditions, after which $\Delta^M_{i,\ell}$ and $\Delta^F_{i,\ell}$ are computed as defined in the main text.

B.2 Probe Objectives

The correctness probes in the main text take the standardized state $\tilde{h}_i = \mathrm{Std}(h^u_{i,\ell,p})$ as input, where $u \in \{S, C\}$ and p denotes one of the three semantic positions. The correctness label $z^u_i \in 0, 1$ is transformed into $t_i = 2z^u_i - 1 \in \{-1, +1\}$. Given inverse regularization coefficient C, the probe fits w, b by minimizing

$$\min_{w,b} \sum_i \omega_{t_i} \log\left(1 + \exp[-t_i(w^\top \tilde{h}_i + b)]\right) + \frac{1}{2C}\|w\|_2^2.$$

Here, ω_{t_i} is the class-balancing weight. A smaller C corresponds to stronger L_2 regularization, and the intercept b is not regularized. The transition-direction probe uses the same objective. We first compute the state difference between the two context conditions, $d_{i,\ell,p} = h^C_{i,\ell,p} - h^S_{i,\ell,p}$, and then use the standardized difference $\mathrm{Std}(d_{i,\ell,p})$ as input to the probe. Rescue is treated as the positive class and Regression as the negative class. Each of the nine probe groups is trained separately at every residual location, with AUROC as the primary evaluation metric.

For each probe group, we use five-fold stratified cross-validation on the HotpotQA training set to select C from $10^{-7}, 10^{-6}, \ldots, 10^{2}$ according to the highest mean AUROC across folds. Within each cross-validation fold, the mean and standard deviation used for feature standardization are computed only from the training split and then applied unchanged to the validation split. After selecting C, these statistics are recomputed on the full HotpotQA training set and the final probe is trained. All settings are then kept fixed when evaluating on HotpotQA dev, MS MARCO, and NQ.

The Metadata-Only Baseline is trained on the HotpotQA training set and evaluated on its independent validation set. Its inputs include one-hot encodings of question type and difficulty, together with the token lengths of the gold answer, supporting evidence, and full context, and the number of supporting facts. All four numerical features are transformed using $\log(1 + x)$. Separate classifiers are trained for answer correctness under the two context conditions and for response-transition direction, using the same cross-validation and standardization procedures as the representation probes.

B.3 Normalized Effects of State Patching

$F(x^+)$ and $F(x^-)$ denote the unpatched First-Divergence Margins under the correct-answer and incorrect-answer conditions, respectively, and $G = F(x^+) - F(x^-)$ denotes the score difference between these two baselines. Repair uses x^- as the receiving input, whereas damage uses x^+ as the

receiving input. Dividing the raw directional effects defined in Section 3.3 by G gives the normalized effects:

$$\rho^{\text{repair}} = \frac{F^{\text{repair}} - F(x^-)}{G}, \qquad \rho^{\text{damage}} = \frac{F(x^+) - F^{\text{damage}}}{G}.$$

ρ is the signed ratio between the intervention-induced score change and the baseline score difference G. For example, $\rho = 0$ indicates no change in the score of the receiving condition. For repair, $\rho = 1$ indicates that the patched score reaches the correct-answer baseline $F(x^+)$; for damage, $\rho = 1$ indicates that the patched score reaches the incorrect-answer baseline $F(x^-)$. Whether repair or damage changes the score in the expected direction is determined by the sign of the raw directional effects E^{repair} and E^{damage} defined in Section 3.3.

B.4 Bootstrap Resampling and Multiple Testing

Uncertainty in intervention effects is estimated by resampling independent samples. Given effects $e_1, \ldots, e_n$ from n samples, each bootstrap replicate draws n samples with replacement and computes their mean. This procedure is repeated 10,000 times, and the 2.5th and 97.5th percentiles of the bootstrap means form the 95% CIs. CIs for probe AUROC are estimated using stratified bootstrap resampling by the true class labels so that both classes are represented in every resample. The random seed is 2026.

We choose the statistical test according to the quantity being evaluated. To test whether probe AUROC exceeds chance level, we randomly permute the sample labels. To compare the AUROC of two probes evaluated on the same samples, we randomly swap their prediction scores within each sample. State-patching and component analyses use sign-flip tests on sample-level effects or effect differences. In the free-form generation experiments, we compare whether the expected correctness flip occurs for the same sample under the primary intervention and a control intervention. HotpotQA uses the exact McNemar test, whereas MS MARCO and NQ use sign-flip tests on paired outcomes. Permutation and sign-flip tests are each repeated 10,000 times.

To control false positives from multiple comparisons, we apply Holm correction to p-values within each predefined family of tests and regard a Holm-corrected $p < 0.05$ as statistically significant. Effect sizes and their uncertainty are reported using means and 95% CIs. With 10,000 randomization trials, the smallest raw p-value that can be reported is approximately $1/10001$. If all four tests within the same family reach this lower bound, their Holm-corrected values can all appear as approximately 0.0004; this does not imply that the four effect sizes are equal.

For presentation, percentages are rounded to one decimal place in the main text and two decimal places in the appendix. AUROC values are reported to four decimal places, and effects measured in nats to two decimal places. Rounding is applied after all statistical calculations.

C Dataset Construction, Context Assembly, and Sample Accounting

C.1 Context Construction and Correctness Evaluation

For each question, we compare a supporting-evidence-only input with a full-context input. The former contains the evidence units designated as supporting; the latter preserves all candidate units in their source order and includes the same supporting evidence. For HotpotQA, supporting sentences are selected using the sentence-level annotations, grouped by title in first-occurrence order, and kept in their original within-title order. MS MARCO uses passages annotated as relevant as supporting evidence, and NQ uses the annotated top-level Wikipedia paragraph; their construction rules are detailed in Appendices C.2 and C.3.

Correctness evaluation first applies normalized EM, which lowercases the answer, removes punctuation and articles, and collapses whitespace. HotpotQA handles `yes`, `no`, and `noanswer` according to its official evaluation rules. A response with EM=1 is labeled correct directly; only an EM=0 response is sent to the semantic judge, which checks equivalence to an acceptable gold answer using

the question and supporting evidence. For MS MARCO, any official gold answer is acceptable. For NQ, multiple short-answer spans that jointly form an answer are combined in document order, with equivalent joining forms retained; any complete form is acceptable.

The sample accounting follows five stages: mapping the source records, retaining questions answered as `UNKNOWN` in all 10 closed-book runs, validating the supporting evidence, finalizing correctness labels under both context conditions, and applying the high-confidence criteria in Section 3.1. Table 2 reports the counts after each stage. Table 3 breaks down the high-confidence samples by behavior. Regression and Rescue samples form the candidate set for bidirectional state patching.

The evaluated splits are HotpotQA train and dev-distractor, MS MARCO train, dev, and test, and NQ train and dev. HotpotQA train is used for probe fitting and intervention-location selection; its dev split provides independent validation. All MS MARCO and NQ splits are evaluated with these settings frozen. The closed-book screen retains 18,714 of 90,447 HotpotQA training questions for Llama, compared with 66,535–66,882 for the two Qwen models (Table 2). No request errors were recorded in Llama's ten probing runs; the lower retention reflects fewer questions meeting the requirement that every response be `UNKNOWN`.

C.2 MS MARCO Construction and Filtering

We evaluate the train, validation (denoted dev), and test splits of MS MARCO QA v1.1. For each question, the supporting-evidence-only input contains passages annotated as relevant, while the full context contains all candidate passages in their original order. A record is retained when it has at least one valid gold answer, one relevant passage, and one non-relevant passage. When multiple official gold answers are available, any one is accepted for correctness evaluation, and all are supplied to the semantic judge.

The raw and mapped sample counts for all three splits are reported in Table 2. On the train split, 79,311 of 82,326 records remain after mapping; exclusions result from missing valid answers (2,183), missing relevant passages (518), missing non-relevant passages (4), or conflicting relevance labels for duplicate passages (310). All splits are evaluated with probe parameters and intervention locations fixed on HotpotQA.

C.3 Natural Questions Construction and Filtering

We evaluate the train and dev splits of NQ v1.0. For each record, we use the first human annotation and construct supporting evidence from its top-level Wikipedia paragraph; the full context contains all non-empty candidate paragraphs in their original order. Multiple short-answer spans are combined in document order, with equivalent complete forms retained. For yes/no questions, the gold answer is `yes` or `no`. Any valid form is accepted for correctness evaluation, and all valid answers are provided to the semantic judge.

The raw and mapped counts for both splits are reported in Table 2. On train, 83,472 of 307,373 records are retained, comprising 80,509 short-answer and 2,963 yes/no questions. The principal exclusion reasons are a missing valid long answer (155,225), a missing usable short or yes/no answer (27,014), a table-form long answer (28,850), and a long answer that is not a top-level paragraph (7,698); 283 additional records exceed the 32,768-token prompt limit. The dev mapping applies the same evidence criteria with a 30,000-token prompt limit. Both splits use probe parameters and intervention locations fixed on HotpotQA.

D Evaluation Protocol, Sample Integrity, and Computational Resources

The HotpotQA train split is used to fit the probes and select candidate and fixed intervention locations, while its dev split provides independent validation. MS MARCO train/dev/test and NQ train/dev are evaluated under these frozen settings. Across datasets, we compare supporting-evidence-only and full-context inputs and apply the same behavioral grouping and mechanistic analyses; probe parameters, network locations, and primary evaluation metrics are not reselected on the target splits. Statistical analyses use independent samples as the resampling unit with 10,000

Table 2: Sample counts across mapping and behavioral screening. Raw/Map gives the source and mapped record counts, respectively. CB, EV, 2C, and HC denote the counts after the 10-run closed-book screen, evidence validation, final correctness labeling in both context conditions, and high-confidence screening, respectively. Correctness is determined by normalized EM; the semantic judge is used only for responses with EM=0.

