ANSWERLESS BRIDGES: RELEVANT TO WHOM? ONE MODEL NEEDS THEM, ANOTHER DOES WITHOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrievers for language models are trained and evaluated with relevance labels: a passage is relevant when it contains the answer, resembles the question, or is annotated as supporting evidence. Each label reads only the question and the passage, so it gives every model the same verdict and never asks: relevant to whom? Multi-hop questions make this concrete, because a model can depend on an *answerless bridge*, a passage without the answer that links the question to the passage containing it. For each model, we delete either the bridge or a control passage from the same context and call the bridge necessary when deleting it alone breaks a correct answer. In each of eleven open-weight models on HotpotQA and MuSiQue, bridge deletion breaks more answers than control deletion, yet for most questions evaluated with several models, the same bridge is necessary for some models and dispensable for others. Lexical, dense and cross-encoder relevance scores give these models one value and predict need no better than chance. ContextCite attribution, fitted to each model's responses to context ablations, predicts need in all ten models tested, and on a shared question it separates the models that need the bridge from those that do not, partly through each model's base rate. The same test recovers the operands of DROP arithmetic questions, whose answers appear in no sentence, and on synthetic prompts shown identically to base and instruction-tuned models it shows that post-training strengthens dependence on the passages holding the answer. The answer to "relevant to whom?" is a particular model, whose own responses measure its need.

1 INTRODUCTION

Retrieval-augmented language models answer questions from retrieved passages, and both the retriever and the retrieved context are judged with relevance labels. Dense passage retrieval treats passages that contain the answer string as positives and evaluates rankings by answer recall (Karpukhin et al., 2020). Multi-hop benchmarks annotate the paragraphs that support each answer (Yang et al., 2018; Trivedi et al., 2022), and lexical and dense retrievers score how closely a passage matches the question (Robertson & Zaragoza, 2009; Reimers & Gurevych, 2019). These labels depend only on the question and the passage, so every model that reads the same context is evaluated against the same target. Utility-based evaluation brings in the model by asking whether it can answer from each passage on its own (Salemi & Zamani, 2024), and on single-hop questions the passages that help a generator depend on the generator (Zhang et al., 2025). Both score a passage by what it adds on its own, which gives little credit to a passage that helps only together with another. For such passages, neither kind of label answers the question this paper asks: relevant to whom? Put concretely: when several models read the same context, which of them need a given passage to answer?

Multi-hop questions make the choice of label consequential. The question in Figure 1a asks when the novel that Lourenço Mutarelli based "Nina" on was first published. One retrieved document names the novel, *Crime and Punishment*, and another gives its publication year, which a reader can connect to the question only through the novel's title. We call a passage of the first kind an *answerless bridge*: it contains no answer string, and it links the question to the passage that does. Answer containment labels the bridge a negative, word overlap with the question favours it over a non-supporting passage from the same context, and supporting-paragraph annotations mark it as evidence for every model. Whether a particular model needs the bridge is a question about that model, and answering it requires intervening on the model's input.

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.

Measuring this need raises two difficulties. Deleting any passage shortens the context and changes its layout, so an answer can break even when the deleted text carried nothing the model used. A test of need therefore has to compare each deletion with a comparable deletion from the same input. Need is also defined per model, and if two models that read an identical context depend on different passages, no single corpus annotation or question-based score can serve as the target for both. Per-model need matters wherever relevance labels are used to train retrievers, to evaluate retrieved context, or to judge whether an attribution method explains a model.

We label necessity separately for each model with a deletion test that carries its own control (Figure 1c). For each question that a model answers correctly with the full context and incorrectly with no context, we delete the bridge and, separately, a non-supporting passage from the same context, the control. The bridge is *necessary* for that model when deleting it alone breaks the correct answer, and *dispensable* otherwise. Both deletions share the question, answer, model and remaining context, so comparing them separates the effect of the bridge from the cost of deleting text, with no fitted parameter or threshold.

