REQUIRED OR REDUNDANT? TRACING EVIDENCE COLLABORATION WITH FACTORIZED CAUSAL ACTIVATION PATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often receive multiple pieces of evidence, but standard analyses measure each piece's marginal influence. This cannot tell us whether evidence is integrated independently, redundantly, or collaboratively. We introduce **Factorized Causal Activation Patching**, a framework that tests whether one evidence statement's causal contribution depends on the state of another. We build a controlled 400-claim fact-verification dataset with matched conditions in which two evidence statements are either *jointly required* or *individually sufficient*. A 2×2 factorial intervention manipulates both statements independently, yielding a dependency contrast between their non-additive interactions. In Qwen3.5-9B and Gemma-4-12B-it, semantic counterfactuals produce strong dependency contrasts of $+3.81$ and $+4.19$ gold-margin units. Whole-word permutations produce smaller positive contrasts of $+0.90$ and $+0.32$, showing that collaboration is detectable under structural as well as semantic perturbation; we do not treat their magnitudes as calibrated. Bidirectional patching shows that the source evidence causally affects the decision, and at layer 0 the patching estimator exactly recovers the behavioral interaction. Position-wise patching reveals a staged profile: partner-conditioned signal is causally recoverable at evidence spans early and at the final decision state late, while claim-endpoint effects are weaker and label-heterogeneous. Impossible reverse autoregressive paths remain exactly null. These results distinguish partner-conditioned collaboration from marginal evidence sensitivity.

1 INTRODUCTION

LLMs increasingly use external evidence after forming an initial answer distribution (Lewis et al., 2020; Gao et al., 2024; Kawada & Kellis, 2026). A model may represent a fact correctly but fail to use it (Kawada & Kellis, 2026), or use one piece of evidence well but miscombine it with another (Longpre et al., 2021; Xu et al., 2024). The core problem is that most behavioral and mechanistic analyses measure each evidence statement's marginal influence (Vig et al., 2020; Meng et al., 2022; Geva et al., 2023). That tells us whether a statement matters, but not whether its effect depends on the other statement.

Consider the claim **"The bridge is in the United States."** In the **jointly required** condition, E1 ("The bridge crosses the Mississippi River") and E2 ("The Mississippi River flows entirely within the United States") entail the claim only together. In the **individually sufficient** condition, either E1 ("The bridge is in New York") or E2 ("The bridge is operated by the Port Authority of New York and New Jersey") entails it alone. We manipulate each statement in a 2×2 factorial: a **semantic counterfactual** changes meaning, while a **word-order permutation** changes structure but preserves words. Crossing these interventions with dependency condition lets us ask whether one statement's causal contribution depends on the other's state.

We introduce **Factorized Causal Activation Patching** to answer this question. The method embeds activation patching in a factorial design. At the behavioral level, we estimate the non-additive interaction between the two evidence statements and contrast it across dependency conditions. At the mechanistic level, we patch each evidence representation in both disruption and restoration di-

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.

rections while varying the partner's state. At layer 0, this patching contrast exactly recovers the behavioral factorial interaction. Applying the same estimand at downstream readout positions connects behavioral dependency to internal information flow.

Across Qwen3.5-9B and Gemma-4-12B-it, multi-evidence decisions are not formed by simply accumulating independent evidence contributions. Semantic counterfactuals produce strong dependency contrasts, while whole-word permutations produce smaller but positive contrasts. Marginal bidirectional patching confirms that source representations causally affect the decision; the factorized contrast isolates the partner-conditioned component and tracks where it remains causally recoverable across depth. Position-wise interventions reveal a staged profile: partner-conditioned signal is recoverable at evidence spans in early layers and at the final decision state in late layers, with weaker and label-heterogeneous claim-endpoint effects. Impossible reverse autoregressive paths remain exactly null. Patch-specificity controls rule out generic explanations based on activation replacement, boundary tokens, or donor length.

This work makes four contributions. First, we formalize evidence collaboration as a partner-conditioned causal effect and connect it exactly to a behavioral 2×2 interaction. Second, we introduce Factorized Causal Activation Patching, which turns activation patching from a marginal influence test into a test of evidence collaboration. Third, we build a controlled 400-claim fact-verification dataset with matched dependency regimes and evaluate it on two model families under semantic and structural interventions. Fourth, we provide a mechanistic account in which dependency-sensitive information is distributed across source evidence representations, intermediate claim readouts, and the final decision state.

2 Related Work

Multi-evidence verification and input attribution. FEVER and AVeriTeC require models to use several evidence items (Thorne et al., 2018; Schlichtkrull et al., 2023), motivating architectures that exchange information across evidence sentences (Zhou et al., 2019; Liu et al., 2020). Predictive gains and attention patterns do not establish that one item's causal contribution depends on another. Perturbation and attribution methods identify influential inputs (Ribeiro et al., 2016; Sundararajan et al., 2017; Lundberg & Lee, 2017), rationale benchmarks test whether selected evidence is sufficient or necessary (DeYoung et al., 2020), and related work tests whether generated explanations causally support answers (Yeo et al., 2025). These approaches measure each item's marginal effect and cannot distinguish independent, redundant, and collaborative integration. We study that interaction directly.