Dataset	Model	Raw / Map	CB	EV	2C	HC
HotpotQA (train)	Qwen2.5-7B	90,447/90,447	66,882	66,869	66,869	56,353
	Qwen2.5-0.5B	90,447/90,447	66,535	66,519	66,519	49,226
	Llama-3.1-8B	90,447/90,447	18,714	18,709	18,709	15,401
	Gemma-4-E4B	90,447/90,447	52,269	52,255	52,255	42,183
HotpotQA (dev)	Qwen2.5-7B	7,405/7,405	5,840	5,839	5,839	4,292
	Qwen2.5-0.5B	7,405/7,405	5,683	5,683	5,683	3,875
	Llama-3.1-8B	7,405/7,405	1,661	1,661	1,661	1,143
	Gemma-4-E4B	7,405/7,405	4,607	4,607	4,607	3,167
MS MARCO (train)	Qwen2.5-7B	82,326/79,311	56,258	56,258	56,258	31,675
	Qwen2.5-0.5B	82,326/79,311	42,267	42,267	42,267	24,535
	Llama-3.1-8B	82,326/79,311	14,236	14,236	14,236	10,209
	Gemma-4-E4B	82,326/79,311	34,154	34,154	34,154	18,530
MS MARCO (dev)	Qwen2.5-7B	10,047/9,663	6,844	6,844	6,844	3,843
	Qwen2.5-0.5B	10,047/9,663	5,136	5,136	5,136	2,972
	Llama-3.1-8B	10,047/9,663	1,710	1,710	1,710	1,230
	Gemma-4-E4B	10,047/9,663	4,180	4,180	4,180	2,207
MS MARCO (test)	Qwen2.5-7B	9,650/9,295	6,579	6,579	6,579	3,638
	Qwen2.5-0.5B	9,650/9,295	4,863	4,863	4,863	2,813
	Llama-3.1-8B	9,650/9,295	1,613	1,613	1,613	1,171
	Gemma-4-E4B	9,650/9,295	3,919	3,919	3,919	2,074
NQ (train)	Qwen2.5-7B	307,373/83,472	50,324	50,324	50,266	33,354
	Qwen2.5-0.5B	307,373/83,472	39,533	39,533	39,499	23,719
	Llama-3.1-8B	307,373/83,472	15,980	15,980	15,977	11,404
	Gemma-4-E4B	307,373/83,472	38,822	38,822	38,791	22,522
NQ (dev)	Qwen2.5-7B	7,830/2,385	1,437	1,437	1,437	1,015
	Qwen2.5-0.5B	7,830/2,385	1,057	1,057	1,057	620
	Llama-3.1-8B	7,830/2,385	413	413	413	335
	Gemma-4-E4B	7,830/2,385	1,115	1,115	1,115	665

bootstrap iterations, and Holm correction is applied within predefined test families at a significance level of 0.05. Detailed definitions are provided in Appendix B.4.

Table 4 reports answer accuracy under the two evidence conditions on HotpotQA, providing an overall behavioral context for the subsequent grouping based on response changes. Accuracy is computed over responses for which the corresponding correctness evaluation is completed and includes answers confirmed as correct by either EM or the semantic judge.

Table 3: Behavioral group counts in the high-confidence subset. Regression denotes a correct supporting-evidence-only answer and an incorrect full-context answer; Rescue denotes the reverse. Persistent Failure and Robust Control denote incorrect and correct answers, respectively, under both conditions. Each row sums to the HC count in Table 2; Regression and Rescue together form the candidate set for bidirectional state patching.

Dataset	Model	Regression	Rescue	Persistent Failure	Robust Control
HotpotQA (train)	Qwen2.5-7B	8,816	1,048	3,979	42,510
	Qwen2.5-0.5B	18,175	1,887	16,680	12,484
	Llama-3.1-8B	2,214	233	606	12,348
	Gemma-4-E4B	8,383	1,023	4,418	28,359
HotpotQA (dev)	Qwen2.5-7B	787	111	599	2,795
	Qwen2.5-0.5B	1,242	112	1,858	663
	Llama-3.1-8B	192	26	80	845
	Gemma-4-E4B	718	93	578	1,778
MS MARCO (train)	Qwen2.5-7B	4,690	826	19,579	6,580
	Qwen2.5-0.5B	3,795	523	17,569	2,648
	Llama-3.1-8B	1,653	419	5,611	2,526
	Gemma-4-E4B	3,370	206	12,674	2,280
MS MARCO (dev)	Qwen2.5-7B	587	96	2,394	766
	Qwen2.5-0.5B	466	66	2,123	317
	Llama-3.1-8B	192	50	666	322
	Gemma-4-E4B	381	22	1,529	275
MS MARCO (test)	Qwen2.5-7B	515	93	2,245	785
	Qwen2.5-0.5B	426	48	2,035	304
	Llama-3.1-8B	185	47	667	272
	Gemma-4-E4B	334	25	1,450	265
NQ (train)	Qwen2.5-7B	5,483	946	7,759	19,166
	Qwen2.5-0.5B	5,425	668	13,726	3,900
	Llama-3.1-8B	1,745	423	2,547	6,689
	Gemma-4-E4B	3,488	782	6,586	11,666
NQ (dev)	Qwen2.5-7B	201	37	123	654
	Qwen2.5-0.5B	171	28	298	123
	Llama-3.1-8B	57	11	33	234
	Gemma-4-E4B	127	22	108	408

The sample sets used in different mechanistic analyses are not identical. Layer-wise scoring and probe feature extraction cover the high-confidence subset, whereas the three types of state-patching experiments draw samples separately from the Regression and Rescue groups. Table 5 reports the completion status of each analysis across all evaluated dataset splits, clarifying the sample counts used in downstream statistical analyses.

Table 4: Semantic accuracy (%) under the two evidence conditions on HotpotQA. The train and dev columns correspond to the training set and independent validation set, respectively. Each cell reports Supporting Evidence Only / Full Context.

Model	Supporting Evidence Only / Full Context	
	train	dev
Qwen2.5-7B	91.87 / 78.41	87.05 / 71.25
Qwen2.5-0.5B	69.24 / 39.71	61.75 / 33.33
Llama-3.1-8B	94.92 / 82.34	92.96 / 78.15
Gemma-4-E4B	88.72 / 72.13	84.11 / 65.44

Unsuccessful records occur only for Gemma on NQ, in both the train and dev splits; Table 5 gives the counts for each analysis. Each analysis uses only its successfully completed records. Appendix K discusses the NQ memory failures. Intervention-specific denominators for HotpotQA are reported in Table 12 of Appendix H.

The primary free-generation intervention is compared with early-layer, question-end, and identity write-back controls on every target split. Table 6 summarizes the observed control flip rates and the paired comparison results.

Across the five target splits, 236 of 240 paired comparisons are significant after Holm correction. The four exceptions are Rescue/Damage versus the early-layer control for Gemma on MS MARCO test ($p = 0.0654$), and Rescue/Damage versus each of the three controls for Gemma on NQ dev ($p = 0.0882$ for each comparison).

The maximum length for free-form generation is 1,024 tokens. On MS MARCO, Qwen2.5-0.5B and Llama reach this limit in 10 and 2 generations, respectively. On NQ, Qwen2.5-7B, Qwen2.5-0.5B, and Llama reach the limit in 1, 66, and 5 generations, respectively, while no such case occurs for Gemma. Length-limited responses are still evaluated for semantic correctness based on the text actually generated.

The GPU server used in this study is equipped with eight NVIDIA A800-SXM4-80GB GPUs and CUDA 13.0. This configuration does not represent the minimum hardware requirement for reproduction. Using fewer GPUs increases processing time, and long-context samples may fail to complete under more limited GPU memory. For the multi-GPU batch processing used in our experiments, GPU-hours are estimated as the number of GPUs running in parallel multiplied by the overall wall-clock time. The resulting costs are approximately 147 GPU-hours for HotpotQA, 77 GPU-hours for MS MARCO, and 309 GPU-hours for NQ, for a total of approximately 533 GPU-hours.

Linear probes are fitted on a compute node with dual Intel Xeon Max 9468 CPUs and 420 GB of memory. On HotpotQA, probe training takes 9 hours for Qwen2.5-0.5B, 23 hours for Qwen2.5-7B, 22 hours for Llama-3.1-8B, and 25 hours for Gemma-4-E4B, for a total of 79 hours.

E Layerwise Context Effects on Answer-Score Margins

For each sample, an answer-score margin is the correct candidate's score minus the incorrect candidate's score. The whole-answer margin uses the candidates' mean token log-probabilities, while the first-divergence margin compares their log-probabilities at the first differing token. Table 7 reports the context effect, defined as the full-context margin minus the supporting-evidence-only margin, at the final residual location on the HotpotQA dev set. Entries are means with 95% bootstrap CIs over independent samples. Positive effects indicate a shift toward the correct candidate; negative effects indicate a shift toward the incorrect candidate.

Table 5: Numbers of successful and failed samples for each mechanistic analysis across the evaluated dataset splits. Each cell reports Success / Error. The same sample may appear in multiple analyses. RSP: Residual Stream Patching; FGP: Free-Generation Patching; CP: Component Patching.

Dataset	Model	Layer-Wise Scoring	Probe Features	RSP	FGP	CP
HotpotQA (train)	Qwen2.5-7B	56,353 / 0	56,353 / 0	9,864 / 0	9,864 / 0	9,864 / 0
	Qwen2.5-0.5B	49,226 / 0	49,226 / 0	20,062 / 0	20,062 / 0	20,062 / 0
	Llama-3.1-8B	15,401 / 0	15,401 / 0	2,447 / 0	2,447 / 0	2,447 / 0
	Gemma-4-E4B	42,183 / 0	42,183 / 0	9,406 / 0	9,406 / 0	9,406 / 0
HotpotQA (dev)	Qwen2.5-7B	4,292 / 0	4,292 / 0	898 / 0	898 / 0	898 / 0
	Qwen2.5-0.5B	3,875 / 0	3,875 / 0	1,354 / 0	1,354 / 0	1,354 / 0
	Llama-3.1-8B	1,143 / 0	1,143 / 0	218 / 0	218 / 0	218 / 0
	Gemma-4-E4B	3,167 / 0	3,167 / 0	811 / 0	811 / 0	811 / 0
MS MARCO (train)	Qwen2.5-7B	31,675 / 0	31,675 / 0	5,516 / 0	5,516 / 0	5,516 / 0
	Qwen2.5-0.5B	24,535 / 0	24,535 / 0	4,318 / 0	4,318 / 0	4,318 / 0
	Llama-3.1-8B	10,209 / 0	10,209 / 0	2,072 / 0	2,072 / 0	2,072 / 0
	Gemma-4-E4B	18,530 / 0	18,530 / 0	3,576 / 0	3,576 / 0	3,576 / 0
MS MARCO (dev)	Qwen2.5-7B	3,843 / 0	3,843 / 0	683 / 0	683 / 0	683 / 0
	Qwen2.5-0.5B	2,972 / 0	2,972 / 0	532 / 0	532 / 0	532 / 0
	Llama-3.1-8B	1,230 / 0	1,230 / 0	242 / 0	242 / 0	242 / 0
	Gemma-4-E4B	2,207 / 0	2,207 / 0	403 / 0	403 / 0	403 / 0
MS MARCO (test)	Qwen2.5-7B	3,638 / 0	3,638 / 0	608 / 0	608 / 0	608 / 0
	Qwen2.5-0.5B	2,813 / 0	2,813 / 0	474 / 0	474 / 0	474 / 0
	Llama-3.1-8B	1,171 / 0	1,171 / 0	232 / 0	232 / 0	232 / 0
	Gemma-4-E4B	2,074 / 0	2,074 / 0	359 / 0	359 / 0	359 / 0
NQ (train)	Qwen2.5-7B	33,354 / 0	33,354 / 0	6,429 / 0	6,429 / 0	6,429 / 0
	Qwen2.5-0.5B	23,719 / 0	23,719 / 0	6,093 / 0	6,093 / 0	6,093 / 0
	Llama-3.1-8B	11,404 / 0	11,404 / 0	2,168 / 0	2,168 / 0	2,168 / 0
	Gemma-4-E4B	22,470 / 52	22,474 / 48	4,050 / 220	4,016 / 254	4,033 / 237
NQ (dev)	Qwen2.5-7B	1,015 / 0	1,015 / 0	238 / 0	238 / 0	238 / 0
	Qwen2.5-0.5B	620 / 0	620 / 0	199 / 0	199 / 0	199 / 0
	Llama-3.1-8B	335 / 0	335 / 0	68 / 0	68 / 0	68 / 0
	Gemma-4-E4B	664 / 1	664 / 1	139 / 10	139 / 10	139 / 10

Table 7 summarizes the answer-score changes induced by context expansion at the final residual location. Figure 3 shows how the whole-answer margin change develops across relative network depth for Regression, Rescue, and Persistent Failure samples.