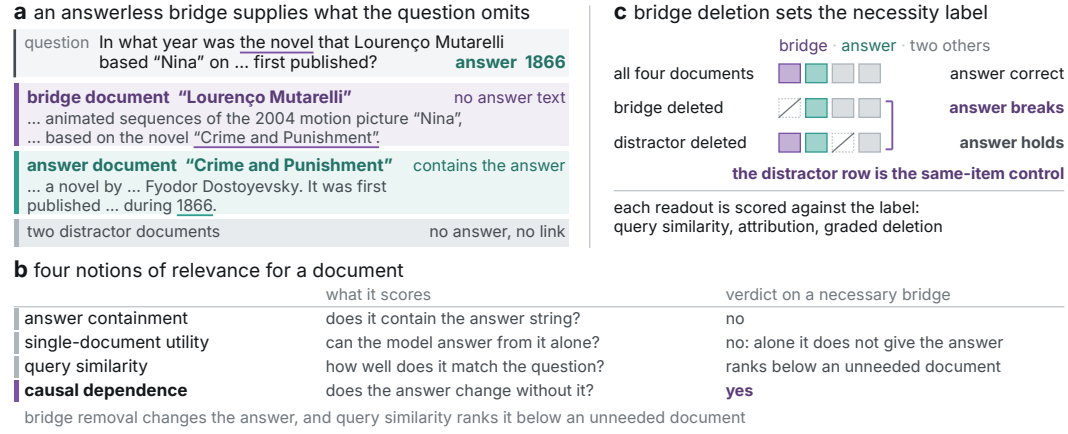

Figure 1: **(a)** One HotpotQA question has four retrieved documents: one names the omitted novel, another contains the answer, and the other two do neither. **(b)** Each relevance measure asks a different question about the bridge. **(c)** For Gemma-2-9B, bridge deletion breaks the correct answer, giving a positive necessity label. Deleting a distractor, the same-item control, preserves the answer. The paired test compares the two deletions across all rows (Section 3).

On eleven open-weight models with HotpotQA and MuSiQue, bridge deletion breaks more correct answers than control deletion in every model, and, as Section 4 shows, the same test finds the operand sentences of DROP arithmetic questions (Dua et al., 2019). Most questions answered by several models have a bridge that some models need and others do not, so the answer to "relevant to whom?" is a particular model. Lexical, dense and cross-encoder relevance scores give all of these models one value and predict necessity no better than chance. ContextCite attribution (Cohen-Wang et al., 2024) separates necessary from dispensable bridges in every model tested, and on a shared question, where every relevance score ties, it also separates the models that need the bridge from those that do not, partly through how often each model needs bridges (Section 5). On synthetic prompts shown identically to base and instruction-tuned models, post-training strengthens dependence on the short notes that hold the answer words (Section 6). Attribution can also be read directly from attention outputs, which requires a choice of how signed contributions are aggregated, and on DROP questions two common choices credit the needed operands in disjoint sets of models (Section 7).

Contributions. We define a per-model necessity label and test it against a control deletion from the same item (Section 3). We show that passages without the answer are necessary beyond the cost of deletion and that which of them are necessary depends on the model (Sections 4 and 5). We use the label as a reference for relevance scores and attribution, within each model and across models on the same question (Section 5), and controlled deletion to measure the effect of post-training (Section 6) and aggregation choices in attention-output attribution (Section 7).

2 RELATED WORK

Multi-hop question answering. HotpotQA places supporting facts across documents and makes their joint role available for evaluation (Yang et al., 2018). MuSiQue composes connected single-hop questions so that an intermediate answer links the reasoning chain (Trivedi et al., 2022), and multi-hop corpora can also expose a final answer together with its individual reasoning steps (Ho et al., 2020). Because later passages in a chain are often reached through earlier ones, retrieval methods follow chains of passages (Xiong et al., 2021) or alternate retrieval with reasoning (Trivedi et al., 2023).

What retrieval evaluation counts as relevant. Dense Passage Retrieval uses answer containment to select positive passages and evaluates the resulting ranking with answer-recall@k (Karpukhin et al., 2020). Classical term specificity and BM25 score matches between query and document terms (Sparck Jones, 1972; Robertson & Zaragoza, 2009). Single-document utility scores whether a generator can answer from one retrieved document (Salemi & Zamani, 2024), and sufficiency asks whether the available context contains enough information to answer (Joren et al., 2025). On single-hop questions, Zhang et al. (2025) define utility as the generator's gain from adding a passage over answering without evidence, and find that useful evidence depends on the generator. Our label is also specific to the model, and it is defined by deleting one passage from the full multi-hop context, with a matched deletion from the same item as the control.