Knowledge conflict and evidence arbitration. A parallel line studies arbitration between parametric memory and supplied context (Longpre et al., 2021; Xu et al., 2024), showing that uptake varies with prior strength (Mallen et al., 2023) and source cues such as provenance and wording (Tan et al., 2024). Recent work detects and addresses conflicting retrieved sources (Cattan et al., 2025; Lee et al., 2025), and mechanistic analyses localize where conflicting knowledge is stored (Pham et al., 2026; Li et al., 2025). Concurrent work formalizes evidence integration as a distributional shift governed by a receiver prior weight and an evidence tilt (Kawada & Kellis, 2026). That framework identifies the functional form of the shift but treats it as a single scalar and does not localize where in the network it is computed. We instead hold the form fixed and use a factorial causal intervention to trace the dependency signal through the network.

Feature interaction and activation patching. Feature-interaction methods quantify departures from additive attribution. Shapley–Taylor indices give an axiomatic interaction decomposition (Sundararajan et al., 2020), and Integrated Hessians extend attribution to pairwise neural interactions (Janizek et al., 2021). Our four-condition comparison is the order-two interaction for two evidence-coherence factors, but an output score alone cannot show which internal states causally carry that dependence. Causal mediation, interchange interventions, and activation patching distinguish encoded from used information (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022), supporting circuit localization (Wang et al., 2023), path-specific analysis (Goldowsky-Dill et al., 2023), and automated circuit discovery (Conmy et al., 2023); conclusions can depend on the corruption and output metric (Zhang & Nanda, 2024). These approaches normally patch in one context. Our central step applies the same patch under two controlled partner states and subtracts, changing the question from

"does this state matter?" to "how much more does it matter when the partner is coherent?" We patch in both directions to avoid dependence on the choice of corruption.

Computation across depth. Mechanistic studies show that factual recall, entity binding, compositional reasoning, and fact verification develop through several layers (Geva et al., 2023; Feng & Steinhardt, 2024; Yang et al., 2025; Biran et al., 2024). For example, factual recall is typically attributed to mid-layer MLP writes, compositional binding to earlier attention heads, and verification signals to both ends of the network. These analyses trace a single clean-corrupted contrast, parametric knowledge, synthetic binding, or bridge entities, and do not isolate how an internal causal effect changes with the state of a second supplied item. Factorized patching connects output-level interaction to internal intervention and follows that dependence across positions and depth.

Positioning. Whereas standard activation patching measures the marginal causal influence of individual components (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Geva et al., 2023), we introduce a factorial estimand that measures how an evidence representation's causal contribution changes with the state of its partner. An exact layer-zero bridge links this partner-conditioned patching contrast to the behavioral 2×2 interaction, unifying output-level evidence interaction with internal causal intervention. Tracing the estimand across layers and positions reveals an architecture-dependent process in which collaboration moves from early source representations through intermediate claim readouts to the late decision state, providing a mechanistic account rather than another marginal attribution map.

2.1 Problem Formulation

We study binary fact verification with a claim c, two evidence statements E_1 and E_2, and a gold label $y \in \{\text{SUPPORTS}, \text{REFUTES}\}$. Our objective is not merely to measure whether each statement affects the prediction, but whether the causal contribution of one statement depends on the state of the other. We refer to this partner-dependent contribution as *evidence collaboration*.

Each claim is instantiated under a dependency regime $r \in \{\text{req}, \text{suf}\}$. In the *jointly required* regime (req), neither statement is sufficient alone and the decision requires their conjunction. In the *individually sufficient* regime (suf), either statement independently establishes the same gold label. The claim and label are held constant across regimes, isolating evidence dependency from claim content.

Let $z_i \in \{0, 1\}$ denote the state of E_i, where $z_i = 0$ is the original evidence and $z_i = 1$ is an intervened version. The intervention instantiates one of the families described in Section **??** (semantic counterfactual or word-order permutation), specified per experiment, provided it preserves the matched construction required by the bridge in Section 2.4. Independently varying the two states produces four factorial cells, written $x^{00}, x^{10}, x^{01}, x^{11}$, where the superscript denotes (z_1, z_2). We suppress the regime, order, and answer-mapping subscripts until they are needed and average over the latter two in all claim-level estimands.

The model is evaluated through the gold-oriented logit margin

$$G(x^{z_1 z_2}) = \ell_y(x^{z_1 z_2}) - \ell_{\bar{y}}(x^{z_1 z_2}),$$

where ℓ_y and $\ell_{\bar{y}}$ are the logits assigned to the gold and opposite labels after undoing the answer-token mapping. This orientation makes effects comparable across labels and mappings.

2.2 Behavioral Factorial Interaction

A marginal intervention effect measures what happens when one evidence statement is changed while fixing the other. Evidence collaboration instead requires a difference-in-differences: whether the effect of changing E_1 depends on the state of E_2, equivalently whether the effect of changing E_2 depends on E_1. We define the behavioral factorial interaction as

$$\mathcal{I}_r = G(x^{00}) - G(x^{10}) - G(x^{01}) + G(x^{11}),$$

where the right-hand side is evaluated within regime r. Equivalently, $\mathcal{I}_r$ is the difference between the effect of corrupting E_1 with an original partner and the effect of corrupting E_1 with an intervened partner:

$$\mathcal{I}_r = \big[G(x^{00}) - G(x^{10})\big] - \big[G(x^{01}) - G(x^{11})\big].$$

Thus, $\mathcal{I}_r$ is positive when corrupting E_1 reduces the gold margin more when E_2 is original than when E_2 is intervened. The same quantity is obtained by exchanging E_1 and E_2, so the estimand measures mutual interaction rather than a directional dependence of one item on the other. We therefore average over evidence roles without conflating two distinct quantities.

