LLM EXTRACTOR ERRORS ARE NOT RANDOM: PROTAGONIST BIAS IN KNOWLEDGE GRAPH EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models have become the prevalent tool for building knowledge graphs from scientific literature. It turns unstructured text into entity-relation graphs which enhance literature-based discovery, biomedical question answering, and drug repurposing. However, the resulting graphs remain far from complete and accurate, and common remedies such as sampling and voting implicitly treat extraction errors as independent noise. Through cross-corpus error analyses, counterfactual experiments, and mechanistic analyses, we show that these errors are structured: relations between a document's supporting entities are extracted systematically worse than those involving its most-mentioned entity, its protagonist. We trace this to document-level salience: an entity's prominence across the whole document, not just in the sentence that states the relation, shapes whether the model opens a relation with it, and also inflates the model's confidence, so frequency- or confidence-based selection reinforces the bias. We therefore separate proposing relations from verifying them: an evidence-anchored extractor re-reads each document sentence by sentence to propose candidates, and a selector keeps only those supported by local evidence. On BioRED, this improves Qwen3.5-27B by 0.07 F1 over standard fine-tuning and raises supporting-entity precision from 0.54 to 0.68 without loss of recall.

1 INTRODUCTION

Knowledge graphs extracted from scientific literature underpin key applications such as literature-based discovery, biomedical question answering, and drug repurposing (Luo et al., 2022; Xue et al., 2024; Lu et al., 2026; Edge et al., 2024). To build these graphs at scale, large language models (LLMs) have become the default tool for document-level relation extraction. However, as models frequently miss valid relation from complex scientific text, comprehensive relation extraction using LLM still remains a major challenge. Recently, existing pipelines typically treat these missing relations as unstructured, random errors, relying on post-hoc remedies such as self-consistency sampling (Wang et al., 2022; Zubillaga et al., 2026), extractor voting (Dong et al., 2014), or retrieval-augmented second passes (Singhania et al., 2024). Nevertheless, these remedies only work if omissions are truly random. When given that human attention concentrates heavily on prominent entities (Edwards et al., 2011; Stoeger et al., 2018), a fundamental question arises: do LLM extraction errors occur at random, or do models suffer from a systematic attention bias inside each document?

To answer this question, we evaluate standard document-level extraction pipelines across model scales, testing open-weight models and closed-source models on several benchmarks (Luo et al., 2022; Li et al., 2016; Tan et al., 2022). In order to measure extraction bias, we define a document's **protagonist** as its most frequently mentioned entity in this document. We then split ground-truth relations into two groups, the one group involving the protagonist and the others that connect two supporting entities excluding the protagonist. Across models tested, we observe a severe protagonist bias: extractors recall protagonist relations nearly twice as often as supporting-entity relations.

This protagonist bias fundamentally distorts the extracted knowledge graph into a star network centered around main entities. Consequently, extractors systematically drop secondary connections like variant-phenotype relations, but literature-based discovery relies precisely on these non-obvious

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.

links. Controlled entity-swapping experiments confirm that this omission is driven by document-level entity salience rather than a lack of local sentence evidence. Because the bias stems from global prominence, single-stage sampling or specific post-hoc filtering cannot fix it without introducing additional reporting errors.

Guided by this diagnosis, we propose a two-stage extraction framework that decouples relation coverage from precision. First, a candidate generator combines evidence-anchored fine-tuning with a sentence-level pass, so that every sentence gets its own opportunity to propose relations. Second, a local-evidence selector keeps or drops each candidate using only its two entities and their evidence sentence, without access to the title, the rest of the document, or how often the candidate was proposed. On BioRED (Luo et al., 2022), the framework improves F1 over standard fine-tuning on the same data at both open-weight scales, and the gain lands where the diagnosis points: precision and F1 on supporting-entity relations rise in every configuration we test. The same sentence-level pass and selector raise the precision of a base closed-source extractor on those relations.

Our contributions are as follows.

- **Benchmark** We evaluate open-weight and closed-source extractors on three datasets under one graph-level protocol: they find 56–88% of the entities on BioRED and Re-DocRED but recover at most 44% of the annotated relations on BioRED and 18% on Re-DocRED, and neither model scale, closed-source access, in-context exemplars, nor the gold entity list closes the gap.
- **Causal and mechanistic analysis** We separate the protagonist effect from co-occurrence frequency with matched counterfactual documents, locate it at the decision of which entity opens a relation, and trace it to a distributed computation at the output position that aggregates title and out-of-sentence mentions. Steering this computation changes whom the model reports but does not preferentially increase supported relations.
- **Methodology** The mechanism separates ”what is proposed” from ”what is kept”. Candidate generation combines evidence-anchored fine-tuning with a sentence-level pass, so that a supporting entity can open a relation in a prompt that contains no other part of the document. A local-evidence selector then decides from the two entities and their evidence sentence alone, without the title, the rest of the document, or how often the candidate was proposed. This method reaches 0.53 F1 with Qwen3.5-27B on BioRED, 0.07 above standard fine-tuning on the same data, and raises precision on supporting-entity relations from 0.54 to 0.68; on a base closed-source extractor it adds 0.03 F1.

2 The Protagonist Bias: From Error Pattern to Mechanism

2.1 The protagonist organizes extraction errors

We benchmark document-level relation extraction across models and datasets, then examine whether missed relations concentrate away from each document's protagonist. Each extractor receives a document and a schema and returns entities and typed relations. We align predicted entities to annotated concepts through mention-span matching and score the resulting graph at the level of (entity pair, relation type). This evaluation yields both micro precision, recall, and F1 and a recovery outcome for every annotated relation, allowing us to quantify the protagonist gap and identify where relations are lost.

Datasets BioRED (Luo et al., 2022) provides 100 test abstracts with 990 gold relations over six entity types and eight relation types, and is our primary benchmark. BC5CDR (Li et al., 2016) provides 500 test abstracts with 1,042 chemical–disease relations; we use it to check that the effect is not tied to one annotation scheme. Re-DocRED (Tan et al., 2022) provides 500 Wikipedia paragraphs with 17,448 relations over 96 types; we use it to check that the effect is not tied to the biomedical domain. All models in this section are base models, without any post-training on these datasets: Qwen3.5-9B and Qwen3.5-27B and DeepSeek-v4-pro, each prompted 0-shot and few-shot with static per-relation exemplars, greedy decoding, thinking disabled.

Results Document-level graph extraction is an unsolved problem for base LLMs. Table 1 and Appendix B, Table 10, show that they recover 8–44% of the annotated relations on BioRED and Re-DocRED and 16–71% on BC5CDR. The loss is not at the entity stage. On BioRED the models

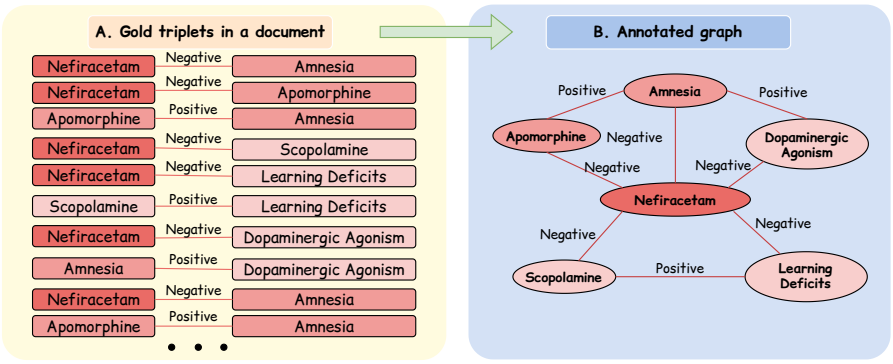


Figure 1: One BioRED sample as the relation statements of its sentences (A) and as its annotated graph (B). Nefiracetam is mentioned 9 times, more than any other entity, and is the document's protagonist; 5 of the 8 annotated relations touch it and 3 join supporting entities.

write down 56–84% of the concepts that take part in a gold relation (76–84% for Qwen3.5-27B and DeepSeek-v4-pro), and 81–89% of the entity surface forms they emit resolve to an annotated concept, whereas relation recall is 0.18–0.44; on Re-DocRED entity recall is 0.68–0.88 against relation recall of 0.08–0.18. The models are not confused about what they output, and they find most of the entities; they fail to connect them. None of the usual remedies closes the gap. Exemplars raise recall (on BioRED, Qwen3.5-27B goes from 0.30 to 0.41 F1 and from 0.28 to 0.44 recall, against 0.32 to 0.38 precision), but the best cell on BioRED still recovers 44% of the relations and the best on Re-DocRED 18%; scale adds 5–34 F1 points from Qwen3.5-9B to Qwen3.5-27B and closed-source access 0–8 points more, and every one of these models misses at least 56% of BioRED's relations. Supplying the gold entity list makes the point directly: entity recall rises to 0.82–0.86, relation F1 moves by at most 0.035. Only DeepSeek-v4-pro on BC5CDR, a dataset with a single relation type, recovers more (relation recall 0.63 0-shot, 0.71 few-shot).

		BioRED		**BC5CDR**		**Re-DocRED**	
Model	Prompt	Entity F1	Relation F1	Entity F1	Relation F1	Entity F1	Relation F1
Qwen3.5-9B	0-shot	0.66	0.24	0.53	0.31	0.79	0.13
Qwen3.5-9B	few-shot	0.73	0.30	0.37	0.25	0.82	0.19
Qwen3.5-27B	0-shot	0.80	0.30	0.69	0.51	0.86	0.19
Qwen3.5-27B	few-shot	0.83	0.41	0.77	0.59	0.83	0.24
DeepSeek-v4-pro	0-shot	0.84	0.38	0.86	0.59	0.86	0.19
DeepSeek-v4-pro	few-shot	0.86	0.42	0.88	0.63	0.93	0.26

Table 1: Knowledge extractors on three datasets (0-shot, or few-shot with one static exemplar per relation type). Relation F1 counts a predicted relation only if pair and type both match; entity F1 scores the entities the extractor writes down against the annotated concepts that take part in a gold relation. Precision and recall for every cell are in Appendix B, Table 10.

Extraction errors follow mention count The benchmark counts the missing relations; we next identify them. The errors concentrate on the same axis. Of the 432 false positives of 0-shot DeepSeek-v4-pro, 42% touch the document's most-mentioned entity and 40% join an entity mentioned once to one mentioned three or more times, and the misses are the mirror image: the share of gold entities the extractor writes down at all rises from 0.67 for entities mentioned once to 0.96 for entities mentioned five or more times. Every document has one entity it is about, and the extracted graph is organized around it. We make that entity a variable, show that the gap it opens holds for every model and corpus, and show how a supporting-entity relation is lost.

Defining the protagonist We make this precise with a variable that any document defines. For a document d with annotated entity set E_d, let $m_d(e)$ be the number of mentions of entity e in d. The protagonist of d is its most-mentioned entity, $p(d) = \arg\max_{e \in E_d} m_d(e)$ which is shown in Figure 1. Every other annotated entity of d is a *supporting entity*. The gold relation set R_d

splits into protagonist relations $R_d^{\mathrm{Pro}} = \{(h, r, t) \in R_d : p(d) \in \{h, t\}\}$ and supporting-entity relations $R_d^{\mathrm{Sup}} = R_d \setminus R_d^{\mathrm{Pro}}$, whose two endpoints are both supporting entities. For an extractor with output $\hat{R}_d$ we report $\mathrm{Rec}^{\mathrm{Pro}} = \sum_d |\hat{R}_d \cap R_d^{\mathrm{Pro}}| / \sum_d |R_d^{\mathrm{Pro}}|$ and $\mathrm{Rec}^{\mathrm{Sup}}$ likewise, together with the precision and F_1 of each class, $\mathrm{Prec}^{\mathrm{Pro}}$, $\mathrm{Prec}^{\mathrm{Sup}}$, F_1^{Pro} and F_1^{Sup}. We write $\Delta = F_1^{\mathrm{Pro}} - F_1^{\mathrm{Sup}}$ in percentage points, the protagonist gap, and say *recall gap* when the same difference is taken on recall alone. On BioRED test the protagonist is a title entity in 93 of 100 abstracts, accounts for 28% of all mentions, and touches 44% of the gold relations (62% on BC5CDR, 25% on Re-DocRED).

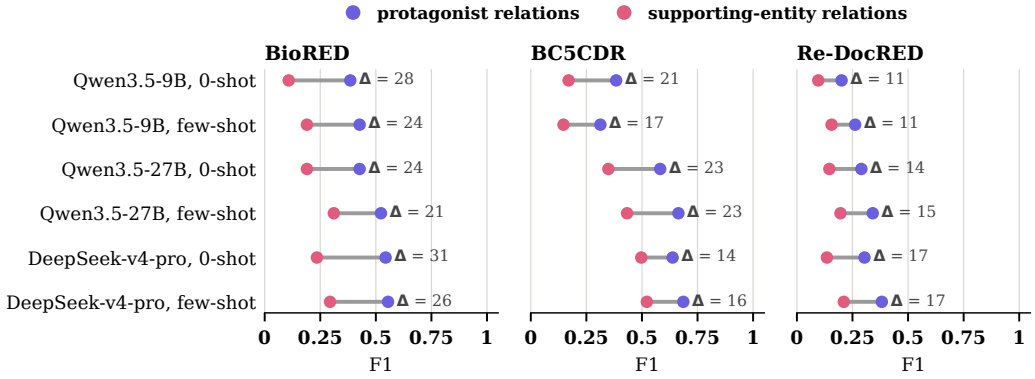


Figure 2: Recall of protagonist relations and of supporting-entity relations for the base extractors of Table 1, with $\Delta = F_1^{\mathrm{Pro}} - F_1^{\mathrm{Sup}}$ in percentage points beside each pair. F_1^{Sup} is below F_1^{Pro} in every cell, so the gap is not an artefact of precision. The numbers are in Table 3.

The imbalance survives scale, domain, and exemplars In every cell of Figure 2, F_1^{Sup} is below F_1^{Pro}. The gap is 21 to 31 points on BioRED, 14 to 23 on BC5CDR, and 11 to 17 on Re-DocRED, and exemplars close it in no cell (details in Appendix A). Holding co-occurrence sentence count, mention count, position, and relation type fixed, involving the protagonist multiplies the odds that a relation is recovered by 2.0 to 3.6 in the regressions of Table 6. The effect is not a property of one model family, scale, annotation scheme, or domain, and it is present before any training on the target corpus. A missed supporting-entity relation is usually not a relation the extractor overlooked: in most cases it emits both entities and links the rarer one to a different, more prominent partner instead, so the loss is a choice of partner rather than a failure to read (Appendix A).

2.2 Document role changes which entity opens a relation

The preceding results identify where extraction errors concentrate, but they do not show that document role itself causes a relation to be omitted. Protagonist status co-varies with mention count, title position, and other properties of the evidence. We therefore ask two questions under controlled inputs: does changing only an entity's document role alter the write-out of a relation whose local evidence is held fixed, and, if it does, at which choice in the serialized output does the change first appear? Both questions are asked of the base extractors of Table 1 and of Qwen3.5-9B and 27B fine-tuned with LoRA on the BioRED training set against the annotated triples, not the evidence-anchored target of Section 3; the fine-tunes are included because their output format is fixed enough to cut at the token boundaries the second question requires.

Changing document role changes whether fixed evidence is reported Let H be a document's protagonist, x a same-type supporting entity, and s a sentence that states the scored relation. We exchange the names of H and x everywhere outside s, including the title, while keeping s byte-identical. The scored relation, its local evidence, entity types, and model are therefore unchanged; only which entity occupies the document-level role is exchanged. Matched-null swaps change the same number of mentions without changing either scored entity, bystander records score an edge unrelated to both, and identity pairs estimate run-to-run noise. On BioRED test, removing protagonist status lowers the write-out rate of the same supported relation by 31 points for the fine-tuned 27B extractor and 30 points for its base counterpart; the independently rerun 9B fine-tune shows a 23.6-point reduction. Granting protagonist status raises write-out by 22–26 points in the fine-tuned

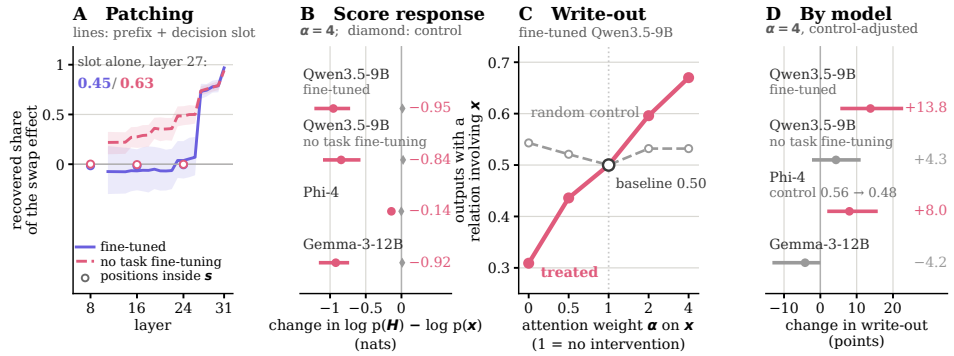


Figure 3: Locating and steering relation initiation. (A) Activation patching on BioRED test: recovered share of the swap effect by layer when the output-side states (assistant prefix and decision slot) are restored; open circles are entity positions inside the evidence sentence s. (B) Score-mode change in $\log p(H) - \log p(x)$ at $\alpha = 4$ with 95% CIs; diamonds are random-position controls. (C) Share of outputs containing a relation that involves x for the fine-tuned Qwen3.5-9B, treated and random control on the same 94 records from 57 documents; $\alpha = 1$ is the unmodified model. (D) Control-adjusted change in write-out at $\alpha = 4$ (94, 94, 50 and 48 records); grey intervals include 0.

extractors but not reliably in the base models. Document role therefore causally changes whether fixed local evidence is reported under this intervention, and the near-identical effect in the fine-tuned and base models shows that it is not a product of that training.

The role intervention primarily changes entity selection The counterfactual establishes an effect on complete outputs; we next ask where that effect enters the serialized relation. Holding the preceding output fixed, we probe three choices exposed by the format: whether the model closes the relation list or opens another entry (D_{stop}), whether H or x opens that entry (D_{head}), and how the relation and tail continue once the head is supplied (D_{cond}). Under these fixed prefixes, demoting H lowers its head log-odds against x by 3.66 nats in the base 9B model and 1.99 nats in the standard-target fine-tune; its mean rank worsens by 0.90 in the former. By contrast, the stop effects are indistinguishable from zero apart from one six-pair prefix cell, and conditional completion after supplying the head does not fall ($+0.149$ and $+0.037$ nats). In the fine-tuned model, learned output ordering saturates the top-1 head under the empty-list prefix, so the primary readout is the pairwise H–x log-odds. The supported conclusion is therefore specific: document role changes the model's relative preference over which entity opens the next relation, rather than its measured propensity to continue or the measured support for the relation-and-tail suffix once the head is fixed. Appendix D.2 formalizes these grammar-aligned readouts and reports prefix variants and the output-order audit.

Simple alternatives do not account for the gap The matched swap is the primary causal test. Additional interventions constrain its interpretation: masking protagonist mentions removes protagonist relations without raising the recall of supporting-entity relations; an explicit instruction not to organize the graph around the main entity changes Rec^{Sup} by at most 0.03; supplying the gold entity list raises Rec^{Sup} by at most 0.10 and leaves a recall gap of 19 to 28 points; and a PubMedBERT pair classifier (Gu et al., 2021) supplied with gold entities retains a gap. These controls do not identify a unique internal mechanism, but they make output-budget competition, instruction following, entity recognition, and autoregressive decoding alone insufficient explanations, as collected in Table 5 and Appendix A. Together with the decision readouts, they motivate the next question: where inside the model is this entity-selection difference recoverable, and can it be manipulated directly?

2.3 Locating and steering relation initiation

The counterfactuals establish that document role changes whether fixed evidence is reported. We now locate the internal states that carry this effect in the standard-target Qwen3.5-9B fine-tune and test whether attention interventions can alter which entities appear in the output.

The role effect is recoverable at the output position Following activation-patching work (Vig et al., 2020; Meng et al., 2022), we replace counterfactual activations with their original-document counterparts and measure recovery of the head log-odds, normalized by the text-swap effect. Patching entity positions inside the unchanged evidence sentence yields near-zero recovery; patching the full sentence does not restore the effect either. Recovery is substantially larger when we replace the states at both the assistant prefix and the decision slot. On test pairs held out from localization, output-side recovery is near zero through the middle layers and near-complete by layer 31, rising at layer 27 in the fine-tuned model and more gradually in the base model, which is shown in Figure 3A. A validation-set decomposition attributes the larger part of the layer-27 recovery to the decision slot and the smaller part to the assistant prefix. This identifies relation initiation as a site where the role effect can be causally recovered, without treating near-complete final-layer recovery as evidence of its origin. Appendix D.3 reports the layer curves and output-site decomposition in Tables 17–19.

Document cues contribute to the initiation decision At layer 27 of the fine-tuned 9B model, blocking attention from the decision slot to protagonist mentions outside the evidence sentence removes 0.19 of the swap effect; blocking the title removes 0.12. On the swapped document, blocking the new protagonist's out-of-sentence mentions restores 0.35. Masking an entity's output-side list entry instead reduces its head score in both treated and matched-control documents, indicating a more general contribution to accessing candidate names. These interventions implicate document cues in the role effect, while no tested source or individual attention head accounts for most of it. Appendix D.4 gives source and head ablations with matched controls.

Attention to the supporting entity controls whether it is written We add $\log \alpha$ to the attention logits targeting mentions of the supporting entity x, with $\alpha = 1$ as baseline and $\alpha = 0$ masking those positions. Applied across all query positions during prompt processing and decoding, the reweighting changes what the fine-tuned 9B extractor emits: the share of outputs containing a relation involving x falls under suppression ($\alpha = 0$) and rises under enhancement ($\alpha = 4$), monotonically in the dose, while a random-position control stays flat, as shown in Figure 3C. The baseline- and random-control-adjusted increase at $\alpha = 4$ is 13.8 percentage points (95% document-clustered CI: 5.5–22.8; 94 matched records from 57 documents), with the document and the model parameters held fixed. Restricted to the relation-decision positions, the same reweighting moves the teacher-forced margin $m = \log p(H) - \log p(x)$ in the same direction, while random-position and bystander controls do not, which is shown in Figure 3B and specified in Appendix D.5. The score response reproduces in the base 9B, in Phi-4 and Gemma-3-12B, and across four evidence-format adapters, while the generation response differs in sign across families, as shown in Figure 3D and detailed in Appendices D.8 and D.7; Tables 25 and 26 give the raw and paired generation panels. From what we have discussed above, we can conclude that steering changes whom the model reports, but it does not make its reports more correct. Therefore, it motivates us to separate proposal from verification in the next section.

3 METHODOLOGY

The mechanism imposes two constraints on the remedy. First, document-level salience acts when the extractor chooses which entity will open a relation, so increasing candidate coverage requires another opportunity to initiate relations around supporting entities. Second, manipulating that gate does not determine whether a proposed relation is true: direct steering changes whom the model reports without preferentially increasing gold relations. We therefore separate *candidate generation* from *local-evidence selection*. Candidate generation trains an evidence-anchored output interface and unions a whole-document pass with a second pass that reads one sentence per prompt; the selector then keeps or drops each candidate using only its local evidence. This choice matches BioRED, where 71% of annotated relations are stated within a single sentence. The whole-document pass is retained because it is the only route for pairs that never share a sentence; the sentence pass creates an alternative route into the candidate pool, while the selector prevents higher proposal coverage from automatically becoming a larger final graph. Figure 4 shows the two parts and the whole pipeline.

3.1 CANDIDATE GENERATION

Evidence-anchored fine-tuning The extractor is Qwen3.5-9B or Qwen3.5-27B fine-tuned on the BioRED training set with LoRA (Hu et al., 2021); the only changes from the baseline fine-tune are

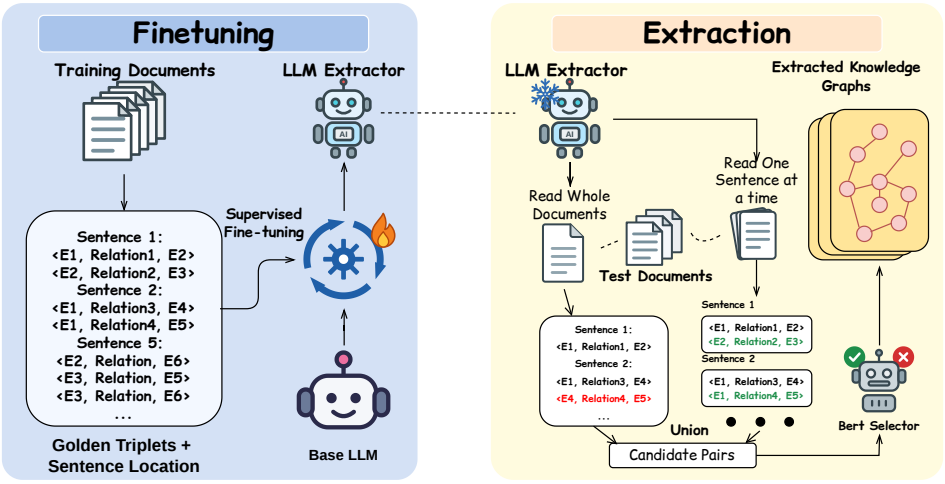


Figure 4: The pipeline. Finetuning: the extractor is fine-tuned once on targets in which every relation carries the sentence that states it. Extraction: the fine-tuned extractor reads the whole document and then one sentence at a time, and the two outputs form the candidate pool. Local-evidence selector: a classifier that sees only a candidate pair and its sentence decides which candidates enter the graph.

the output format the model learns to produce and the prompt rule that describes it. The standard output lists the entities and then the relations as (head, relation, tail) triples in annotation order. In our format every relation carries a fourth element, (head, relation, tail, S), an evidence anchor derived heuristically from entity co-occurrence: S names the sentences in which both entities are mentioned, or the nearest two sentences mentioning one entity each when they never share one, and the relations are listed in the order of their sentences. BioRED does not annotate evidence, so S is derived from entity mentions, stated to the model as a rule, and not scored at evaluation. This target gives the later sentence pass and the selector a shared, sentence-indexed interface. Its role in the pipeline is to raise candidate coverage and attach candidates to evidence locations, not to solve the bias by target formatting alone.