Deletion, rationales, and context attribution. Erasure measures how predictions change when part of a model's input is removed (Zeiler & Fergus, 2014; Li et al., 2016). Rationale benchmarks test selected evidence for sufficiency and comprehensiveness (DeYoung et al., 2020), and Sun et al. (2026) evaluate highlight explanations of context use. ContextCite reads passage contributions from a surrogate fitted to context ablations (Cohen-Wang et al., 2024), AttriBoT approximates leave-one-out context attribution (Liu et al., 2025), and SelfCite rewards citations using necessity and sufficiency under context ablation (Chuang et al., 2025). We use single-passage deletion against a matched same-item control as the reference for evaluating such attributions.

Closest work. Hu et al. (2026) show that hard prompt compression can leave the answer path of many bridge examples incomplete across three multi-hop datasets, and that reinserting the missing supporting paragraph under a matched token budget recovers most of the lost accuracy for Qwen3-8B. Both studies start from supporting-paragraph annotations. Testing whether deleting the annotated bridge breaks each model's correct answer, we find that a bridge one model needs can be dispensable for another (Section 5). Similarity can also miss statements that reasoning requires (BehnamGhader et al., 2023). Appendix I extends these connections.

3 A DELETION LABEL WITH A SAME-ITEM CONTROL

The label. We treat each released checkpoint as a separate model and call one question evaluated with one checkpoint an item–model row. In the cross-document setting, the bridge and the answer lie in different documents. A row is eligible when greedy decoding answers correctly with the full context and incorrectly closed-book under a normalised substring scorer (Appendix B), so the model's answer depends on the context. On an eligible row, the bridge is necessary when deleting it alone breaks the correct answer, and dispensable otherwise. The control enters through the paired test below, which asks whether bridge deletions break more answers than control deletions on the same rows. For full context C, gold answer a and passage b_i, the graded deletion score is

$$\Delta_i = \log p(a \mid C) - \log p(a \mid C \setminus b_i),$$

where $\log p$ sums teacher-forced token log-probabilities (Zeiler & Fergus, 2014; Li et al., 2016). The bridge's Δ_i is the continuous form of the deletion that defines the label. For a target passage, which is the bridge here and an operand or a note in later settings, the paired contrast $\Delta_{\text{target}} - \Delta_{\text{control}}$ measures how much more its deletion costs than the control's. The primary test compares bridge and control answer breaks with a one-sided exact McNemar test (McNemar, 1947), which uses the rows where exactly one of the two deletions breaks the answer and gives each row equal weight. Bootstrap intervals resample items (Efron, 1979), with draw counts in Appendix H.

Bridges and controls. Supporting-paragraph annotations identify the bridge. We use HotpotQA bridge-type validation questions in the distractor setting, which have exactly two supporting paragraphs, and answerable MuSiQue questions, which have at least two. The answer string must occur as a case-insensitive substring in exactly one paragraph of the benchmark context, and that paragraph must be supporting. The bridge is the first other supporting paragraph, so it never contains the answer string (Appendix C reports shared answer words). The presented context contains four passages of at most 700 characters each: the answer passage, the bridge and two non-supporting distractors, shuffled per item. The first distractor in dataset order serves as the control, so the paired comparison holds the question, answer, model and remaining context fixed. The control is on average 11% longer than the bridge, so its deletion removes more text, and restricting the comparison to controls at least as long as their bridges increases the graded contrast from $+1.59$ to $+1.76$ nats (Table 9).

Readouts. A readout assigns each passage a score, and we compare three kinds of readout with deletion. Relevance scores read only the question and the passage: TF-IDF cosine over content words, BM25, the MiniLM bi-encoder, and the MS MARCO and BGE cross-encoder rerankers (Robertson & Zaragoza, 2009; Reimers & Gurevych, 2019; Wang et al., 2020; Nogueira & Cho, 2019; Bajaj et al., 2016; Chen et al., 2024). ContextCite fits a surrogate to each checkpoint's responses to 48 context ablations per item (Cohen-Wang et al., 2024), and its attribution share divides a passage's positive score by the item's summed positive passage scores. Writer mass, used in the DROP, source-count and synthetic settings, reads attention output in one forward pass: it projects each head's output onto the answer token's unembedding direction, as in direct logit attribution (Elhage et al., 2021), and splits it by source position. For answer token y with unit unembedding direction d_y, let $c_{\ell hs}$ be the output of attention head h in layer ℓ from source position s at the position that predicts y, projected onto d_y. A role is a labelled group of source positions, such as a bridge, an operand, a note or the control. The writer mass of role r keeps the positive part of each contribution, which we call the positive-half convention, and normalises over roles,