Table 6: Free-generation control outcomes on the target datasets. Early-layer and question-end columns report the maximum semantic flip rate (%) across Regression/Repair, Regression/Damage, Rescue/Repair, and Rescue/Damage for each model and split. Identity write-back yields 0% in every condition. The last column counts comparisons in which the primary intervention has a significantly higher paired flip rate than a control after Holm correction; each model and split has 12 comparisons.

Dataset	Split	Model	Early-layer max	Question-end max	Significant / 12
MS MARCO	train	Qwen2.5-0.5B	4.21%	0.77%	12/12
		Qwen2.5-7B	1.37%	0.89%	12/12
		Llama-3.1-8B	4.41%	0.25%	12/12
		Gemma-4-E4B	7.92%	0.00%	12/12
	dev	Qwen2.5-0.5B	6.12%	3.39%	12/12
		Qwen2.5-7B	1.24%	1.11%	12/12
		Llama-3.1-8B	4.17%	0.00%	12/12
		Gemma-4-E4B	4.55%	0.00%	12/12
	test	Qwen2.5-0.5B	2.78%	2.44%	12/12
		Qwen2.5-7B	2.39%	0.20%	12/12
		Llama-3.1-8B	2.17%	0.00%	12/12
		Gemma-4-E4B	20.00%	0.00%	11/12
NQ	train	Qwen2.5-0.5B	4.87%	0.72%	12/12
		Qwen2.5-7B	2.24%	0.67%	12/12
		Llama-3.1-8B	4.15%	0.24%	12/12
		Gemma-4-E4B	2.59%	0.00%	12/12
	dev	Qwen2.5-0.5B	4.35%	1.23%	12/12
		Qwen2.5-7B	6.06%	3.03%	12/12
		Llama-3.1-8B	5.26%	0.00%	12/12
		Gemma-4-E4B	5.22%	0.00%	9/12

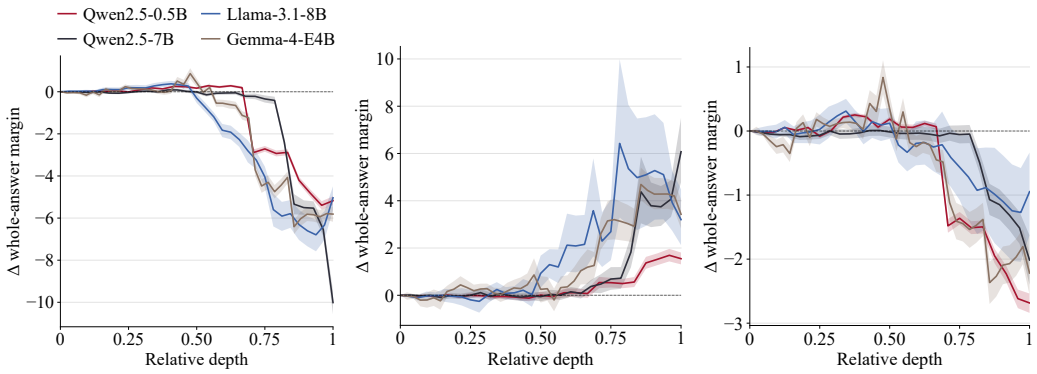
F Probe Performance and Training-Selected Residual Locations

For each model, we train six types of correctness probes and three types of transition-direction probes. The correctness probes predict whether the model answers correctly under either the supporting-evidence-only or full-context condition, using hidden states at the final tokens of the supporting facts, the final token of the question, or immediately before answer generation. The transition-direction probes use the state difference between the two context conditions at the corresponding position to distinguish Regression from Rescue. For each probe type, the residual location is selected according to five-fold cross-validation AUROC on the HotpotQA training set. The probe and selected location are then kept fixed for evaluation on the validation set.

Table 8 combines validation-set AUROC with the corresponding residual location selected on the training set. Each entry gives AUROC followed by the residual-location index in parentheses. Supporting-Fact End, Question End, and Pre-Generation refer to the mean state at the final token of each supporting fact, the final question-token state, and the pre-generation state, respectively. Across all four models and both correctness conditions, probes at the pre-generation state achieve higher AUROC than probes at the supporting-fact end state; for Gemma under full context, the values are 0.8815 and 0.6122. Figure 4 shows the layerwise trajectories of the six correctness-probe types. Their position-dependent differences emerge through network depth, while training and dev-set tra-

Table 7: Context-induced changes in answer-score margins at the final residual location on the HotpotQA dev set. Values are means with 95% bootstrap CIs over independent samples, measured in nats and rounded to two decimal places. Robust Control is omitted because only the correct-answer score is computed for this group, so no correct-versus-incorrect margin is defined.

Model	Group	Whole-Answer Effect [95% CI]	First-Divergence Effect [95% CI]
Qwen2.5-7B	Regression	-10.01 [-10.58, -9.46]	-23.57 [-24.30, -22.85]
	Rescue	+6.07 [4.79, 7.50]	+13.23 [11.67, 14.90]
	Persistent Failure	-2.01 [-2.40, -1.63]	-4.79 [-5.62, -3.96]
Qwen2.5-0.5B	Regression	-5.18 [-5.39, -4.97]	-10.43 [-10.76, -10.11]
	Rescue	+1.55 [1.30, 1.81]	+4.44 [3.83, 5.06]
	Persistent Failure	-2.68 [-2.84, -2.53]	-4.72 [-4.99, -4.46]
Llama-3.1-8B	Regression	-5.04 [-5.63, -4.49]	-11.97 [-12.94, -11.02]
	Rescue	+3.20 [2.14, 4.46]	+6.26 [4.84, 7.76]
	Persistent Failure	-0.95 [-1.63, -0.33]	-2.09 [-3.49, -0.77]
Gemma-4-E4B	Regression	-5.81 [-6.18, -5.45]	-15.78 [-16.39, -15.17]
	Rescue	+3.43 [2.84, 4.07]	+9.48 [8.36, 10.62]
	Persistent Failure	-2.22 [-2.56, -1.89]	-4.44 [-5.11, -3.78]



(a) Answer regression (b) Answer rescue (c) Persistent failure

Figure 3: Whole-answer margin changes (Full Context minus Supporting Evidence Only) on the HotpotQA dev set. Panels (a)–(c) show Regression, Rescue, and Persistent Failure, respectively. Lines denote models, shaded bands indicate 95% bootstrap CIs, and the dashed horizontal line marks zero effect. The x-axis shows relative residual-stream depth; effects on the y-axis are measured in nats.

jectories show broadly similar trends. The three transition-direction probe trajectories are shown in Figure 2(e–h) in the main text.

To test whether the hidden-state probes exploit information beyond question type, difficulty, text length, and the number of supporting facts, Table 9 reports a Metadata-Only Baseline on the HotpotQA validation set. This baseline uses no model hidden states and follows the same training procedure as the representation probes.

Table 8: Validation-set AUROC and training-selected residual locations for the three probe tasks on HotpotQA. Each cell reports AUROC followed by the selected residual-location index in parentheses. Supporting-Fact End denotes the mean hidden state at the final token of each supporting fact; Question End denotes the state at the final question token; Pre-Generation denotes the state immediately before answer generation.

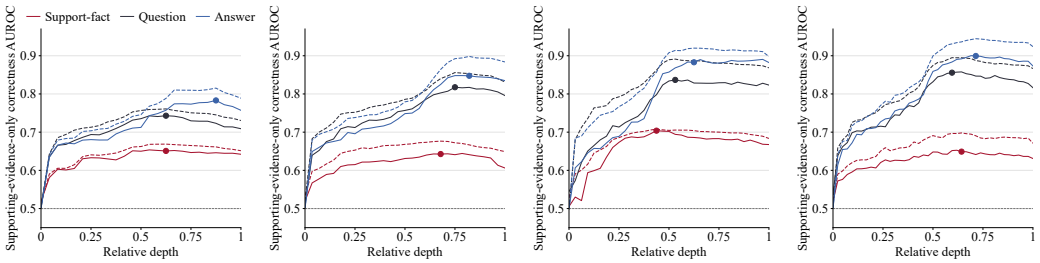
Model	Task	Supporting-Fact End	Question End	Pre-Generation
Qwen2.5-7B	Supporting-evidence-only correctness	.6427 (19)	.8176 (21)	.8476 (23)
	Full-context correctness	.5806 (16)	.7618 (21)	.7998 (23)
	Transition direction	.7312 (19)	.8286 (21)	.8305 (23)
Qwen2.5-0.5B	Supporting-evidence-only correctness	.6508 (15)	.7430 (15)	.7831 (21)
	Full-context correctness	.5877 (13)	.7055 (13)	.7338 (16)
	Transition direction	.7186 (13)	.7223 (14)	.7872 (21)
Llama-3.1-8B	Supporting-evidence-only correctness	.7039 (14)	.8368 (17)	.8829 (20)
	Full-context correctness	.5875 (14)	.7736 (16)	.8270 (19)
	Transition direction	.6731 (14)	.7740 (17)	.8065 (31)
Gemma-4-E4B	Supporting-evidence-only correctness	.6490 (27)	.8557 (25)	.8994 (30)
	Full-context correctness	.6122 (26)	.8266 (26)	.8815 (30)
	Transition direction	.6313 (24)	.8758 (25)	.9071 (30)

Table 9: Metadata-Only Baseline performance on the HotpotQA validation set. Each cell reports AUROC / balanced accuracy, computed on the independent validation set. S-E, F-C, and T-D denote supporting-evidence-only correctness, full-context correctness, and Regression-versus-Rescue transition direction, respectively. Probe wins reports the number (out of 9) of representation probes with significantly higher AUROC than the corresponding baseline after paired testing and Holm correction.

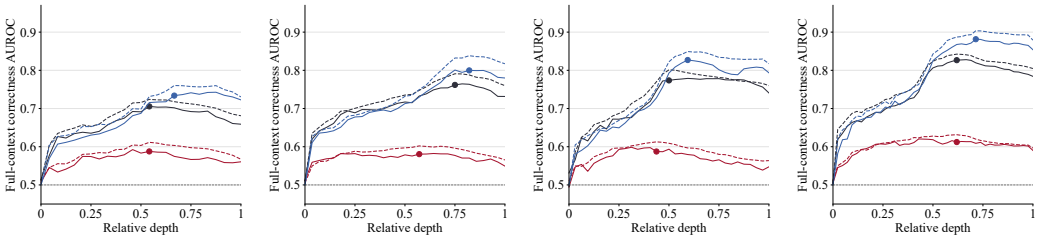
Model	S-E	F-C	T-D	Probe wins
Qwen2.5-7B	.5654 / .5003	.5483 / .5018	.6341 / .6356	9/9
Qwen2.5-0.5B	.6284 / .5210	.5561 / .5110	.6370 / .6077	7/9
Llama-3.1-8B	.6503 / .5295	.5661 / .4991	.6084 / .6755	6/9
Gemma-4-E4B	.5374 / .4993	.5391 / .4998	.5690 / .5367	8/9

Comparing Table 8 with Table 9 shows that pre-generation probes achieve higher AUROC than the corresponding metadata baselines. Across the nine paired probe-baseline comparisons, 9, 7, 6, and 8 are significant after Holm correction for Qwen2.5-7B, Qwen2.5-0.5B, Llama-3.1-8B, and Gemma-4-E4B, respectively. Thus, most representation probes provide predictive information beyond these observable sample attributes.