A second pass, one sentence per prompt The same extractor then reads the document again, one sentence at a time: each prompt contains the title, one sentence, and the unchanged schema and instructions, and nothing else of the document. A relation enters the candidate pool only if both of its entities appear in the sentence, the title being context rather than source text, and the candidate pool is the union of these outputs with the output of the whole-document pass. This is an input-side intervention on the cues identified in Section 2.3: it removes out-of-sentence mentions and repeated document context from the relation-initiation decision while leaving the title and the local evidence intact. The title can still supply a salience cue, so we do not claim that this pass erases the mechanism. Instead, it gives every sentence a fresh decision point at which a supporting entity can enter the candidate pool. Keeping the whole document and merely instructing the model to focus on one sentence would retain the same document-level cues at that decision point.

3.2 Local-evidence selection

A classifier that sees only the pair and its sentence The pool trades precision for recall, so a selector decides what to keep. The usual choice, self-consistency (Wang et al., 2022), counts how often a pair was proposed, which is what the reporting preference already inflates. We score each candidate on its own evidence instead: a BioLinkBERT-large classifier (Yasunaga et al., 2022) reads the two entities with their types and the sentence in which they co-occur, and sees no vote counts, mention counts, title, or anything else from the document. It is trained on the BioRED training set without new annotation and decides only whether a pair is supported; a kept pair takes its type from the extractor that proposed it with the evidence in view. Appendix B gives the selection rule (Algorithm 1) and the training details, including the threshold $\tau = 0.53$, fixed once on the validation pool of the sampling extractor and never re-tuned.

System	P	R	F1	Pair F1	$\mathrm{Prec}^{\mathrm{Sup}}$	F_1^{Sup}
Qwen3.5-27B						
Base, 0-shot	0.32	0.28	0.30	0.51	0.43	0.36
Base, few-shot	0.38	0.44	0.41	0.56	0.43	0.45
Fine-tuned, standard target	0.50	0.42	0.46	0.60	0.54	0.47
Fine-tuned, evidence-anchored target	0.49	**0.52**	0.50	0.62	0.51	0.50
+ sentence pass + local-evidence selector	**0.59**	0.47	**0.53**	**0.65**	**0.68**	**0.53**
Qwen3.5-9B						
Base, 0-shot	0.35	0.18	0.24	0.41	0.46	0.27
Base, few-shot	0.34	0.28	0.30	0.49	0.48	0.38
Fine-tuned, standard target	0.42	0.37	0.39	0.51	0.42	0.38
Fine-tuned, evidence-anchored target	0.44	**0.43**	0.44	0.56	0.47	0.43
+ sentence pass + local-evidence selector	**0.56**	0.39	**0.46**	**0.59**	**0.66**	**0.48**
DeepSeek-v4-pro (base; the same selector)						
Few-shot, one pass	0.41	**0.42**	0.42	0.55	0.46	0.44
+ sentence pass + local-evidence selector	**0.54**	0.38	**0.45**	**0.60**	**0.64**	**0.45**

Table 2: BioRED test, every row scored by the same scorer: a predicted relation whose endpoints do not resolve to annotated concepts is discarded for every system alike. P, R and F1 require the pair and the relation type to match; Pair F1 ignores the relation type; $\mathrm{Prec}^{\mathrm{Sup}}$ and F_1^{Sup} are pair-level and restricted to relations whose two endpoints are both supporting entities. The selector is BioLinkBERT-large (four seeds averaged) with $\tau = 0.53$ frozen on validation and the schema validator applied. Per-class recall, precision and F1 for every row are in Table 11.

Taken together, the two parts change the unit the model reads and the basis on which a relation is kept, not what the model prefers to report. The pipeline does not claim to remove the relation-initiation gate: the whole-document pass can retain its reporting preference, and the title remains visible in the sentence pass. The sentence pass instead gives every sentence its own opportunity to propose a relation with the evidence in view, and the selector makes final inclusion depend on local support rather than proposal frequency or document-level salience. A supporting-entity relation can therefore enter through its sentence without being forced into the final graph merely because the gate was opened. The selector is applied unchanged to every pool, including that of DeepSeek-v4-pro, a base model.

4 Experimental Results

Settings BioRED provides 400 training, 100 validation, and 100 test abstracts (990 test relations). The extractors are fine-tuned with LoRA on the training set, with the checkpoint chosen on the validation set, and decoded greedily; the selector is BioLinkBERT-large (Yasunaga et al., 2022) trained on 15,466 pairs from the training set, with its threshold frozen on the validation pool. Scoring follows Table 1. Hyperparameters, the construction of the candidate pools, and the scoring of pool-based rows are given in Appendix B.

The ladder Table 2 adds the parts one at a time. On Qwen3.5-27B, anchoring the targets to their evidence sentences adds 0.04 F1 over the standard target, from 0.46 to 0.50. Adding the sentence pass raises $\mathrm{Rec}^{\mathrm{Sup}}$ from 0.48 to 0.61 and, left unselected, gives 0.54 F1; selecting the pool by agreement of the two passes instead collapses $\mathrm{Rec}^{\mathrm{Sup}}$ to 0.19, as in Table 13. The local-evidence selector keeps 0.53 F1, 0.03 above the single anchored pass and 0.07 above the standard fine-tune, and it is the only selection rule that raises precision on supporting-entity relations without collapsing their recalls, from 0.51 to 0.68. The unselected pool and the graph filtered by the selector are two operating points on the same candidates. We report the selected graph as the default configuration: the two are statistically indistinguishable under the main convention, the sel+ected graph leads under the stricter convention and on validation, and it removes most unresolvable predictions (302 to 39). The same recipe carries Qwen3.5-9B from 0.39 to 0.46, and the pipeline built on the anchored 9B exceeds the same pipeline on the un-anchored one by 0.04. Reading twice beats reading many times: the 17-output sampling pool of the anchored 27B, selected by the same selector, reaches 0.49. The variants that did not help are listed in Appendix C. The gain is confined to relations whose entities share a sentence; the two strata are scored separately in Appendix C. On the base DeepSeek-v4-pro,

the sentence pass and the selector gain 0.03 F1 both 0-shot, from 0.38 to 0.41, and few-shot, from 0.42 to 0.45, with precision rising by 0.10 and 0.13.

Does the gain land where the diagnosis says? The two parts act on the two quantities the diagnosis names. The sentence pass raises the recall of supporting-entity relations over the single anchored pass by 0.12 for Qwen3.5-27B (95% CI 0.07 to 0.18) and by 0.13 for Qwen3.5-9B (0.08 to 0.17). The selector then raises precision on the same class over the unselected pool by 0.16 (0.12 to 0.21) and 0.20 (0.13 to 0.26). For DeepSeek-v4-pro, the full pipeline raises the same precision over its single pass by 0.18 (0.08 to 0.27). Every one of these intervals excludes zero (2,000 document-clustered bootstrap resamples over test documents).

5 Related Work

LLM-based knowledge graph construction LLMs are now the standard front end for turning documents into graphs, whether the graph serves retrieval (Edge et al., 2024; Gutiérrez et al., 2024; Guo et al., 2025), enrichment of a curated resource (Lu et al., 2026), or extraction from long scientific documents (Khalighinejad et al., 2024; Xue et al., 2024). When recall matters, these pipelines generate more than once and keep what recurs: Knowledge Vault fuses extractors by agreement (Dong et al., 2014), self-consistency votes over samples (Wang et al., 2022), sampling-and-selection schemes score document-level candidates by how often they are generated (Zubillaga et al., 2026), and L3X pairs recall-oriented retrieval-augmented generation with a scrutiny stage (Singhania et al., 2024). All of them treat what a single pass misses as noise that repetition averages out, and treat generation frequency as a proxy for reliability. We show that the missing edges have a document-level structure—they are the ones that do not touch the protagonist—and that generation frequency tracks that same structure, which is why vote-based selection cannot repair it (Section 4).

Document-level relation extraction Supervised document-level RE classifies every type-compatible entity pair with the whole document as context, on DocRED and its re-annotation (Yao et al., 2019; Tan et al., 2022; Zhou et al., 2021) and, in the biomedical domain, on BioRED (Luo et al., 2022), where PubMedBERT-style pair classifiers (Gu et al., 2021) remain the strongest systems (Lai et al., 2023; Islamaj et al., 2024). LLM variants either prompt for the full graph (Xue et al., 2024) or use the LLM to refine a classifier's candidates (Li et al., 2026; Zhang et al., 2025a;b). The biomedical error analyses come closest to our finding: in a manual sample of 40 BioREx errors, 83% involve an entity mentioned only once, and of the BioCreative VIII test relations that at least half of the participating systems missed, 52% contain such an entity (Lai et al., 2023; Islamaj et al., 2024). Both stop at mention frequency. We show that frequency is a proxy for a document role—once the protagonist is controlled, how often a relation is stated no longer predicts recall—and that the discriminative classifier also exhibits a gap as the generative extractors, so the cause is not the decoding scheme.

6 Conclusion

In summary, this work identifies document-level salience as a systematic influence on LLM relation extraction. Across datasets and model scales, relations between supporting entities are recovered less reliably than those involving the protagonist. Matched counterfactuals show that changing document-level entity roles alters relation reporting while the local evidence remains fixed. Decision-level probes and activation interventions implicate entity selection at relation initiation, while steering experiments show that increasing entity participation does not reliably improve correct-relation recovery. Motivated by this distinction, our pipeline separates evidence-anchored candidate generation from local-evidence selection. On BioRED, it improves Qwen3.5-27B from 0.46 to 0.53 relation F1 over standard fine-tuning and raises supporting-entity pair precision from 0.54 to 0.68 without reducing its recall. These findings support separating candidate generation from evidence-based selection in LLM knowledge graph extraction. Extending these gains to cross-sentence relations remains an open challenge.

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools for editing a research paper to improve readability. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpreting results, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, and generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, are not applicable to this work. We have reviewed all parts polished by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 states the experimental setting, and Appendix B gives the fine-tuning hyperparameters, the selector's backbone, training data and frozen threshold, the construction of the candidate pools, and the scoring convention, including which predictions are discarded and how the per-class and pair-level columns are defined. Every selection choice was frozen on the validation split before the single test run, and the decoding noise floor is stated in the same appendix. Appendix A documents the controlled interventions behind Section 2, and Appendix D the intervention sites, uncertainty estimates and per-model panels behind the mechanistic analysis. The prompt templates, the target format of the evidence-anchored fine-tune, and the definition of every reported metric are given in the appendices, so that each number can be traced to the procedure that produced it.

REFERENCES

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610, 2014.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Aled M Edwards, Ruth Isserlin, Gary D Bader, Stephen V Frye, Timothy M Willson, and Frank H Yu. Too many roads not taken. *Nature*, 470(7333):163–165, 2011.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *EMNLP (Findings)*, pp. 10746–10761, 2025.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. volume 37, pp. 59532–59569, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard A A Jonker, Sofia I R Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, et al. The overview of the biored (biomedical relation extraction dataset) track at biocreative viii. *Database*, 2024: baae069, 2024.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. Genie: Generative information extraction. In *Proceedings of the 2022 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4626–4643, 2022.

Ghazal Khalighinejad, Defne Circi, L Brinson, and Bhuwan Dhingra. Extracting polymer nanocomposite samples from full-length documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13163–13175, 2024.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. Biorex: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487, 2023.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. Llm with relation classifier for document-level relation extraction. *IEEE Transactions on Big Data*, 2026.

Yuxing Lu, Wei Wu, Xukai Zhao, Rui Peng, and Jinzhuo Wang. Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. volume 38, pp. 56421–56453, 2026.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. volume 35, pp. 17359–17372, 2022.

Sneha Singhania, Simon Razniewski, and Gerhard Weikum. Recall them all: Retrieval-augmented language models for long object list extraction from long documents. *arXiv preprint arXiv:2405.02732*, 2024.

Thomas Stoeger, Martin Gerlach, Richard I Morimoto, and Luís A Nunes Amaral. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology*, 16(9): e2006643, 2018.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 8472–8487, 2022.

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. volume 33, pp. 12388–12401, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. 2022.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. Autore: Document-level relation extraction with large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: System demonstrations)*, pp. 211–220, 2024.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 764–777, 2019.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8003–8016, 2022.

Fu Zhang, Xinlong Jin, Jingwei Cheng, Hongsen Yu, and Huangming Xu. Rethinking the role of llms for document-level relation extraction: a refiner with task distribution and probability fusion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6293–6312, 2025a.

Fu Zhang, Hongsen Yu, Jingwei Cheng, and Huangming Xu. Entity pair-guided relation summarization and retrieval in llms for document-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4022–4037, 2025b.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14612–14620, 2021.

Mikel Zubillaga, Oscar Sainz, Oier Lopez de Lacalle, and Eneko Agirre. Do not be greedy, think twice: Sampling and selection for document-level information extraction. *arXiv preprint arXiv:2601.18395*, 2026.

A ADDITIONAL STATISTICAL ANALYSIS

		BioRED			BC5CDR			Re-DocRED		
Model	Prompt	F_1^{Pro}	F_1^{Sup}	Δ	F_1^{Pro}	F_1^{Sup}	Δ	F_1^{Pro}	F_1^{Sup}	Δ
Qwen3.5-9B	0-shot	0.39	0.11	28	0.38	0.17	21	0.20	0.10	11
Qwen3.5-9B	few-shot	0.43	0.19	24	0.31	0.15	17	0.26	0.16	11
Qwen3.5-27B	0-shot	0.43	0.19	24	0.58	0.35	23	0.29	0.15	14
Qwen3.5-27B	few-shot	0.52	0.31	21	0.66	0.43	23	0.34	0.20	15
DeepSeek-v4-pro	0-shot	0.54	0.24	31	0.64	0.50	14	0.30	0.14	17
DeepSeek-v4-pro	few-shot	0.56	0.29	26	0.69	0.52	16	0.38	0.21	17
		$\mathrm{Rec}^{\mathrm{Pro}}$	$\mathrm{Rec}^{\mathrm{Sup}}$	Δ_R	$\mathrm{Rec}^{\mathrm{Pro}}$	$\mathrm{Rec}^{\mathrm{Sup}}$	Δ_R	$\mathrm{Rec}^{\mathrm{Pro}}$	$\mathrm{Rec}^{\mathrm{Sup}}$	Δ_R
Qwen3.5-9B	0-shot	0.32	0.08	24	0.28	0.10	18	0.16	0.06	10
Qwen3.5-9B	few-shot	0.42	0.16	26	0.20	0.09	12	0.21	0.10	11
Qwen3.5-27B	0-shot	0.43	0.17	27	0.50	0.24	26	0.23	0.09	14
Qwen3.5-27B	few-shot	0.59	0.33	27	0.63	0.34	28	0.28	0.13	15
DeepSeek-v4-pro	0-shot	0.54	0.19	34	0.71	0.50	21	0.23	0.08	15
DeepSeek-v4-pro	few-shot	0.61	0.28	33	0.80	0.55	24	0.31	0.14	18

Table 3: The two classes of relation scored separately for the extractors of Table 1: F_1 in the upper block, which is what Figure 2 plots, and recall in the lower block. Δ is the F_1 gap and Δ_R the recall gap, both in percentage points. Every cell of both blocks is positive. On BioRED, few-shot Qwen3.5-9B and 0-shot Qwen3.5-27B coincide in the upper block by chance: their per-class precision and recall are almost exchanged.

	Qwen3.5-9B		Qwen3.5-27B		DeepSeek-v4-pro	
	0-shot	few-shot	0-shot	few-shot	0-shot	few-shot
Supporting-entity relations ($n = 105$)						
Recall	0.18	0.30	0.32	0.59	0.35	0.50
Entity never emitted	0.42	0.40	0.24	0.19	0.34	0.26
Entities emitted, pair not linked	0.30	0.22	0.46	0.42	0.32	0.30
rarer entity linked to protagonist instead	0.09	0.18	0.27	0.28	0.19	0.21
Pair linked, wrong relation type	0.28	0.38	0.30	0.40	0.34	0.43
Protagonist relations ($n = 155$)						
Recall	0.28	0.37	0.39	0.58	0.48	0.59
Entity never emitted	0.29	0.25	0.18	0.11	0.19	0.17
Entities emitted, pair not linked	0.35	0.32	0.40	0.43	0.43	0.40
Pair linked, wrong relation type	0.36	0.43	0.41	0.46	0.38	0.43

Table 4: Why relations are missed (BioRED test, relations whose entities co-occur in at least two sentences). Values in the miss rows are fractions of the missed relations; the indented row is the share of misses in which the rarer entity was linked to the protagonist instead; n is the number of gold relations in each class after this restriction, Recall is over those n, and the remaining rows partition the missed ones.

The role, not the number of sentences the two entities share Two counts must be kept apart. The *mention count* of an entity is how often that entity appears in the document; it is a property of an entity, and it defines the protagonist ($p(d) = \arg\max_e m_d(e)$). The *co-occurrence sentence count* of a relation is the number of sentences in which its two entities both appear; it is a property of a relation, and it measures how much textual evidence the document offers for that one relation. The two are distinct: the two supporting entities of one relation can co-occur in three sentences, while the two entities of a relation involving a protagonist mentioned nine times co-occur in only one. Repetition of the second kind is the obvious alternative explanation of the gap, and it fails. Protagonist recall does not rise with the co-occurrence sentence count (0.54, 0.43, and 0.51 for relations whose entities co-occur in one, two, and three or more sentences), whereas the recall of supporting-entity relations does (0.17, 0.29, and 0.43; protagonist $\times$ co-occurrence interaction $p = 0.03$ for 0-shot DeepSeek-v4-pro), so the recall gap is widest where the evidence is thinnest:

37 points for relations whose entities co-occur once and 8 points for relations stated three or more times. A logistic regression on the 704 same-sentence gold relations of BioRED test that holds co-occurrence sentence count, the mention count of the rarer entity, position, and relation type fixed keeps the protagonist term significant in every fit and the co-occurrence term in none; the fits are in Table 6. Repetition is what a supporting-entity relation needs in order to be extracted, and what a protagonist relation does not need.

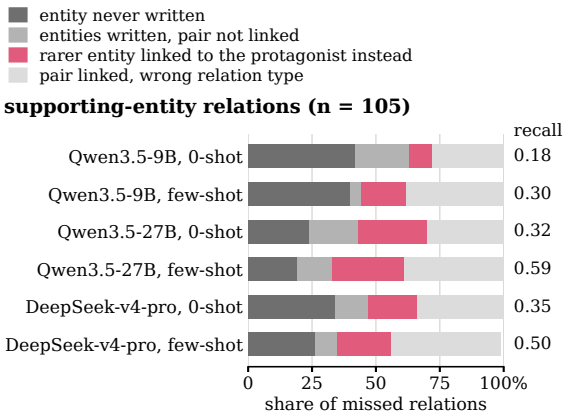


Figure 5: Why a supporting-entity relation is missed on BioRED test, among relations whose entities co-occur in at least two sentences: each bar partitions the missed relations of one extractor by failure type, and the right column is that extractor's recall over these relations. The numbers, including those for protagonist relations, are in Table 4.

How a supporting-entity relation is lost Figure 5 classifies the misses on BioRED among relations whose entities co-occur in at least two sentences, so that a miss cannot be blamed on the relation being hard to find. A supporting-entity relation is lost in one of three ways: its rarer entity is never emitted (0.19–0.42 of misses), the two entities are emitted but not linked (0.22–0.46), or the pair is linked with the wrong type (0.28–0.43). In the middle case the extractor almost always gives the rarer entity a different partner: it does so in 0.97, 0.99, and 0.82 of these misses for the anchored Qwen3.5-27B, the standard-target Qwen3.5-9B, and 0-shot DeepSeek-v4-pro. That substitute partner sits near the top of the document's mention ranking, on average at position 1.8, 2.3, and 1.9, where the true partner sits at 4.9, 4.1, and 3.9 and an entity drawn at random from those the extractor emitted sits at 9.7, 8.9, and 6.9; it is the protagonist itself in 0.59, 0.63, and 0.61 of the substitutions. Protagonist relations are substituted at the same rate (0.96, 0.98, 0.78), but their substitutes sit lower (3.6, 3.2, 4.5); the comparison is one-sided there, because the true partner of a protagonist relation already occupies position 1, so no substitute can be more prominent. What the two classes share is that the substitute is drawn from the prominent end of the document rather than at random.

Details of the error profile The numbers behind Figures 2 and 5 are in Tables 3 and 4. Across models, recall of a gold relation rises with the number of sentences in which its entities co-occur (for 0-shot DeepSeek-v4-pro, 0.31 for pairs that never share a sentence, 0.32 for one shared sentence, 0.36 for two, and 0.49 for three or more), and the share of gold entities written down rises with mention count (0.67, 0.80, 0.86, and 0.96 for one, two, three to four, and five or more mentions). Exemplars raise both recalls in every cell of Table 3 except Qwen3.5-9B on BC5CDR, where both fall; within a corpus, higher recall does not close the gap: on BioRED, few-shot DeepSeek-v4-pro has the highest $\mathrm{Rec}^{\mathrm{Pro}}$ (0.61) but also the largest recall gap (33 points, against 27 for few-shot Qwen3.5-27B). In Table 4, the share of not-linked misses in which the rarer entity was linked to the protagonist is 58% (Qwen3.5-27B 0-shot), 67% (few-shot), 59% (DeepSeek-v4-pro 0-shot), and 69% (few-shot), and for Qwen3.5-9B, which never emits the rarer entity in 40–42% of its misses, 31% and 81%. Exemplars raise the recall of supporting-entity relations by 12–27 points and protagonist recall by 9–19, and move misses from omission toward mislabeling (the wrong-type share rises by 9–10 points for supporting-entity relations and 5–7 for protagonist relations). Protagonist relations have their entity emitted more often (never emitted in 11–29% of their misses, against 19–42% for supporting-entity relations).

Controlled interventions Table 5 collects the interventions summarized in the main text; each row changes one factor and leaves everything else fixed. *Attention crowding*: for DeepSeek-v4-pro, every mention of the protagonist is replaced by a placeholder of equal length (`[MASK]` padded to the mention's length). The model then almost stops proposing protagonist relations (413 down to 59 0-shot and 498 down to 21 few-shot; $\mathrm{Rec}^{\mathrm{Pro}}$ 0.54 down to 0.07 and 0.61 down to 0.01) and proposes more supporting-entity relations instead (359 up to 469, and 496 up to 644), but the number of correct ones does not rise (109 against 98, and 156 against 154), so the recall of supporting-entity relations is unchanged 0-shot (0.19 and 0.18) and few-shot (0.28 and 0.28) while its precision falls (0.30 to 0.21, and 0.31 to 0.24). *Prompt misreading*: an instruction not to organize the graph around the main entity is added to the prompt; the recall of supporting-entity relations moves by $+0.03$ 0-shot and -0.03 few-shot. *Entity recognition*: supplying the gold entity list raises the recall of supporting-entity relations by at most 10 points (Qwen3.5-9B from 0.08 to 0.18, Qwen3.5-27B from 0.17 to 0.23, DeepSeek-v4-pro from 0.19 to 0.21) and leaves a gap of 19–28 points, against 24–34 without it; for Qwen3.5-27B the gap does not narrow at all (27 points before and 28 after), and for DeepSeek-v4-pro it narrows only because protagonist recall falls from 0.54 to 0.47, not because the recall of supporting-entity relations rises. *Autoregressive decoding*: a PubMedBERT pair classifier (Gu et al., 2021), trained on the BioRED training set in the style of BioREx (Lai et al., 2023), given the gold entities and scoring every type-compatible pair, shows $\mathrm{Rec}^{\mathrm{Pro}}$ 0.66 and $\mathrm{Rec}^{\mathrm{Sup}}$ 0.47 on BioRED test (a recall gap of 19 points), against 29 points for both the standard-target Qwen3.5-9B and the anchored Qwen3.5-27B fine-tuned on the same data and scored on the same split; the classifier's gap is narrower than theirs, not absent. *Training signal*: a Qwen3.5-9B student fine-tuned on the gold relations that a Qwen3.5-9B teacher recovered when prompted once per sentence (the recovered set is 50% of the training gold, and it is short of relations on rarely mentioned entities) recovers fewer supporting-entity relations than the model fine-tuned on all of the gold ($\mathrm{Rec}^{\mathrm{Sup}}$ 0.15 against 0.24) at a similar gap (27 against 29 points); the evidence that the gap follows the training targets is the controlled deletion of Table 9, not this run. Two further regression results belong here: in Table 6, mention count reaches significance in 2 of the 5 fits and position in 3, all of them fine-tuned Qwen3.5 fits and neither for 0-shot DeepSeek-v4-pro; and in a separate specification on the fine-tuned models, how common a concept is in the training corpus does not favor recall (odds ratio per log-unit of corpus frequency 0.76 for Qwen3.5-9B, $p = 0.01$, and 0.82 for Qwen3.5-27B, $p = 0.12$, the wrong direction for a frequency account) and the novelty label is non-significant.