Our primary behavioral estimand is the *dependency contrast*

$$\Delta_{\text{beh}} = \mathbb{E}[\mathcal{I}_{\text{req}}] - \mathbb{E}[\mathcal{I}_{\text{suf}}] \,.$$

This contrast imposes a stricter criterion than detecting non-additivity alone: collaboration must be stronger when the task requires both statements than when either statement is independently sufficient.

2.3 Factorized Causal Activation Patching

Conventional activation patching measures the marginal effect of replacing an internal representation (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023). We retain the same patching operator but embed it in the factorial design, thereby testing whether an internal causal contribution depends on the partner evidence. The order-two interaction has the same functional form as the Shapley–Taylor index for two binary factors (Sundararajan et al., 2020), but is defined on causal interventions rather than model attributions, which is what allows an internal locus to be tested rather than an output score alone.

Let $h_S^{(\ell)}(x)$ denote the residual-stream activation at layer ℓ and token set S for prompt x. Write $\text{swap}_S^{(\ell)}(x \leftarrow x')$ for a forward pass on x in which $h_S^{(\ell)}(x)$ is replaced by $h_S^{(\ell)}(x')$, leaving all other activations unchanged. The token set S_i denotes the evidence positions that differ between the two cells being patched; it occupies the same absolute positions in donor and recipient, guaranteed by the matched factorial construction.

For evidence role i, regime r, and partner state $s \in \{0, 1\}$, we patch the changed span of E_i in both directions. The disruption effect injects the intervened representation into the original prompt, and the restoration effect injects the original representation into the intervened prompt:

$$D_{i,r,s}^{(\ell)} = G\Big(\text{swap}_{S_i}^{(\ell)}\big[x^{s,0} \leftarrow x^{s,1}\big]\Big) - G(x^{s,0}), \quad R_{i,r,s}^{(\ell)} = G\Big(\text{swap}_{S_i}^{(\ell)}\big[x^{s,1} \leftarrow x^{s,0}\big]\Big) - G(x^{s,1}).$$

For $s = 0$ this uses the pair (x^{00}, x^{10}); for $s = 1$, (x^{01}, x^{11}). The partner state and all remaining prompt factors are fixed across donor and recipient.

Because disruption and restoration use different recipient baselines, we combine them into the symmetric magnitude

$$M_{i,r,s}^{(\ell)} = \frac{R_{i,r,s}^{(\ell)} - D_{i,r,s}^{(\ell)}}{2}.$$

The halving is not arbitrary. At layer 0, substituting the definitions above cancels the double-counting and yields $M_{i,r,s}^{(0)} = G(x^{s,0}) - G(x^{s,1})$, so M coincides with the marginal effect of the intervention on the gold margin. At deeper layers the identity is approximate but the scale remains comparable to the behavioral interaction.

The partner-conditioned causal effect is $P_{i,r}^{(\ell)} = M_{i,r,0}^{(\ell)} - M_{i,r,1}^{(\ell)}$. It measures how much the causal contribution of E_i changes when its partner moves from the original to the intervened state. Averaging over evidence roles, evidence orders, and answer mappings within a single intervention family gives $P_r^{(\ell)}$, and the layer-specific dependency contrast is

$$\Delta_{\text{patch}}^{(\ell)} = \mathbb{E}\Big[P_{\text{req}}^{(\ell)}\Big] - \mathbb{E}\Big[P_{\text{suf}}^{(\ell)}\Big] \,.$$

Factorized patching is not a new replacement operator; it is a causal design that converts marginal activation patching into a test of partner-conditioned, non-additive interaction.

2.4 Behavioral–Mechanistic Bridge

The behavioral and mechanistic estimands are connected exactly at the input to the first transformer layer. Under the matched construction of the factorial cells, the donor and recipient share identical

token IDs at every position outside the changed span S_i. Layer-0 patching therefore reproduces the donor's input token IDs at every position, and the patched activation is numerically identical to the donor's to within floating-point tolerance.

For E_1 with an original partner, substituting the layer-0 definitions cancels the halving and gives $M^{(0)}_{1,r,0} = G(x^{00}) - G(x^{10})$. With an intervened partner, $M^{(0)}_{1,r,1} = G(x^{01}) - G(x^{11})$. Their partner-conditioned difference is therefore

$$P^{(0)}_{1,r} = M^{(0)}_{1,r,0} - M^{(0)}_{1,r,1} = G(x^{00}) - G(x^{10}) - G(x^{01}) + G(x^{11}) = \mathcal{I}_r.$$

The same identity holds when patching E_2. Consequently, $P^{(0)}_r = \mathcal{I}_r$ and $\Delta^{(0)}_{\text{patch}} = \Delta_{\text{beh}}$. This is the *bridge identity*: at layer 0 the mechanistic patching estimand and the behavioral factorial interaction are the same quantity. At deeper layers equality is no longer guaranteed: $\Delta^{(\ell)}_{\text{patch}}$ instead measures how the factorial interaction is preserved, transformed, or lost within the network.

2.5 Position-Conditioned Information Flow

Source-span patching identifies where collaboration is causally present in an evidence representation, but not where that information becomes accessible for downstream computation. We therefore generalize the patching target from the source span S_i to an arbitrary token position t. Donor and recipient prompts still differ only in the state of E_i, but we replace the activation at t rather than at the full changed span. The construction requires that the surface token at t is identical across donor and recipient, so that layer-0 activations at t coincide and any divergence at deeper layers must arise through contextual propagation from the manipulated evidence.