Tables 8 and 9 summarize probe performance and the baseline comparison. Figure 4 plots out-of-fold training AUROC and dev-set AUROC across residual locations for the six correctness-probe types, showing how predictive information develops with network depth and how the two trajectories compare.



(a) Qwen2.5-0.5B (b) Qwen2.5-7B (c) Llama-3.1-8B (d) Gemma-4-E4B



(e) Qwen2.5-0.5B (f) Qwen2.5-7B (g) Llama-3.1-8B (h) Gemma-4-E4B

Figure 4: Layerwise AUROC of correctness probes across residual locations on HotpotQA. Panels (a)–(d) show supporting-evidence-only correctness and (e)–(h) full-context correctness; within each row, models are ordered Qwen2.5-0.5B, Qwen2.5-7B, Llama-3.1-8B, and Gemma-4-E4B. Support-fact denotes the mean hidden state at the final token of each supporting fact, Question the state at the final question token, and Answer the pre-generation state. Solid lines show dev-set AUROC of probes fitted on the full training set; dashed lines show five-fold out-of-fold training AUROC. Circles mark training-selected residual locations, and the horizontal dashed line indicates chance-level AUROC (0.5).

G Intervention-Location Selection and Layer-Wise Effects

To examine how intervention effects vary with network depth, we perform bidirectional patching at the locations selected by the nine training-set probes, together with a predefined early-layer control and the final residual location. Each probe selects one location, so the union in Table 10 may contain several distinct locations per model. The fixed location maximizes the equally weighted average of the four training-set mean effects across the two behavioral groups and intervention directions; it is then used for the primary validation-set analyses.

The repair intervention starts from the condition that produces an incorrect response and writes in the same question's hidden state from the condition that produces a correct response. Its effect is positive when the correct answer's score increases relative to the incorrect answer. Damage uses the reverse replacement and is positive when the correct answer's relative score decreases. Table 11 reports both directional effects for the Regression and Rescue groups across the four models.

Table 11 reports validation-set effects at each model's fixed intervention location selected on the HotpotQA training set. Figure 5 shows how these effects vary across the probe-selected candidates, early-layer control, and final residual location. Each circle marks the fixed location listed in Table 10 and used for Table 11.

All 16 raw effects at the fixed locations are positive (Table 11). For Qwen2.5-7B in the Regression group, for example, repair and damage effects are 14.93 and 14.14 nats. Writing the correct-condition state into an incorrect-condition run raises the correct answer's relative score; the reverse

Table 10: Candidate and fixed intervention locations for residual-stream patching on HotpotQA. The probe-selected column gives the deduplicated union of locations selected by the nine probes; the early-layer control and final location are also tested. The fixed intervention location is selected using training-set intervention effects.

Model	Unique Probe-Selected Locations	Early-Layer Control	Final Location	Fixed Intervention Location
Qwen2.5-7B	16, 19, 21, 23	7	28	23
Qwen2.5-0.5B	13, 14, 15, 16, 21	6	24	21
Llama-3.1-8B	14, 16, 17, 19, 20, 31	8	32	31
Gemma-4-E4B	24, 25, 26, 27, 30	11	42	30

Table 11: Directional effects on the First-Divergence Margin at the fixed intervention locations on the HotpotQA validation set. RR, RN, DR, and DN denote Repair Raw, Repair Normalized, Damage Raw, and Damage Normalized, respectively. Raw columns report mean effects and 95% CIs in nats. Normalized columns report means of per-sample ratios between the raw effect and that sample's unperturbed score gap, as defined in Appendix B.3. Values are rounded to two decimal places.

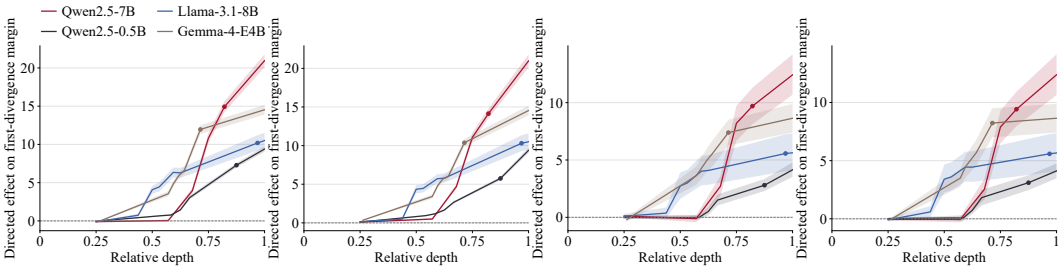
Model	Group	RR: [95% CI]	RN: [95% CI]	DR: [95% CI]	DN: [95% CI]
Qwen2.5-7B	Regression	14.93 [14.20, 15.65]	.67 [.63, .70]	14.14 [13.51, 14.79]	.62 [.58, .65]
	Rescue	9.70 [8.11, 11.30]	.75 [.63, .90]	9.42 [7.96, 10.93]	.69 [.58, .79]
Qwen2.5-0.5B	Regression	7.28 [6.99, 7.57]	.71 [.66, .75]	5.75 [5.53, 5.98]	.59 [.56, .62]
	Rescue	2.80 [2.29, 3.35]	.70 [.57, .82]	3.10 [2.47, 3.74]	.73 [.56, .90]
Llama-3.1-8B	Regression	10.19 [9.19, 11.20]	.86 [.81, .90]	10.30 [9.28, 11.35]	.87 [.82, .91]
	Rescue	5.57 [4.01, 7.24]	.88 [.75, .99]	5.59 [3.99, 7.28]	.87 [.73, .98]
Gemma-4-E4B	Regression	11.96 [11.39, 12.52]	.78 [.76, .81]	10.37 [9.87, 10.86]	.69 [.67, .72]
	Rescue	7.39 [6.27, 8.51]	.76 [.67, .84]	8.23 [7.00, 9.50]	.86 [.76, .96]

replacement lowers it in a correct-condition run. The normalized means are .67 and .62, respectively, corresponding to about 67% and 62% of the unperturbed score gap. Across models, Figure 5 shows stronger effects at some middle-to-late candidate locations, demonstrating that the choice of residual location shapes how strongly the replacement changes the correct-versus-incorrect score margin.

H Semantic Flips in Free Generation and Control Results

In the free-generation experiments, we regenerate the unperturbed baseline and the response after state patching, then measure whether correctness changes in the expected direction. Answers not confirmed by EM are evaluated by the semantic judge. A sample enters a direction-specific denominator only when both generations and their correctness evaluations are complete and the unperturbed baseline reproduces the original correctness label. Across the two context conditions, Qwen2.5-7B, Qwen2.5-0.5B, Llama, and Gemma have 58, 236, 7, and 20 baseline responses, respectively, that fail to reproduce their original correctness labels; these are excluded from the corresponding denominators.

Table 12 reports the flip counts, valid sample sizes, and rates at the fixed intervention locations. Repair counts incorrect answers that become correct; Damage counts correct answers that become incorrect.



(a) Regression / Repair (b) Regression / Damage (c) Rescue / Repair (d) Rescue / Damage

Figure 5: Layerwise directional effects on the First-Divergence Margin on the HotpotQA dev set. Regression and Rescue denote behavioral groups; Repair and Damage denote the two state-replacement directions. Panels (a)–(d) show Regression/Repair, Regression/Damage, Rescue/Repair, and Rescue/Damage, respectively. The y-axis is the intervention-induced change in the First-Divergence Margin, measured in nats. Effects are direction-coded: positive values indicate an increased correct-versus-incorrect margin for Repair and a decreased margin for Damage. Lines denote models, shaded bands show 95% CIs, circles mark the training-selected fixed intervention locations, and the dashed horizontal line marks zero effect. The x-axis shows relative residual-stream depth.

Table 12: Free-generation semantic flip rates at the fixed intervention locations on the HotpotQA dev set. Each row reports the number of expected correctness flips, the valid sample count, and the flip rate with its 95% CI. The flip rate is the flip count divided by the valid sample count; rates and intervals are percentages rounded to two decimal places.

Model	Group / Direction	Flips	Valid Samples	Flip Rate [95% CI]
Qwen2.5-7B	Regression / Repair	483	752	64.23 [60.77, 67.55]
	Regression / Damage	345	782	44.12 [40.66, 47.57]
	Rescue / Repair	64	97	65.98 [56.70, 75.26]
	Rescue / Damage	66	107	61.68 [52.34, 71.03]
Qwen2.5-0.5B	Regression / Repair	613	1,128	54.34 [51.51, 57.27]
	Regression / Damage	424	1,159	36.58 [33.82, 39.34]
	Rescue / Repair	53	84	63.10 [52.38, 73.81]
	Rescue / Damage	52	101	51.49 [41.58, 61.39]
Llama-3.1-8B	Regression / Repair	148	187	79.14 [73.26, 85.03]
	Regression / Damage	134	191	70.16 [63.87, 76.44]
	Rescue / Repair	21	26	80.77 [65.38, 96.15]
	Rescue / Damage	21	25	84.00 [68.00, 96.00]
Gemma-4-E4B	Regression / Repair	507	707	71.71 [68.32, 74.96]
	Regression / Damage	360	714	50.42 [46.78, 54.06]
	Rescue / Repair	58	89	65.17 [55.06, 75.28]
	Rescue / Damage	56	92	60.87 [51.09, 70.65]

Table 13 compares the main intervention at the fixed location with three controls: early-layer patching, question-end patching, and identity write-back. Each column uses samples that satisfy its own baseline-reproduction and answer-evaluation criteria, so valid sample sizes may differ across columns.

Table 13: Semantic flip rates (%) for the main intervention and controls on the HotpotQA dev set. A value of 0.00 indicates that no expected correctness flip was observed under that condition.