Hypothesis	Intervention	Model	Before → after
Attention crowding	mask every protagonist mention	DeepSeek-v4-pro, 0-shot / few-shot	$\mathrm{Rec}^{\mathrm{Sup}}$ 0.19→0.18 / 0.28→0.28
Prompt misread	instruct: do not build around the main entity	DeepSeek-v4-pro, 0-shot / few-shot	Δ_R 34→30 / 33→33
Entity recognition	supply the gold entity list	Qwen3.5-9B / Qwen3.5-27B / DeepSeek-v4-pro, 0-shot	Δ_R 24→19 / 27→28 / 34→26
Autoregressive decoding	pair classifier, gold entities, all pairs	PubMedBERT	Δ_R = 19
Training signal	distill from a sentence-by-sentence teacher	Qwen3.5-9B	Δ_R 29→27
Training distribution	remove 25% of targets: at random	Qwen3.5-9B	Δ_R 14→12
	only protagonist relations	Qwen3.5-9B	Δ_R 14→1
	only supporting-entity relations	Qwen3.5-9B	Δ_R 14→15
Training corpus	fine-tune on Re-DocRED (25% protagonist)	Qwen3.5-9B	Δ_R 10→7

Table 5: Controlled interventions on BioRED. Each row changes one factor; $\mathrm{Rec}^{\mathrm{Sup}}$ is the recall of supporting-entity relations and $\Delta_R = \mathrm{Rec}^{\mathrm{Pro}} - \mathrm{Rec}^{\mathrm{Sup}}$ in percentage points, the recall gap of the lower block of Table 3. The upper block leaves the gap in place or narrows it by at most a quarter; in the lower block the gap moves with the training targets: it changes when their protagonist share changes and stays put under random deletion.

Score readout and training distribution Table 7 gives the next-token readout for both fine-tuned Qwen3.5 models, with the whole document or only the evidence sentence as input. In every row a fabricated protagonist relation and a true supporting-entity relation are indistinguishable (AUC 0.46–0.49), and the model separates true from fabricated relations less well among protagonist pairs (AUC 0.55–0.60) than among the others (0.67–0.74); in a regression of the log-probability on protagonist and truth indicators, touching the protagonist is worth $+0.32$ (Qwen3.5-27B) and $+0.37$ (Qwen3.5-9B) against $+0.36$ and $+0.40$ for being true. Table 8 gives the annotation base rates on all three splits. Table 9 gives the full training-distribution intervention: the three runs share text, prompts, entity lists, and hyperparameters and differ only in which 25% of the target relations are deleted; the protagonist-only deletion removes 55% of the protagonist targets (protagonist share 0.46 before the deletion and 0.27 after) and is the only run in which the protagonist odds ratio falls below 1 (0.92, against 1.65–1.86 in the other three) and the only one in which the gap nearly disappears (14 points down to 1, against 12 under random deletion and 15 when supporting-entity targets are deleted instead); these odds ratios refit the specification of Table 6 on the 823 same-sentence gold relations of the validation split, where no single run's odds ratio is significant on its own, so the recall gap is the primary reading. The protagonist $\times$ co-occurrence interaction of Table 6 is significant for 0-shot DeepSeek-v4-pro ($p = 0.03$) and not for fine-tuned Qwen3.5-27B with anchors ($p = 0.67$), although in that fit too the number of shared sentences stops predicting recall once the protagonist term is included. Across corpora: on Re-DocRED, where 25% of gold relations touch the protagonist, Qwen3.5-9B goes from a 10-point gap 0-shot to 7 points after fine-tuning on its training set; fine-tuning on BioRED, where 46% of training targets touch the protagonist, leaves the gap at 23–30 points on test for the four fine-tuned Qwen3.5 configurations and at 14 points on validation for the Qwen3.5-9B baseline of Table 9. The two corpora differ in overall recall, so the two gaps are reported side by side rather than compared. The base models, which saw neither training set, show the gap on both corpora before any training on the target corpus: 24–34 points on BioRED and 10–18 on Re-DocRED, where recall is much lower throughout ($\mathrm{Rec}^{\mathrm{Pro}}$ 0.15–0.31 against 0.32–0.61). These experiments do not identify the training origin of that tendency.

Extractor	Protagonist	Co-occurrences	Mentions	Position	Association	R^2
DeepSeek-v4-pro, 0-shot	3.60*	0.80	1.51	0.89	0.31*	0.13
Qwen3.5-9B, standard target	3.43*	1.03	1.19	0.39*	0.46*	0.13
Qwen3.5-27B, standard target	1.99*	0.95	2.69*	0.48	0.71	0.20
Qwen3.5-27B, anchored target	2.95*	0.72	1.95*	0.27*	0.31*	0.16
Qwen3.5-27B, anchored, sentence pass	2.32*	1.25	1.60	0.34*	0.28*	0.17

Table 6: Logistic regression of recall on the 704 same-sentence gold relations of BioRED test, one fit per extractor (the Qwen3.5 rows are fine-tuned on BioRED train); odds ratios with document-clustered standard errors (*: $p < 0.05$). The last column is the pseudo-R^2. Co-occurrence sentence count and mention count enter as logarithms (odds ratio per factor e). Position is the relative position of the first co-occurrence; Association is the relation type indicator. Without the protagonist term (mention count retained) the co-occurrence odds ratio is 0.9–1.5 and never significant. With a protagonist $\times$ co-occurrence interaction, 0-shot DeepSeek-v4-pro gives protagonist 14.9*, co-occurrences 2.9, interaction 0.20* ($p = 0.03$); for the anchored Qwen3.5-27B the interaction is 1.3 ($p = 0.67$).

B TRAINING AND SCORING DETAILS

Extractor fine-tuning Qwen3.5-9B and Qwen3.5-27B are fine-tuned with LoRA (Hu et al., 2021) of rank 32 and $\alpha = 64$ on all projection matrices, batch 4 with 2-step gradient accumulation, learning rate 10^{-4}, warm-up over 3% of the steps, and a maximum sequence length of 8,192 tokens, for 3 epochs (Qwen3.5-9B) or 4 epochs (Qwen3.5-27B); the standard-target and evidence-anchored fine-tunes of a model share every setting except the target. The checkpoint is chosen by validation F1, and decoding is greedy with a budget of 2,048 new tokens. The sentence pass issues one prompt per sentence, 11.5 per abstract on average (1,147 prompts for the 100 test abstracts), each containing the title, the sentence, and the unchanged schema and instructions.

Selector The selector is BioLinkBERT-large (Yasunaga et al., 2022), trained for 4 epochs with learning rate 1×10^{-5}, batch 32, and 256 input tokens on 15,466 pairs from the BioRED training

		True		Fabricated		AUC, true vs. fabricated		
Model (fine-tuned)	Input	prot.	other	prot.	other	prot.	other	cross
Qwen3.5-9B	whole document	−0.94	−1.05	−1.03	−1.42	0.58	0.74	0.49
Qwen3.5-9B	one sentence	−0.74	−0.91	−0.84	−1.21	0.60	0.71	0.46
Qwen3.5-27B	whole document	−1.02	−1.12	−1.11	−1.45	0.55	0.72	0.49
Qwen3.5-27B	one sentence	−0.64	−0.79	−0.73	−1.02	0.59	0.67	0.47

Table 7: Next-token log-probability per token of a candidate relation supplied as the model's own output (BioRED validation, 100 documents; models fine-tuned on BioRED train). An OLS of the log-probability on protagonist and truth indicators with document-clustered errors gives +0.37 (protagonist) and +0.40 (true) for Qwen3.5-9B and +0.32 and +0.36 for Qwen3.5-27B, all $p < 0.001$. Decoys are type-compatible co-mentioned pairs that are not annotated. AUC is the probability that a true relation outscores a decoy within the same group; the last column compares true supporting-entity relations against fabricated protagonist relations. Cross: AUC of a true relation between two supporting-entity entities against a fabricated relation involving the protagonist.

	pair shares a sentence		pair never shares a sentence	
Split	protagonist	other	protagonist	other
train	0.65 (n=2155)	0.27 (n=5791)	0.25 (n=1009)	0.05 (n=13016)
validation	0.67 (632)	0.21 (1937)	0.21 (261)	0.06 (3991)
test	0.57 (580)	0.19 (1952)	0.37 (268)	0.04 (4569)

Table 8: Annotation base rate: probability that a type-compatible entity pair is annotated as related, BioRED.

Training targets	prot. share	F1	$\mathrm{Rec}^{\mathrm{Pro}}$	$\mathrm{Rec}^{\mathrm{Sup}}$	Δ_R	prot. odds ratio [95% CI]
all relations	0.46	0.44	0.48	0.34	14	1.86 [0.89, 3.89]
25% removed at random	0.46	0.39	0.40	0.28	12	1.65 [0.74, 3.67]
25% removed, protagonist only	0.27	0.36	0.32	0.30	1	0.92 [0.50, 1.69]
25% removed, supporting-entity only	0.61	0.37	0.41	0.25	15	1.86 [0.83, 4.18]

Table 9: Training-distribution intervention (BioRED validation, Qwen3.5-9B; the fine-tuned baseline and three further runs with identical text, prompts, and hyperparameters, each with 25% of the target relations removed). Protagonist odds ratios refit the specification of Table 6 on the 823 same-sentence gold relations of the validation split, with 95% intervals in brackets.

set, of which 3,906 are candidates whose endpoints resolve to no annotated concept and are added as negatives. Four seeds are trained and their scores averaged; the four validation pair-F1 values are 0.750, 0.743, 0.727, and 0.736 (mean 0.739, standard deviation 0.010). Its input renders the rarer entity first, each entity bracketed with its type, followed by the local evidence: the first sentence in which both entities are mentioned, or, for a candidate whose entities never share a sentence, the first sentence mentioning each, concatenated. At training and at inference the evidence sentence is located by the extractor's own entity strings in the document text; no annotation is used at inference. The threshold $\tau = 0.53$ is the median of the band of thresholds that are indistinguishable on the validation pool of the standard-target Qwen3.5-27B (17 sampled outputs per document), and it is reused unchanged on the two-pass pools of every extractor, including DeepSeek-v4-pro. A schema validator, which checks entity types, relation names, and type-pair signatures against the schema, is a further label-free step of the pipeline and is applied in every pipeline row. BioRED annotates exactly one relation type per entity pair, so a kept pair needs only one type; the two passes disagree about the type of 48 kept pairs on test, and the whole-document pass, which read the full context, decides them. The pair classifier of Table 5 is a separate model and remains PubMedBERT-base (Gu et al., 2021).

Algorithm 1 Candidate selection for one document

Require: triples A from the whole-document pass and F from the sentence pass; selector q; threshold τ

$\mathcal{C} \leftarrow A \cup F$; for each triple in $\mathcal{C}$, $s(h, r, t) \leftarrow$ number of passes among A, F that proposed it
$G \leftarrow \emptyset$
for each entity pair $\{h, t\}$ that appears in some triple of $\mathcal{C}$ **do**
  $e \leftarrow$ local evidence of $\{h, t\}$: the first sentence mentioning both, else the first sentence of each
  if $q(h, t, e) < \tau$ **then continue** ▷ not related; q sees neither r nor s
  end if
  $R \leftarrow \{\, r : (h, r, t) \in \mathcal{C},\ s(h, r, t)$ maximal for $\{h, t\} \,\}$; $R_A \leftarrow \{\, r : (h, r, t) \in A \,\}$
  if $|R| > 1$ and $R \cap R_A \neq \emptyset$ **then** $R \leftarrow R \cap R_A$ ▷ passes disagree: A decides
  end if
  $G \leftarrow G \cup \{(h, r, t) : r \in R\}$
end for
return G

		BioRED				BC5CDR				Re-DocRED			
		Entity		Relation		Entity		Relation		Entity		Relation	
Model	Prompt	P	R	P	R	P	R	P	R	P	R	P	R
Qwen3.5-9B	0-shot	0.82	0.56	0.35	0.18	0.95	0.37	0.57	0.22	0.95	0.68	0.29	0.08
Qwen3.5-9B	few-shot	0.81	0.66	0.34	0.28	0.94	0.23	0.63	0.16	0.96	0.72	0.36	0.13
Qwen3.5-27B	0-shot	0.85	0.76	0.32	0.28	0.96	0.54	0.68	0.40	0.97	0.77	0.41	0.12
Qwen3.5-27B	few-shot	0.83	0.83	0.38	0.44	0.96	0.64	0.68	0.52	0.98	0.72	0.43	0.17
DeepSeek-v4-pro	0-shot	0.89	0.79	0.43	0.34	0.93	0.79	0.55	0.63	0.94	0.80	0.43	0.12
DeepSeek-v4-pro	few-shot	0.88	0.84	0.41	0.42	0.94	0.83	0.57	0.71	0.98	0.88	0.49	0.18

Table 10: Precision and recall behind Table 1. Entity precision is the share of emitted entity strings that resolve to an annotated concept; entity recall the share of the annotated concepts that take part in a gold relation which the extractor writes down; relation precision and recall are pair-and-type exact.

Scoring We score the 990 of BioRED's 1,163 test relations whose two endpoints are each denoted by a mention of their own; the remainder attach to spans that the corpus normalizes to two concepts at once, which a span-grounded extractor cannot be credited for (1,042 of 1,066 on BC5CDR). Every row of Tables 1, 2 and 13 is scored by the same scorer. Predicted entity strings are resolved to annotated concepts by span overlap; a prediction whose head or tail resolves to no annotated concept is discarded before scoring, for single generations and for pipeline rows alike, so no row gains precision from a resolution failure that another row pays for. The final triples of a pipeline row are serialized in the format of a single generation and scored by that same code path. F1 requires the

entity pair and the relation type to match. Pair F1 ignores the relation type and scores the unordered pair alone. The per-class columns split gold and predicted pairs by whether the protagonist is one of the two endpoints, and report recall, precision, and F1 within each class, which are reported in Table 11. Because discarding unresolvable predictions is one of two defensible conventions, Table 12 repeats the pair-level column with those predictions kept in the precision denominator. Two greedy runs of one checkpoint differ by up to 0.03 F1 with the composition of the decoding batch, and we report paired bootstrap CIs for all method comparisons.

System	$\mathrm{Rec}^{\mathrm{Pro}}$	$\mathrm{Rec}^{\mathrm{Sup}}$	$\mathrm{Prec}^{\mathrm{Pro}}$	$\mathrm{Prec}^{\mathrm{Sup}}$	F_1^{Pro}	F_1^{Sup}
Qwen3.5-27B						
Base, 0-shot	0.67	0.31	0.68	0.43	0.67	0.36
Base, few-shot	0.74	0.46	0.64	0.43	0.69	0.45
Fine-tuned, standard target	0.71	0.42	0.84	0.54	0.77	0.47
Fine-tuned, evidence-anchored target	0.80	0.48	0.74	0.51	0.77	0.50
+ sentence pass + local-evidence selector	0.75	0.44	0.82	0.68	0.79	0.53
Qwen3.5-9B						
Base, 0-shot	0.46	0.19	0.77	0.46	0.57	0.27
Base, few-shot	0.58	0.31	0.64	0.48	0.61	0.38
Fine-tuned, standard target	0.65	0.34	0.71	0.42	0.68	0.38
Fine-tuned, evidence-anchored target	0.74	0.39	0.69	0.47	0.71	0.43
+ sentence pass + local-evidence selector	0.65	0.38	0.80	0.66	0.72	0.48
DeepSeek-v4-pro						
Few-shot, one pass	0.74	0.41	0.64	0.46	0.69	0.44
+ sentence pass + local-evidence selector	0.71	0.34	0.85	0.64	0.77	0.45

Table 11: Per-class pair-level metrics for every row of Table 2, BioRED test. Gold and predicted pairs are split by whether the protagonist is one of the two endpoints (431 protagonist pairs and 559 supporting-entity pairs of 990). The relation type is ignored, as in the Pair F1 column of Table 2.

System	Pair F1	Pair F1, unresolved as FP	unresolved predictions
Qwen3.5-27B, anchored target, one pass	0.62	0.58	125
+ schema validator	0.63	0.60	116
Qwen3.5-27B, union of the two passes	0.66	0.58	302
Qwen3.5-27B, full pipeline	0.65	0.64	39
Qwen3.5-9B, full pipeline	0.59	0.58	28
DeepSeek-v4-pro few-shot, one pass	0.55	0.49	263
DeepSeek-v4-pro few-shot, full pipeline	0.60	0.59	42
DeepSeek-v4-pro 0-shot, one pass	0.54	0.48	213
DeepSeek-v4-pro 0-shot, full pipeline	0.58	0.57	29

Table 12: Robustness of the pair-level column to the treatment of predictions whose endpoints resolve to no annotated concept. The main convention discards them; the middle column keeps them in the precision denominator. Under that convention the full pipeline is ahead of the unselected pool by 0.06 pair F1 (95% CI 0.02 to 0.10), because most of what the selector removes is unresolvable.

C Additional results

BC5CDR On BC5CDR, the anchored Qwen3.5-27B trained on BioRED and applied without adaptation has $\mathrm{Rec}^{\mathrm{Pro}}$ 0.81 and $\mathrm{Rec}^{\mathrm{Sup}}$ 0.67 (a recall gap of 14 points); the union with the sentence pass raises $\mathrm{Rec}^{\mathrm{Sup}}$ to 0.78 (a gap of 13) and recall of relations whose entities co-occur in one sentence from 0.73 to 0.88, and DeepSeek-v4-pro's recall on that tier rises from 0.60 to 0.74 0-shot and 0.67 to 0.79 few-shot. No selector was trained for BC5CDR, so these are pool numbers, and F1 falls with precision, from 0.476 to 0.442 for the 27B.

Same-sentence and cross-sentence relations The method is built on sentences, and 29% of BioRED's test relations (286 of 990) join entities that never share one. Scored separately, the gain

Selection over the pool of the anchored Qwen3.5-27B	F1	Pair F1	$\mathrm{Rec}^{\mathrm{Sup}}$	$\mathrm{Prec}^{\mathrm{Sup}}$	F_1^{Sup}
Whole-document pass alone	0.50	0.62	0.48	0.51	0.50
Local-evidence selector (full pipeline)	0.53	0.65	0.44	0.68	0.53
Selection by agreement of the two passes	0.36	0.43	0.19	0.67	0.30

Table 13: How the two-pass pool is selected, on BioRED test, under the scorer of Table 2. F1 is pair-and-type exact; the other columns are pair-level, and the last three are restricted to relations whose two endpoints are both supporting entities. The first and third rows are the anchored and full-pipeline rows of Table 2.

falls on the same-sentence relations and the cross-sentence subset loses: for the anchored Qwen3.5-27B, F1 goes from 0.56 to 0.59 within a sentence and from 0.40 to 0.35 across sentences; for the anchored Qwen3.5-9B, from 0.48 to 0.52 and from 0.36 to 0.28; for few-shot DeepSeek-v4-pro, from 0.47 to 0.51 and from 0.29 to 0.24. The method's gain is therefore confined to relations stated within a sentence.

Variants that did not help The variants below were evaluated under the previous, oracle-localized protocol, in which the selector could see the annotated mentions when locating evidence. Every variant was chosen on the validation set and run once on test. Distilling a Qwen3.5-9B from a teacher prompted sentence by sentence gives 0.335 F1 with $\mathrm{Rec}^{\mathrm{Sup}}$ 0.15, against 0.39 and 0.24 for the standard fine-tune. Protagonist-aware negatives for the selector cost 0.021 F1 and 0.05 the recall of supporting-entity relations; per-source thresholds gain 0.025 on validation and 0.003 on test; relabeling kept pairs with a 4-class relation classifier (validation accuracy 0.69, against 0.84 for the extractor's own type on the kept annotated pairs) costs 0.023; the selector shown the whole abstract, in the input format of the supervised systems (Lai et al., 2023), scores 0.53 against 0.55 for the local-evidence selector retrained on the same machine; and filling in, after selection, the pairs that a BioREx-style exhaustive-pair classifier (Lai et al., 2023) labels as related among the entities the extractor recognized raises recall from 0.50 to 0.55 and the recall of supporting-entity relations from 0.38 to 0.43 but leaves F1 unchanged, because the 199 pairs it adds are correct 28% of the time. Each failure is of the kind the diagnosis predicts: a selector cannot recover what the proposer never offered, and whole-document decisions are again exposed to the salience cues that shape proposal.

D MECHANISTIC EXPERIMENTS AND CROSS-MODEL RESULTS

This appendix supplies the protocols, uncertainty estimates, controls, and model-specific results underlying Section 2.3. Recovery of an internal score, intervention on attention, and changes in freely generated relations are distinct outcomes.

D.1 MODELS, UNITS, AND INTERVENTION SITES

M0 denotes Qwen3.5-9B without task fine-tuning; M1′ denotes its standard-target task fine-tune. M1–M4 denote evidence-format adapters trained with natural, random-drop, low-hub, and high-hub targets. Phi-4 and Gemma denote Phi-4 and Gemma-3-12B in the cross-family experiments. Layer numbers are zero-based. The intervention layers are 27 for the 9B models and Phi-4, and 41 for Gemma. Layer exploration uses validation data; score and generation comparisons are reported separately on test data.

Units and populations A record is a scored swap or generation target, not necessarily a distinct document. Each $n = N/D$ reports records and unique document clusters. We retain each panel's eligible population; denominators can differ across sites, arms, and controls. In particular, the cross-family 50-record generation panels do not contain 50 independent documents. Unless stated otherwise, tables use all eligible records. The clean decision panel additionally requires one evidence sentence, both entry surfaces in that sentence, an unchanged rendered evidence line, judge confidence at least 0.9, and no alias flattening. Treated clean swaps must transfer the document hub to the swap-in entity without conflicting other gold relations; clean controls preserve their required hub identity. Gold-degree matching is a separate, stricter stratum and is not imposed on the all-record intervention panels.

Uncertainty Intervals are 95% document-clustered bootstrap intervals, retaining the estimates and denominators from the archived reports. Fixed-prefix decisions use 2,000 draws and report archived Holm-adjusted p values; 9B test patching uses 1,000 draws. Cross-family input-site patching and teacher-forced contrasts use the archived 500-draw estimates. Cross-family decision-slot curves below are recomputed from existing raw records with 2,000 document-clustered draws (seed 0), without rerunning models. Other intervention intervals are pointwise: scanning layers, heads, arms, and doses does not constitute a multiplicity-adjusted discovery test. An interval containing zero does not establish equivalence.

Table 14: Intervention sites and measurement scope.

Symbol / panel	Definition
S_H, S_x	Last token of the relevant entity mention inside the unchanged evidence sentence. The input-site curve pools the applicable entity site across demotion and promotion pairs; it is not a simultaneous joint patch.
S_s	All aligned tokens of the unchanged evidence sentence.
S_{slot}	Entry-opener position whose output predicts the first token of the next relation's head name.
S_{pre}	Assistant-prefix positions preceding the decision slot.
S_{gen}	$S_{\text{pre}} \cup S_{\text{slot}}$, aligned by offset from the assistant start.
Score intervention	Attention reweighting at fixed relation-decision query positions.
Generation intervention	Reweighting across query positions during prompt processing and decoding; this is broader than score mode.

D.2 FIXED-PREFIX DECISION DECOMPOSITION

Generative information extraction linearizes structured triples and emits them autoregressively (Josifoski et al., 2022). For our serialization, once the preceding output prefix z is fixed, the probability of a particular next relation follows the chain rule:

$$p(\text{continue}, h, r, t \mid z, d) = p(\text{continue} \mid z, d)\, p(h \mid \text{continue}, z, d)\, p(r, t \mid h, \text{continue}, z, d), \quad (1)$$

where d is the document and (h, r, t) the next serialized relation. This grammar induces three operational readouts: closing the relation list (D_{stop}), selecting a head entity (D_{head}), and completing the

relation after a supplied head (D_{cond}). They are format-aligned conditional measurements, not universal cognitive modules or an additive causal mediation decomposition. $P_\emptyset$ is the empty-relation prefix; P_{+x} additionally includes x in the candidate list; P_{by} follows a bystander relation. Each reported D is the counterfactual-minus-original change with its paired control subtracted. A negative head contrast reduces the protagonist's advantage over x; it is not the same sign convention as the original-minus-counterfactual recovery denominator below. Conditional completion uses mean teacher-forced relation-and-tail log probability.

Tables 15 and 16 include both the all and clean populations, pooled over validation and test records. Missing or single-document intervals are shown explicitly. The standard-target fine-tune also has a strong output-order constraint: in the audited demotion panel the top-scoring first head is the first listed candidate in 582/582 original inputs, with mean top-1 probability 0.994. A movable head margin therefore need not change the selected head under this prefix.

Table 15: 9B M0: paired-control-adjusted decision contrasts. p_H is the archived Holm-adjusted p value; units are nats except probabilities and ranks.