$$p_r = \frac{\sum_{s \in S_r} \sum_{\ell,h} \max(c_{\ell hs}, 0)}{\sum_{r'} \sum_{s \in S_{r'}} \sum_{\ell,h} \max(c_{\ell hs}, 0)},$$

where S_r is the set of token positions in role r (Appendix D gives the full definition). We score readouts by ROC-AUC for separating necessary from dispensable bridges, where 0.5 indicates no discrimination, and use graded deletion as the reference.

Settings and models. Every setting deletes a target against a same-item control (Table 3). The cross-document setting covers HotpotQA and MuSiQue with eleven models. A DROP setting with seven models tests deletion on computed answers that appear in no sentence, where the needed sentences are known by construction. A source-count setting varies the number of sentences that jointly supply an answer from two to five. A synthetic setting shows identical inputs to the base and instruction-tuned releases of four models, to isolate post-training. The readout set is smaller than the paired set, because ContextCite runs on a subset of the items, keeps the rows where its own response is correct, and could not be fitted for Qwen2.5-7B (Table 4).

4 Bridge and operand deletions exceed same-item controls

Deleting the bridge breaks more correct answers than deleting the control in both corpora and in all eleven models (Figure 2b). Pooled over 1729 rows, bridge deletion breaks 556 answers and control deletion breaks 126. The difference holds in each corpus separately, with 282 versus 49 breaks across 983 HotpotQA rows and 274 versus 77 across 746 MuSiQue rows. One-sided exact McNemar tests give $p < 0.05$ in ten of eleven models in each corpus.

DROP arithmetic questions test deletion in a setting where answer containment marks every sentence negative. Each question combines quantities stated in different sentences, so the computed answer appears in no sentence. An operand is a sentence that supplies one quantity in the computation, and the control is the earliest remaining sentence that contains a number. Operands are needed by construction, so this setting checks that deletion recovers evidence known to be necessary. Deleting either operand costs more than deleting the control, with bootstrap intervals above zero for both operands in all seven models (Figure 3c). The gaps remain positive when the control is re-chosen to maximise question overlap (Table 9).