Model	Group	Direction	Main Intervention	Early Layer	Question-End Control	Identity Write-Back Control
Qwen2.5-0.5B	Regression	Repair	54.34	2.13	0.71	0.00
	Regression	Damage	36.58	0.86	0.09	0.00
	Rescue	Repair	63.10	0.00	1.19	0.00
	Rescue	Damage	51.49	3.96	0.00	0.00
Qwen2.5-7B	Regression	Repair	64.23	0.93	0.53	0.00
	Regression	Damage	44.12	0.13	0.13	0.00
	Rescue	Repair	65.98	4.12	1.03	0.00
	Rescue	Damage	61.68	1.87	0.93	0.00
Llama-3.1-8B	Regression	Repair	79.14	3.21	0.53	0.00
	Regression	Damage	70.16	0.00	0.00	0.00
	Rescue	Repair	80.77	0.00	0.00	0.00
	Rescue	Damage	84.00	0.00	0.00	0.00
Gemma-4-E4B	Regression	Repair	71.71	3.54	0.00	0.00
	Regression	Damage	50.42	0.84	0.00	0.00
	Rescue	Repair	65.17	3.37	0.00	0.00
	Rescue	Damage	60.87	2.17	0.00	0.00

Table 12 shows bidirectional correctness flips for all four models in both behavioral groups. In Table 13, the main-intervention flip rates range from 36.58% to 84.00%; the maximum rates are 4.12% for early-layer patching and 1.19% for question-end patching, while identity write-back produces no expected flips. These results show that the generated-answer changes are concentrated at the training-selected pre-generation locations, consistent with the candidate-score effects above. Results for MS MARCO and NQ are reported in Appendices J.3 and K.3, respectively.

I Component Intervention Effects on the HotpotQA Validation Set

To test whether the effect of state patching depends on using a state source from the same question, we compute the difference between paired same-question and equal-norm random-source effects for the full residual, attention, and MLP outputs. Table 14 reports the mean differences on the HotpotQA validation set with 95% bootstrap CIs. Positive values indicate a larger directional effect for the paired source; negative values indicate a larger effect for the random source.

All 48 source comparisons (12 per model) are significant after Holm correction. The differences for the full residual and attention outputs are positive under all conditions, indicating that same-question states consistently produce stronger answer-related effects than equal-norm random sources for these outputs. MLP effects vary by model: all four differences are negative for Llama and positive for Gemma. Thus, source specificity depends on both the model and the component being patched.

To compare component effects under paired same-question patching, Table 15 reports the attention-output effect minus the MLP-output effect. Positive values indicate a stronger attention intervention; negative values indicate a stronger MLP intervention.

Table 14: Component source specificity in the First-Divergence Margin on the HotpotQA validation set. Entries report the mean directional effect difference (paired same-question source minus equal-norm random source) and its 95% CI, in nats. Full Residual, Attention, and MLP denote the output states of the full residual stream, attention module, and multilayer perceptron, respectively. Values are rounded to two decimal places.

Model	Group	Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression	Repair	14.87 [14.11, 15.64]	4.78 [4.47, 5.10]	1.43 [1.18, 1.69]
	Regression	Damage	10.87 [10.22, 11.51]	3.83 [3.53, 4.13]	1.32 [1.08, 1.59]
	Rescue	Repair	9.10 [7.41, 10.84]	2.48 [1.88, 3.11]	1.48 [0.85, 2.17]
	Rescue	Damage	9.71 [8.23, 11.24]	2.86 [2.20, 3.56]	1.62 [1.01, 2.29]
Qwen2.5-0.5B	Regression	Repair	6.56 [6.29, 6.83]	3.45 [3.30, 3.60]	0.33 [0.25, 0.41]
	Regression	Damage	4.41 [4.21, 4.61]	2.78 [2.64, 2.93]	0.44 [0.37, 0.51]
	Rescue	Repair	2.51 [2.06, 2.98]	1.39 [1.14, 1.65]	0.37 [0.22, 0.52]
	Rescue	Damage	2.96 [2.45, 3.49]	1.57 [1.31, 1.84]	0.43 [0.28, 0.59]
Llama-3.1-8B	Regression	Repair	10.34 [9.34, 11.36]	0.81 [0.65, 0.99]	-0.57 [-0.79, -0.36]
	Regression	Damage	9.20 [8.29, 10.13]	0.88 [0.70, 1.07]	-0.46 [-0.71, -0.22]
	Rescue	Repair	4.73 [3.29, 6.25]	0.29 [0.02, 0.63]	-1.17 [-1.82, -0.59]
	Rescue	Damage	5.17 [3.67, 6.75]	0.37 [0.11, 0.69]	-1.13 [-1.81, -0.51]
Gemma-4-E4B	Regression	Repair	12.29 [11.63, 12.94]	6.15 [5.77, 6.53]	7.65 [7.22, 8.08]
	Regression	Damage	8.30 [7.78, 8.82]	5.11 [4.76, 5.47]	5.11 [4.76, 5.46]
	Rescue	Repair	7.46 [6.05, 8.96]	3.62 [2.74, 4.54]	4.93 [3.91, 6.02]
	Rescue	Damage	8.09 [6.71, 9.50]	3.99 [3.22, 4.80]	4.91 [4.02, 5.87]

Table 15 shows that Qwen2.5-0.5B and Llama exhibit an attention advantage in both behavioral groups and intervention directions. For Gemma, all four differences are negative, indicating stronger paired interventions for the MLP output. Qwen2.5-7B shows a significant attention advantage for both Regression directions; its two Rescue differences, marked †, are not significant after Holm correction. Together with Table 14, these results show that component interventions vary both in sensitivity to same-question state sources and in the relative effects of attention and MLP across models. For Llama's MLP output, a larger random-source effect refers to the directional answer-score margin. The corresponding source-specificity results for MS MARCO and NQ are reported in Appendices J.4 and K.4, respectively.

J Cross-Dataset Validation Results on MS MARCO

This appendix reports results on the MS MARCO train, dev, and test splits. Tables 16 and 17 summarize answer accuracy and final-position score-margin changes across all three splits; the frozen probe, residual-patching, and free-generation analyses also cover all three splits.

Table 16 reports answer accuracy under the supporting-evidence-only and full-context conditions. Each rate is computed over samples with finalized correctness labels under both conditions.

Across train, dev, and test, all four models have lower accuracy under full context than under supporting evidence only. Table 17 reports the change in the First-Divergence Margin at the final residual location, computed as Full Context minus Supporting Evidence Only for the log-probability advantage of the correct candidate over the incorrect candidate. The margin decreases for Regression and Persistent Failure and increases for Rescue across all three splits.

Table 15: Difference between paired attention and MLP intervention effects on the HotpotQA validation set. Entries report the mean difference and its 95% CI, in nats; positive values favor attention and negative values favor MLP. Values are rounded to two decimal places. † marks a difference that is not significant after Holm correction; unmarked differences are significant.

Model	Group	Direction	Difference [95% CI]
Qwen2.5-7B	Regression	Repair	+3.41 [3.04, 3.78]
	Regression	Damage	+2.40 [2.03, 2.76]
	Rescue	Repair	+0.72 [-0.23, 1.63]†
	Rescue	Damage	+0.73 [-0.19, 1.62]†
Qwen2.5-0.5B	Regression	Repair	+3.14 [2.97, 3.31]
	Regression	Damage	+2.53 [2.37, 2.68]
	Rescue	Repair	+1.14 [0.88, 1.39]
	Rescue	Damage	+1.33 [1.07, 1.62]
Llama-3.1-8B	Regression	Repair	+1.39 [1.14, 1.66]
	Regression	Damage	+1.30 [1.03, 1.59]
	Rescue	Repair	+1.47 [0.87, 2.09]
	Rescue	Damage	+1.44 [0.79, 2.19]
Gemma-4-E4B	Regression	Repair	-1.04 [-1.28, -0.80]
	Regression	Damage	-1.10 [-1.31, -0.88]
	Rescue	Repair	-2.14 [-2.85, -1.45]
	Rescue	Damage	-1.27 [-1.88, -0.66]

Across all three splits, the correct-over-incorrect margin decreases for Regression and Persistent Failure, and increases for Rescue. These consistent directions show that context expansion shifts answer-candidate scores differently across behavioral groups.

J.1 Frozen Transition-Direction Probes

To test transfer across datasets and data splits, Table 18 reports the AUROC and 95% CIs obtained by applying the transition-direction probes fixed on HotpotQA to the MS MARCO train, dev, and test splits. Probe parameters and residual locations remain fixed at their HotpotQA selections.

Across all three splits, the pre-generation AUROC exceeds the corresponding supporting-fact end AUROC for every model. The question-end and pre-generation tests are significant after Holm correction throughout; supporting-fact end significance is less consistent, with only Qwen2.5-7B significant on train and none of the dev/test tests significant (Table 18). This pattern indicates that the transferable transition signal is strongest near answer generation.

J.2 Frozen Residual Main Effects

Whether the predictive information can influence answer scores is tested through bidirectional state patching. Table 19 reports first-divergence directional effects on MS MARCO at the main intervention locations selected on HotpotQA, separately for train, dev, and test. Repair writes the same question's correct-response state into its incorrect-response condition; damage applies the reverse replacement.

Table 16: Answer accuracy (%) under the two evidence conditions on MS MARCO.

Split	Model	Supporting Evidence Only	Full Context
Train	Qwen2.5-7B	60.08%	37.32%
	Qwen2.5-0.5B	52.39%	31.89%
	Llama-3.1-8B	54.97%	36.40%
	Gemma-4-E4B	58.86%	33.35%
Dev	Qwen2.5-7B	59.98%	37.03%
	Qwen2.5-0.5B	52.90%	31.58%
	Llama-3.1-8B	55.56%	37.37%
	Gemma-4-E4B	58.97%	33.95%
Test	Qwen2.5-7B	61.15%	37.63%
	Qwen2.5-0.5B	52.72%	31.50%
	Llama-3.1-8B	53.38%	34.97%
	Gemma-4-E4B	58.59%	34.60%

Table 17: Changes in the First-Divergence Margin at the final residual location on MS MARCO (nats).

Split	Model	Regression	Rescue	Persistent Failure
Train	Qwen2.5-7B	-24.34	+16.19	-7.63
	Qwen2.5-0.5B	-7.78	+4.15	-2.71
	Llama-3.1-8B	-12.10	+6.35	-4.26
	Gemma-4-E4B	-19.24	+8.91	-7.69
Dev	Qwen2.5-7B	-24.33	+14.96	-7.68
	Qwen2.5-0.5B	-7.50	+4.08	-2.75
	Llama-3.1-8B	-12.65	+5.45	-4.33
	Gemma-4-E4B	-19.41	+8.51	-7.83
Test	Qwen2.5-7B	-23.94	+17.19	-7.71
	Qwen2.5-0.5B	-7.92	+3.31	-2.64
	Llama-3.1-8B	-12.13	+5.95	-3.95
	Gemma-4-E4B	-19.00	+8.10	-7.83

All 48 effects across the three splits are positive, with 95% CIs above zero. Thus, states from the same question's correct-response condition continue to shift the First-Divergence Margin in the expected direction on both official MS MARCO splits.

J.3 Frozen Free Generation

Whether the changes in answer scores extend to complete generated responses is tested using the free-generation results in Table 20. Each cell reports the flip rate and its 95% CI.

Both repair and reverse-direction damage occur in both behavioral groups across all models and splits. The main intervention exceeds the early-layer, question-end, and identity controls in all 12

Table 18: AUROC and 95% CIs of the frozen transition-direction probes on MS MARCO. The final column gives the Holm-corrected p value for testing whether the Supporting-Fact End probe exceeds chance (AUROC $= 0.5$).