Population	Decision	Demotion	Promotion
all	$D_{\text{stop}}(P_{\text{by}})$	-0.022 $[-0.227, +0.188]$ $n = 65/25$ $p_H = 1$	-0.046 $[-0.757, +0.638]$ $n = 19/5$ $p_H = 1$
all	$D_{\text{stop}}(P_\emptyset)$	$+0.092$ $[-0.089, +0.286]$ $n = 578/126$ $p_H = 0.888$	$+0.047$ $[-0.554, +0.567]$ $n = 114/36$ $p_H = 0.915$
all	$D_{\text{head}}(P_\emptyset) : H - x$	-3.007 $[-3.814, -2.187]$ $n = 384/79$ $p_H = 6.24 \times 10^{-13}$	-2.690 $[-4.824, -1.122]$ $n = 98/32$ $p_H = 0.000813$
all	$D_{\text{head}}(P_{+x}) : H - x$	-3.895 $[-5.172, -2.320]$ $n = 165/56$ $p_H = 1.58 \times 10^{-8}$	–
all	$D_{\text{head}}(P_{\text{by}}) : H - x$	-4.012 $[-5.440, -2.184]$ $n = 42/14$ $p_H = 9.96 \times 10^{-8}$	-3.280 $[-8.985, -0.419]$ $n = 15/4$
all	$\Delta P_H(P_\emptyset)$	-0.151 $[-0.200, -0.104]$ $n = 578/126$ $p_H = 3.14 \times 10^{-11}$	-0.095 $[-0.210, -0.011]$ $n = 99/32$ $p_H = 0.0741$
all	$\Delta\text{rank}_H(P_\emptyset)$	$+0.647$ $[+0.383, +0.926]$ $n = 578/126$ $p_H = 3.57 \times 10^{-9}$	$+1.939$ $[+0.734, +3.509]$ $n = 99/32$ $p_H = 0.00981$
all	D_{cond}	$+0.061$ $[-0.005, +0.133]$ $n = 578/126$ $p_H = 0.165$	-0.044 $[-0.132, +0.039]$ $n = 114/36$ $p_H = 0.507$
clean	$D_{\text{stop}}(P_{\text{by}})$	$+0.063$ $[-0.246, +0.625]$ $n = 6/3$	-0.950 $[-1.250, +0.250]$ $n = 5/2$
clean	$D_{\text{stop}}(P_\emptyset)$	$+0.314$ $[-0.090, +0.656]$ $n = 175/53$ $p_H = 0.189$	-0.572 $[-1.890, +0.662]$ $n = 34/15$ $p_H = 0.477$
clean	$D_{\text{head}}(P_\emptyset) : H - x$	-3.664 $[-5.346, -2.202]$ $n = 107/34$ $p_H = 2.58 \times 10^{-5}$	-2.856 $[-5.494, -0.649]$ $n = 20/11$ $p_H = 0.0318$

Continued on next page

Table 15 (continued)

Population	Decision	Demotion	Promotion
clean	$D_{\mathrm{head}}(P_{+x}) : H - x$	-5.846 $[-7.590, -3.238]$ $n = 62/20$ $p_H = 1.7 \times 10^{-11}$	–
clean	$D_{\mathrm{head}}(P_{\mathrm{by}}) : H - x$	–	-0.923 [NA, NA] $n = 1/1$
clean	$\Delta P_H(P_\emptyset)$	-0.231 $[-0.336, -0.141]$ $n = 175/53$ $p_H = 7.36 \times 10^{-6}$	-0.080 $[-0.177, +0.002]$ $n = 20/11$ $p_H = 0.0811$
clean	$\Delta\mathrm{rank}_H(P_\emptyset)$	$+0.897$ $[+0.539, +1.333]$ $n = 175/53$ $p_H = 6.9 \times 10^{-6}$	$+0.850$ $[+0.231, +1.857]$ $n = 20/11$ $p_H = 0.00459$
clean	D_{cond}	$+0.149$ $[+0.043, +0.268]$ $n = 175/53$ $p_H = 0.0268$	$+0.016$ $[-0.034, +0.090]$ $n = 34/15$ $p_H = 0.28$

Table 16: 9B M1′: paired-control-adjusted decision contrasts. p_H is the archived Holm-adjusted p value; units are nats except probabilities and ranks.

Population	Decision	Demotion	Promotion
all	$D_{\mathrm{stop}}(P_{\mathrm{by}})$	$+0.184$ $[-0.329, +0.618]$ $n = 65/25$ $p_H = 0.879$	$+0.332$ $[-0.544, +0.917]$ $n = 19/5$ $p_H = 0.879$
all	$D_{\mathrm{stop}}(P_\emptyset)$	$+0.023$ $[-0.109, +0.162]$ $n = 578/126$ $p_H = 0.952$	-0.243 $[-0.540, +0.015]$ $n = 114/36$ $p_H = 0.0998$
all	$D_{\mathrm{head}}(P_\emptyset) : H - x$	-0.839 $[-1.638, -0.035]$ $n = 384/79$ $p_H = 0.00774$	-1.139 $[-1.955, -0.388]$ $n = 98/32$ $p_H = 0.00164$
all	$D_{\mathrm{head}}(P_{+x}) : H - x$	-2.224 $[-3.277, -1.361]$ $n = 165/56$ $p_H = 4.58 \times 10^{-6}$	–
all	$D_{\mathrm{head}}(P_{\mathrm{by}}) : H - x$	-1.421 $[-2.377, -0.890]$ $n = 42/14$ $p_H = 0.000154$	-0.366 $[-0.892, +0.219]$ $n = 15/4$
all	$\Delta P_H(P_\emptyset)$	$+0.000$ $[-0.000, +0.001]$ $n = 578/126$ $p_H = 0.828$	-0.000 $[-0.000, +0.000]$ $n = 99/32$ $p_H = 0.828$
all	$\Delta\mathrm{rank}_H(P_\emptyset)$	-0.237 $[-0.530, +0.029]$ $n = 578/126$ $p_H = 0.506$	$+0.222$ $[-0.333, +0.733]$ $n = 99/32$ $p_H = 0.506$
all	D_{cond}	$+0.017$ $[-0.005, +0.040]$ $n = 578/126$ $p_H = 0.367$	$+0.002$ $[-0.056, +0.059]$ $n = 114/36$ $p_H = 0.949$
clean	$D_{\mathrm{stop}}(P_{\mathrm{by}})$	$+0.208$ $[+0.063, +0.624]$ $n = 6/3$	-0.050 $[-0.750, +0.125]$ $n = 5/2$

Continued on next page

Table 16 (continued)

Population	Decision	Demotion	Promotion
clean	$D_{\text{stop}}(P_\emptyset)$	$+0.110$ $[-0.164, +0.374]$ $n = 175/53$ $p_H = 0.473$	-0.337 $[-0.987, +0.224]$ $n = 34/15$ $p_H = 0.473$
clean	$D_{\text{head}}(P_\emptyset) : H - x$	-1.993 $[-3.077, -1.012]$ $n = 107/34$ $p_H = 0.000328$	-1.012 $[-2.191, +0.121]$ $n = 20/11$ $p_H = 0.00789$
clean	$D_{\text{head}}(P_{+x}) : H - x$	-3.329 $[-5.289, -1.979]$ $n = 62/20$ $p_H = 2.88 \times 10^{-5}$	–
clean	$D_{\text{head}}(P_{\text{by}}) : H - x$	–	$+1.273$ [NA, NA] $n = 1/1$
clean	$\Delta P_H(P_\emptyset)$	$+0.001$ $[-0.002, +0.004]$ $n = 175/53$ $p_H = 0.835$	$+0.000$ $[-0.001, +0.001]$ $n = 20/11$ $p_H = 0.835$
clean	$\Delta\text{rank}_H(P_\emptyset)$	$+0.034$ $[-0.581, +0.646]$ $n = 175/53$ $p_H = 1$	-0.250 $[-0.945, +0.190]$ $n = 20/11$ $p_H = 1$
clean	D_{cond}	$+0.037$ $[+0.004, +0.073]$ $n = 175/53$ $p_H = 0.0899$	$+0.018$ $[-0.037, +0.081]$ $n = 34/15$ $p_H = 0.628$

D.3 Activation patching and output-site decomposition

Let $m = \log p(H) - \log p(x)$ and $e = m_{\text{orig}} - m_{\text{cf}}$ be the natural text-swap effect. Denoising replaces a counterfactual state by its original-document state and measures $(m_{\text{patch}} - m_{\text{cf}})/e$. Noising replaces an original state by its counterfactual counterpart and measures $(m_{\text{orig}} - m_{\text{patch}})/e$. We include treated demotion and promotion pairs with $|e| \geq 0.5$ nats, pool record-level ratios, and bootstrap by document. Ratios are not clipped to $[0, 1]$.

Residual patches replace the layer output at aligned sites. The tables retain every sampled residual layer in the test report, including supplementary input-site layers with smaller denominators. The output-site split is a separate validation experiment. Large late-layer recovery identifies a state carrying the decision, not necessarily the point where the causal signal arose; normalized recovery alone does not prove a unique circuit.

Table 17: 9B M0: test residual-state recovery across all sampled layers.

Site	Layer	Denoising recovery	Noising loss
S_H/S_x	2	$+0.000$ $[-0.005, +0.007]$ $n = 39/17$	-0.009 $[-0.028, +0.010]$ $n = 39/17$
S_H/S_x	4	$+0.007$ $[-0.002, +0.021]$ $n = 39/17$	-0.001 $[-0.011, +0.010]$ $n = 39/17$
S_H/S_x	8	$+0.001$ $[-0.007, +0.009]$ $n = 172/61$	-0.000 $[-0.006, +0.006]$ $n = 172/61$
S_H/S_x	16	-0.004 $[-0.008, -0.000]$ $n = 172/61$	-0.001 $[-0.006, +0.006]$ $n = 172/61$

Continued on next page

Table 17 (continued)

Site	Layer	Denoising recovery	Noising loss
S_H/S_x	24	-0.001 $[-0.005, +0.003]$ $n = 172/61$	-0.002 $[-0.006, +0.003]$ $n = 172/61$
S_s	8	-0.024 $[-0.042, -0.008]$ $n = 172/61$	-0.000 $[-0.015, +0.013]$ $n = 172/61$
S_s	16	-0.019 $[-0.034, -0.003]$ $n = 172/61$	-0.004 $[-0.022, +0.016]$ $n = 172/61$
S_s	24	$+0.001$ $[-0.009, +0.011]$ $n = 172/61$	-0.006 $[-0.019, +0.006]$ $n = 172/61$
S_{gen}	11	$+0.219$ $[+0.100, +0.325]$ $n = 172/61$	$+0.336$ $[+0.216, +0.457]$ $n = 172/61$
S_{gen}	14	$+0.221$ $[+0.104, +0.321]$ $n = 172/61$	$+0.339$ $[+0.224, +0.443]$ $n = 172/61$
S_{gen}	15	$+0.270$ $[+0.170, +0.373]$ $n = 172/61$	$+0.402$ $[+0.301, +0.504]$ $n = 172/61$
S_{gen}	16	$+0.275$ $[+0.164, +0.375]$ $n = 172/61$	$+0.410$ $[+0.309, +0.513]$ $n = 172/61$
S_{gen}	17	$+0.289$ $[+0.169, +0.393]$ $n = 172/61$	$+0.411$ $[+0.304, +0.518]$ $n = 172/61$
S_{gen}	18	$+0.293$ $[+0.176, +0.395]$ $n = 172/61$	$+0.418$ $[+0.313, +0.521]$ $n = 172/61$
S_{gen}	19	$+0.361$ $[+0.254, +0.467]$ $n = 172/61$	$+0.493$ $[+0.376, +0.598]$ $n = 172/61$
S_{gen}	20	$+0.352$ $[+0.231, +0.471]$ $n = 172/61$	$+0.483$ $[+0.363, +0.597]$ $n = 172/61$
S_{gen}	21	$+0.354$ $[+0.241, +0.465]$ $n = 172/61$	$+0.482$ $[+0.353, +0.601]$ $n = 172/61$
S_{gen}	22	$+0.358$ $[+0.236, +0.465]$ $n = 172/61$	$+0.488$ $[+0.362, +0.602]$ $n = 172/61$
S_{gen}	23	$+0.485$ $[+0.380, +0.581]$ $n = 172/61$	$+0.617$ $[+0.509, +0.720]$ $n = 172/61$
S_{gen}	24	$+0.485$ $[+0.392, +0.583]$ $n = 172/61$	$+0.615$ $[+0.496, +0.722]$ $n = 172/61$
S_{gen}	25	$+0.499$ $[+0.398, +0.594]$ $n = 172/61$	$+0.630$ $[+0.532, +0.724]$ $n = 172/61$
S_{gen}	26	$+0.500$ $[+0.408, +0.589]$ $n = 172/61$	$+0.634$ $[+0.534, +0.728]$ $n = 172/61$
S_{gen}	27	$+0.738$ $[+0.693, +0.785]$ $n = 172/61$	$+0.817$ $[+0.758, +0.867]$ $n = 172/61$
S_{gen}	28	$+0.756$ $[+0.709, +0.798]$ $n = 172/61$	$+0.830$ $[+0.766, +0.877]$ $n = 172/61$

Continued on next page

Table 17 (continued)

Site	Layer	Denoising recovery	Noising loss
S_{gen}	29	$+0.773$ $[+0.730, +0.814]$ $n = 172/61$	$+0.845$ $[+0.790, +0.889]$ $n = 172/61$
S_{gen}	30	$+0.791$ $[+0.744, +0.834]$ $n = 172/61$	$+0.851$ $[+0.799, +0.894]$ $n = 172/61$
S_{gen}	31	$+0.943$ $[+0.896, +0.978]$ $n = 172/61$	$+0.941$ $[+0.891, +0.975]$ $n = 172/61$

Table 18: 9B M1′: test residual-state recovery across all sampled layers.

Site	Layer	Denoising recovery	Noising loss
S_H/S_x	2	$+0.004$ $[-0.024, +0.028]$ $n = 33/15$	-0.016 $[-0.043, +0.013]$ $n = 33/15$
S_H/S_x	4	$+0.015$ $[-0.023, +0.048]$ $n = 33/15$	-0.005 $[-0.035, +0.023]$ $n = 33/15$
S_H/S_x	8	-0.015 $[-0.042, +0.008]$ $n = 157/59$	$+0.006$ $[-0.012, +0.023]$ $n = 157/59$
S_H/S_x	16	-0.017 $[-0.038, +0.001]$ $n = 157/59$	$+0.011$ $[-0.011, +0.032]$ $n = 157/59$
S_H/S_x	24	-0.005 $[-0.014, +0.006]$ $n = 157/59$	$+0.001$ $[-0.018, +0.021]$ $n = 157/59$
S_s	8	-0.057 $[-0.096, -0.023]$ $n = 157/59$	-0.009 $[-0.036, +0.021]$ $n = 157/59$
S_s	16	-0.066 $[-0.103, -0.034]$ $n = 157/59$	-0.050 $[-0.085, -0.016]$ $n = 157/59$
S_s	24	-0.044 $[-0.067, -0.021]$ $n = 157/59$	-0.047 $[-0.085, -0.005]$ $n = 157/59$
S_{gen}	11	-0.074 $[-0.300, +0.182]$ $n = 157/59$	$+0.188$ $[-0.026, +0.429]$ $n = 157/59$
S_{gen}	14	-0.076 $[-0.295, +0.166]$ $n = 157/59$	$+0.188$ $[-0.038, +0.450]$ $n = 157/59$
S_{gen}	15	-0.065 $[-0.272, +0.158]$ $n = 157/59$	$+0.100$ $[-0.090, +0.315]$ $n = 157/59$
S_{gen}	16	-0.067 $[-0.262, +0.150]$ $n = 157/59$	$+0.087$ $[-0.108, +0.312]$ $n = 157/59$
S_{gen}	17	-0.061 $[-0.272, +0.171]$ $n = 157/59$	$+0.095$ $[-0.103, +0.293]$ $n = 157/59$
S_{gen}	18	-0.062 $[-0.284, +0.179]$ $n = 157/59$	$+0.117$ $[-0.078, +0.329]$ $n = 157/59$
S_{gen}	19	-0.050 $[-0.250, +0.174]$ $n = 157/59$	$+0.142$ $[-0.035, +0.364]$ $n = 157/59$

Continued on next page

Table 18 (continued)

Site	Layer	Denoising recovery	Noising loss
S_{gen}	20	-0.067 $[-0.266, +0.163]$ $n = 157/59$	$+0.111$ $[-0.083, +0.306]$ $n = 157/59$
S_{gen}	21	-0.068 $[-0.264, +0.146]$ $n = 157/59$	$+0.115$ $[-0.068, +0.327]$ $n = 157/59$
S_{gen}	22	-0.064 $[-0.250, +0.155]$ $n = 157/59$	$+0.121$ $[-0.068, +0.339]$ $n = 157/59$
S_{gen}	23	$+0.038$ $[-0.138, +0.233]$ $n = 157/59$	$+0.249$ $[+0.084, +0.435]$ $n = 157/59$
S_{gen}	24	$+0.037$ $[-0.144, +0.232]$ $n = 157/59$	$+0.249$ $[+0.067, +0.439]$ $n = 157/59$
S_{gen}	25	$+0.054$ $[-0.131, +0.266]$ $n = 157/59$	$+0.270$ $[+0.094, +0.436]$ $n = 157/59$
S_{gen}	26	$+0.070$ $[-0.111, +0.274]$ $n = 157/59$	$+0.285$ $[+0.111, +0.448]$ $n = 157/59$
S_{gen}	27	$+0.732$ $[+0.657, +0.808]$ $n = 157/59$	$+0.824$ $[+0.768, +0.877]$ $n = 157/59$
S_{gen}	28	$+0.744$ $[+0.680, +0.806]$ $n = 157/59$	$+0.835$ $[+0.781, +0.884]$ $n = 157/59$
S_{gen}	29	$+0.778$ $[+0.719, +0.840]$ $n = 157/59$	$+0.856$ $[+0.808, +0.904]$ $n = 157/59$
S_{gen}	30	$+0.788$ $[+0.723, +0.850]$ $n = 157/59$	$+0.869$ $[+0.818, +0.914]$ $n = 157/59$
S_{gen}	31	$+0.968$ $[+0.920, +1.005]$ $n = 157/59$	$+0.967$ $[+0.926, +1.005]$ $n = 157/59$

Table 19: Validation decomposition of S_{gen} into decision slot and assistant prefix. Components need not add linearly.

Model / layer	Direction	S_{slot}	S_{pre}	S_{gen}
9B M0 / L23	denoise	$+0.375$ $[+0.307, +0.447]$ $n = 168/64$	$+0.202$ $[+0.115, +0.293]$ $n = 168/64$	$+0.569$ $[+0.475, +0.659]$ $n = 168/64$
9B M0 / L23	noise	$+0.401$ $[+0.307, +0.491]$ $n = 168/64$	$+0.274$ $[+0.169, +0.353]$ $n = 168/64$	$+0.660$ $[+0.538, +0.765]$ $n = 168/64$
9B M0 / L27	denoise	$+0.632$ $[+0.561, +0.705]$ $n = 168/64$	$+0.082$ $[+0.041, +0.119]$ $n = 168/64$	$+0.742$ $[+0.683, +0.810]$ $n = 168/64$
9B M0 / L27	noise	$+0.751$ $[+0.686, +0.820]$ $n = 168/64$	$+0.155$ $[+0.106, +0.199]$ $n = 168/64$	$+0.819$ $[+0.755, +0.872]$ $n = 168/64$
9B M1′ / L23	denoise	-0.146 $[-0.382, +0.090]$ $n = 163/60$	$+0.086$ $[-0.214, +0.357]$ $n = 163/60$	$+0.167$ $[-0.060, +0.390]$ $n = 163/60$

Continued on next page

Table 19 (continued)

Model / layer	Direction	S_{slot}	S_{pre}	S_{gen}
9B M1′ / L23	noise	$+0.251$ $[-0.058, +0.635]$ $n = 163/60$	$+0.451$ $[+0.251, +0.665]$ $n = 163/60$	$+0.354$ $[+0.193, +0.539]$ $n = 163/60$
9B M1′ / L27	denoise	$+0.448$ $[+0.316, +0.590]$ $n = 163/60$	$+0.168$ $[+0.026, +0.274]$ $n = 163/60$	$+0.725$ $[+0.639, +0.807]$ $n = 163/60$
9B M1′ / L27	noise	$+0.640$ $[+0.498, +0.807]$ $n = 163/60$	$+0.318$ $[+0.227, +0.409]$ $n = 163/60$	$+0.835$ $[+0.763, +0.902]$ $n = 163/60$

D.4 Attention knockout and individual heads

At the decision slot, knockout masks attention to a class of key positions. The original-side loss is $(m_{\text{base}} - m_{\text{knock}})/e$; counterfactual-side recovery is $(m_{\text{knock}} - m_{\text{base}})/e$. Matched-control columns instead report the unnormalized change Δm in nats. They must not be read as recovery fractions or mechanically subtracted from ratios computed on different records.

The title, out-of-sentence mentions ($H_{\text{out}}, x_{\text{out}}$), evidence sentence ($s$), and remaining document text (other) test document sources. The pre-entity classes refer to the output-side candidate list. Effects at pre-H and pre-x also occur in matched controls and therefore do not specifically identify document-role mediation. Joint or whole-document masks need not equal the sum of constituent masks. Individual heads are ablated at layer 27; all-head and selected-head combinations are reported alongside them.

Identity tests within one forward pass are near numerical zero (approximately 7×10^{-5} and 2×10^{-5} nats for M0 and M1′). Cross-batch identity differences are larger (approximately 0.064 and 0.0017 nats, respectively). The former is an implementation check; the latter is a raw-score numerical floor, not a normalized recovery threshold.

Table 20: 9B validation source knockouts at layer 27 and available joint-layer masks. Each statistic retains its own record/document denominator.

Model / layer	Source	Original loss	CF recovery	Matched Δm
9B M0 / L27	H_out	$+0.028$ $[-0.014, +0.069]$ $n = 167/63$	-0.006 $[-0.035, +0.029]$ $n = 153/60$	-0.168 $[-0.260, -0.071]$ $n = 60/40$
9B M0 / L27	all_cues	$+0.025$ $[-0.026, +0.087]$ $n = 167/63$	-0.001 $[-0.113, +0.080]$ $n = 167/63$	$+0.006$ $[-0.191, +0.152]$ $n = 60/40$
9B M0 / L27	doc	$+0.103$ $[+0.032, +0.191]$ $n = 167/63$	$+0.038$ $[-0.035, +0.109]$ $n = 167/63$	-0.377 $[-0.628, -0.146]$ $n = 60/40$
9B M0 / L27	not_s	$+0.046$ $[-0.006, +0.097]$ $n = 167/63$	$+0.071$ $[-0.022, +0.145]$ $n = 167/63$	-0.190 $[-0.459, +0.045]$ $n = 60/40$
9B M0 / L27	other	$+0.046$ $[+0.011, +0.079]$ $n = 167/63$	$+0.015$ $[-0.029, +0.059]$ $n = 167/63$	-0.238 $[-0.381, -0.102]$ $n = 60/40$
9B M0 / L27	pre_H	$+0.155$ $[+0.068, +0.241]$ $n = 164/61$	-0.230 $[-0.330, -0.133]$ $n = 164/61$	-1.034 $[-1.389, -0.687]$ $n = 59/39$
9B M0 / L27	pre_all	-0.221 $[-0.383, -0.085]$ $n = 167/63$	-0.084 $[-0.162, +0.005]$ $n = 167/63$	$+0.420$ $[-0.100, +0.943]$ $n = 60/40$
9B M0 / L27	pre_by	-0.009 $[-0.040, +0.018]$ $n = 39/17$	-0.009 $[-0.037, +0.010]$ $n = 39/17$	$+0.019$ $[-0.123, +0.161]$ $n = 15/12$

Continued on next page

Table 20 (continued)

Model / layer	Source	Original loss	CF recovery	Matched Δm
9B M0 / L27	pre_ents	-0.218	-0.084	$+0.398$
		$[-0.380, -0.077]$	$[-0.164, +0.006]$	$[-0.136, +0.904]$
		$n = 167/63$	$n = 167/63$	$n = 60/40$
9B M0 / L27	pre_x	-0.356	$+0.166$	$+1.580$
		$[-0.533, -0.180]$	$[+0.058, +0.288]$	$[+1.283, +1.938]$
		$n = 167/63$	$n = 167/63$	$n = 60/40$
9B M0 / L27	s	$+0.032$	-0.021	-0.078
		$[+0.006, +0.071]$	$[-0.060, +0.008]$	$[-0.153, -0.011]$
		$n = 167/63$	$n = 167/63$	$n = 60/40$
9B M0 / L27	title	$+0.046$	-0.023	-0.150
		$[+0.010, +0.097]$	$[-0.076, +0.016]$	$[-0.240, -0.068]$
		$n = 158/61$	$n = 158/61$	$n = 55/38$
9B M0 / L27	tmpl	$+0.015$	$+0.011$	-0.022
		$[+0.003, +0.030]$	$[-0.007, +0.029]$	$[-0.080, +0.034]$
		$n = 167/63$	$n = 167/63$	$n = 60/40$
9B M0 / L27	x_out	-0.035	$+0.025$	$+0.379$
		$[-0.071, +0.009]$	$[-0.080, +0.091]$	$[+0.244, +0.511]$
		$n = 153/60$	$n = 167/63$	$n = 55/37$
9B M1′ / L27	H_out	$+0.186$	-0.045	-0.697
		$[+0.066, +0.303]$	$[-0.154, +0.058]$	$[-1.012, -0.446]$
		$n = 160/61$	$n = 149/60$	$n = 60/40$
9B M1′ / L27	all_cues	$+0.129$	$+0.309$	-0.506
		$[+0.005, +0.256]$	$[+0.156, +0.467]$	$[-0.866, -0.186]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	doc	$+0.172$	$+0.286$	-0.584
		$[+0.023, +0.341]$	$[+0.143, +0.434]$	$[-0.945, -0.237]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	not_s	$+0.135$	$+0.328$	-0.360
		$[+0.026, +0.263]$	$[+0.192, +0.476]$	$[-0.734, +0.017]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	other	$+0.018$	-0.048	$+0.046$
		$[-0.040, +0.081]$	$[-0.102, +0.008]$	$[-0.122, +0.201]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	pre_H	$+0.312$	-0.377	-1.224
		$[+0.133, +0.524]$	$[-0.572, -0.194]$	$[-1.658, -0.872]$
		$n = 156/58$	$n = 156/58$	$n = 59/39$
9B M1′ / L27	pre_all	-0.154	-0.321	$+0.807$
		$[-0.487, +0.238]$	$[-0.562, -0.115]$	$[+0.071, +1.573]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	pre_by	-0.013	$+0.006$	$+0.154$
		$[-0.072, +0.048]$	$[-0.051, +0.045]$	$[-0.013, +0.387]$
		$n = 37/18$	$n = 37/18$	$n = 15/12$
9B M1′ / L27	pre_ents	-0.165	-0.308	$+0.826$
		$[-0.489, +0.197]$	$[-0.529, -0.106]$	$[-0.010, +1.565]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	pre_x	-0.280	$+0.166$	$+1.498$
		$[-0.562, +0.031]$	$[-0.062, +0.360]$	$[+1.076, +2.020]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	s	$+0.013$	-0.054	-0.111
		$[-0.047, +0.069]$	$[-0.131, +0.020]$	$[-0.292, +0.086]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	title	$+0.124$	$+0.090$	-0.570
		$[+0.019, +0.235]$	$[+0.011, +0.180]$	$[-0.839, -0.319]$
		$n = 153/59$	$n = 153/59$	$n = 55/38$
9B M1′ / L27	tmpl	$+0.003$	-0.030	$+0.081$
		$[-0.043, +0.051]$	$[-0.072, +0.013]$	$[-0.043, +0.204]$
		$n = 160/61$	$n = 160/61$	$n = 60/40$
9B M1′ / L27	x_out	-0.103	$+0.354$	$+0.425$
		$[-0.223, -0.000]$	$[+0.218, +0.519]$	$[+0.241, +0.640]$
		$n = 149/60$	$n = 160/61$	$n = 55/37$

Table 21: All individual layer-27 head ablations and tested combinations (validation).