Using the same bidirectional construction, we define $M^{(\ell)}_{i,r,s,t}$, followed by

$$P^{(\ell)}_{i,r,t} = M^{(\ell)}_{i,r,0,t} - M^{(\ell)}_{i,r,1,t}, \qquad \Delta^{(\ell)}_{\text{flow},t} = \mathbb{E}\left[P^{(\ell)}_{\text{req},t}\right] - \mathbb{E}\left[P^{(\ell)}_{\text{suf},t}\right].$$

The resulting position-by-layer profile indicates where dependency-sensitive information is causally readable as computation proceeds from the evidence representations toward the output state.

These effects are not assumed to be conserved across positions and should not be interpreted as classical mediation proportions. A signal that is causally read out at the final token may have been computed at the source span and then overwritten; the magnitude at the final token is not the fraction of the source-span effect that survives, but rather the causal influence of the final-token state considered independently. The specific source spans, downstream positions, layers, intervention families, and architectural falsification controls used to instantiate these estimands are described separately in the experimental setup.

3 Experimental Setup

Controlled factorial dataset. Our primary evaluation uses 400 manually constructed and audited claims spanning eight domains: geography, history, medicine, physics, politics, psychology, religion, and world wars, with 50 claims per domain. The dataset contains 280 SUPPORTS and 120 REFUTES claims. Each claim is paired with two matched evidence conditions: one in which both statements are jointly required and one in which they are individually sufficient. The claim and gold label remain fixed across conditions. For each evidence statement, we provide an original and a meaning-changing counterfactual version, and crossing the two evidence states yields four factorial cells. We additionally cross both evidence orders ($E_1 \to E_2$ and $E_2 \to E_1$) and both assignments of SUPPORTS/REFUTES to answer tokens A/B. The complete design contains $400 \times 2 \times 4 \times 2 \times 2 = 12{,}800$ prompts. All evidence pairs, counterfactual spans, labels, orders, and factorial relations were validated before model inference.

Models and decision scoring. We evaluate Qwen3.5-9B, a 32-layer hybrid architecture, and Gemma-4-12B-it, a 48-layer decoder-only transformer. Models are evaluated in their instruction-tuned chat formats using `bfloat16` activations. We use forced-choice scoring rather than free-form generation: each model assigns logits to the verified single-token verbalizers A and B. After reversing the randomized answer mapping, the primary outcome is the gold-oriented logit margin

$\ell_y - \ell_{\bar{y}}$. Counterbalancing answer mappings prevents fixed token preferences from being confounded with the factual label.

Evidence interventions. We use two complementary intervention families. *Semantic counterfactuals* replace a localized proposition-bearing span with a fluent alternative that changes the evidence's implication while preserving the surrounding sentence. *Whole-word order permutations* disrupt compositional structure while preserving lexical content. Permutations operate on word blocks defined from tokenizer offsets: all subword tokens belonging to a word move together, so words are never fragmented. We generate deterministic derangements using four seeds (42–45), reject permutations with insufficient displacement, and audit sampled outputs for fragmentation, disruption, and interpretability. Behavioral analyses aggregate across all four seeds; mechanistic word-order experiments use the frozen seed-42 intervention. We do not claim token-level semantic correspondence for permuted spans.

Activation-patching protocol. We patch the `resid_pre` stream at the token positions altered by the intervention: the aligned counterfactual span for semantic interventions and the complete evidence span for word-order interventions. Every edge is patched in both disruption and restoration directions, for both evidence roles and both partner states. We evaluate layers $\{0, 8, 16, 24, 31\}$ for Qwen and $\{0, 12, 24, 36, 47\}$ for Gemma, covering the input, early, middle, late, and final portions of each network. Before full inference, we validate hook placement, tensor shapes, unpatched forward equivalence, and stored-prompt replay. At layer 0, a complete source-span patch must reconstruct the donor output; we verify this against the full-vocabulary logits and treat any reconstruction failure as an invalid run. All analyses use raw gold-margin effects rather than recovery ratios.

Position-flow protocol. To trace collaboration beyond the source span, we patch one-token readout positions at the endpoint of the later evidence statement, the claim endpoint, and the final prompt token immediately preceding prediction. We reuse the same factorial edges, bidirectional effects, sampled layers, and partner-conditioned estimand as in source-span patching. Two architectural invariants serve as hard falsification controls. First, downstream positions must be exactly null at layer 0, before contextual information can propagate from the manipulated evidence. Second, an evidence statement occurring later in the autoregressive sequence cannot affect the earlier partner endpoint. This impossible reverse path must remain null at every layer. Site-to-site effects are interpreted as causal accessibility, not as conserved mediation proportions.

Patch-specificity controls. We test whether the word-order patching results reflect evidence content rather than arbitrary activation replacement. Controls are evaluated at the strongest early collaboration layer: layer 8 for Qwen and layer 12 for Gemma. *Dose-response patches* interpolate donor activations at $\alpha \in \{0.25, 0.50, 0.75\}$, with the full patch providing $\alpha = 1$. *Boundary controls* patch only the first and last evidence tokens. *Self-patches* replace an activation with itself and constitute a hard sham null. Finally, *wrong-content controls* use an exact-length donor from a different example while matching the factorial cell, regime, label, evidence order, and answer mapping. Donors are frozen before inference. Exact matching yields complete control cohorts of 380 Qwen claims and 384 Gemma claims, corresponding to 95% and 96% coverage of the controlled set.