Split	Model	Supporting-Fact End	Question End	Pre-Generation	p
Train	Qwen2.5-7B	.5485 [.5284, .5689]	.6322 [.6123, .6522]	.7471 [.7298, .7637]	.0009
	Qwen2.5-0.5B	.5032 [.4767, .5301]	.5321 [.5059, .5580]	.6414 [.6158, .6667]	.4040
	Llama-3.1-8B	.5169 [.4862, .5478]	.5564 [.5253, .5877]	.6996 [.6726, .7261]	.1434
	Gemma-4-E4B	.5269 [.4863, .5669]	.6349 [.5974, .6716]	.8252 [.7999, .8498]	.1030
Dev	Qwen2.5-7B	.4999 [.4365, .5611]	.6270 [.5656, .6865]	.7564 [.7078, .8017]	.5079
	Qwen2.5-0.5B	.5258 [.4454, .6051]	.5058 [.4347, .5765]	.6524 [.5851, .7178]	.5179
	Llama-3.1-8B	.5518 [.4644, .6394]	.5304 [.4454, .6108]	.6288 [.5500, .7039]	.2608
	Gemma-4-E4B	.4931 [.3599, .6258]	.7641 [.6841, .8403]	.8420 [.7801, .8985]	.5427
Test	Qwen2.5-7B	.5502 [.4849, .6163]	.6035 [.5474, .6599]	.7763 [.7266, .8215]	.0602
	Qwen2.5-0.5B	.4493 [.3626, .5377]	.4823 [.3944, .5747]	.6874 [.6112, .7611]	1.0000
	Llama-3.1-8B	.5156 [.4189, .6098]	.4659 [.3794, .5561]	.7120 [.6323, .7862]	.7403
	Gemma-4-E4B	.4335 [.3272, .5417]	.5416 [.4325, .6531]	.8103 [.7399, .8749]	.8622

Table 19: Mean First-Divergence Margin effects and 95% CIs at the frozen main intervention locations on MS MARCO (nats).

Split	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	9.12 [8.83, 9.40]	8.54 [8.29, 8.80]	10.58 [9.92, 11.26]	10.40 [9.73, 11.09]
	Qwen2.5-0.5B	2.78 [2.67, 2.90]	2.59 [2.49, 2.70]	2.15 [1.92, 2.40]	2.27 [2.03, 2.53]
	Llama-3.1-8B	5.57 [5.25, 5.90]	5.69 [5.36, 6.01]	4.78 [4.32, 5.24]	4.85 [4.39, 5.31]
	Gemma-4-E4B	10.28 [9.97, 10.59]	8.74 [8.48, 9.00]	4.95 [4.24, 5.71]	5.12 [4.39, 5.89]
Dev	Qwen2.5-7B	8.91 [8.11, 9.75]	8.39 [7.68, 9.12]	8.84 [7.18, 10.62]	8.88 [7.28, 10.61]
	Qwen2.5-0.5B	2.77 [2.44, 3.09]	2.55 [2.27, 2.85]	1.84 [1.24, 2.50]	1.89 [1.22, 2.62]
	Llama-3.1-8B	5.67 [4.68, 6.71]	5.82 [4.80, 6.84]	3.81 [2.72, 5.00]	3.88 [2.79, 5.07]
	Gemma-4-E4B	10.50 [9.53, 11.53]	8.61 [7.84, 9.41]	5.19 [3.08, 7.44]	5.79 [3.49, 8.34]
Test	Qwen2.5-7B	8.67 [7.84, 9.53]	8.20 [7.44, 8.94]	11.73 [9.59, 13.94]	11.69 [9.49, 13.98]
	Qwen2.5-0.5B	3.14 [2.78, 3.52]	2.96 [2.64, 3.30]	1.45 [0.91, 2.05]	1.54 [0.95, 2.20]
	Llama-3.1-8B	5.06 [4.21, 5.94]	5.17 [4.31, 6.07]	4.13 [2.87, 5.45]	4.11 [2.89, 5.45]
	Gemma-4-E4B	9.80 [8.80, 10.83]	8.16 [7.33, 8.99]	4.49 [2.65, 6.50]	4.16 [2.42, 6.22]

Holm-corrected comparisons per model on train and dev; on test, 11/12 comparisons are significant for Gemma and all 12 are significant for the other models. Thus, the frozen intervention transfers to generated responses, with one control contrast remaining inconclusive for Gemma on the test split.

J.4 Component Directional Specificity

Tables 21–23 replace the component-source plot with its underlying estimates for MS MARCO train, dev, and test, respectively. Each entry is the mean first-divergence directional-effect difference

Table 20: Free-generation semantic flip rates (%) and 95% CIs under the frozen main intervention on MS MARCO.

Split	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	28.44[27.15,29.73]	18.45[17.31,19.54]	56.71[53.16,60.13]	53.73[50.37,57.21]
	Qwen2.5-0.5B	26.04[24.53,27.53]	21.41[20.08,22.78]	39.29[34.44,44.39]	27.49[23.50,31.71]
	Llama-3.1-8B	37.99[35.65,40.33]	41.05[38.67,43.42]	63.73[59.07,68.38]	63.11[58.50,67.72]
	Gemma-4-E4B	25.97[24.50,27.48]	15.30[14.11,16.55]	52.00[45.00,59.00]	44.55[37.62,51.49]
Dev	Qwen2.5-7B	30.09[26.37,33.98]	19.21[16.12,22.47]	61.96[52.17,71.74]	55.56[45.56,65.56]
	Qwen2.5-0.5B	25.36[21.33,29.62]	20.55[16.89,24.43]	30.61[18.37,42.86]	30.51[18.64,42.37]
	Llama-3.1-8B	36.32[29.47,43.16]	37.70[30.89,44.50]	56.25[41.67,70.83]	52.08[37.50,66.67]
	Gemma-4-E4B	25.60[21.33,30.13]	15.22[11.55,18.90]	59.09[36.36,77.27]	52.38[33.33,71.43]
Test	Qwen2.5-7B	29.88[25.90,34.06]	20.36[17.00,23.91]	60.67[50.56,70.79]	60.44[50.55,70.33]
	Qwen2.5-0.5B	29.38[25.00,33.76]	24.87[20.60,29.15]	33.33[19.44,50.00]	29.27[17.07,43.90]
	Llama-3.1-8B	37.50[30.43,44.57]	39.89[32.79,46.99]	56.52[41.30,69.57]	56.52[43.42,69.57]
	Gemma-4-E4B	25.85[21.23,30.77]	12.69[9.37,16.31]	56.00[36.00,76.00]	44.00[24.00,64.00]

Table 21: Component source-specificity effects on the MS MARCO train split (mean [95% CI], nats).

Model	Group / Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression / Repair	9.21 [8.91, 9.51]	3.09 [2.97, 3.21]	1.57 [1.48, 1.67]
	Regression / Damage	6.43 [6.22, 6.65]	2.57 [2.46, 2.67]	1.35 [1.26, 1.45]
	Rescue / Repair	8.06 [7.41, 8.71]	0.99 [0.85, 1.14]	2.74 [2.43, 3.04]
	Rescue / Damage	8.59 [7.91, 9.29]	1.19 [1.03, 1.36]	2.98 [2.66, 3.31]
Qwen2.5-0.5B	Regression / Repair	2.71 [2.59, 2.83]	1.01 [0.96, 1.06]	0.42 [0.39, 0.45]
	Regression / Damage	2.22 [2.12, 2.31]	0.97 [0.91, 1.02]	0.41 [0.38, 0.44]
	Rescue / Repair	1.98 [1.75, 2.22]	0.82 [0.70, 0.93]	0.34 [0.27, 0.42]
	Rescue / Damage	2.33 [2.07, 2.60]	0.88 [0.76, 1.01]	0.38 [0.31, 0.46]
Llama-3.1-8B	Regression / Repair	5.54 [5.20, 5.86]	0.40 [0.36, 0.43]	-0.35 [-0.41, -0.28]
	Regression / Damage	5.01 [4.72, 5.31]	0.55 [0.50, 0.60]	-0.35 [-0.42, -0.28]
	Rescue / Repair	3.84 [3.40, 4.29]	0.21 [0.16, 0.25]	-0.83 [-1.00, -0.67]
	Rescue / Damage	4.01 [3.55, 4.48]	0.28 [0.22, 0.34]	-0.97 [-1.15, -0.79]
Gemma-4-E4B	Regression / Repair	8.61 [8.30, 8.92]	4.53 [4.35, 4.71]	5.51 [5.31, 5.72]
	Regression / Damage	5.52 [5.31, 5.74]	3.65 [3.50, 3.79]	3.45 [3.30, 3.59]
	Rescue / Repair	4.50 [3.80, 5.23]	2.39 [1.93, 2.89]	2.72 [2.27, 3.19]
	Rescue / Damage	4.72 [3.93, 5.54]	2.60 [2.14, 3.09]	3.05 [2.53, 3.60]

between the paired same-question source and the equal-norm random source, followed by its 95% bootstrap CI, in nats. Positive values favor the paired source; negative values favor the random source. A dagger marks a comparison that is not significant after Holm correction within the model.

Table 22: Component source-specificity effects on the MS MARCO dev split (mean [95% CI], nats).

Model	Group / Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression / Repair	8.47 [7.65, 9.28]	2.94 [2.61, 3.28]	1.33 [1.08, 1.59]
	Regression / Damage	6.10 [5.49, 6.75]	2.29 [2.02, 2.58]	1.31 [1.05, 1.58]
	Rescue / Repair	5.35 [3.78, 7.01]	0.72 [0.40, 1.07]	1.80 [1.11, 2.57]
	Rescue / Damage	6.09 [4.37, 7.85]	0.81 [0.45, 1.19]	1.86 [1.11, 2.64]
Qwen2.5-0.5B	Regression / Repair	2.66 [2.34, 2.99]	1.00 [0.84, 1.17]	0.40 [0.31, 0.49]
	Regression / Damage	2.13 [1.86, 2.40]	0.96 [0.81, 1.12]	0.35 [0.28, 0.44]
	Rescue / Repair	1.53 [1.00, 2.12]	0.81 [0.50, 1.17]	0.31 [0.16, 0.48]
	Rescue / Damage	1.85 [1.21, 2.55]	0.93 [0.58, 1.32]	0.40 [0.21, 0.62]
Llama-3.1-8B	Regression / Repair	5.94 [4.90, 7.04]	0.48 [0.36, 0.61]	-0.16 [-0.35, 0.01]†
	Regression / Damage	5.53 [4.54, 6.53]	0.64 [0.49, 0.82]	-0.20 [-0.42, 0.01]†
	Rescue / Repair	3.12 [2.08, 4.27]	0.06 [-0.03, 0.15]†	-0.45 [-0.71, -0.19]
	Rescue / Damage	3.18 [2.09, 4.37]	0.25 [0.12, 0.42]	-0.71 [-1.12, -0.34]
Gemma-4-E4B	Regression / Repair	9.05 [8.09, 10.02]	4.39 [3.88, 4.93]	5.72 [5.13, 6.35]
	Regression / Damage	5.76 [5.09, 6.44]	3.52 [3.12, 3.95]	3.56 [3.11, 4.02]
	Rescue / Repair	5.75 [3.18, 8.64]	2.65 [1.23, 4.39]	3.63 [1.73, 5.66]
	Rescue / Damage	5.56 [3.23, 8.15]	2.33 [1.08, 3.81]	3.23 [1.89, 4.59]

The full-residual source advantage is positive in every model, condition, and split. Llama's MLP differences are negative throughout, while Gemma's are positive; the Qwen models and the remaining component contrasts show condition- and split-dependent magnitudes (Tables 21–23).