Model	Head(s)	Original loss	CF recovery	Matched Δm
9B M0	head0	-0.009	$+0.006$	$+0.078$
		$[-0.040, +0.011]$	$[-0.008, +0.027]$	$[-0.018, +0.250]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head1	$+0.035$	$+0.012$	-0.022
		$[+0.010, +0.064]$	$[-0.011, +0.034]$	$[-0.114, +0.070]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head2	-0.014	-0.003	$+0.053$
		$[-0.029, -0.002]$	$[-0.011, +0.006]$	$[+0.002, +0.115]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head3	$+0.001$	$+0.009$	$+0.034$
		$[-0.015, +0.016]$	$[-0.002, +0.019]$	$[-0.027, +0.095]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head4	$+0.039$	-0.041	-0.064
		$[-0.032, +0.126]$	$[-0.123, +0.030]$	$[-0.241, +0.071]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head5	$+0.109$	-0.098	-0.108
		$[-0.051, +0.321]$	$[-0.263, +0.056]$	$[-0.798, +0.609]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head6	-0.010	$+0.017$	$+0.051$
		$[-0.072, +0.044]$	$[-0.035, +0.081]$	$[-0.299, +0.347]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head7	$+0.007$	-0.001	-0.013
		$[-0.010, +0.026]$	$[-0.014, +0.012]$	$[-0.062, +0.038]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head8	$+0.006$	$+0.002$	-0.022
		$[-0.027, +0.039]$	$[-0.025, +0.033]$	$[-0.125, +0.072]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head9	$+0.008$	$+0.009$	-0.032
		$[-0.002, +0.017]$	$[-0.002, +0.024]$	$[-0.073, +0.011]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head10	$+0.001$	$+0.019$	-0.094
		$[-0.175, +0.137]$	$[-0.099, +0.170]$	$[-0.584, +0.396]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head11	-0.051	$+0.013$	$+0.223$
		$[-0.146, +0.034]$	$[-0.076, +0.101]$	$[-0.121, +0.643]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head12	$+0.063$	-0.041	$+0.057$
		$[-0.082, +0.250]$	$[-0.221, +0.110]$	$[-0.185, +0.339]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head13	$+0.027$	$+0.009$	-0.049
		$[+0.007, +0.046]$	$[-0.009, +0.028]$	$[-0.122, +0.027]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head14	$+0.010$	-0.004	-0.003
		$[-0.000, +0.022]$	$[-0.012, +0.002]$	$[-0.024, +0.021]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head15	$+0.002$	$+0.009$	$+0.097$
		$[-0.019, +0.021]$	$[-0.002, +0.025]$	$[+0.013, +0.237]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	head_all	$+0.064$	-0.155	$+0.507$
		$[-0.159, +0.293]$	$[-0.385, +0.055]$	$[-0.269, +1.266]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	top2	$+0.145$	-0.124	-0.070
		$[+0.001, +0.362]$	$[-0.286, +0.007]$	$[-0.816, +0.707]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M0	top3	$+0.175$	-0.171	-0.141
		$[-0.003, +0.398]$	$[-0.336, -0.024]$	$[-0.818, +0.635]$
		$n = 115/55$	$n = 115/55$	$n = 40/29$
9B M1′	head0	$+0.023$	-0.008	$+0.080$
		$[-0.030, +0.097]$	$[-0.084, +0.044]$	$[-0.027, +0.267]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$

Continued on next page

Table 21 (continued)

Model	Head(s)	Original loss	CF recovery	Matched Δm
9B M1′	head1	-0.037	$+0.020$	$+0.110$
		$[-0.120, +0.043]$	$[-0.056, +0.094]$	$[-0.073, +0.280]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head2	-0.015	$+0.030$	$+0.029$
		$[-0.047, +0.017]$	$[+0.007, +0.057]$	$[-0.057, +0.095]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head3	$+0.012$	$+0.014$	-0.079
		$[-0.016, +0.042]$	$[-0.008, +0.037]$	$[-0.175, +0.001]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head4	-0.076	$+0.074$	-0.124
		$[-0.194, +0.029]$	$[-0.026, +0.198]$	$[-0.376, +0.129]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head5	$+0.031$	-0.034	-0.141
		$[-0.213, +0.261]$	$[-0.257, +0.193]$	$[-0.897, +0.619]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head6	-0.037	$+0.065$	$+0.056$
		$[-0.157, +0.063]$	$[-0.036, +0.193]$	$[-0.367, +0.421]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head7	$+0.026$	$+0.010$	-0.011
		$[-0.001, +0.061]$	$[-0.013, +0.030]$	$[-0.108, +0.084]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head8	$+0.015$	-0.020	$+0.087$
		$[-0.053, +0.088]$	$[-0.098, +0.050]$	$[-0.080, +0.246]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head9	$+0.023$	-0.001	-0.013
		$[-0.000, +0.053]$	$[-0.023, +0.022]$	$[-0.077, +0.048]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head10	$+0.144$	-0.071	-0.273
		$[-0.046, +0.353]$	$[-0.270, +0.133]$	$[-0.844, +0.233]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head11	$+0.178$	-0.052	$+0.069$
		$[-0.064, +0.458]$	$[-0.339, +0.186]$	$[-0.344, +0.536]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head12	$+0.074$	-0.028	$+0.100$
		$[-0.093, +0.282]$	$[-0.247, +0.134]$	$[-0.295, +0.545]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head13	$+0.005$	-0.010	$+0.020$
		$[-0.075, +0.097]$	$[-0.091, +0.067]$	$[-0.135, +0.166]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head14	$+0.017$	$+0.003$	-0.033
		$[+0.001, +0.038]$	$[-0.018, +0.020]$	$[-0.082, +0.010]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head15	$+0.006$	$+0.001$	$+0.057$
		$[-0.010, +0.025]$	$[-0.020, +0.021]$	$[-0.030, +0.195]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	head_all	$+0.265$	-0.024	-0.315
		$[-0.238, +0.845]$	$[-0.540, +0.429]$	$[-1.594, +1.169]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	top2	$+0.266$	-0.124	-0.210
		$[+0.040, +0.548]$	$[-0.398, +0.123]$	$[-0.747, +0.439]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
9B M1′	top3	$+0.347$	-0.176	-0.124
		$[+0.018, +0.720]$	$[-0.528, +0.131]$	$[-0.724, +0.553]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$

D.5 SCORE STEERING: DOSE CURVES AND CONTROLS

For selected key positions, the intervention adds $\log \alpha$ to the attention logits before normalization. Thus $\alpha = 1$ leaves the attention distribution unchanged and $\alpha = 0$ masks the keys. The x arm targets document mentions of x; the H_{out} arm targets protagonist mentions outside the evidence sentence. Despite the archived name `lower_H`, the latter enhances H_{out} for $\alpha > 1$ and suppresses

it for $\alpha < 1$. The joint arm enhances x and suppresses H_{out} by reciprocal factors. Random-position controls use the same number of random document positions outside the target class; bystander controls target the bystander's mentions. In the joint arm the random control uses two disjoint sets with opposite logit shifts, whereas the bystander control uses a single set with the positive shift.

All score deltas compare the intervention with its own $\alpha = 1$ baseline on the same record and side. The principal dose panel uses $\alpha \in \{0, 0.5, 2, 4\}$; validation and adapter extensions also include other recorded doses. A negative Δm favors x relative to H. Validation curves are separated from test curves; a small positive control effect is not treated as identically zero.

Table 22: Test head-margin changes (Δm, nats): complete dose curves for targets, random positions, and bystanders.

Model / arm / control	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
Gemma / L41 H_{out}, bystander $n = 69/22$	$+0.030$ $[-0.083, +0.195]$	$+0.004$ $[-0.072, +0.107]$	$+0.056$ $[-0.097, +0.233]$	$+0.104$ $[-0.126, +0.349]$
Gemma / L41 H_{out}, random $n = 177/63$	$+0.023$ $[-0.019, +0.070]$	$+0.001$ $[-0.035, +0.037]$	$+0.026$ $[-0.017, +0.071]$	$+0.051$ $[-0.015, +0.122]$
Gemma / L41 H_{out}, target $n = 177/63$	-0.542 $[-0.722, -0.367]$	-0.186 $[-0.250, -0.125]$	$+0.288$ $[+0.186, +0.412]$	$+0.632$ $[+0.395, +0.951]$
Gemma / L41 x mentions, bystander $n = 69/22$	$+0.030$ $[-0.086, +0.175]$	$+0.004$ $[-0.067, +0.101]$	$+0.056$ $[-0.100, +0.219]$	$+0.104$ $[-0.175, +0.339]$
Gemma / L41 x mentions, random $n = 178/63$	$+0.017$ $[-0.020, +0.058]$	$+0.006$ $[-0.028, +0.055]$	-0.000 $[-0.040, +0.038]$	$+0.011$ $[-0.021, +0.050]$
Gemma / L41 x mentions, target $n = 178/63$	$+0.511$ $[+0.381, +0.665]$	$+0.246$ $[+0.173, +0.318]$	-0.370 $[-0.455, -0.280]$	-0.920 $[-1.157, -0.733]$
9B M0 / L27 H_{out}, bystander $n = 69/22$	-0.018 $[-0.096, +0.053]$	$+0.022$ $[-0.021, +0.063]$	$+0.022$ $[-0.064, +0.111]$	$+0.037$ $[-0.157, +0.237]$
9B M0 / L27 H_{out}, random $n = 174/62$	-0.002 $[-0.033, +0.028]$	$+0.005$ $[-0.013, +0.023]$	$+0.030$ $[+0.004, +0.060]$	$+0.058$ $[-0.014, +0.135]$
9B M0 / L27 H_{out}, target $n = 174/62$	-0.380 $[-0.555, -0.218]$	-0.158 $[-0.236, -0.087]$	$+0.258$ $[+0.161, +0.362]$	$+0.568$ $[+0.384, +0.768]$
9B M0 / L27 x mentions, bystander $n = 69/22$	-0.018 $[-0.099, +0.056]$	$+0.022$ $[-0.022, +0.064]$	$+0.022$ $[-0.063, +0.102]$	$+0.037$ $[-0.149, +0.224]$
9B M0 / L27 x mentions, random $n = 175/62$	-0.001 $[-0.015, +0.014]$	$+0.013$ $[-0.002, +0.028]$	$+0.016$ $[-0.000, +0.032]$	$+0.023$ $[-0.006, +0.051]$
9B M0 / L27 x mentions, target $n = 175/62$	$+0.527$ $[+0.334, +0.740]$	$+0.254$ $[+0.165, +0.351]$	-0.373 $[-0.520, -0.245]$	-0.845 $[-1.099, -0.574]$
9B M1′ / L27 H_{out}, bystander $n = 69/22$	$+0.015$ $[-0.078, +0.111]$	$+0.002$ $[-0.049, +0.055]$	-0.034 $[-0.090, +0.027]$	-0.064 $[-0.201, +0.078]$
9B M1′ / L27 H_{out}, random $n = 174/62$	$+0.047$ $[+0.017, +0.076]$	$+0.024$ $[-0.007, +0.052]$	-0.007 $[-0.042, +0.027]$	-0.043 $[-0.098, +0.008]$
9B M1′ / L27 H_{out}, target $n = 174/62$	-0.613 $[-0.783, -0.439]$	-0.267 $[-0.342, -0.193]$	$+0.472$ $[+0.339, +0.605]$	$+1.080$ $[+0.793, +1.368]$

Continued on next page

Table 22 (continued)

Model / arm / control	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
9B M1′ / L27 x mentions, bystander $n = 69/22$	$+0.015$ $[-0.081, +0.108]$	$+0.002$ $[-0.051, +0.055]$	-0.034 $[-0.088, +0.031]$	-0.064 $[-0.193, +0.077]$
9B M1′ / L27 x mentions, random $n = 175/62$	$+0.016$ $[-0.012, +0.044]$	$+0.017$ $[-0.003, +0.037]$	$+0.011$ $[-0.014, +0.039]$	$+0.014$ $[-0.025, +0.057]$
9B M1′ / L27 x mentions, target $n = 175/62$	$+0.446$ $[+0.331, +0.557]$	$+0.211$ $[+0.161, +0.268]$	-0.354 $[-0.459, -0.253]$	-0.954 $[-1.220, -0.719]$
Phi-4 / L27 H_{out}, bystander $n = 69/22$	$+0.034$ $[-0.012, +0.086]$	-0.008 $[-0.038, +0.024]$	-0.045 $[-0.080, -0.005]$	-0.044 $[-0.108, +0.029]$
Phi-4 / L27 H_{out}, random $n = 177/63$	$+0.000$ $[-0.018, +0.017]$	$+0.005$ $[-0.010, +0.022]$	-0.007 $[-0.022, +0.007]$	-0.032 $[-0.054, -0.011]$
Phi-4 / L27 H_{out}, target $n = 177/63$	-0.107 $[-0.142, -0.073]$	-0.056 $[-0.078, -0.034]$	$+0.056$ $[+0.025, +0.086]$	$+0.163$ $[+0.106, +0.229]$
Phi-4 / L27 x mentions, bystander $n = 69/22$	$+0.034$ $[-0.013, +0.082]$	-0.008 $[-0.042, +0.022]$	-0.045 $[-0.082, -0.003]$	-0.044 $[-0.113, +0.029]$
Phi-4 / L27 x mentions, random $n = 178/63$	-0.016 $[-0.034, +0.001]$	-0.010 $[-0.027, +0.007]$	-0.012 $[-0.027, +0.005]$	-0.002 $[-0.017, +0.016]$
Phi-4 / L27 x mentions, target $n = 178/63$	$+0.047$ $[+0.031, +0.064]$	$+0.022$ $[+0.007, +0.036]$	-0.057 $[-0.078, -0.034]$	-0.141 $[-0.177, -0.104]$

Table 23: Validation target-dose exploration, including Phi-4's additional layers. Test claims use the held-out panel above.

Model / layer	Arm	α	Δm	Records / docs
9B M0 / L23	H_{out}	0	-0.079 $[-0.120, -0.040]$	148/65
9B M0 / L23	H_{out}	0.25	-0.053 $[-0.083, -0.025]$	148/65
9B M0 / L23	H_{out}	0.5	-0.033 $[-0.058, -0.008]$	148/65
9B M0 / L23	H_{out}	2	$+0.043$ $[+0.012, +0.078]$	148/65
9B M0 / L23	H_{out}	4	$+0.124$ $[+0.054, +0.199]$	148/65
9B M0 / L23	H_{out}	8	$+0.166$ $[+0.043, +0.298]$	148/65
9B M0 / L23	x mentions	0	$+0.094$ $[+0.050, +0.147]$	148/65
9B M0 / L23	x mentions	0.25	$+0.085$ $[+0.046, +0.131]$	148/65
9B M0 / L23	x mentions	0.5	$+0.045$ $[+0.014, +0.085]$	148/65
9B M0 / L23	x mentions	2	-0.101 $[-0.136, -0.074]$	148/65
9B M0 / L23	x mentions	4	-0.219 $[-0.286, -0.158]$	148/65
9B M0 / L23	x mentions	8	-0.410 $[-0.538, -0.299]$	148/65
9B M0 / L27	H_{out}	0	-0.269 $[-0.400, -0.173]$	148/65

Continued on next page

Table 23 (continued)

Model / layer	Arm	α	Δm	Records / docs
9B M0 / L27	H_{out}	0.25	-0.184 $[-0.258, -0.119]$	148/65
9B M0 / L27	H_{out}	0.5	-0.114 $[-0.156, -0.076]$	148/65
9B M0 / L27	H_{out}	2	$+0.184$ $[+0.128, +0.248]$	148/65
9B M0 / L27	H_{out}	4	$+0.470$ $[+0.349, +0.606]$	148/65
9B M0 / L27	H_{out}	8	$+0.863$ $[+0.662, +1.095]$	148/65
9B M0 / L27	x mentions	0	$+0.380$ $[+0.279, +0.483]$	148/65
9B M0 / L27	x mentions	0.25	$+0.275$ $[+0.204, +0.348]$	148/65
9B M0 / L27	x mentions	0.5	$+0.169$ $[+0.128, +0.221]$	148/65
9B M0 / L27	x mentions	2	-0.276 $[-0.351, -0.203]$	148/65
9B M0 / L27	x mentions	4	-0.698 $[-0.882, -0.539]$	148/65
9B M0 / L27	x mentions	8	-1.229 $[-1.555, -0.940]$	148/65
9B M1′ / L27	H_{out}	0	-0.822 $[-1.094, -0.572]$	148/65
9B M1′ / L27	H_{out}	0.25	-0.570 $[-0.754, -0.408]$	148/65
9B M1′ / L27	H_{out}	0.5	-0.365 $[-0.476, -0.261]$	148/65
9B M1′ / L27	H_{out}	2	$+0.512$ $[+0.369, +0.666]$	148/65
9B M1′ / L27	H_{out}	4	$+1.170$ $[+0.897, +1.449]$	148/65
9B M1′ / L27	H_{out}	8	$+1.856$ $[+1.467, +2.258]$	148/65
9B M1′ / L27	x mentions	0	$+0.456$ $[+0.299, +0.654]$	148/65
9B M1′ / L27	x mentions	0.25	$+0.309$ $[+0.196, +0.440]$	148/65
9B M1′ / L27	x mentions	0.5	$+0.187$ $[+0.120, +0.259]$	148/65
9B M1′ / L27	x mentions	2	-0.347 $[-0.443, -0.259]$	148/65
9B M1′ / L27	x mentions	4	-0.840 $[-1.044, -0.654]$	148/65
9B M1′ / L27	x mentions	8	-1.518 $[-1.858, -1.230]$	148/65
Gemma / L41	H_{out}	0	-0.278 $[-0.485, -0.066]$	128/60
Gemma / L41	H_{out}	0.5	-0.120 $[-0.234, +0.017]$	128/60
Gemma / L41	H_{out}	2	$+0.169$ $[-0.038, +0.358]$	128/60
Gemma / L41	H_{out}	4	$+0.408$ $[+0.065, +0.743]$	128/60
Gemma / L41	H_{out}	8	$+0.744$ $[+0.252, +1.205]$	128/60
Gemma / L41	x mentions	0	$+0.616$ $[+0.447, +0.842]$	128/60
Gemma / L41	x mentions	0.5	$+0.279$ $[+0.180, +0.376]$	128/60

Continued on next page

Table 23 (continued)

Model / layer	Arm	α	Δm	Records / docs
Gemma / L41	x mentions	2	-0.443 $[-0.583, -0.300]$	128/60
Gemma / L41	x mentions	4	-1.056 $[-1.384, -0.748]$	128/60
Gemma / L41	x mentions	8	-1.786 $[-2.334, -1.377]$	128/60
Phi-4 / L23	H_{out}	0	$+0.010$ $[-0.017, +0.035]$	128/60
Phi-4 / L23	H_{out}	0.5	$+0.023$ $[+0.000, +0.042]$	128/60
Phi-4 / L23	H_{out}	2	$+0.009$ $[-0.009, +0.028]$	128/60
Phi-4 / L23	H_{out}	4	$+0.025$ $[-0.011, +0.060]$	128/60
Phi-4 / L23	H_{out}	8	$+0.022$ $[-0.030, +0.087]$	128/60
Phi-4 / L23	x mentions	0	$+0.017$ $[-0.010, +0.040]$	128/60
Phi-4 / L23	x mentions	0.5	$+0.015$ $[-0.004, +0.034]$	128/60
Phi-4 / L23	x mentions	2	$+0.009$ $[-0.009, +0.025]$	128/60
Phi-4 / L23	x mentions	4	$+0.019$ $[-0.005, +0.043]$	128/60
Phi-4 / L23	x mentions	8	-0.017 $[-0.049, +0.015]$	128/60
Phi-4 / L27	H_{out}	0	-0.087 $[-0.120, -0.055]$	128/60
Phi-4 / L27	H_{out}	0.5	-0.027 $[-0.053, -0.003]$	128/60
Phi-4 / L27	H_{out}	2	$+0.076$ $[+0.056, +0.096]$	128/60
Phi-4 / L27	H_{out}	4	$+0.179$ $[+0.136, +0.229]$	128/60
Phi-4 / L27	H_{out}	8	$+0.307$ $[+0.244, +0.383]$	128/60
Phi-4 / L27	x mentions	0	$+0.100$ $[+0.070, +0.136]$	128/60
Phi-4 / L27	x mentions	0.5	$+0.053$ $[+0.033, +0.074]$	128/60
Phi-4 / L27	x mentions	2	-0.059 $[-0.085, -0.033]$	128/60
Phi-4 / L27	x mentions	4	-0.162 $[-0.231, -0.105]$	128/60
Phi-4 / L27	x mentions	8	-0.317 $[-0.419, -0.227]$	128/60
Phi-4 / L31	H_{out}	0	-0.008 $[-0.038, +0.027]$	128/60
Phi-4 / L31	H_{out}	0.5	-0.002 $[-0.020, +0.015]$	128/60
Phi-4 / L31	H_{out}	2	$+0.039$ $[+0.013, +0.064]$	128/60
Phi-4 / L31	H_{out}	4	$+0.063$ $[+0.025, +0.104]$	128/60
Phi-4 / L31	H_{out}	8	$+0.108$ $[+0.048, +0.181]$	128/60
Phi-4 / L31	x mentions	0	$+0.070$ $[+0.040, +0.114]$	128/60
Phi-4 / L31	x mentions	0.5	$+0.025$ $[+0.005, +0.046]$	128/60

Continued on next page

Table 23 (continued)

Model / layer	Arm	α	Δm	Records / docs
Phi-4 / L31	x mentions	2	-0.013 $[-0.038, +0.009]$	128/60
Phi-4 / L31	x mentions	4	-0.093 $[-0.150, -0.045]$	128/60
Phi-4 / L31	x mentions	8	-0.180 $[-0.273, -0.094]$	128/60
Phi-4 / L35	H_{out}	0	-0.027 $[-0.047, -0.006]$	128/60
Phi-4 / L35	H_{out}	0.5	-0.011 $[-0.028, +0.004]$	128/60
Phi-4 / L35	H_{out}	2	$+0.059$ $[+0.036, +0.082]$	128/60
Phi-4 / L35	H_{out}	4	$+0.115$ $[+0.089, +0.142]$	128/60
Phi-4 / L35	H_{out}	8	$+0.213$ $[+0.169, +0.257]$	128/60
Phi-4 / L35	x mentions	0	$+0.069$ $[+0.041, +0.097]$	128/60
Phi-4 / L35	x mentions	0.5	$+0.036$ $[+0.019, +0.049]$	128/60
Phi-4 / L35	x mentions	2	-0.043 $[-0.064, -0.022]$	128/60
Phi-4 / L35	x mentions	4	-0.087 $[-0.125, -0.054]$	128/60
Phi-4 / L35	x mentions	8	-0.181 $[-0.248, -0.121]$	128/60

The dose slope is computed for each record by regressing Δm on $\log_2 \alpha$ over positive doses, including the baseline $(\alpha, \Delta m) = (1, 0)$; record slopes are then averaged with document-clustered intervals. The masked $\alpha = 0$ condition is excluded from the log-dose fit.

Table 24: Test score slopes for all four models and all target/control arms, in nats per doubling of dose.

Model / layer	Arm	Control	Slope	Records / docs
9B M0 / L27	H_{out}	bystander	$+0.007$ $[-0.068, +0.085]$	69/22
9B M0 / L27	H_{out}	random	$+0.019$ $[-0.008, +0.048]$	174/62
9B M0 / L27	H_{out}	target	$+0.244$ $[+0.160, +0.349]$	174/62
9B M0 / L27	x mentions	bystander	$+0.007$ $[-0.073, +0.093]$	69/22
9B M0 / L27	x mentions	random	$+0.004$ $[-0.007, +0.015]$	175/62
9B M0 / L27	x mentions	target	-0.367 $[-0.505, -0.249]$	175/62
9B M1′ / L27	H_{out}	bystander	-0.023 $[-0.077, +0.037]$	69/22
9B M1′ / L27	H_{out}	random	-0.021 $[-0.040, -0.003]$	174/62
9B M1′ / L27	H_{out}	target	$+0.451$ $[+0.328, +0.570]$	174/62
9B M1′ / L27	x mentions	bystander	-0.023 $[-0.079, +0.036]$	69/22
9B M1′ / L27	x mentions	random	$+0.000$ $[-0.015, +0.017]$	175/62

Continued on next page

Table 24 (continued)

Model / layer	Arm	Control	Slope	Records / docs
9B M1′ / L27	x mentions	target	-0.385 $[-0.490, -0.293]$	175/62
Gemma / L41	H_{out}	bystander	$+0.035$ $[-0.078, +0.122]$	69/22
Gemma / L41	H_{out}	random	$+0.018$ $[-0.008, +0.042]$	177/63
Gemma / L41	H_{out}	target	$+0.274$ $[+0.186, +0.370]$	177/63
Gemma / L41	x mentions	bystander	$+0.035$ $[-0.078, +0.126]$	69/22
Gemma / L41	x mentions	random	$+0.001$ $[-0.010, +0.013]$	178/63
Gemma / L41	x mentions	target	-0.387 $[-0.480, -0.302]$	178/63
Phi-4 / L27	H_{out}	bystander	-0.015 $[-0.046, +0.013]$	69/22
Phi-4 / L27	H_{out}	random	-0.012 $[-0.019, -0.004]$	177/63
Phi-4 / L27	H_{out}	target	$+0.071$ $[+0.048, +0.100]$	177/63
Phi-4 / L27	x mentions	bystander	-0.015 $[-0.046, +0.013]$	69/22
Phi-4 / L27	x mentions	random	$+0.001$ $[-0.004, +0.008]$	178/63
Phi-4 / L27	x mentions	target	-0.055 $[-0.068, -0.042]$	178/63

D.6 Free generation: participation, controls, and quality

Generation uses greedy decoding with the intervention active across query positions at the selected layer, including prompt processing and decoding. It therefore tests control under a broader protocol than fixed-prefix scoring. For each condition, the baseline is the unmodified generation from the same run and chunk.