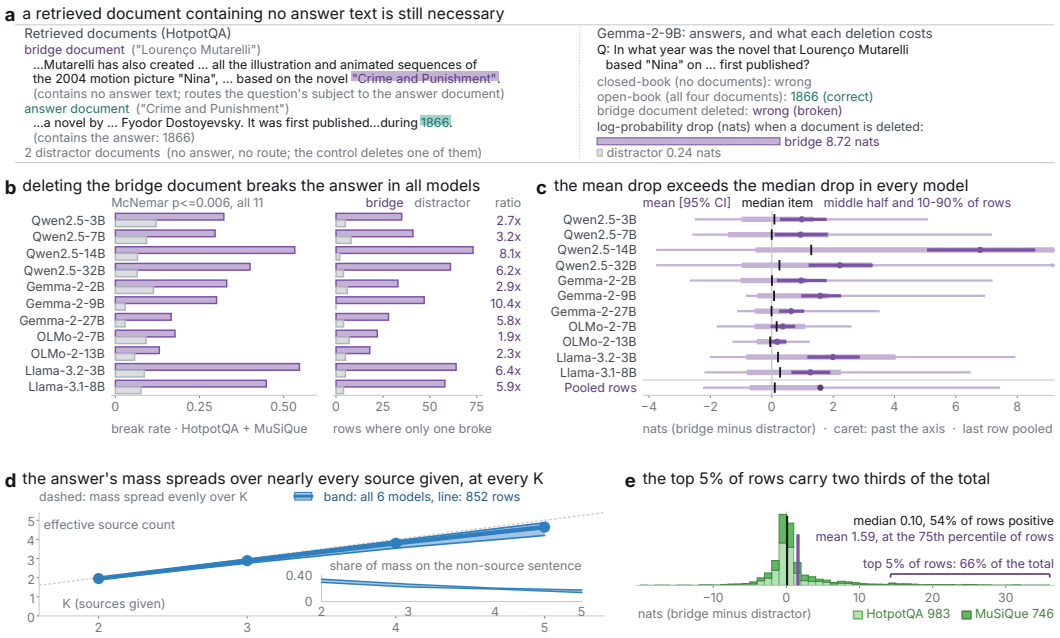

Figure 2: Bridge deletion breaks more answers than control deletion across both natural corpora. **(a)** In one HotpotQA item, bridge deletion breaks Gemma-2-9B's correct answer, while control deletion costs 0.24 nats. **(b)** Paired answer-break counts across 1729 rows. Discordances count answer breaks unique to either deletion. **(c)** Bridge-minus-control log-probability drops: means with bootstrap 95% intervals, medians and row quantiles. Every interquartile range spans zero. **(d)** Effective number of attributed sources against K in the source-count setting, with a count $= K$ diagonal. The line pools 852 rows with correct and incorrect answers, the band spans six models, and the inset gives the control's share before renormalisation. **(e)** Pooled bridge-minus-control drops.

Single-document utility scores each source on its own, so it cannot credit sentences that supply an answer only together. The source-count setting asks how writer mass divides credit among such sentences: a question asks for the sum of single-digit yard numbers stated in K DROP sentences, and one control sentence is added. We remove the control, renormalise the shares over the K sources, and divide the effective number of attributed sources, the exponential of the writer-mass entropy, by K. This coverage is 1 when the mass is spread evenly over the sources and $1/K$ when one source holds all of it. Coverage stays between 0.84 and 0.99 across six models and $K \in \{2, 3, 4, 5\}$ (Figure 2d). Before renormalisation, the control's share falls from 0.30–0.35 at $K = 2$ to 0.14–0.18 at $K = 5$, close to an even split over all $K + 1$ sentences, so writer mass credits the control like a source.

5 Attribution predicts model-specific necessity that relevance scores miss

Necessity varies by model. The necessary-bridge fraction is 0.314 over the 1127 readout rows and ranges from 0.131 to 0.539 across models (Figure 4d). Of the 198 items in this set, 180 are evaluated with at least two models, and 120 of these 180 have a bridge that some models need and others do not. Every model compared on an item answers it correctly with the full context and incorrectly closed-book, so the disagreement arises among models that all solve the same input. Over all pairs of models that share an item, the two labels disagree at a rate of 0.284, between the zero of a fixed item annotation and the 0.436 expected from independent draws at each model's own rate (Appendix E).

What a readout must track. A relevance score reads only the question and the passage, so it gives a bridge the same value for every model. Every bridge is an annotated supporting paragraph, so a retriever trained on these annotations, such as multi-hop dense retrieval (Xiong et al., 2021), treats the bridges a model needs and those it can do without as the same positives. A score that reproduced