Table 24 gives the paired-source Attention-minus-MLP effect differences on the MS MARCO train, dev, and test splits. Positive differences favor attention and negative differences favor MLP. A dagger marks a difference that is not significant after Holm correction; unmarked differences are significant.

The train results show significant attention advantages for Qwen2.5-0.5B and Llama and significant MLP advantages for Qwen2.5-7B in Rescue and Gemma in all four conditions. Dev/test retain the attention advantage for Qwen2.5-0.5B and Llama, while Qwen2.5-7B favors attention in Regression and MLP in Rescue. Gemma favors MLP throughout; its two Rescue contrasts on test are not significant after Holm correction (daggered entries).

K Cross-Dataset Validation Results on Natural Questions

This appendix reports results on the NQ train and dev splits. Tables 25 and 26 summarize answer accuracy and final-position score-margin changes across both splits; the frozen probe, residual-patching, and free-generation analyses also cover both splits.

Table 25 reports answer accuracy under the supporting-evidence-only and full-context conditions. Each rate is computed over samples with finalized correctness labels under both conditions.

For all four models, full-context accuracy is lower than supporting-evidence-only accuracy on both splits. Table 26 reports the change in the First-Divergence Margin at the final residual location, computed as Full Context minus Supporting Evidence Only for the log-probability advantage of the correct candidate over the incorrect candidate. The margin decreases for Regression and Persistent Failure and increases for Rescue on both splits.

Table 23: Component source-specificity effects on the MS MARCO test split (mean [95% CI], nats).

Model	Group / Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression / Repair	8.63 [7.78, 9.52]	2.90 [2.55, 3.26]	1.43 [1.16, 1.72]
	Regression / Damage	6.07 [5.45, 6.70]	2.33 [2.04, 2.63]	1.17 [0.93, 1.43]
	Rescue / Repair	8.95 [6.86, 11.20]	1.18 [0.72, 1.73]	3.17 [2.24, 4.13]
	Rescue / Damage	9.34 [7.21, 11.53]	1.11 [0.62, 1.75]	3.64 [2.61, 4.72]
Qwen2.5-0.5B	Regression / Repair	3.09 [2.72, 3.47]	1.13 [0.95, 1.30]	0.45 [0.36, 0.55]
	Regression / Damage	2.55 [2.23, 2.88]	1.15 [0.97, 1.34]	0.51 [0.41, 0.61]
	Rescue / Repair	1.60 [0.99, 2.30]	0.48 [0.23, 0.75]	0.15 [0.06, 0.26]
	Rescue / Damage	1.82 [1.10, 2.66]	0.69 [0.33, 1.11]	0.22 [0.07, 0.39]
Llama-3.1-8B	Regression / Repair	5.12 [4.25, 6.04]	0.32 [0.23, 0.41]	-0.38 [-0.59, -0.18]
	Regression / Damage	4.46 [3.72, 5.23]	0.38 [0.27, 0.50]	-0.54 [-0.80, -0.31]
	Rescue / Repair	3.00 [1.80, 4.31]	0.15 [0.04, 0.29]	-0.33 [-0.65, -0.01]†
	Rescue / Damage	3.04 [1.79, 4.39]	0.22 [0.12, 0.34]	-0.78 [-1.22, -0.36]
Gemma-4-E4B	Regression / Repair	8.00 [7.07, 8.95]	4.01 [3.49, 4.55]	5.04 [4.43, 5.65]
	Regression / Damage	4.89 [4.26, 5.55]	3.35 [2.95, 3.77]	3.03 [2.59, 3.48]
	Rescue / Repair	3.72 [2.08, 5.56]	1.76 [0.92, 2.70]	2.22 [1.20, 3.33]
	Rescue / Damage	4.00 [2.15, 6.29]	1.74 [0.75, 2.80]	2.25 [1.15, 3.53]

Table 24: Attention-minus-MLP paired-source effect differences on MS MARCO ([95% CIs], nats). A dagger marks a difference that is not significant after Holm correction.

	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	1.47 [1.33, 1.62]	1.13 [0.99, 1.27]	-2.58 [-2.91, -2.24]	-2.86 [-3.26, -2.48]
	Qwen2.5-0.5B	0.58 [0.53, 0.63]	0.60 [0.54, 0.65]	0.46 [0.35, 0.56]	0.47 [0.35, 0.59]
	Llama-3.1-8B	0.77 [0.70, 0.84]	0.88 [0.79, 0.97]	1.19 [1.04, 1.34]	1.48 [1.31, 1.66]
	Gemma-4-E4B	-1.44 [-1.56, -1.32]	-1.41 [-1.51, -1.31]	-0.66 [-0.90, -0.41]	-0.60 [-0.86, -0.36]
Dev	Qwen2.5-7B	1.56 [1.17, 1.95]	0.99 [0.64, 1.34]	-2.05 [-2.90, -1.25]	-2.24 [-3.13, -1.42]
	Qwen2.5-0.5B	0.63 [0.47, 0.79]	0.69 [0.53, 0.86]	0.46 [0.20, 0.74]	0.48 [0.21, 0.77]
	Llama-3.1-8B	0.60 [0.43, 0.78]	0.84 [0.60, 1.10]	0.76 [0.47, 1.06]	1.23 [0.78, 1.72]
	Gemma-4-E4B	-1.57 [-1.91, -1.24]	-1.43 [-1.70, -1.17]	-0.95 [-1.71, -0.27]	-0.93 [-1.80, -0.08]
Test	Qwen2.5-7B	1.39 [0.97, 1.82]	1.02 [0.63, 1.40]	-3.10 [-4.15, -2.09]	-3.75 [-5.03, -2.51]
	Qwen2.5-0.5B	0.71 [0.55, 0.87]	0.77 [0.61, 0.95]	0.34 [0.10, 0.61]	0.43 [0.10, 0.82]
	Llama-3.1-8B	0.77 [0.55, 1.01]	0.92 [0.67, 1.19]	0.87 [0.53, 1.24]	1.49 [0.96, 2.04]
	Gemma-4-E4B	-1.52 [-1.90, -1.15]	-1.39 [-1.72, -1.08]	-0.77 [-1.63, -0.04]†	-0.66 [-1.57, 0.14]†

Across both splits, the correct-over-incorrect margin decreases for Regression and Persistent Failure, and increases for Rescue. The same group-specific directions across models and splits provide a behavioral reference for the frozen mechanistic analyses that follow.

Table 25: Answer accuracy (%) under the two evidence conditions on NQ.

Split	Model	Supporting Evidence Only	Full Context
Train	Qwen2.5-7B	79.87%	66.73%
	Qwen2.5-0.5B	57.16%	38.94%
	Llama-3.1-8B	79.09%	67.82%
	Gemma-4-E4B	77.31%	66.70%
Dev	Qwen2.5-7B	86.64%	72.51%
	Qwen2.5-0.5B	63.39%	43.33%
	Llama-3.1-8B	88.38%	73.85%
	Gemma-4-E4B	85.29%	73.18%

Table 26: Changes in the First-Divergence Margin at the final residual location on NQ (nats) .

Split	Model	Regression	Rescue	Persistent Failure
Train	Qwen2.5-7B	-24.61	+16.09	-6.80
	Qwen2.5-0.5B	-9.83	+5.28	-5.03
	Llama-3.1-8B	-12.11	+7.86	-3.75
	Gemma-4-E4B	-18.54	+11.25	-6.51
Dev	Qwen2.5-7B	-25.05	+16.75	-6.64
	Qwen2.5-0.5B	-9.86	+5.56	-5.01
	Llama-3.1-8B	-12.65	+7.11	-3.32
	Gemma-4-E4B	-17.57	+13.68	-7.33

K.1 Frozen Transition-Direction Probes

Table 27 reports the AUROC and 95% CIs obtained by applying the transition-direction probes fixed on HotpotQA to the NQ train and dev splits. Neither probe parameters nor residual locations are adjusted based on NQ.

Across both splits, the pre-generation AUROC exceeds the corresponding supporting-fact end AUROC for every model. All pre-generation tests are significant after Holm correction, while supporting-fact end significance varies by split: three of four models are significant on train, and none are significant on dev (Table 27). The transferable transition-direction signal is therefore most consistently detected near answer generation.

K.2 Frozen Residual Main Effects

Table 28 tests whether these states can influence answer scores. It reports the mean First-Divergence Margin effects and 95% CIs at the HotpotQA-selected main locations, separately for NQ train and dev. Repair writes the same question's correct-response state into its incorrect-response condition; damage applies the reverse replacement.

All 32 effects across the two splits are positive, with 95% CIs above zero. The frozen same-question state replacement therefore shifts the First-Divergence Margin in the expected direction on both NQ train and dev.

Table 27: AUROC and 95% CIs of the frozen transition-direction probes on NQ. The final column gives the Holm-corrected p value for testing whether the Supporting-Fact End probe exceeds chance (AUROC $= 0.5$).

Split	Model	Supporting-Fact End	Question End	Pre-Generation	p
Train	Qwen2.5-7B	.5242 [.5041, .5436]	.5985 [.5792, .6174]	.7347 [.7178, .7511]	.0092
	Qwen2.5-0.5B	.5611 [.5385, .5835]	.5265 [.5036, .5495]	.7094 [.6900, .7287]	.0009
	Llama-3.1-8B	.5052 [.4750, .5356]	.6524 [.6242, .6799]	.8032 [.7799, .8257]	.3650
	Gemma-4-E4B	.5449 [.5235, .5669]	.7143 [.6951, .7341]	.8825 [.8696, .8950]	.0009
Dev	Qwen2.5-7B	.4721 [.3676, .5782]	.5958 [.5028, .6882]	.7457 [.6630, .8232]	.7004
	Qwen2.5-0.5B	.4868 [.3663, .6074]	.4424 [.3235, .5558]	.7145 [.6159, .8043]	1.0000
	Llama-3.1-8B	.3780 [.2185, .5455]	.5582 [.3620, .7528]	.7225 [.5646, .8596]	.9014
	Gemma-4-E4B	.5372 [.3930, .6779]	.7885 [.6657, .8948]	.9767 [.9524, .9936]	.2843

Table 28: Mean First-Divergence Margin effects and 95% CIs at the frozen main intervention locations on NQ (nats).