$P_x = P(x \in \text{rel})$ is the fraction of target records whose output contains at least one relation with x as an endpoint; P_H is defined analogously. These are participation rates, not relation precision or recall. For a metric Y, the paired random-control-adjusted change is

$$\Delta_{\text{adj}} Y = (Y_{\text{target},\alpha} - Y_{\text{target},1}) - (Y_{\text{random},\alpha} - Y_{\text{random},1}).$$

Only records with both contrasts enter this paired estimate. Consequently, the adjusted estimate cannot in general be reconstructed by subtracting the full-panel averages in Table 25. Rates and their deltas are proportions; changes in gold and non-gold relations are counts per target record.

The M1′ x-arm raw panel contains 100 records from 57 documents; the adjusted panel retains 94 records from the same 57 documents. At $\alpha = 4$, the adjusted participation increase is 0.138 [0.055, 0.228], whereas the gold-relation count change is 0.064 [-0.047, 0.182]. Phi-4's 0.080 [0.019, 0.158] adjusted participation contrast coexists with an unchanged raw treated rate (0.56) and a random-control decline (0.56 to 0.48). Its entity F1 decreases at this dose. Gemma's adjusted participation contrast is negative and its interval reaches zero. These results support controllability in the tested fine-tuned extractor, but neither universal monotonic generation control nor a consistent extraction-quality gain.

Table 25: Test generation: same-run baseline → intervention participation rates. All tested doses and controls are retained.

Model	Arm	Control	α	Records (docs)	P_x	P_H
9B M0	H_{out}	target	0	100 (44)	0.410→0.440	0.930→0.800
9B M0	H_{out}	target	0.5	100 (44)	0.410→0.420	0.930→0.910
9B M0	H_{out}	target	2	100 (44)	0.410→0.420	0.930→0.860
9B M0	H_{out}	target	4	100 (44)	0.410→0.430	0.930→0.790
9B M0	H_{out}	bystander	0	25 (22)	0.280→0.360	0.920→0.960
9B M0	H_{out}	bystander	0.5	25 (22)	0.280→0.280	0.920→0.880
9B M0	H_{out}	bystander	2	25 (22)	0.280→0.280	0.920→0.880
9B M0	H_{out}	bystander	4	25 (22)	0.280→0.320	0.920→0.920
9B M0	H_{out}	random	0	100 (44)	0.400→0.410	0.940→0.900
9B M0	H_{out}	random	0.5	100 (44)	0.400→0.420	0.940→0.950
9B M0	H_{out}	random	2	100 (44)	0.400→0.400	0.940→0.930
9B M0	H_{out}	random	4	100 (44)	0.400→0.380	0.940→0.910
9B M0	x mentions	target	0	100 (57)	0.380→0.340	0.910→0.930
9B M0	x mentions	target	0.5	100 (57)	0.380→0.360	0.910→0.900
9B M0	x mentions	target	2	100 (57)	0.380→0.410	0.910→0.880
9B M0	x mentions	target	4	100 (57)	0.380→0.430	0.910→0.890
9B M0	x mentions	bystander	0	25 (22)	0.280→0.320	0.920→0.880
9B M0	x mentions	bystander	0.5	25 (22)	0.280→0.280	0.920→0.880
9B M0	x mentions	bystander	2	25 (22)	0.280→0.320	0.920→0.960
9B M0	x mentions	bystander	4	25 (22)	0.280→0.280	0.920→0.960
9B M0	x mentions	random	0	100 (61)	0.410→0.390	0.930→0.950
9B M0	x mentions	random	0.5	100 (61)	0.410→0.410	0.930→0.930
9B M0	x mentions	random	2	100 (61)	0.410→0.400	0.930→0.920
9B M0	x mentions	random	4	100 (61)	0.410→0.430	0.930→0.900
9B M1′	H_{out}	target	0	100 (44)	0.360→0.370	1.000→0.870
9B M1′	H_{out}	target	0.5	100 (44)	0.360→0.390	1.000→1.000
9B M1′	H_{out}	target	2	100 (44)	0.360→0.330	1.000→1.000
9B M1′	H_{out}	target	4	100 (44)	0.360→0.260	1.000→0.980
9B M1′	H_{out}	bystander	0	25 (22)	0.400→0.360	1.000→0.960
9B M1′	H_{out}	bystander	0.5	25 (22)	0.400→0.400	1.000→0.960
9B M1′	H_{out}	bystander	2	25 (22)	0.400→0.360	1.000→0.960
9B M1′	H_{out}	bystander	4	25 (22)	0.400→0.320	1.000→0.960
9B M1′	H_{out}	random	0	100 (44)	0.360→0.310	1.000→0.980
9B M1′	H_{out}	random	0.5	100 (44)	0.360→0.350	1.000→1.000
9B M1′	H_{out}	random	2	100 (44)	0.360→0.350	1.000→0.970
9B M1′	H_{out}	random	4	100 (44)	0.360→0.380	1.000→0.970
9B M1′	x mentions	target	0	100 (57)	0.500→0.310	1.000→0.990
9B M1′	x mentions	target	0.5	100 (57)	0.500→0.440	1.000→0.990
9B M1′	x mentions	target	2	100 (57)	0.500→0.590	1.000→0.990
9B M1′	x mentions	target	4	100 (57)	0.500→0.660	1.000→1.000
9B M1′	x mentions	bystander	0	25 (22)	0.400→0.360	1.000→0.960
9B M1′	x mentions	bystander	0.5	25 (22)	0.400→0.400	1.000→1.000
9B M1′	x mentions	bystander	2	25 (22)	0.400→0.400	1.000→1.000
9B M1′	x mentions	bystander	4	25 (22)	0.400→0.320	1.000→0.960
9B M1′	x mentions	random	0	100 (61)	0.510→0.540	1.000→0.990
9B M1′	x mentions	random	0.5	100 (61)	0.510→0.520	1.000→1.000
9B M1′	x mentions	random	2	100 (61)	0.510→0.540	1.000→1.000
9B M1′	x mentions	random	4	100 (61)	0.510→0.540	1.000→1.000
Phi-4	H_{out}	target	0	50 (23)	0.640→0.560	0.960→0.920
Phi-4	H_{out}	target	0.5	50 (23)	0.640→0.580	0.960→0.920
Phi-4	H_{out}	target	2	50 (23)	0.640→0.540	0.960→0.960
Phi-4	H_{out}	target	4	50 (23)	0.640→0.480	0.960→0.940
Phi-4	H_{out}	bystander	0	22 (19)	0.318→0.318	0.864→0.864

Continued on next page

Table 25 (continued)

Model	Arm	Control	α	Records (docs)	P_x	P_H
Phi-4	H_{out}	bystander	0.5	22 (19)	0.318→0.364	0.864→0.864
Phi-4	H_{out}	bystander	2	22 (19)	0.318→0.318	0.864→0.909
Phi-4	H_{out}	bystander	4	22 (19)	0.318→0.227	0.864→0.909
Phi-4	H_{out}	random	0	50 (23)	0.640→0.560	0.960→0.960
Phi-4	H_{out}	random	0.5	50 (23)	0.640→0.560	0.960→0.960
Phi-4	H_{out}	random	2	50 (23)	0.640→0.600	0.960→0.940
Phi-4	H_{out}	random	4	50 (23)	0.640→0.580	0.960→0.960
Phi-4	x mentions	target	0	50 (29)	0.560→0.460	0.880→0.880
Phi-4	x mentions	target	0.5	50 (29)	0.560→0.500	0.880→0.880
Phi-4	x mentions	target	2	50 (29)	0.560→0.560	0.880→0.880
Phi-4	x mentions	target	4	50 (29)	0.560→0.560	0.880→0.860
Phi-4	x mentions	bystander	0	22 (19)	0.318→0.318	0.864→0.864
Phi-4	x mentions	bystander	0.5	22 (19)	0.318→0.318	0.864→0.864
Phi-4	x mentions	bystander	2	22 (19)	0.318→0.273	0.864→0.909
Phi-4	x mentions	bystander	4	22 (19)	0.318→0.227	0.864→0.909
Phi-4	x mentions	random	0	50 (29)	0.560→0.540	0.880→0.880
Phi-4	x mentions	random	0.5	50 (29)	0.560→0.540	0.880→0.880
Phi-4	x mentions	random	2	50 (29)	0.560→0.560	0.880→0.880
Phi-4	x mentions	random	4	50 (29)	0.560→0.480	0.880→0.860
Gemma	H_{out}	target	0	50 (23)	0.460→0.520	0.960→0.980
Gemma	H_{out}	target	0.5	50 (23)	0.460→0.500	0.960→0.980
Gemma	H_{out}	target	2	50 (23)	0.460→0.480	0.960→1.000
Gemma	H_{out}	target	4	50 (23)	0.460→0.460	0.960→1.000
Gemma	H_{out}	bystander	0	22 (19)	0.364→0.500	0.909→0.955
Gemma	H_{out}	bystander	0.5	22 (19)	0.364→0.455	0.909→0.909
Gemma	H_{out}	bystander	2	22 (19)	0.364→0.409	0.909→0.955
Gemma	H_{out}	bystander	4	22 (19)	0.364→0.409	0.909→1.000
Gemma	H_{out}	random	0	50 (23)	0.460→0.460	0.960→0.980
Gemma	H_{out}	random	0.5	50 (23)	0.460→0.460	0.960→0.980
Gemma	H_{out}	random	2	50 (23)	0.460→0.460	0.960→0.980
Gemma	H_{out}	random	4	50 (23)	0.460→0.440	0.960→0.960
Gemma	x mentions	target	0	50 (29)	0.460→0.460	0.920→0.940
Gemma	x mentions	target	0.5	50 (29)	0.460→0.460	0.920→0.940
Gemma	x mentions	target	2	50 (29)	0.460→0.460	0.920→0.960
Gemma	x mentions	target	4	50 (29)	0.460→0.420	0.920→0.960
Gemma	x mentions	bystander	0	22 (19)	0.364→0.455	0.909→0.955
Gemma	x mentions	bystander	0.5	22 (19)	0.364→0.455	0.909→0.909
Gemma	x mentions	bystander	2	22 (19)	0.364→0.409	0.909→0.955
Gemma	x mentions	bystander	4	22 (19)	0.364→0.409	0.909→1.000
Gemma	x mentions	random	0	50 (31)	0.460→0.460	0.920→0.940
Gemma	x mentions	random	0.5	50 (31)	0.460→0.460	0.920→0.940
Gemma	x mentions	random	2	50 (31)	0.460→0.460	0.920→0.940
Gemma	x mentions	random	4	50 (31)	0.460→0.460	0.920→0.960

Table 26: Test generation: paired random-control-adjusted participation changes, with 95% document-clustered intervals. Denominators are records (documents).

Model / arm / dose / sample	Adjusted ΔP_x	Adjusted ΔP_H
9B M0 H_{out}, $\alpha = 0$ 99 (44)	$+0.020$ $[-0.074, +0.113]$	-0.091 $[-0.200, +0.011]$

Continued on next page

Table 26 (continued)

Model / arm / dose / sample	Adjusted ΔP_x	Adjusted ΔP_H
9B M0 H_{out}, $\alpha = 0.5$ 99 (44)	-0.010 $[-0.094, +0.085]$	-0.030 $[-0.084, +0.000]$
9B M0 H_{out}, $\alpha = 2$ 99 (44)	$+0.010$ $[-0.063, +0.074]$	-0.061 $[-0.139, +0.000]$
9B M0 H_{out}, $\alpha = 4$ 99 (44)	$+0.040$ $[-0.100, +0.178]$	-0.111 $[-0.221, -0.028]$
9B M0 x mentions, $\alpha = 0$ 94 (57)	-0.021 $[-0.098, +0.052]$	-0.011 $[-0.051, +0.024]$
9B M0 x mentions, $\alpha = 0.5$ 94 (57)	-0.032 $[-0.110, +0.044]$	-0.021 $[-0.058, +0.000]$
9B M0 x mentions, $\alpha = 2$ 94 (57)	$+0.053$ $[-0.018, +0.131]$	-0.032 $[-0.072, +0.000]$
9B M0 x mentions, $\alpha = 4$ 94 (57)	$+0.043$ $[-0.023, +0.111]$	$+0.011$ $[-0.042, +0.066]$
9B M1′ H_{out}, $\alpha = 0$ 99 (44)	$+0.061$ $[-0.028, +0.161]$	-0.111 $[-0.198, -0.039]$
9B M1′ H_{out}, $\alpha = 0.5$ 99 (44)	$+0.040$ $[-0.011, +0.094]$	$+0.000$ $[+0.000, +0.000]$
9B M1′ H_{out}, $\alpha = 2$ 99 (44)	-0.020 $[-0.080, +0.036]$	$+0.030$ $[+0.000, +0.080]$
9B M1′ H_{out}, $\alpha = 4$ 99 (44)	-0.121 $[-0.210, -0.036]$	$+0.000$ $[-0.044, +0.053]$
9B M1′ x mentions, $\alpha = 0$ 94 (57)	-0.234 $[-0.316, -0.159]$	$+0.000$ $[-0.028, +0.030]$
9B M1′ x mentions, $\alpha = 0.5$ 94 (57)	-0.085 $[-0.153, -0.022]$	-0.011 $[-0.034, +0.000]$
9B M1′ x mentions, $\alpha = 2$ 94 (57)	$+0.064$ $[+0.021, +0.116]$	-0.011 $[-0.034, +0.000]$
9B M1′ x mentions, $\alpha = 4$ 94 (57)	$+0.138$ $[+0.055, +0.228]$	$+0.000$ $[+0.000, +0.000]$
Phi-4 H_{out}, $\alpha = 0$ 50 (23)	$+0.000$ $[-0.073, +0.077]$	-0.040 $[-0.120, +0.000]$
Phi-4 H_{out}, $\alpha = 0.5$ 50 (23)	$+0.020$ $[-0.052, +0.079]$	-0.040 $[-0.098, +0.000]$
Phi-4 H_{out}, $\alpha = 2$ 50 (23)	-0.060 $[-0.235, +0.059]$	$+0.020$ $[+0.000, +0.061]$
Phi-4 H_{out}, $\alpha = 4$ 50 (23)	-0.100 $[-0.290, +0.070]$	-0.020 $[-0.068, +0.000]$

Continued on next page

Table 26 (continued)

Model / arm / dose / sample	Adjusted ΔP_x	Adjusted ΔP_H
Phi-4 x mentions, $\alpha = 0$ 50 (29)	-0.080 $[-0.182, +0.000]$	$+0.000$ $[+0.000, +0.000]$
Phi-4 x mentions, $\alpha = 0.5$ 50 (29)	-0.040 $[-0.098, +0.000]$	$+0.000$ $[+0.000, +0.000]$
Phi-4 x mentions, $\alpha = 2$ 50 (29)	$+0.000$ $[-0.059, +0.055]$	$+0.000$ $[+0.000, +0.000]$
Phi-4 x mentions, $\alpha = 4$ 50 (29)	$+0.080$ $[+0.019, +0.158]$	$+0.000$ $[+0.000, +0.000]$
Gemma H_{out}, $\alpha = 0$ 50 (23)	$+0.060$ $[-0.109, +0.213]$	$+0.000$ $[+0.000, +0.000]$
Gemma H_{out}, $\alpha = 0.5$ 50 (23)	$+0.040$ $[+0.000, +0.121]$	$+0.000$ $[+0.000, +0.000]$
Gemma H_{out}, $\alpha = 2$ 50 (23)	$+0.020$ $[-0.043, +0.098]$	$+0.020$ $[+0.000, +0.073]$
Gemma H_{out}, $\alpha = 4$ 50 (23)	$+0.020$ $[-0.114, +0.135]$	$+0.040$ $[+0.000, +0.110]$
Gemma x mentions, $\alpha = 0$ 48 (29)	$+0.000$ $[-0.059, +0.068]$	$+0.000$ $[+0.000, +0.000]$
Gemma x mentions, $\alpha = 0.5$ 48 (29)	$+0.000$ $[+0.000, +0.000]$	$+0.000$ $[+0.000, +0.000]$
Gemma x mentions, $\alpha = 2$ 48 (29)	$+0.000$ $[-0.053, +0.059]$	$+0.021$ $[+0.000, +0.074]$
Gemma x mentions, $\alpha = 4$ 48 (29)	-0.042 $[-0.132, +0.000]$	$+0.000$ $[-0.050, +0.063]$

Table 27: Test generation: paired adjusted changes in gold relations involving each entity and entity F1.

Model / arm / dose / sample	Δx gold count	ΔH gold count	Δ entity F1
9B M0 H_{out}, $\alpha = 0$ 99 (44)	-0.010 $[-0.101, +0.073]$	$+0.000$ $[-0.229, +0.256]$	-0.011 $[-0.061, +0.032]$
9B M0 H_{out}, $\alpha = 0.5$ 99 (44)	$+0.020$ $[-0.050, +0.093]$	-0.172 $[-0.553, +0.069]$	$+0.018$ $[-0.009, +0.048]$
9B M0 H_{out}, $\alpha = 2$ 99 (44)	-0.010 $[-0.061, +0.022]$	$+0.051$ $[-0.165, +0.271]$	-0.016 $[-0.044, +0.006]$
9B M0 H_{out}, $\alpha = 4$ 99 (44)	-0.020 $[-0.099, +0.069]$	-0.192 $[-0.441, +0.038]$	-0.047 $[-0.098, -0.004]$
9B M0 x mentions, $\alpha = 0$ 94 (57)	-0.011 $[-0.064, +0.033]$	$+0.000$ $[-0.138, +0.159]$	-0.008 $[-0.024, +0.012]$

Continued on next page

Table 27 (continued)

Model / arm / dose / sample	Δx gold count	ΔH gold count	Δ entity F1
9B M0 x mentions, $\alpha = 0.5$ 94 (57)	-0.021 $[-0.063, +0.021]$	-0.287 $[-0.852, +0.036]$	$+0.007$ $[-0.009, +0.023]$
9B M0 x mentions, $\alpha = 2$ 94 (57)	-0.021 $[-0.066, +0.020]$	-0.138 $[-0.319, +0.029]$	-0.019 $[-0.038, -0.002]$
9B M0 x mentions, $\alpha = 4$ 94 (57)	$+0.011$ $[-0.035, +0.073]$	-0.223 $[-0.426, -0.049]$	-0.009 $[-0.031, +0.013]$
9B M1′ H_{out}, $\alpha = 0$ 99 (44)	$+0.020$ $[-0.053, +0.094]$	-0.010 $[-0.274, +0.236]$	$+0.028$ $[-0.003, +0.065]$
9B M1′ H_{out}, $\alpha = 0.5$ 99 (44)	$+0.051$ $[-0.009, +0.124]$	$+0.232$ $[-0.088, +0.633]$	$+0.015$ $[+0.000, +0.034]$
9B M1′ H_{out}, $\alpha = 2$ 99 (44)	$+0.000$ $[-0.058, +0.056]$	$+0.495$ $[-0.033, +1.320]$	$+0.012$ $[-0.019, +0.048]$
9B M1′ H_{out}, $\alpha = 4$ 99 (44)	-0.081 $[-0.167, -0.011]$	$+0.141$ $[-0.221, +0.670]$	-0.020 $[-0.044, +0.002]$
9B M1′ x mentions, $\alpha = 0$ 94 (57)	-0.043 $[-0.136, +0.051]$	-0.032 $[-0.290, +0.195]$	-0.015 $[-0.033, +0.003]$
9B M1′ x mentions, $\alpha = 0.5$ 94 (57)	-0.021 $[-0.112, +0.060]$	$+0.170$ $[+0.034, +0.326]$	-0.004 $[-0.016, +0.007]$
9B M1′ x mentions, $\alpha = 2$ 94 (57)	$+0.032$ $[-0.022, +0.090]$	$+0.085$ $[-0.097, +0.250]$	$+0.011$ $[-0.000, +0.023]$
9B M1′ x mentions, $\alpha = 4$ 94 (57)	$+0.064$ $[-0.047, +0.182]$	$+0.043$ $[-0.295, +0.366]$	$+0.006$ $[-0.010, +0.025]$
Phi-4 H_{out}, $\alpha = 0$ 50 (23)	-0.080 $[-0.176, +0.000]$	-0.100 $[-0.547, +0.311]$	$+0.014$ $[-0.001, +0.031]$
Phi-4 H_{out}, $\alpha = 0.5$ 50 (23)	$+0.100$ $[+0.000, +0.236]$	$+0.040$ $[-0.104, +0.218]$	$+0.005$ $[-0.002, +0.012]$
Phi-4 H_{out}, $\alpha = 2$ 50 (23)	-0.080 $[-0.170, +0.000]$	$+0.100$ $[-0.041, +0.255]$	$+0.002$ $[-0.006, +0.013]$
Phi-4 H_{out}, $\alpha = 4$ 50 (23)	-0.040 $[-0.186, +0.112]$	$+0.140$ $[-0.043, +0.319]$	-0.014 $[-0.037, +0.010]$
Phi-4 x mentions, $\alpha = 0$ 50 (29)	-0.080 $[-0.155, -0.019]$	-0.020 $[-0.125, +0.077]$	-0.006 $[-0.019, +0.005]$
Phi-4 x mentions, $\alpha = 0.5$ 50 (29)	-0.020 $[-0.080, +0.046]$	$+0.000$ $[-0.115, +0.163]$	-0.007 $[-0.030, +0.008]$
Phi-4 x mentions, $\alpha = 2$ 50 (29)	$+0.000$ $[-0.059, +0.055]$	-0.080 $[-0.218, +0.063]$	-0.003 $[-0.007, +0.003]$
Phi-4 x mentions, $\alpha = 4$ 50 (29)	$+0.000$ $[-0.049, +0.055]$	-0.160 $[-0.354, -0.020]$	-0.021 $[-0.045, -0.005]$

Continued on next page

Table 27 (continued)

Model / arm / dose / sample	Δx gold count	ΔH gold count	Δ entity F1
Gemma H_{out}, $\alpha = 0$ 50 (23)	$+0.000$ $[-0.095, +0.096]$	$+0.180$ $[-0.224, +0.609]$	$+0.037$ $[+0.007, +0.078]$
Gemma H_{out}, $\alpha = 0.5$ 50 (23)	$+0.020$ $[+0.000, +0.067]$	-0.040 $[-0.275, +0.211]$	$+0.004$ $[-0.003, +0.013]$
Gemma H_{out}, $\alpha = 2$ 50 (23)	-0.020 $[-0.064, +0.000]$	$+0.140$ $[+0.022, +0.315]$	$+0.007$ $[-0.004, +0.025]$
Gemma H_{out}, $\alpha = 4$ 50 (23)	-0.020 $[-0.169, +0.139]$	$+0.140$ $[-0.170, +0.479]$	$+0.027$ $[-0.003, +0.083]$
Gemma x mentions, $\alpha = 0$ 48 (29)	$+0.000$ $[-0.089, +0.078]$	-0.146 $[-0.375, +0.086]$	-0.003 $[-0.011, +0.006]$
Gemma x mentions, $\alpha = 0.5$ 48 (29)	-0.021 $[-0.070, +0.000]$	$+0.062$ $[+0.000, +0.174]$	$+0.001$ $[-0.002, +0.003]$
Gemma x mentions, $\alpha = 2$ 48 (29)	$+0.021$ $[-0.071, +0.114]$	$+0.208$ $[+0.040, +0.394]$	$+0.004$ $[-0.000, +0.011]$
Gemma x mentions, $\alpha = 4$ 48 (29)	$+0.021$ $[-0.065, +0.117]$	$+0.000$ $[-0.062, +0.062]$	$+0.001$ $[-0.008, +0.013]$

Table 28: Test generation: paired adjusted non-gold relation counts and parseability after closing incomplete JSON.