Natural external-validity evaluation. We complement the controlled study with natural two-evidence examples from FEVER, AVeriTeC, and SciFact. Because these datasets do not provide matched dependency regimes, we first behaviorally screen 8,649 examples using both evidence statements, each statement alone, both evidence orders, both answer mappings, and a claim-only condition. An example enters a model-specific individually sufficient cohort only when the model reliably predicts the gold label from either evidence statement alone and from their combination, while the claim-only condition does not already solve the task. This procedure yields 698 Qwen and 980 Gemma sufficient-evidence examples. After enforcing whole-word permutation eligibility, the analyzed cohorts contain 657 Qwen examples (48 AVeriTeC, 595 FEVER, and 14 SciFact) and 916 Gemma examples (61 AVeriTeC, 839 FEVER, and 16 SciFact). We apply the complete behavioral 2×2 permutation design with four seeds, both orders, and both answer mappings. Because natural data contain only the sufficient regime, this experiment evaluates external behavioral robustness rather than a natural required-minus-sufficient dependency contrast.

Table 1: Controlled behavioral factorial interactions under semantic counterfactuals (Sem.) and whole-word order permutations (Perm.). $\mathcal{I}_{\text{req}}$ and $\mathcal{I}_{\text{suf}}$ are regime-specific interactions; $\Delta_{\text{dep}} = \mathcal{I}_{\text{req}} - \mathcal{I}_{\text{suf}}$. Brackets: 95% stratified claim-bootstrap intervals. $P(\Delta_i > 0)$ is the fraction of claims with positive contrast.

Model	Int.	$\mathcal{I}_{\text{req}}$	$\mathcal{I}_{\text{suf}}$	Δ_{dep}	$P(\Delta_i > 0)$
Qwen	Sem.	$+2.57\,[+2.43, +2.71]$	$-1.24\,[-1.46, -1.01]$	$\mathbf{+3.81}\,[+3.55, +4.07]$	91.3%
Gemma	Sem.	$+3.73\,[+3.55, +3.92]$	$-0.46\,[-0.79, -0.13]$	$\mathbf{+4.19}\,[+3.80, +4.57]$	85.5%
Qwen	Perm.	$+0.94\,[+0.86, +1.01]$	$+0.03\,[-0.01, +0.08]$	$\mathbf{+0.90}\,[+0.81, +1.00]$	82.0%
Gemma	Perm.	$+0.96\,[+0.85, +1.06]$	$+0.64\,[+0.56, +0.74]$	$\mathbf{+0.32}\,[+0.18, +0.44]$	50.3%

Table 2: Natural sufficient-evidence cohorts under whole-word permutation. Pooled estimates are claim weighted and dominated by FEVER.

Model	Dataset	n	Interaction
Qwen	AVeriTeC	48	$-0.09\,[-0.33, +0.14]$
	FEVER	595	$\mathbf{+0.31}\,[+0.24, +0.37]$
	SciFact	14	$+0.23\,[-0.07, +0.50]$
	Pooled	657	$\mathbf{+0.28}\,[+0.21, +0.34]$
Gemma	AVeriTeC	61	$-0.17\,[-0.56, +0.19]$
	FEVER	839	$\mathbf{+0.37}\,[+0.28, +0.46]$
	SciFact	16	$\mathbf{-0.71}\,[-1.27, -0.26]$
	Pooled	916	$\mathbf{+0.32}\,[+0.23, +0.41]$

Statistical analysis. The claim is the unit of analysis. We first average intervention effects over answer mappings, evidence orders, evidence roles, directions, and permutation seeds as appropriate, preventing repeated prompt variants from being treated as independent observations. We report means with 95% percentile intervals from 10,000 stratified claim-level bootstrap resamples. Controlled analyses are stratified by domain; natural-data analyses are stratified by dataset and gold label, with dataset-specific intervals also reported. Label-stratified and equal-label summaries are used where label imbalance could determine a pooled trajectory. All inference includes prompt-replay audits, choice-flip counts, and no-choice-flip sensitivity analyses when replay is not exact. Transition, relocation, and patch-specificity analyses formulated after observing the primary profiles are explicitly treated as post-completion descriptive analyses; we report bootstrap intervals without introducing post hoc significance tests.

4 Results

4.1 Joint Requirement Produces Stronger Behavioral Collaboration

Table 1 reports the behavioral factorial interactions. Under semantic counterfactuals, both models separated the two regimes as predicted: positive required interactions and negative sufficient interactions, yielding dependency contrasts of $+3.81$ for Qwen and $+4.19$ for Gemma, positive in 91.3% and 85.5% of claims respectively. Both models therefore treat jointly required evidence as conjunctive and individually sufficient evidence as compensatory.

Whole-word order permutation produced the same directional result with smaller, more model-dependent contrasts. Qwen showed a positive required interaction and a sufficient interaction near zero, giving a contrast of $+0.90$; Gemma's sufficient interaction remained positive, leaving a smaller contrast of $+0.32$ with only 50.3% of claim-level contrasts above zero. This asymmetry suggests that Qwen's structural sensitivity is regime-gated: structural corruption matters only when both statements are logically required. Gemma's is not, indicating that its evidence integration is less conditioned on logical role.

The sufficient-regime interaction also flips sign between intervention families in both models (negative under semantic counterfactuals, non-negative under permutation). Compensatory processing therefore appears specific to semantic corruption; structural disruption invokes a different response.