Split	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	12.75 [12.47, 13.03]	11.10 [10.87, 11.34]	9.66 [9.08, 10.26]	9.86 [9.27, 10.46]
	Qwen2.5-0.5B	6.84 [6.72, 6.96]	6.12 [6.01, 6.23]	3.14 [2.91, 3.38]	3.75 [3.50, 4.01]
	Llama-3.1-8B	9.26 [8.92, 9.61]	9.46 [9.11, 9.81]	6.70 [6.15, 7.25]	6.73 [6.19, 7.28]
	Gemma-4-E4B	11.08 [10.77, 11.39]	8.63 [8.39, 8.87]	8.55 [8.05, 9.06]	9.09 [8.58, 9.59]
Dev	Qwen2.5-7B	13.18 [11.73, 14.64]	11.21 [9.96, 12.47]	12.29 [9.16, 15.71]	11.83 [8.44, 15.32]
	Qwen2.5-0.5B	6.73 [6.04, 7.42]	6.20 [5.60, 6.83]	4.40 [3.00, 5.96]	4.79 [3.17, 6.50]
	Llama-3.1-8B	9.39 [7.58, 11.21]	9.54 [7.70, 11.38]	6.63 [3.83, 9.60]	6.57 [3.86, 9.65]
	Gemma-4-E4B	10.01 [8.54, 11.48]	7.81 [6.67, 8.98]	12.25 [9.10, 15.84]	12.37 [9.37, 15.50]

K.3 Frozen Free Generation

Table 29 reports free-generation outcomes separately for NQ train and dev. Each cell gives the flip percentage and its 95% CI. Eligibility requires successful generation, reproduction of the original correctness by the unperturbed baseline, and completed answer evaluation.

Both repair and reverse-direction damage occur in both behavioral groups across all models on train and dev. Rescue repair rates range from 54.59% to 81.82%. The main intervention is significantly stronger than all 12 controls for every model on train and for Qwen2.5-7B, Qwen2.5-0.5B, and Llama on dev; Gemma passes 9/12 dev comparisons. These results show that the frozen score-level intervention can alter correctness of complete generated responses on both NQ splits.

K.4 Component Directional Specificity

Tables 30–31 report the values represented by the NQ component-source plot for train and dev, respectively. Each entry is the mean first-divergence directional-effect difference between the paired same-question source and the equal-norm random source, followed by its 95% bootstrap CI, in nats. Positive values favor the paired source; negative values favor the random source. A dagger marks a comparison that is not significant after Holm correction within the model.

Table 29: Free-generation semantic flip rates (%) and 95% CIs under the frozen main intervention on NQ.

Split	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	44.23[42.90,45.58]	25.12[23.97,26.28]	54.59[51.23,57.83]	43.14[39.96,46.32]
	Qwen2.5-0.5B	47.89[46.45,49.35]	29.13[27.90,30.38]	57.99[53.86,62.12]	45.09[41.10,49.08]
	Llama-3.1-8B	62.35[60.07,64.64]	62.69[60.45,64.88]	76.83[72.68,80.98]	66.91[62.35,71.46]
	Gemma-4-E4B	52.78[51.03,54.50]	28.29[26.72,29.83]	72.60[69.32,75.89]	53.33[49.66,57.01]
Dev	Qwen2.5-7B	44.68[37.77,52.13]	18.50[13.00,24.00]	63.64[45.45,78.79]	47.22[30.56,63.89]
	Qwen2.5-0.5B	51.03[42.76,59.31]	31.90[25.15,39.26]	70.00[50.00,90.00]	47.83[26.09,69.57]
	Llama-3.1-8B	59.65[47.37,71.93]	55.36[42.86,67.86]	81.82[54.55,100.00]	81.82[54.55,100.00]
	Gemma-4-E4B	54.78[45.22,64.35]	18.64[11.86,26.27]	80.00[60.00,95.00]	30.00[10.00,50.00]

Table 30: Component source-specificity effects on the NQ train split (mean [95% CI], nats).

Model	Group / Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression / Repair	12.19 [11.90, 12.48]	4.47 [4.35, 4.60]	1.49 [1.41, 1.56]
	Regression / Damage	8.32 [8.10, 8.54]	3.28 [3.17, 3.38]	1.13 [1.05, 1.20]
	Rescue / Repair	8.71 [8.13, 9.30]	2.53 [2.33, 2.75]	1.58 [1.38, 1.79]
	Rescue / Damage	8.75 [8.17, 9.34]	2.70 [2.47, 2.95]	1.62 [1.42, 1.83]
Qwen2.5-0.5B	Regression / Repair	6.12 [6.01, 6.24]	2.82 [2.75, 2.89]	0.61 [0.57, 0.65]
	Regression / Damage	4.48 [4.38, 4.58]	2.03 [1.97, 2.09]	0.52 [0.48, 0.55]
	Rescue / Repair	2.82 [2.58, 3.06]	1.58 [1.46, 1.71]	0.27 [0.19, 0.34]
	Rescue / Damage	3.55 [3.29, 3.82]	1.90 [1.75, 2.04]	0.34 [0.25, 0.42]
Llama-3.1-8B	Regression / Repair	9.32 [8.97, 9.67]	0.79 [0.74, 0.84]	-0.39 [-0.47, -0.31]
	Regression / Damage	8.47 [8.16, 8.78]	1.12 [1.05, 1.20]	-0.30 [-0.38, -0.21]
	Rescue / Repair	5.59 [5.09, 6.11]	0.34 [0.27, 0.41]	-0.49 [-0.62, -0.36]
	Rescue / Damage	5.56 [5.07, 6.05]	0.50 [0.39, 0.63]	-0.50 [-0.64, -0.37]
Gemma-4-E4B	Regression / Repair	10.91 [10.57, 11.25]	5.74 [5.55, 5.93]	6.91 [6.68, 7.13]
	Regression / Damage	6.23 [6.01, 6.44]	4.41 [4.24, 4.57]	3.65 [3.50, 3.81]
	Rescue / Repair	7.54 [6.98, 8.11]	4.06 [3.72, 4.39]	5.16 [4.75, 5.57]
	Rescue / Damage	6.96 [6.47, 7.48]	3.69 [3.37, 4.01]	4.19 [3.86, 4.52]

The full-residual source advantage is positive in every model, condition, and split. Llama's MLP differences are negative throughout, whereas Gemma's are positive; several Qwen and Llama component contrasts on NQ dev do not pass Holm correction (Table 31). This preserves the overall same-question source pattern while making the component-level exceptions explicit.

Table 32 reports the paired-source Attention-minus-MLP effect differences on the NQ train and dev splits. Positive values favor attention and negative values favor MLP. A dagger marks a difference that is not significant after Holm correction; unmarked differences are significant.

All 16 train-split differences are significant after Holm correction. Their signs indicate attention advantages for Qwen2.5-7B, Qwen2.5-0.5B, and Llama, and MLP advantages for Gemma. The dev

Table 31: Component source-specificity effects on the NQ dev split (mean [95% CI], nats).

Model	Group / Direction	Full Residual	Attention	MLP
Qwen2.5-7B	Regression / Repair	12.28 [10.85, 13.74]	4.51 [3.98, 5.06]	1.30 [0.95, 1.69]
	Regression / Damage	8.59 [7.43, 9.76]	3.75 [3.22, 4.31]	0.88 [0.49, 1.29]
	Rescue / Repair	12.37 [9.01, 16.03]	3.22 [1.97, 4.65]	2.64 [1.67, 3.71]
	Rescue / Damage	11.07 [7.57, 14.88]	3.14 [1.84, 4.59]	1.73 [0.82, 2.75]
Qwen2.5-0.5B	Regression / Repair	5.68 [5.04, 6.33]	2.54 [2.23, 2.88]	0.51 [0.28, 0.74]
	Regression / Damage	4.18 [3.66, 4.69]	1.76 [1.44, 2.09]	0.46 [0.29, 0.63]
	Rescue / Repair	3.70 [2.43, 4.96]	1.56 [1.03, 2.15]	0.29 [-0.11, 0.71]†
	Rescue / Damage	4.08 [2.56, 5.54]	2.27 [1.45, 3.25]	0.44 [0.06, 0.87]†
Llama-3.1-8B	Regression / Repair	9.50 [7.62, 11.37]	0.68 [0.50, 0.88]	-0.14 [-0.53, 0.25]†
	Regression / Damage	8.52 [6.86, 10.21]	1.22 [0.82, 1.66]	-0.14 [-0.58, 0.30]†
	Rescue / Repair	6.73 [3.93, 9.75]	0.47 [0.25, 0.68]	-0.30 [-0.72, 0.10]†
	Rescue / Damage	6.20 [3.49, 9.24]	0.77 [0.34, 1.27]†	-0.29 [-0.96, 0.37]†
Gemma-4-E4B	Regression / Repair	9.61 [8.12, 11.14]	5.57 [4.61, 6.59]	6.57 [5.49, 7.64]
	Regression / Damage	5.65 [4.62, 6.72]	4.15 [3.38, 4.98]	3.27 [2.61, 3.96]
	Rescue / Repair	11.22 [7.47, 15.50]	5.90 [3.90, 7.96]	7.40 [4.40, 11.26]
	Rescue / Damage	8.48 [5.68, 11.56]	4.69 [2.77, 6.99]	5.37 [3.49, 7.40]

Table 32: Attention-minus-MLP paired-source effect differences on NQ (mean [95% CI], measured in nats). A dagger marks a difference that is not significant after Holm correction.

Split	Model	Regression Repair	Regression Damage	Rescue Repair	Rescue Damage
Train	Qwen2.5-7B	2.95 [2.83, 3.07]	2.11 [2.00, 2.22]	0.81 [0.54, 1.07]	0.86 [0.55, 1.16]
	Qwen2.5-0.5B	2.04 [1.97, 2.12]	1.51 [1.44, 1.59]	1.32 [1.20, 1.45]	1.63 [1.47, 1.78]
	Llama-3.1-8B	1.18 [1.09, 1.28]	1.45 [1.34, 1.57]	0.87 [0.72, 1.02]	0.98 [0.81, 1.14]
	Gemma-4-E4B	-0.70 [-0.81, -0.59]	-0.72 [-0.82, -0.62]	-2.21 [-2.49, -1.94]	-1.20 [-1.40, -1.00]
Dev	Qwen2.5-7B	3.01 [2.48, 3.58]	2.54 [2.04, 3.03]	0.91 [-0.63, 2.38]†	1.52 [0.06, 3.06]†
	Qwen2.5-0.5B	1.78 [1.42, 2.12]	1.14 [0.77, 1.53]	1.20 [0.75, 1.66]	2.00 [1.22, 2.83]
	Llama-3.1-8B	0.86 [0.48, 1.28]	1.45 [0.88, 2.08]	0.73 [0.38, 1.08]	1.10 [0.43, 1.95]
	Gemma-4-E4B	-0.93 [-1.52, -0.38]	-0.55 [-1.07, -0.03]	-3.38 [-5.97, -1.11]	-2.16 [-3.82, -0.72]

split preserves these directions for Qwen2.5-0.5B, Llama, and Gemma; Qwen2.5-7B's two Rescue differences are daggered and therefore do not reach significance after correction.

In the NQ mechanistic analyses, out-of-memory failures occur only for Gemma. On train, 52 samples fail during layer-wise scoring and 48 during probe-feature extraction; among 4,270 transition samples, 220 (5.15%) fail in residual patching, 254 (5.95%) in free generation, and 237 (5.55%) in component patching. On dev, 10 of 149 transition samples fail in each of these three intervention analyses. The longer full-context inputs require intermediate activations for state extraction and patching, whose additional memory allocations exceed the available single-GPU memory for

these cases. All estimates use successfully completed samples, and free-generation denominators additionally require baseline correctness reproduction and answer evaluation.