Model / arm / dose / sample	Δx non-gold	ΔH non-gold	Δ parse-ok
9B M0 H_{out}, $\alpha = 0$ 99 (44)	-0.051 $[-0.208, +0.101]$	-0.960 $[-2.833, +0.165]$	$+0.000$ $[+0.000, +0.000]$
9B M0 H_{out}, $\alpha = 0.5$ 99 (44)	-0.242 $[-0.667, +0.010]$	$+4.030$ $[-0.124, +10.063]$	$+0.000$ $[+0.000, +0.000]$
9B M0 H_{out}, $\alpha = 2$ 99 (44)	$+0.495$ $[-0.134, +1.671]$	$+1.111$ $[-0.176, +3.281]$	$+0.000$ $[+0.000, +0.000]$
9B M0 H_{out}, $\alpha = 4$ 99 (44)	$+0.131$ $[-0.580, +1.020]$	$+1.273$ $[-1.253, +5.288]$	$+0.000$ $[+0.000, +0.000]$
9B M0 x mentions, $\alpha = 0$ 94 (57)	$+0.032$ $[-0.077, +0.143]$	-0.511 $[-3.163, +1.361]$	$+0.000$ $[+0.000, +0.000]$
9B M0 x mentions, $\alpha = 0.5$ 94 (57)	-0.223 $[-0.677, +0.073]$	-0.660 $[-1.867, +0.280]$	$+0.000$ $[+0.000, +0.000]$
9B M0 x mentions, $\alpha = 2$ 94 (57)	$+1.883$ $[+0.059, +4.414]$	$+1.362$ $[-0.169, +3.467]$	$+0.000$ $[+0.000, +0.000]$
9B M0 x mentions, $\alpha = 4$ 94 (57)	$+0.489$ $[-1.262, +3.129]$	$+1.255$ $[-0.348, +3.677]$	$+0.011$ $[+0.000, +0.034]$

Continued on next page

Table 28 (continued)

Model / arm / dose / sample	Δx non-gold	ΔH non-gold	Δ parse-ok
9B M1′	$+0.192$	-0.071	$+0.000$
H_{out}, $\alpha = 0$	$[+0.021, +0.387]$	$[-0.421, +0.278]$	$[-0.025, +0.030]$
99 (44)			
9B M1′	$+0.030$	$+0.030$	$+0.000$
H_{out}, $\alpha = 0.5$	$[-0.102, +0.196]$	$[-0.402, +0.490]$	$[+0.000, +0.000]$
99 (44)			
9B M1′	-0.131	$+0.293$	$+0.000$
H_{out}, $\alpha = 2$	$[-0.402, +0.032]$	$[-0.333, +0.957]$	$[+0.000, +0.000]$
99 (44)			
9B M1′	-0.182	-0.010	$+0.000$
H_{out}, $\alpha = 4$	$[-0.371, -0.050]$	$[-0.600, +0.513]$	$[+0.000, +0.000]$
99 (44)			
9B M1′	-0.234	-0.298	-0.011
x mentions, $\alpha = 0$	$[-0.443, -0.031]$	$[-0.981, +0.236]$	$[-0.032, +0.000]$
94 (57)			
9B M1′	-0.074	-0.426	$+0.000$
x mentions, $\alpha = 0.5$	$[-0.266, +0.082]$	$[-0.955, -0.033]$	$[+0.000, +0.000]$
94 (57)			
9B M1′	$+0.181$	$+0.245$	$+0.000$
x mentions, $\alpha = 2$	$[-0.245, +0.489]$	$[-0.139, +0.613]$	$[+0.000, +0.000]$
94 (57)			
9B M1′	$+0.213$	$+0.053$	$+0.000$
x mentions, $\alpha = 4$	$[-0.245, +0.671]$	$[-0.579, +0.800]$	$[+0.000, +0.000]$
94 (57)			
Phi-4	$+0.080$	$+0.080$	$+0.000$
H_{out}, $\alpha = 0$	$[-0.036, +0.211]$	$[-0.175, +0.364]$	$[+0.000, +0.000]$
50 (23)			
Phi-4	$+0.040$	$+0.160$	$+0.000$
H_{out}, $\alpha = 0.5$	$[-0.047, +0.141]$	$[-0.041, +0.416]$	$[+0.000, +0.000]$
50 (23)			
Phi-4	-0.100	$+0.260$	$+0.000$
H_{out}, $\alpha = 2$	$[-0.475, +0.204]$	$[-0.137, +0.723]$	$[+0.000, +0.000]$
50 (23)			
Phi-4	-0.060	$+0.160$	$+0.000$
H_{out}, $\alpha = 4$	$[-0.431, +0.261]$	$[-0.204, +0.563]$	$[+0.000, +0.000]$
50 (23)			
Phi-4	-0.020	$+0.040$	$+0.000$
x mentions, $\alpha = 0$	$[-0.125, +0.055]$	$[-0.061, +0.163]$	$[+0.000, +0.000]$
50 (29)			
Phi-4	-0.040	-0.040	$+0.000$
x mentions, $\alpha = 0.5$	$[-0.130, +0.000]$	$[-0.153, +0.101]$	$[+0.000, +0.000]$
50 (29)			
Phi-4	-0.040	-0.060	$+0.000$
x mentions, $\alpha = 2$	$[-0.099, +0.000]$	$[-0.180, +0.083]$	$[+0.000, +0.000]$
50 (29)			
Phi-4	$+0.060$	-0.020	$+0.000$
x mentions, $\alpha = 4$	$[+0.000, +0.135]$	$[-0.146, +0.093]$	$[+0.000, +0.000]$
50 (29)			
Gemma	$+0.060$	-0.120	$+0.000$
H_{out}, $\alpha = 0$	$[-0.086, +0.200]$	$[-0.660, +0.354]$	$[+0.000, +0.000]$
50 (23)			
Gemma	$+0.040$	$+0.040$	$+0.000$
H_{out}, $\alpha = 0.5$	$[+0.000, +0.135]$	$[-0.122, +0.238]$	$[+0.000, +0.000]$
50 (23)			
Gemma	$+0.000$	-0.240	$+0.000$
H_{out}, $\alpha = 2$	$[-0.079, +0.078]$	$[-0.520, -0.023]$	$[+0.000, +0.000]$
50 (23)			

Continued on next page

Table 28 (continued)

Model / arm / dose / sample	Δx non-gold	ΔH non-gold	Δ parse-ok
Gemma H_{out}, $\alpha = 4$ 50 (23)	$+0.000$ $[-0.137, +0.106]$	-0.160 $[-0.832, +0.457]$	$+0.020$ $[+0.000, +0.065]$
Gemma x mentions, $\alpha = 0$ 48 (29)	-0.104 $[-0.224, +0.024]$	-0.042 $[-0.394, +0.278]$	$+0.000$ $[+0.000, +0.000]$
Gemma x mentions, $\alpha = 0.5$ 48 (29)	$+0.021$ $[+0.000, +0.069]$	-0.062 $[-0.183, +0.000]$	$+0.000$ $[+0.000, +0.000]$
Gemma x mentions, $\alpha = 2$ 48 (29)	$+0.000$ $[-0.112, +0.111]$	-0.208 $[-0.480, -0.019]$	$+0.000$ $[+0.000, +0.000]$
Gemma x mentions, $\alpha = 4$ 48 (29)	-0.104 $[-0.202, -0.021]$	-0.104 $[-0.370, +0.114]$	$+0.000$ $[+0.000, +0.000]$

Parsing diagnostics Strict JSON success and success after closure repair are distinct outcomes. We retain both raw rates for the target arms below; the adjusted parseability panel above uses the latter. A valid or repairable output is not necessarily a correct extraction.

Table 29: Target-arm generation parsing rates: same-run baseline → intervention. Strict and repaired parsing are kept separate.

Model	Arm	α	Records (docs)	Strict JSON	After closure repair
9B M0	H_{out}	0	100	0.840→0.760	1.000→1.000
9B M0	H_{out}	0.5	100	0.840→0.740	1.000→1.000
9B M0	H_{out}	2	100	0.840→0.710	1.000→1.000
9B M0	H_{out}	4	100	0.840→0.690	1.000→1.000
9B M0	x mentions	0	100	0.840→0.830	1.000→1.000
9B M0	x mentions	0.5	100	0.840→0.770	1.000→1.000
9B M0	x mentions	2	100	0.840→0.780	1.000→1.000
9B M0	x mentions	4	100	0.840→0.760	1.000→1.000
9B M1′	H_{out}	0	100	0.980→0.960	1.000→0.990
9B M1′	H_{out}	0.5	100	0.980→1.000	1.000→1.000
9B M1′	H_{out}	2	100	0.980→1.000	1.000→1.000
9B M1′	H_{out}	4	100	0.980→0.990	1.000→1.000
9B M1′	x mentions	0	100	0.980→0.990	1.000→0.990
9B M1′	x mentions	0.5	100	0.980→1.000	1.000→1.000
9B M1′	x mentions	2	100	0.980→1.000	1.000→1.000
9B M1′	x mentions	4	100	0.980→1.000	1.000→1.000
Phi-4	H_{out}	0	50	1.000→1.000	1.000→1.000
Phi-4	H_{out}	0.5	50	1.000→1.000	1.000→1.000
Phi-4	H_{out}	2	50	1.000→1.000	1.000→1.000
Phi-4	H_{out}	4	50	1.000→1.000	1.000→1.000
Phi-4	x mentions	0	50	1.000→1.000	1.000→1.000
Phi-4	x mentions	0.5	50	1.000→1.000	1.000→1.000
Phi-4	x mentions	2	50	1.000→1.000	1.000→1.000
Phi-4	x mentions	4	50	1.000→1.000	1.000→1.000
Gemma	H_{out}	0	50	1.000→1.000	1.000→1.000
Gemma	H_{out}	0.5	50	1.000→1.000	1.000→1.000
Gemma	H_{out}	2	50	1.000→1.000	1.000→1.000
Gemma	H_{out}	4	50	1.000→1.000	1.000→1.000
Gemma	x mentions	0	50	1.000→1.000	1.000→1.000
Gemma	x mentions	0.5	50	1.000→1.000	1.000→1.000

Continued on next page

Table 29 (continued)

Model	Arm	α	Records (docs)	Strict JSON	After closure repair
Gemma	x mentions	2	50	1.000→1.000	1.000→1.000
Gemma	x mentions	4	50	1.000→1.000	1.000→1.000

D.7 Training variants and adapter ablations

The four evidence-format training variants have protagonist-target shares 0.456 (M1, natural), 0.459 (M2, random-drop), 0.274 (M3, low-hub), and 0.607 (M4, high-hub). Their fixed-prefix contrasts use paired controls; these must not be confused with the unadjusted original-minus-counterfactual margins in the adapter-ablation panel. The validation knockout scan evaluates layers 23, 27, and 31. The test steering slopes for the x arm have overlapping intervals; overlap alone is not an equivalence test.

For the LoRA experiment we zero the selected trained `lora_B` increments while keeping the base weights fixed. Removing attention increments is not removal of the base attention module. The tested `linear_attn` selection has no applicable trained LoRA tensors and is a no-op control. Configurations with a relation-F1 loss greater than 0.05 or parseability below 0.95 are marked degraded; the M3 baseline itself has parseability 0.947. Broad MLP or late-layer ablations that degrade extraction cannot be used as selective mechanistic evidence.

Table 30: Evidence-format adapters: fixed-prefix, paired-control-adjusted head contrasts, validation and test pooled.

Adapter	Targets (hub share)	Demotion	Promotion
M1	Natural (0.456)	-3.431 $[-4.501, -2.480]$ $n = 384/79$ $p_H = 1.8 \times 10^{-13}$	-2.376 $[-3.858, -0.974]$ $n = 98/32$ $p_H = 5.35 \times 10^{-5}$
M2	Random-drop (0.459)	-2.821 $[-3.739, -2.017]$ $n = 384/79$ $p_H = 1.52 \times 10^{-12}$	-1.841 $[-3.208, -0.730]$ $n = 98/32$ $p_H = 0.000917$
M3	Low-hub (0.274)	-2.150 $[-2.893, -1.413]$ $n = 384/79$ $p_H = 3.32 \times 10^{-10}$	-1.350 $[-2.538, -0.405]$ $n = 98/32$ $p_H = 0.00342$
M4	High-hub (0.607)	-3.539 $[-4.651, -2.511]$ $n = 384/79$ $p_H = 1.02 \times 10^{-14}$	-2.415 $[-4.072, -1.010]$ $n = 98/32$ $p_H = 0.000106$

Table 31: Validation layer selection: H_{cues} knockout (title and out-of-sentence protagonist cues) for all four adapters.

Adapter	Layer	Original loss	CF recovery	Matched Δm
M1	23	-0.001 $[-0.016, +0.012]$ $n = 100/56$	-0.001 $[-0.013, +0.012]$ $n = 99/56$	-0.018 $[-0.081, +0.041]$ $n = 30/22$
M1	27	$+0.064$ $[+0.019, +0.115]$ $n = 100/56$	$+0.027$ $[-0.014, +0.068]$ $n = 99/56$	-0.177 $[-0.277, -0.075]$ $n = 30/22$
M1	31	$+0.006$ $[-0.014, +0.024]$ $n = 100/56$	$+0.019$ $[+0.001, +0.042]$ $n = 99/56$	-0.066 $[-0.092, -0.038]$ $n = 30/22$

Continued on next page

Table 31 (continued)

Adapter	Layer	Original loss	CF recovery	Matched Δm
M2	23	$+0.010$ $[-0.000, +0.021]$ $n = 96/52$	$+0.004$ $[-0.008, +0.016]$ $n = 95/52$	-0.048 $[-0.106, +0.003]$ $n = 30/22$
M2	27	$+0.090$ $[+0.035, +0.149]$ $n = 96/52$	$+0.016$ $[-0.035, +0.058]$ $n = 95/52$	-0.189 $[-0.293, -0.087]$ $n = 30/22$
M2	31	$+0.024$ $[+0.008, +0.041]$ $n = 96/52$	$+0.016$ $[+0.002, +0.031]$ $n = 95/52$	-0.068 $[-0.120, -0.024]$ $n = 30/22$
M3	23	$+0.013$ $[-0.006, +0.031]$ $n = 100/54$	-0.007 $[-0.024, +0.010]$ $n = 99/54$	-0.075 $[-0.132, -0.012]$ $n = 30/22$
M3	27	$+0.100$ $[+0.055, +0.156]$ $n = 100/54$	$+0.032$ $[-0.018, +0.085]$ $n = 99/54$	-0.164 $[-0.254, -0.083]$ $n = 30/22$
M3	31	$+0.015$ $[-0.009, +0.037]$ $n = 100/54$	$+0.013$ $[-0.010, +0.035]$ $n = 99/54$	-0.073 $[-0.098, -0.045]$ $n = 30/22$
M4	23	-0.005 $[-0.028, +0.011]$ $n = 100/56$	$+0.004$ $[-0.011, +0.022]$ $n = 99/56$	-0.055 $[-0.116, +0.012]$ $n = 30/22$
M4	27	$+0.081$ $[+0.026, +0.140]$ $n = 100/56$	$+0.025$ $[-0.025, +0.075]$ $n = 99/56$	-0.235 $[-0.369, -0.132]$ $n = 30/22$
M4	31	$+0.018$ $[-0.002, +0.037]$ $n = 100/56$	$+0.024$ $[+0.001, +0.055]$ $n = 99/56$	-0.081 $[-0.113, -0.046]$ $n = 30/22$

Table 32: Adapter test steering: suppression endpoint, enhancement endpoints, and fitted slope (nats per doubling), retaining all target and control arms. The joint arm has no $\alpha = 0$ condition.

Adapter / arm / control	$\Delta m, \alpha = 0$	$\Delta m, \alpha = 4$	$\Delta m, \alpha = 8$	Slope
M1 / L27 H_{out}, target $n = 177/63$	-0.201 $[-0.358, -0.103]$	$+0.302$ $[+0.197, +0.410]$	$+0.552$ $[+0.386, +0.727]$	$+0.156$ $[+0.102, +0.210]$
M1 / L27 H_{out}, bystander $n = 69/22$	$+0.024$ $[-0.049, +0.103]$	-0.050 $[-0.218, +0.107]$	-0.062 $[-0.329, +0.222]$	-0.015 $[-0.093, +0.064]$
M1 / L27 H_{out}, random $n = 177/63$	-0.010 $[-0.030, +0.011]$	$+0.008$ $[-0.033, +0.045]$	$+0.040$ $[-0.026, +0.106]$	$+0.011$ $[-0.007, +0.029]$
M1 / L27 x mentions, target $n = 178/63$	$+0.136$ $[+0.097, +0.174]$	-0.355 $[-0.453, -0.270]$	-0.739 $[-0.929, -0.564]$	-0.197 $[-0.245, -0.150]$
M1 / L27 x mentions, bystander $n = 69/22$	$+0.024$ $[-0.050, +0.105]$	-0.050 $[-0.211, +0.099]$	-0.062 $[-0.326, +0.204]$	-0.015 $[-0.098, +0.062]$
M1 / L27 x mentions, random $n = 178/63$	$+0.001$ $[-0.017, +0.017]$	$+0.007$ $[-0.023, +0.041]$	$+0.013$ $[-0.037, +0.076]$	$+0.004$ $[-0.008, +0.018]$
M1 / L27 $x \uparrow, H \downarrow$, target $n = 177/63$	–	-0.488 $[-0.603, -0.371]$	-0.905 $[-1.111, -0.707]$	-0.266 $[-0.324, -0.208]$
M1 / L27 $x \uparrow, H \downarrow$, bystander $n = 69/22$	–	-0.050 $[-0.208, +0.112]$	-0.062 $[-0.345, +0.247]$	-0.015 $[-0.089, +0.064]$

Continued on next page

Table 32 (continued)

Adapter / arm / control	$\Delta m, \alpha = 0$	$\Delta m, \alpha = 4$	$\Delta m, \alpha = 8$	Slope
M1 / L27 $x \uparrow, H \downarrow$, random $n = 177/63$	–	-0.011 $[-0.033, +0.007]$	-0.001 $[-0.030, +0.025]$	-0.006 $[-0.014, +0.002]$
M2 / L27 H_{out}, target $n = 177/63$	-0.206 $[-0.361, -0.102]$	$+0.287$ $[+0.198, +0.386]$	$+0.528$ $[+0.377, +0.688]$	$+0.151$ $[+0.106, +0.196]$
M2 / L27 H_{out}, bystander $n = 69/22$	-0.016 $[-0.064, +0.030]$	$+0.012$ $[-0.059, +0.099]$	$+0.051$ $[-0.098, +0.243]$	$+0.016$ $[-0.027, +0.068]$
M2 / L27 H_{out}, random $n = 177/63$	-0.011 $[-0.032, +0.008]$	$+0.024$ $[-0.003, +0.054]$	$+0.071$ $[+0.014, +0.126]$	$+0.018$ $[+0.002, +0.033]$
M2 / L27 x mentions, target $n = 178/63$	$+0.144$ $[+0.098, +0.204]$	-0.368 $[-0.488, -0.255]$	-0.734 $[-0.944, -0.539]$	-0.197 $[-0.260, -0.142]$
M2 / L27 x mentions, bystander $n = 69/22$	-0.016 $[-0.067, +0.033]$	$+0.012$ $[-0.063, +0.102]$	$+0.051$ $[-0.089, +0.241]$	$+0.016$ $[-0.028, +0.068]$
M2 / L27 x mentions, random $n = 178/63$	-0.014 $[-0.030, +0.002]$	$+0.018$ $[-0.002, +0.037]$	$+0.044$ $[+0.013, +0.078]$	$+0.011$ $[+0.003, +0.022]$
M2 / L27 $x \uparrow, H \downarrow$, target $n = 177/63$	–	-0.496 $[-0.632, -0.377]$	-0.907 $[-1.156, -0.692]$	-0.269 $[-0.334, -0.205]$
M2 / L27 $x \uparrow, H \downarrow$, bystander $n = 69/22$	–	$+0.012$ $[-0.058, +0.101]$	$+0.051$ $[-0.089, +0.260]$	$+0.016$ $[-0.025, +0.066]$
M2 / L27 $x \uparrow, H \downarrow$, random $n = 177/63$	–	-0.012 $[-0.026, +0.002]$	-0.003 $[-0.023, +0.017]$	-0.005 $[-0.012, +0.002]$
M3 / L27 H_{out}, target $n = 177/63$	-0.166 $[-0.323, -0.070]$	$+0.280$ $[+0.199, +0.375]$	$+0.489$ $[+0.348, +0.633]$	$+0.137$ $[+0.096, +0.178]$
M3 / L27 H_{out}, bystander $n = 69/22$	-0.021 $[-0.070, +0.017]$	$+0.047$ $[-0.023, +0.120]$	$+0.086$ $[-0.045, +0.235]$	$+0.022$ $[-0.014, +0.059]$
M3 / L27 H_{out}, random $n = 177/63$	-0.003 $[-0.018, +0.010]$	$+0.044$ $[+0.013, +0.074]$	$+0.092$ $[+0.036, +0.154]$	$+0.022$ $[+0.007, +0.038]$
M3 / L27 x mentions, target $n = 178/63$	$+0.140$ $[+0.095, +0.189]$	-0.312 $[-0.430, -0.216]$	-0.628 $[-0.830, -0.453]$	-0.175 $[-0.233, -0.125]$
M3 / L27 x mentions, bystander $n = 69/22$	-0.021 $[-0.067, +0.016]$	$+0.047$ $[-0.026, +0.122]$	$+0.086$ $[-0.045, +0.220]$	$+0.022$ $[-0.012, +0.061]$
M3 / L27 x mentions, random $n = 178/63$	$+0.015$ $[+0.003, +0.027]$	$+0.024$ $[+0.000, +0.048]$	$+0.045$ $[+0.008, +0.084]$	$+0.010$ $[+0.001, +0.019]$
M3 / L27 $x \uparrow, H \downarrow$, target $n = 177/63$	–	-0.426 $[-0.566, -0.308]$	-0.780 $[-1.003, -0.571]$	-0.235 $[-0.304, -0.182]$
M3 / L27 $x \uparrow, H \downarrow$, bystander $n = 69/22$	–	$+0.047$ $[-0.026, +0.126]$	$+0.086$ $[-0.043, +0.231]$	$+0.022$ $[-0.012, +0.062]$
M3 / L27 $x \uparrow, H \downarrow$, random $n = 177/63$	–	$+0.012$ $[-0.006, +0.028]$	$+0.018$ $[-0.007, +0.041]$	-0.002 $[-0.009, +0.004]$

Continued on next page

Table 32 (continued)

Adapter / arm / control	$\Delta m, \alpha = 0$	$\Delta m, \alpha = 4$	$\Delta m, \alpha = 8$	Slope
M4 / L27 H_{out}, target $n = 177/63$	-0.226 $[-0.398, -0.114]$	$+0.332$ $[+0.225, +0.445]$	$+0.602$ $[+0.431, +0.765]$	$+0.172$ $[+0.118, +0.222]$
M4 / L27 H_{out}, bystander $n = 69/22$	-0.023 $[-0.092, +0.043]$	$+0.009$ $[-0.108, +0.146]$	-0.005 $[-0.208, +0.276]$	$+0.004$ $[-0.054, +0.072]$
M4 / L27 H_{out}, random $n = 177/63$	-0.011 $[-0.031, +0.009]$	$+0.007$ $[-0.032, +0.047]$	$+0.034$ $[-0.039, +0.115]$	$+0.008$ $[-0.010, +0.027]$
M4 / L27 x mentions, target $n = 178/63$	$+0.142$ $[+0.100, +0.184]$	-0.387 $[-0.483, -0.292]$	-0.786 $[-0.986, -0.601]$	-0.210 $[-0.258, -0.163]$
M4 / L27 x mentions, bystander $n = 69/22$	-0.023 $[-0.087, +0.048]$	$+0.009$ $[-0.100, +0.148]$	-0.005 $[-0.208, +0.250]$	$+0.004$ $[-0.054, +0.070]$
M4 / L27 x mentions, random $n = 178/63$	-0.002 $[-0.022, +0.019]$	-0.008 $[-0.031, +0.016]$	$+0.004$ $[-0.027, +0.046]$	$+0.001$ $[-0.009, +0.012]$
M4 / L27 $x \uparrow, H \downarrow$, target $n = 177/63$	– 	-0.533 $[-0.661, -0.416]$	-0.986 $[-1.191, -0.786]$	-0.290 $[-0.351, -0.233]$
M4 / L27 $x \uparrow, H \downarrow$, bystander $n = 69/22$	– 	$+0.009$ $[-0.098, +0.149]$	-0.005 $[-0.228, +0.237]$	$+0.004$ $[-0.050, +0.072]$
M4 / L27 $x \uparrow, H \downarrow$, random $n = 177/63$	– 	-0.024 $[-0.046, -0.005]$	-0.027 $[-0.055, -0.001]$	-0.013 $[-0.024, -0.003]$

Table 33: Validation LoRA-increment ablations. The natural swap margin is unadjusted; matched controls are separate raw changes in nats. F1 is relation F1; parse is the archived capability-panel parse rate.

Adapter / removed increments	F1 / parse	Natural swap margin	Matched Δm	Status
M1 Baseline	0.419 / 1.000	$+5.102$ $[+3.964, +6.294]$ $n = 189/68$	$+0.722$ $[+0.082, +1.610]$	Baseline
M1 Attention L27	0.447 / 0.992	$+5.131$ $[+3.962, +6.246]$ $n = 189/68$	$+0.745$ $[+0.092, +1.549]$	Retained
M1 All attention	0.456 / 0.992	$+5.171$ $[+4.001, +6.321]$ $n = 189/68$	$+0.699$ $[+0.046, +1.427]$	Retained
M1 Linear attention	0.419 / 1.000	$+5.102$ $[+3.971, +6.274]$ $n = 189/68$	$+0.722$ $[+0.076, +1.573]$	No-op
M1 All MLP	0.225 / 0.940	$+4.703$ $[+3.569, +6.003]$ $n = 189/68$	$+0.014$ $[-0.492, +0.451]$	Degraded
M1 Layers 22–31	0.321 / 0.992	$+5.223$ $[+4.056, +6.409]$ $n = 189/68$	$+0.721$ $[+0.049, +1.517]$	Degraded
M3 Baseline	0.338 / 0.947	$+3.072$ $[+2.152, +3.900]$ $n = 189/68$	$+0.265$ $[+0.003, +0.541]$	Baseline

Continued on next page

Table 33 (continued)

Adapter / removed increments	F1 / parse	Natural swap margin	Matched Δm	Status
M3 Attention L27	0.353 / 0.985	$+3.124$ $[+2.263, +4.067]$ $n = 189/68$	$+0.252$ $[-0.035, +0.563]$	Retained
M3 All attention	0.424 / 0.970	$+3.686$ $[+2.791, +4.600]$ $n = 189/68$	$+0.380$ $[+0.004, +0.750]$	Retained
M3 Linear attention	0.338 / 0.947	$+3.072$ $[+2.303, +3.997]$ $n = 189/68$	$+0.265$ $[+0.006, +0.551]$	No-op
M3 All MLP	0.236 / 0.947	$+3.857$ $[+2.816, +4.956]$ $n = 189/68$	-0.001 $[-0.501, +0.531]$	Degraded
M3 Layers 22–31	0.286 / 0.985	$+3.305$ $[+2.427, +4.188]$ $n = 189/68$	$+0.332$ $[+0.018, +0.642]$	Degraded
M4 Baseline	0.369 / 0.985	$+5.033$ $[+3.860, +6.186]$ $n = 189/68$	$+0.546$ $[+0.054, +1.105]$	Baseline
M4 Attention L27	0.386 / 1.000	$+5.088$ $[+3.957, +6.270]$ $n = 189/68$	$+0.565$ $[+0.089, +1.082]$	Retained
M4 All attention	0.359 / 1.000	$+4.989$ $[+3.841, +6.086]$ $n = 189/68$	$+0.590$ $[+0.017, +1.218]$	Retained
M4 Linear attention	0.369 / 0.985	$+5.033$ $[+3.947, +6.187]$ $n = 189/68$	$+0.546$ $[+0.039, +1.102]$	No-op
M4 All MLP	0.247 / 0.940	$+4.789$ $[+3.691, +6.023]$ $n = 189/68$	$+0.092$ $[-0.424, +0.591]$	Degraded
M4 Layers 22–31	0.350 / 0.977	$+5.097$ $[+3.969, +6.112]$ $n = 189/68$	$+0.595$ $[+0.097, +1.210]$	Retained

D.8 Cross-family localization and teacher-forced contrasts

Phi-4 and Gemma were evaluated on the same intervention questions with model-specific token positions and validation-selected depths. Their available residual scan is a validation denoising experiment, not a replication of the 9B test noising panel. Table 34 pairs the archived input-site curve with the full decision-slot curve reconstructed from raw records, retaining treated pairs with $|e| \geq 0.5$. Duplicate patch keys retain a completed measurement in preference to an uncomputed placeholder. Decision-slot intervals resample document clusters and preserve record weighting.