Table 3: Semantic-counterfactual dependency contrast at four causal-readout locations. Layer numbers shown above each estimate. Both architectural null controls (layer-0 downstream, impossible reverse autoregressive path) passed exactly in both models.

Model	Layer-0 bridge	Early source	Claim endpoint	Final position
Qwen	L0 $+\mathbf{3.81}$ $[+3.55, +4.07]$	L8 $+1.13$ $[+0.97, +1.29]$	L16 $+0.45$ $[+0.39, +0.51]$	L31 $+\mathbf{2.55}$ $[+2.37, +2.73]$
Gemma	L0 $+\mathbf{4.19}$ $[+3.80, +4.57]$	L12 $+4.19$ $[+3.81, +4.56]$	L24 $+0.99$ $[+0.82, +1.15]$	L47 $+\mathbf{4.19}$ $[+3.81, +4.58]$

Table 4: Patch-specificity controls at the strongest early source layer (Qwen L8; Gemma L12). Full–boundary and full–wrong compare the correct full-span effect with the corresponding control. Self-patching is an exact sham null.

Model	Dose slope	Full–boundary	Full–wrong	Self	Pos. slope / Monotonic
Qwen	$+\mathbf{0.76}$ $[+0.67, +0.86]$	$+\mathbf{0.72}$ $[+0.63, +0.81]$	$+\mathbf{2.02}$ $[+1.82, +2.22]$	0.000	78.9% / 44.7%
Gemma	$+\mathbf{0.38}$ $[+0.18, +0.57]$	$+\mathbf{0.53}$ $[+0.35, +0.71]$	$+\mathbf{1.85}$ $[+1.46, +2.24]$	0.000	51.8% / 23.4%

Because the two families alter different properties of the evidence, their raw magnitudes are not treated as calibrated effect sizes.

4.2 Natural Evidence Shows Dataset-Dependent Interaction

We next tested whole-word permutation on natural examples behaviorally screened as individually sufficient. Pooled interactions were positive for both Qwen ($+0.28$, $n = 657$) and Gemma ($+0.32$, $n = 916$), and stable across permutation seeds, but dataset-specific results concentrated this apparent replication in FEVER (Table 2).

On FEVER both models produced positive interactions ($+0.31$ Qwen, $+0.37$ Gemma). AVeriTeC estimates were negative but inconclusive for both models; SciFact was inconclusive for Qwen and negative for Gemma (-0.71, $n = 16$). FEVER dominated both cohorts (90.6% Qwen, 91.6% Gemma), so pooled estimates should not be read as uniform replication.

Non-additive structural interaction therefore generalizes to natural sufficient-evidence FEVER examples in both models, but its direction is not stable across datasets. It also shows collaboration is not the same as logical necessity, since a positive interaction can persist when either statement is behaviorally sufficient. Seven Qwen and one Gemma screening-replay choice flips further qualify these descriptive estimates.

4.3 Factorized Patching Exactly Recovers the Behavioral Interaction

Factorized patching satisfied the behavioral–mechanistic bridge. At layer 0, patching the full altered source span reconstructed the factorial donor exactly: zero maximum full-vocabulary-logit error, and the patching partner modulation matched the behavioral factorial interaction with zero numerical error across 6,400 role-conditioned comparisons per model (Table 3). The bridge held for both semantic ($+3.81$ Qwen, $+4.19$ Gemma) and word-order interventions ($+0.96$ Qwen, $+0.38$ Gemma).

Source-span profiles then diverged across depth. Qwen's semantic contrast fell from $+3.81$ at layer 0 to $+1.13$ at layer 8 and was near zero by layer 16. Gemma retained its contrast through layer 12, reversed at layer 24 (-0.50), and approached zero thereafter. Word-order interaction followed a smaller related profile: Qwen positive at layer 8 only, Gemma positive through layers 12 and 24.

The two models therefore differ in how long the source representation carries the interaction—Qwen's conjunctive window closes by layer 16, Gemma's by layer 24—but both resolve the interaction near the midpoint of network depth, suggesting the timing is architectural rather than model-specific. Qwen reproduced its stored predictions exactly; Gemma's cross-server drift did not affect the conclusion (excluding stored-choice flips gave a layer-0 contrast of $+4.20$ and the same trajectory).

4.4 Collaboration Becomes Accessible at the Final Decision State

Position-conditioned patching revealed a staged source-to-decision profile. Both hard architectural controls passed exactly: downstream layer-0 patches had zero activation distance and zero margin effect, and later evidence never affected an earlier partner endpoint through the impossible reverse autoregressive path.

For Qwen, little stable interaction accumulated at the later-evidence endpoint. A dependency-sensitive signal became accessible at the claim endpoint at layer 16 ($+0.45$) and appeared strongly at the final prompt token at layer 31 ($+2.55$). Gemma showed a later trajectory: claim endpoint peak at layer 24 ($+0.99$), strong final-token effect at layer 47 ($+4.19$).

Claim-position effects were label-heterogeneous. At the peak claim layer, Qwen's contrast was -0.19 for REFUTES and $+0.72$ for SUPPORTS; Gemma's was -0.94 and $+1.82$. After equal weighting, Qwen retained a positive claim-minus-source contrast at layer 16 ($+0.20$) whereas Gemma's at layer 24 was inconclusive (-0.06). The claim-endpoint effect is therefore a label-specific asymmetry that survives pooling only because SUPPORTS claims dominate. In contrast, the label-balanced final-minus-claim contrast was positive in both models ($+2.38$ Qwen, $+4.97$ Gemma). The only label-balanced relocation signal is the final-token effect, and the robust conclusion is that dependency-sensitive information becomes causally accessible at the final prediction state after disappearing from the late source representations—not that both models share an identical intermediate pathway.

4.5 Collaboration Depends on Patch Strength, Span, and Content

Patch-specificity controls ruled out several generic explanations (Table 4). Dose response was nonlinear and saturated near 75% replacement, with positive slopes for Qwen ($+0.76$) and Gemma ($+0.38$). Saturation below full replacement indicates the dependency signal is distributed across the patched span, so replacing three-quarters already recovers most of the accessible effect. The graded pattern was more consistent for Qwen (78.9% positive slopes vs. 51.8%), and strictly monotonic trajectories were uncommon (44.7% vs. 23.4%); we therefore interpret the dose result as a nonlinear population-average response, not as universally monotonic at the claim level.

The effect also depended on patch extent and donor identity. Boundary-only patches produced small (Qwen) or negative (Gemma) effects, and paired full-minus-boundary contrasts were positive for both models. Exact-length wrong-content donors produced strongly negative contrasts (-1.26 Qwen, -1.48 Gemma) rather than reproducing the full-span effect, and self-patching produced exactly zero error in both models. Collaboration is therefore not explained by merely altering activations, patching evidence boundaries, or inserting any representation of the correct length: recovering the effect requires the appropriate evidence content across the full span.

5 Discussion

Evidence contributions are relational rather than independent. The central lesson of our study is that language models do not necessarily combine multiple evidence statements as independent votes with fixed contributions. Instead, the causal effect of one statement depends on the state of the other: changing E_1 has a different effect when E_2 is intact than when E_2 is also disrupted. This partner dependence was substantially stronger when both statements were jointly required than when either was individually sufficient, producing semantic dependency contrasts of 3.81 in Qwen3.5-9B and 4.19 in Gemma-4-12B-it. Consequently, an evidence statement does not possess a context-independent causal importance that can be fully characterized by a single marginal patching effect.

Collaboration is distributed across a processing trajectory. The interaction was not permanently localized to one layer or component. It was first expressed strongly at the source evidence spans, diminished there as depth increased, appeared at the claim endpoint around mid-depth, and became strongest at the final prompt position in late layers. For Qwen, the claim-endpoint contrast peaked at layer 16 and the final-position contrast reached 2.55 at layer 31; for Gemma, the corresponding effects appeared at layer 24 and reached 4.19 at layer 47. We interpret this pattern as a staged transformation of the causal interaction: the dependency is initially encoded where the

evidence occurs, subsequently becomes available to a claim-level representation, and is ultimately expressed in the state from which the decision is produced. The experiment traces the spatial distribution of the same factorial estimand, although it does not establish that an unchanged representational object is literally transported between these positions.

The processing schedule is architecture-dependent. The overall progression was shared across the two models, but its timing differed. Qwen's source-level interaction declined sharply by the middle of the network, whereas Gemma retained a strong source-span interaction through approximately the first quarter of its layers before the transition at mid-depth. Thus, the important cross-model regularity is not a particular causal layer, but the broader progression from source encoding to downstream decision expression. Mechanistic claims based on a single normalized layer or architecture may therefore miss meaningful differences in how long models retain locally accessible evidence interactions.

Collaboration is not reducible to generic corruption sensitivity. Semantic counterfactuals produced substantially larger dependency contrasts than whole-word permutation, while permutation still reproduced the predicted required-versus-sufficient difference (0.90 for Qwen and 0.31 for Gemma). The weaker and more model-dependent permutation results show that disrupting any feature of the evidence is insufficient to explain the semantic effects. In natural sufficient-evidence cohorts, pooled permutation interactions were positive for both models (0.276 and 0.317), but these estimates were dominated by FEVER and varied across datasets. These natural results support external relevance while also delimiting the claim: because the natural cohorts contain only individually sufficient examples, they test interaction under sufficiency rather than the required-minus-sufficient dependency contrast identified by the controlled experiment.

Implications for mechanistic interpretation. Standard activation patching asks whether replacing a representation changes an output. Our results show that the answer can depend critically on what accompanying evidence remains available. A component that appears weak under one partner state may become causally important under another, and a large marginal effect need not indicate genuine evidence collaboration. Factorizing the intervention over both evidence states separates marginal sensitivity from non-additive integration and connects the resulting internal quantity directly to an observable behavioral interaction. The dose-response, boundary, unrelated-donor, self-patch, reconstruction, and autoregressive-null controls further indicate that this interaction is tied to the manipulated evidence representations rather than arbitrary activation replacement.

Scope and outlook. Our conclusions are based on two instruction-tuned language models, a controlled fact-verification dataset, and selected layers and prompt positions. They should therefore be interpreted as evidence for a shared mechanism in these settings rather than a universal account of all multi-context reasoning. Extending the design to more architectures, naturally occurring jointly necessary evidence, and finer-grained circuit interventions will be required to determine how broadly the progression generalizes. Nevertheless, the present results establish a concrete finding about language-model decision formation: when multiple facts must be combined, their causal contributions are partner-dependent, and this relationship is progressively re-expressed from the evidence context toward the final decision state.