Input-site recovery is small in magnitude, with some intervals excluding zero; we do not claim every input layer is exactly null. Decision-slot recovery increases later, with Phi-4 reaching 0.94 at layer 39 and Gemma reaching 0.95 at layer 47. The depth and shape differ across families. Knockout also differs: document-source effects at the tested Phi-4 and Gemma layers are generally smaller than the highlighted 9B effects, and negative losses or recoveries are retained rather than discarded. The score and generation tests are in Tables 22 and 25–29; their divergent generation responses limit claims of a universal intervention.

The teacher-forced panel at the end measures the scored head's log probability, not the $H - x$ margin, and conditional relation-and-tail completion after supplying the head. It pools validation and test. Phi-4's demotion conditional-completion interval excludes zero, but its archived Holm-adjusted $p = 0.0516$ does not cross 0.05. These measurements should not be collapsed into one claim that conditional completion is either universally affected or universally unchanged.

Table 34: Cross-family validation denoising: every sampled input-site and decision-slot residual layer. Input-site and slot intervals use 500 and 2,000 bootstrap draws, respectively.

Model	Layer	S_H/S_x recovery	S_{slot} recovery
Phi-4	1	$+0.000$	$+0.002$
		$[-0.003, +0.003]$	$[-0.002, +0.007]$
		$n = 122/59$	$n = 122/59$
Phi-4	3	$+0.004$	$+0.003$
		$[-0.003, +0.013]$	$[-0.001, +0.009]$
		$n = 122/59$	$n = 122/59$
Phi-4	5	$+0.004$	$+0.002$
		$[-0.002, +0.010]$	$[-0.003, +0.007]$
		$n = 122/59$	$n = 122/59$
Phi-4	7	$+0.006$	-0.000
		$[-0.001, +0.013]$	$[-0.003, +0.003]$
		$n = 122/59$	$n = 122/59$
Phi-4	9	$+0.005$	$+0.003$
		$[-0.001, +0.011]$	$[-0.002, +0.008]$
		$n = 122/59$	$n = 122/59$
Phi-4	11	$+0.007$	$+0.003$
		$[+0.000, +0.016]$	$[-0.003, +0.009]$
		$n = 122/59$	$n = 122/59$
Phi-4	13	$+0.004$	$+0.003$
		$[-0.000, +0.009]$	$[-0.003, +0.009]$
		$n = 122/59$	$n = 122/59$
Phi-4	15	$+0.003$	$+0.008$
		$[-0.002, +0.008]$	$[+0.003, +0.013]$
		$n = 122/59$	$n = 122/59$
Phi-4	17	$+0.004$	$+0.013$
		$[-0.001, +0.012]$	$[-0.001, +0.025]$
		$n = 122/59$	$n = 122/59$
Phi-4	19	$+0.002$	$+0.018$
		$[-0.002, +0.007]$	$[-0.001, +0.035]$
		$n = 122/59$	$n = 122/59$
Phi-4	21	$+0.000$	$+0.045$
		$[-0.004, +0.005]$	$[+0.004, +0.080]$
		$n = 122/59$	$n = 122/59$
Phi-4	23	-0.001	$+0.084$
		$[-0.005, +0.004]$	$[+0.035, +0.129]$
		$n = 122/59$	$n = 122/59$
Phi-4	25	$+0.000$	$+0.217$
		$[-0.005, +0.005]$	$[+0.151, +0.273]$
		$n = 122/59$	$n = 122/59$
Phi-4	27	$+0.001$	$+0.396$
		$[-0.003, +0.005]$	$[+0.333, +0.457]$
		$n = 122/59$	$n = 122/59$
Phi-4	29	-0.000	$+0.473$
		$[-0.002, +0.002]$	$[+0.410, +0.534]$
		$n = 122/59$	$n = 122/59$
Phi-4	31	$+0.001$	$+0.549$
		$[-0.001, +0.005]$	$[+0.485, +0.611]$
		$n = 122/59$	$n = 122/59$
Phi-4	33	$+0.001$	$+0.635$
		$[-0.001, +0.002]$	$[+0.576, +0.690]$
		$n = 122/59$	$n = 122/59$
Phi-4	35	-0.001	$+0.746$
		$[-0.002, +0.000]$	$[+0.694, +0.796]$
		$n = 122/59$	$n = 122/59$
Phi-4	37	-0.000	$+0.895$
		$[-0.002, +0.001]$	$[+0.841, +0.947]$
		$n = 122/59$	$n = 122/59$

Continued on next page

Table 34 (continued)

Model	Layer	S_H/S_x recovery	S_{slot} recovery
Phi-4	39	$+0.000$ $[+0.000, +0.000]$ $n = 122/59$	$+0.941$ $[+0.887, +0.989]$ $n = 122/59$
Gemma	1	$+0.007$ $[+0.001, +0.015]$ $n = 128/60$	-0.003 $[-0.010, +0.002]$ $n = 128/60$
Gemma	3	$+0.001$ $[-0.006, +0.007]$ $n = 128/60$	-0.002 $[-0.006, +0.003]$ $n = 128/60$
Gemma	5	$+0.000$ $[-0.011, +0.010]$ $n = 128/60$	-0.000 $[-0.005, +0.004]$ $n = 128/60$
Gemma	7	$+0.011$ $[+0.001, +0.022]$ $n = 128/60$	$+0.003$ $[-0.003, +0.011]$ $n = 128/60$
Gemma	9	$+0.015$ $[+0.004, +0.028]$ $n = 128/60$	$+0.001$ $[-0.005, +0.006]$ $n = 128/60$
Gemma	11	$+0.012$ $[+0.002, +0.023]$ $n = 128/60$	$+0.000$ $[-0.005, +0.005]$ $n = 128/60$
Gemma	13	$+0.014$ $[+0.003, +0.027]$ $n = 128/60$	-0.003 $[-0.010, +0.003]$ $n = 128/60$
Gemma	15	$+0.011$ $[+0.000, +0.021]$ $n = 128/60$	$+0.001$ $[-0.007, +0.009]$ $n = 128/60$
Gemma	17	$+0.005$ $[-0.007, +0.015]$ $n = 128/60$	-0.002 $[-0.009, +0.004]$ $n = 128/60$
Gemma	19	-0.003 $[-0.019, +0.010]$ $n = 128/60$	-0.014 $[-0.028, -0.003]$ $n = 128/60$
Gemma	21	-0.002 $[-0.021, +0.013]$ $n = 128/60$	-0.011 $[-0.026, +0.000]$ $n = 128/60$
Gemma	23	-0.006 $[-0.023, +0.005]$ $n = 128/60$	-0.000 $[-0.022, +0.019]$ $n = 128/60$
Gemma	25	-0.005 $[-0.024, +0.009]$ $n = 128/60$	-0.002 $[-0.039, +0.034]$ $n = 128/60$
Gemma	27	-0.005 $[-0.023, +0.008]$ $n = 128/60$	$+0.043$ $[-0.030, +0.127]$ $n = 128/60$
Gemma	29	-0.001 $[-0.005, +0.002]$ $n = 128/60$	$+0.179$ $[+0.058, +0.305]$ $n = 128/60$
Gemma	31	-0.002 $[-0.006, +0.002]$ $n = 128/60$	$+0.192$ $[+0.078, +0.310]$ $n = 128/60$
Gemma	33	-0.001 $[-0.009, +0.004]$ $n = 128/60$	$+0.188$ $[+0.072, +0.310]$ $n = 128/60$
Gemma	35	-0.001 $[-0.005, +0.004]$ $n = 128/60$	$+0.243$ $[+0.119, +0.361]$ $n = 128/60$
Gemma	37	-0.004 $[-0.011, +0.001]$ $n = 128/60$	$+0.285$ $[+0.162, +0.399]$ $n = 128/60$

Continued on next page

Table 34 (continued)

Model	Layer	S_H/S_x recovery	S_{slot} recovery
Gemma	39	-0.003 $[-0.010, +0.002]$ $n = 128/60$	$+0.293$ $[+0.170, +0.406]$ $n = 128/60$
Gemma	41	-0.004 $[-0.010, +0.000]$ $n = 128/60$	$+0.436$ $[+0.341, +0.530]$ $n = 128/60$
Gemma	43	$+0.000$ $[-0.001, +0.002]$ $n = 128/60$	$+0.479$ $[+0.389, +0.565]$ $n = 128/60$
Gemma	45	-0.003 $[-0.010, +0.000]$ $n = 128/60$	$+0.618$ $[+0.547, +0.684]$ $n = 128/60$
Gemma	47	$+0.000$ $[+0.000, +0.000]$ $n = 128/60$	$+0.952$ $[+0.906, +0.986]$ $n = 128/60$

Table 35: Cross-family validation knockouts: three layers and ten source classes per model, including candidate-list controls.

Model / layer	Source	Original loss	CF recovery	Matched Δm
Gemma / L29	title	$+0.016$ $[-0.004, +0.036]$ $n = 123/58$	-0.004 $[-0.029, +0.019]$ $n = 123/58$	-0.033 $[-0.260, +0.169]$ $n = 36/27$
Gemma / L29	H_out	-0.018 $[-0.048, +0.002]$ $n = 128/60$	$+0.020$ $[-0.009, +0.071]$ $n = 114/56$	$+0.022$ $[-0.122, +0.153]$ $n = 40/29$
Gemma / L29	x_out	$+0.015$ $[-0.001, +0.032]$ $n = 114/56$	-0.018 $[-0.033, -0.005]$ $n = 128/60$	-0.013 $[-0.157, +0.104]$ $n = 38/28$
Gemma / L29	s	-0.005 $[-0.044, +0.025]$ $n = 128/60$	-0.022 $[-0.040, -0.007]$ $n = 128/60$	$+0.010$ $[-0.145, +0.211]$ $n = 40/29$
Gemma / L29	other	$+0.010$ $[-0.027, +0.056]$ $n = 128/60$	$+0.010$ $[-0.018, +0.048]$ $n = 128/60$	$+0.057$ $[-0.281, +0.467]$ $n = 40/29$
Gemma / L29	H_cues	-0.003 $[-0.033, +0.023]$ $n = 128/60$	$+0.017$ $[-0.023, +0.064]$ $n = 127/60$	$+0.032$ $[-0.187, +0.211]$ $n = 40/29$
Gemma / L29	all_cues	$+0.009$ $[-0.027, +0.039]$ $n = 128/60$	$+0.003$ $[-0.035, +0.054]$ $n = 128/60$	-0.064 $[-0.310, +0.176]$ $n = 40/29$
Gemma / L29	pre_ents	-0.067 $[-0.132, -0.013]$ $n = 128/60$	-0.050 $[-0.136, +0.003]$ $n = 128/60$	$+0.762$ $[+0.096, +1.561]$ $n = 40/29$
Gemma / L29	pre_H	$+0.024$ $[+0.005, +0.046]$ $n = 124/57$	-0.036 $[-0.066, -0.012]$ $n = 124/57$	-0.615 $[-0.975, -0.243]$ $n = 40/29$
Gemma / L29	pre_x	$+0.025$ $[-0.004, +0.061]$ $n = 128/60$	-0.014 $[-0.063, +0.016]$ $n = 128/60$	$+0.010$ $[-0.167, +0.231]$ $n = 40/29$
Gemma / L35	title	$+0.002$ $[-0.006, +0.009]$ $n = 123/58$	-0.003 $[-0.009, +0.004]$ $n = 123/58$	-0.021 $[-0.105, +0.073]$ $n = 36/27$
Gemma / L35	H_out	-0.010 $[-0.029, +0.003]$ $n = 128/60$	-0.002 $[-0.007, +0.004]$ $n = 114/56$	-0.041 $[-0.144, +0.073]$ $n = 40/29$

Continued on next page

Table 35 (continued)

Model / layer	Source	Original loss	CF recovery	Matched Δm
Gemma / L35	x_out	$+0.003$ $[-0.006, +0.017]$ $n = 114/56$	-0.016 $[-0.038, -0.001]$ $n = 128/60$	-0.009 $[-0.079, +0.064]$ $n = 38/28$
Gemma / L35	s	$+0.002$ $[-0.005, +0.008]$ $n = 128/60$	-0.004 $[-0.012, +0.004]$ $n = 128/60$	-0.018 $[-0.137, +0.089]$ $n = 40/29$
Gemma / L35	other	-0.003 $[-0.028, +0.015]$ $n = 128/60$	-0.001 $[-0.023, +0.026]$ $n = 128/60$	$+0.107$ $[-0.063, +0.308]$ $n = 40/29$
Gemma / L35	H_cues	-0.013 $[-0.031, +0.003]$ $n = 128/60$	$+0.008$ $[+0.001, +0.016]$ $n = 127/60$	$+0.010$ $[-0.112, +0.131]$ $n = 40/29$
Gemma / L35	all_cues	-0.006 $[-0.024, +0.008]$ $n = 128/60$	-0.006 $[-0.026, +0.008]$ $n = 128/60$	-0.027 $[-0.131, +0.085]$ $n = 40/29$
Gemma / L35	pre_ents	$+0.008$ $[-0.034, +0.068]$ $n = 128/60$	$+0.015$ $[-0.031, +0.060]$ $n = 128/60$	-0.345 $[-0.646, -0.011]$ $n = 40/29$
Gemma / L35	pre_H	-0.010 $[-0.022, +0.002]$ $n = 124/57$	$+0.005$ $[-0.002, +0.013]$ $n = 124/57$	$+0.057$ $[-0.047, +0.155]$ $n = 40/29$
Gemma / L35	pre_x	-0.004 $[-0.016, +0.012]$ $n = 128/60$	-0.004 $[-0.014, +0.007]$ $n = 128/60$	-0.023 $[-0.138, +0.101]$ $n = 40/29$
Gemma / L41	title	$+0.010$ $[-0.002, +0.022]$ $n = 123/58$	-0.013 $[-0.022, -0.006]$ $n = 123/58$	-0.098 $[-0.290, +0.109]$ $n = 36/27$
Gemma / L41	H_out	-0.018 $[-0.109, +0.033]$ $n = 128/60$	$+0.017$ $[-0.043, +0.123]$ $n = 114/56$	-0.100 $[-0.352, +0.196]$ $n = 40/29$
Gemma / L41	x_out	-0.040 $[-0.059, -0.023]$ $n = 114/56$	$+0.038$ $[-0.002, +0.081]$ $n = 128/60$	$+0.414$ $[+0.196, +0.620]$ $n = 38/28$
Gemma / L41	s	-0.005 $[-0.038, +0.015]$ $n = 128/60$	$+0.019$ $[-0.022, +0.094]$ $n = 128/60$	-0.114 $[-0.283, +0.044]$ $n = 40/29$
Gemma / L41	other	$+0.037$ $[-0.004, +0.077]$ $n = 128/60$	-0.033 $[-0.067, -0.007]$ $n = 128/60$	-0.203 $[-0.395, -0.017]$ $n = 40/29$
Gemma / L41	H_cues	-0.001 $[-0.084, +0.047]$ $n = 128/60$	$+0.001$ $[-0.052, +0.084]$ $n = 127/60$	-0.203 $[-0.503, +0.134]$ $n = 40/29$
Gemma / L41	all_cues	-0.046 $[-0.147, +0.011]$ $n = 128/60$	$+0.046$ $[-0.017, +0.132]$ $n = 128/60$	$+0.174$ $[-0.189, +0.502]$ $n = 40/29$
Gemma / L41	pre_ents	-0.140 $[-0.218, -0.084]$ $n = 128/60$	-0.027 $[-0.066, +0.016]$ $n = 128/60$	$+1.375$ $[+0.588, +2.265]$ $n = 40/29$
Gemma / L41	pre_H	$+0.075$ $[+0.027, +0.139]$ $n = 124/57$	-0.112 $[-0.170, -0.060]$ $n = 124/57$	-0.826 $[-1.403, -0.317]$ $n = 40/29$
Gemma / L41	pre_x	-0.225 $[-0.317, -0.152]$ $n = 128/60$	$+0.078$ $[+0.026, +0.154]$ $n = 128/60$	$+2.224$ $[+1.655, +2.790]$ $n = 40/29$
Phi-4 / L27	title	$+0.005$ $[-0.002, +0.012]$ $n = 116/57$	-0.001 $[-0.006, +0.003]$ $n = 116/57$	-0.017 $[-0.070, +0.027]$ $n = 36/27$
Phi-4 / L27	H_out	$+0.013$ $[+0.006, +0.020]$ $n = 121/59$	-0.012 $[-0.020, -0.004]$ $n = 108/56$	-0.079 $[-0.111, -0.045]$ $n = 40/29$

Continued on next page

Table 35 (continued)

Model / layer	Source	Original loss	CF recovery	Matched Δm
Phi-4 / L27	x_out	-0.015	$+0.025$	$+0.075$
		$[-0.026, -0.006]$	$[+0.019, +0.032]$	$[+0.045, +0.109]$
		$n = 108/56$	$n = 121/59$	$n = 38/28$
Phi-4 / L27	s	$+0.009$	-0.012	-0.003
		$[+0.003, +0.017]$	$[-0.017, -0.007]$	$[-0.046, +0.033]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L27	other	-0.010	$+0.001$	$+0.055$
		$[-0.023, +0.005]$	$[-0.011, +0.012]$	$[-0.036, +0.145]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L27	H_cues	$+0.017$	-0.012	-0.092
		$[+0.009, +0.026]$	$[-0.020, -0.005]$	$[-0.144, -0.041]$
		$n = 121/59$	$n = 120/59$	$n = 40/29$
Phi-4 / L27	all_cues	$+0.005$	$+0.016$	-0.043
		$[-0.006, +0.017]$	$[+0.010, +0.022]$	$[-0.103, +0.006]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L27	pre_ents	-0.002	-0.011	-0.029
		$[-0.019, +0.017]$	$[-0.024, +0.000]$	$[-0.109, +0.041]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L27	pre_H	$+0.004$	-0.005	-0.033
		$[-0.002, +0.013]$	$[-0.013, +0.002]$	$[-0.091, +0.030]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
Phi-4 / L27	pre_x	-0.013	$+0.006$	$+0.052$
		$[-0.026, -0.002]$	$[-0.001, +0.015]$	$[+0.001, +0.104]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L31	title	-0.000	-0.003	$+0.018$
		$[-0.005, +0.006]$	$[-0.009, +0.002]$	$[-0.016, +0.057]$
		$n = 116/57$	$n = 116/57$	$n = 36/27$
Phi-4 / L31	H_out	$+0.001$	-0.003	$+0.009$
		$[-0.004, +0.006]$	$[-0.008, +0.001]$	$[-0.035, +0.059]$
		$n = 121/59$	$n = 108/56$	$n = 40/29$
Phi-4 / L31	x_out	-0.011	$+0.001$	$+0.066$
		$[-0.019, -0.004]$	$[-0.005, +0.005]$	$[+0.039, +0.096]$
		$n = 108/56$	$n = 121/59$	$n = 38/28$
Phi-4 / L31	s	$+0.001$	-0.003	$+0.047$
		$[-0.004, +0.007]$	$[-0.009, +0.002]$	$[+0.018, +0.071]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L31	other	$+0.012$	$+0.001$	-0.044
		$[-0.000, +0.025]$	$[-0.007, +0.009]$	$[-0.145, +0.042]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L31	H_cues	$+0.001$	-0.004	$+0.019$
		$[-0.006, +0.007]$	$[-0.009, +0.000]$	$[-0.035, +0.084]$
		$n = 121/59$	$n = 120/59$	$n = 40/29$
Phi-4 / L31	all_cues	-0.006	$+0.001$	$+0.038$
		$[-0.014, +0.001]$	$[-0.003, +0.006]$	$[-0.010, +0.100]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L31	pre_ents	-0.041	-0.010	$+0.300$
		$[-0.066, -0.019]$	$[-0.024, +0.003]$	$[+0.110, +0.514]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L31	pre_H	$+0.004$	-0.019	$+0.040$
		$[-0.005, +0.015]$	$[-0.037, -0.004]$	$[-0.038, +0.122]$
		$n = 117/56$	$n = 117/56$	$n = 40/29$
Phi-4 / L31	pre_x	-0.034	$+0.003$	$+0.308$
		$[-0.065, -0.012]$	$[-0.015, +0.020]$	$[+0.125, +0.570]$
		$n = 121/59$	$n = 121/59$	$n = 40/29$
Phi-4 / L35	title	$+0.005$	$+0.000$	-0.020
		$[+0.000, +0.011]$	$[-0.004, +0.004]$	$[-0.072, +0.028]$
		$n = 116/57$	$n = 116/57$	$n = 36/27$
Phi-4 / L35	H_out	$+0.004$	-0.004	-0.017
		$[+0.000, +0.008]$	$[-0.009, -0.000]$	$[-0.043, +0.009]$
		$n = 121/59$	$n = 108/56$	$n = 40/29$

Continued on next page

Table 35 (continued)

Model / layer	Source	Original loss	CF recovery	Matched Δm
Phi-4 / L35	x_out	-0.008 $[-0.014, -0.003]$ $n = 108/56$	$+0.009$ $[+0.003, +0.016]$ $n = 121/59$	$+0.068$ $[+0.028, +0.112]$ $n = 38/28$
Phi-4 / L35	s	$+0.005$ $[+0.001, +0.010]$ $n = 121/59$	-0.004 $[-0.011, +0.000]$ $n = 121/59$	-0.018 $[-0.057, +0.014]$ $n = 40/29$
Phi-4 / L35	other	$+0.014$ $[+0.000, +0.032]$ $n = 121/59$	-0.003 $[-0.012, +0.006]$ $n = 121/59$	-0.071 $[-0.202, +0.031]$ $n = 40/29$
Phi-4 / L35	H_cues	$+0.008$ $[+0.003, +0.014]$ $n = 121/59$	-0.000 $[-0.004, +0.003]$ $n = 120/59$	-0.038 $[-0.106, +0.005]$ $n = 40/29$
Phi-4 / L35	all_cues	$+0.003$ $[-0.002, +0.010]$ $n = 121/59$	$+0.007$ $[+0.000, +0.014]$ $n = 121/59$	$+0.004$ $[-0.046, +0.046]$ $n = 40/29$
Phi-4 / L35	pre_ents	-0.048 $[-0.090, -0.018]$ $n = 121/59$	-0.005 $[-0.014, +0.004]$ $n = 121/59$	$+0.278$ $[+0.081, +0.514]$ $n = 40/29$
Phi-4 / L35	pre_H	$+0.027$ $[+0.008, +0.051]$ $n = 117/56$	-0.037 $[-0.073, -0.010]$ $n = 117/56$	-0.097 $[-0.179, -0.037]$ $n = 40/29$
Phi-4 / L35	pre_x	-0.076 $[-0.144, -0.031]$ $n = 121/59$	$+0.040$ $[+0.008, +0.090]$ $n = 121/59$	$+0.356$ $[+0.160, +0.593]$ $n = 40/29$

Table 36: Cross-family teacher-forced paired-control contrasts, validation and test pooled. These head scores are distinct from the fixed-prefix head margin.

Model	Metric	Demotion	Promotion
Phi-4	$\log p(\text{head})$	-2.482 $[-2.991, -1.896]$ $n = 433/103$ $p_H = 1.32 \times 10^{-21}$	$+2.090$ $[+0.889, +3.186]$ $n = 59/22$ $p_H = 0.000422$
Phi-4	D_{cond}	$+0.071$ $[+0.012, +0.124]$ $n = 578/126$ $p_H = 0.0516$	$+0.055$ $[-0.058, +0.168]$ $n = 114/36$ $p_H = 0.377$
Gemma	$\log p(\text{head})$	-4.793 $[-6.036, -3.653]$ $n = 433/103$ $p_H = 3.82 \times 10^{-14}$	$+5.452$ $[+1.938, +8.657]$ $n = 59/22$ $p_H = 0.00337$
Gemma	D_{cond}	$+0.010$ $[-0.091, +0.121]$ $n = 578/126$ $p_H = 1$	$+0.044$ $[-0.246, +0.256]$ $n = 114/36$ $p_H = 1$

Scope of the mechanistic conclusion Together, these experiments locate recoverable output-side states and show that attention interventions can move initiation scores and, in the standard-target 9B extractor, entity participation in generated relations. They do not establish that one attention head or document source is solely responsible, that task fine-tuning created the effect, that all model families respond identically in generation, or that steering alone improves evidence-supported extraction.