6 Conclusion

We studied how language models combine multiple evidence statements when making a fact-verification decision. Factorized Causal Activation Patching converts activation patching from a marginal influence test into a partner-conditioned causal interaction test and provides an exact layer-zero bridge between the internal patching contrast and the corresponding behavioral factorial interaction. The resulting account is more specific than saying that evidence "matters": Qwen3.5-9B and Gemma-4-12B-it do not treat the two statements as independent votes, but construct a joint representation whose causal signature changes systematically across network depth and prompt position.

At the source spans, semantic counterfactuals produced large early-layer dependency contrasts of 3.81 for Qwen and 4.19 for Gemma. This source-level interaction subsequently diminished, while a downstream interaction emerged at the claim endpoint around mid-depth—0.45 at Qwen layer

16 and 0.99 at Gemma layer 24—and became strongest at the final decision position in the last sampled layer, reaching 2.55 for Qwen and 4.19 for Gemma. Whole-word permutation reproduced the dependency effect behaviorally, although more weakly (0.90 and 0.31), while natural sufficient-evidence cohorts showed positive pooled interactions of 0.276 and 0.317, with important dataset heterogeneity and FEVER dominance. Dose-response, boundary, unrelated-donor, self-patch, reconstruction, and autoregressive-null controls further establish that the observed progression is tied to the manipulated evidence representations rather than generic activation replacement.

Together, these results support a process account of multi-evidence decision formation: partner-conditioned evidence interaction is first encoded where the evidence appears, becomes available at an intermediate claim representation, and is ultimately expressed at the decision token. The principal lesson is therefore not merely that language models use multiple evidence statements, but that they causally transform their relationship as information moves from context to decision.

Reproducibility Statement

All cohorts, intervention protocols, hypotheses, depth windows, aggregation rules, and statistical tests were frozen in machine-readable files before confirmatory outcomes were analyzed. Discovery and validation examples are disjoint. Each run records model and artifact hashes, token IDs, evidence-span boundaries, corruption seeds, evidence order, numerical precision, software versions, and architecture-specific intervention sites. Observational parity, self-patching, endpoint reconstruction, and causal-mask controls are checked automatically; collection aborts upon any mismatch. The supplementary material provides the code, frozen protocols, cohort indices, per-example estimates, and artifact manifests needed to reconstruct the reported analyses. Appendix **??** documents the checkpoints, prompts, deterministic settings, and experimental inventory.

Ethics Statement

This study analyzes existing language models on publicly available fact-verification datasets and introduces no new human-subject data. The corrupted evidence conditions are synthetic experimental interventions and must not be interpreted as factual claims. Our analysis is conditional on examples exhibiting a predefined interaction-positive phenotype and therefore does not support conclusions about all model decisions or general reasoning ability. We report negative localization results and limitations explicitly to avoid overstating the identified mechanisms. The experiments require substantial computation; we reduce unnecessary cost by reusing frozen activations, separating discovery from confirmation, and performing robustness analyses from previously collected boundary-level outputs whenever possible.

References

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130, 2024. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. (D)RAGged into a conflict: Detecting and addressing conflicting sources in search-augmented LLMs. In *arXiv preprint arXiv:2506.08500*, 2025. doi: 10.48550/arXiv.2506.08500. URL https://arxiv.org/abs/2506.08500. To appear in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. doi: 10.48550/arXiv.2312.10997. URL https://arxiv.org/abs/2312.10997.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing context-sensitive information flow in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10697–10714, 2023. doi: 10.18653/v1/2023.findings-emnlp.712. URL https://aclanthology.org/2023.findings-emnlp.712/.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining the outputs of deep networks using integrated Hessians. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=exkxR6F-P1.

Sebastien Kawada and Manolis Kellis. Evidence integration in large language models. *arXiv preprint arXiv:2609.04290*, 2026. doi: 10.48550/arXiv.2609.04290. URL https://arxiv.org/abs/2609.04290.

Jungyeon Lee, Kangmin Lee, and Taeuk Kim. MAGIC: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8783–8803, 2025. doi: 10.18653/v1/2025.findings-emnlp.466. URL https://aclanthology.org/2025.findings-emnlp.466/.

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 *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2503.10996. ICML 2025 Spotlight; arXiv:2503.10996.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7342–7351, 2020. doi: 10.18653/v1/2020.acl-main.655. URL https://aclanthology.org/2020.acl-main.655/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. 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 *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Minh V. T. Pham, Hsuvas Borkakoty, and Yufang Hou. Where knowledge collides: A mechanistic study of intra-memory knowledge conflict in language models. *arXiv preprint arXiv:2601.04327*, 2026. doi: 10.48550/arXiv.2601.04327. URL https://arxiv.org/abs/2601.04327.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65128–65167, 2023. doi: 10.52202/075280-2842. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/65128-Abstract-Datasets_and_Benchmarks.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9259–9268, 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6207–6227, 2024. doi: 10.18653/v1/2024.acl-long.337. URL https://aclanthology.org/2024.acl-long.337/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul. arXiv preprint arXiv:2211.00593.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992, 2025. doi: 10.18653/v1/2025.findings-acl.205. URL https://aclanthology.org/2025.findings-acl.205/.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. Towards faithful natural language explanations: A study using activation patching in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29362–29386, 2025. doi: 10.18653/v1/2025.emnlp-main.1494. URL https://aclanthology.org/2025.emnlp-main.1494/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=exkxR6F-P1.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Chaozhuo Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 892–901, 2019. doi: 10.18653/v1/P19-1085. URL https://aclanthology.org/P19-1085/.