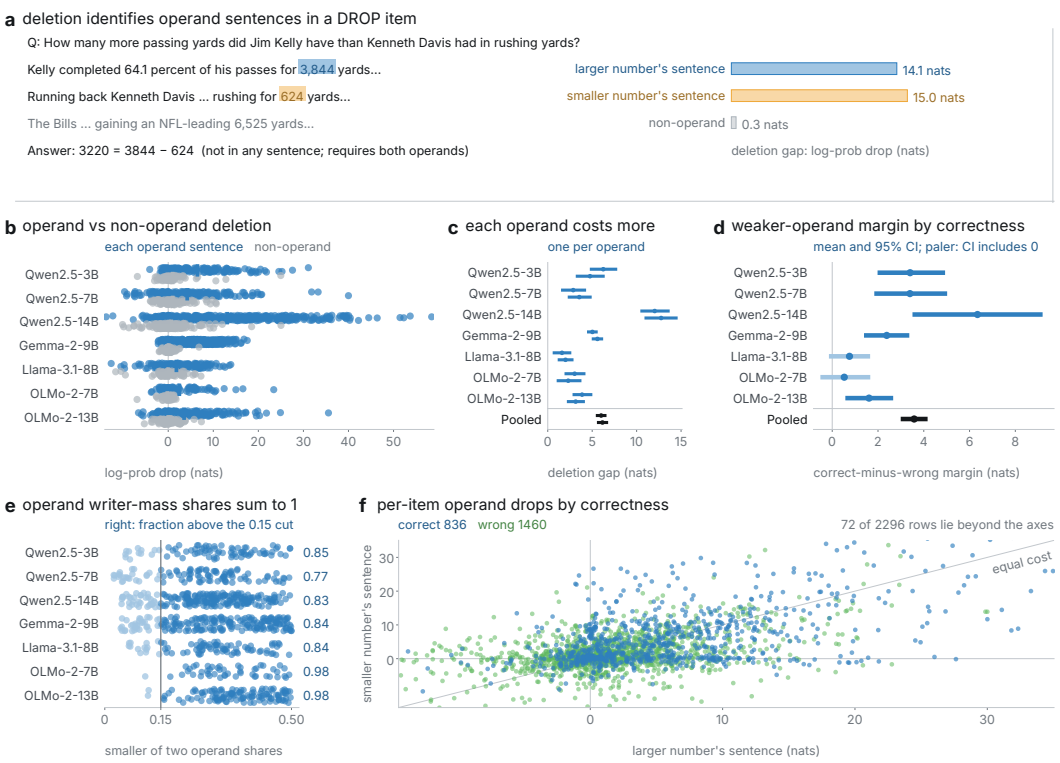

Figure 3: Operand deletion separates necessary sentences from controls in DROP arithmetic. Costs are gold-answer log-probability drops (nats). **(a)** In one item, operand deletion costs 14–15 nats, against 0.27 for the number-bearing control. **(b)** Each of 836 correctly answered rows supplies two operand costs and one control cost across seven models. **(c)** Mean operand-minus-control drops have 95% intervals above zero for both operands in all seven models. **(d)** The margin is the lower operand cost minus control cost. Correct-minus-incorrect mean margins are positive, with intervals including zero for two models. **(e)** Smaller operand share of positive-half writer mass on 2296 rows (836 correct, 1460 wrong) and the fraction meeting 0.15. **(f)** Deletion costs for larger- and smaller-number operand sentences, split by correctness.

these annotations exactly would give every bridge the same value and an AUC of 0.5. A fixed score can still rank one model's necessary bridges above its dispensable ones, so we first evaluate each readout pooled and within each model, asking which bridges a model needs. We then ask to whom a bridge is relevant by comparing the models that read the same item, where a fixed score ties, so any separation between them must come from the models themselves.

Relevance scores and attribution. Relevance scores do no better than chance at separating necessary from dispensable bridges (Figure 4a). Pooled AUC is 0.453 for TF-IDF and 0.535 for MiniLM, and the pooled intervals of all five relevance scores include or lie below 0.5 (Table 8). Bridges share more question content words than controls do (4.16 versus 3.02, Table 9), and TF-IDF, which weights this overlap, does not favour the bridges a model needs. Within models, MiniLM exceeds 0.5 in eight of ten ($p = 0.109$) and each reranker in five of ten, so no relevance score consistently favours necessary bridges. ContextCite's attribution share reaches a pooled AUC of 0.728 $[0.683, 0.771]$ and exceeds every relevance score in all ten models (two-sided sign test, $p = 0.002$). Its advantage over MiniLM, the strongest relevance score, is $+0.194$ $[+0.128, +0.260]$ (paired whole-item bootstrap). Graded deletion, the continuous form of the label's intervention, reaches 0.825 $[0.785, 0.863]$, and in every model the AUCs fall in the order graded deletion, attribution share, TF-IDF (Figure 4a; Table 7). Its AUC stays below 1 because drop size does not decide whether the greedy answer changes.

Across models on the same item. The 120 items with both labels give 984 pairs of a model that needs the bridge and one that does not, and the across-model AUC counts how often a readout scores



Figure 4: Attribution discriminates bridge necessity in each model. **(a)** ROC-AUC of each readout against the necessity label. Attribution exceeds chance in all ten models (two-sided sign test, $p = 0.002$). **(b)** Causal share normalises zero-clipped deletion drops within each item. Spearman correlations average over 353 necessary-bridge rows with defined shares. **(c)** Top-1 retention: a bridge ranks first in its item. TF-IDF retains 0.55 of necessary and 0.50 of dispensable bridges, versus attribution's 0.32 and 0.12, so attribution separates the two kinds more widely. **(d)** Necessary-bridge fractions have 95% Wilson intervals. The pooled interval treats rows as independent.

the needing model higher (Appendix G). Relevance scores tie on every pair (0.5); attribution share reaches 0.628 [0.566, 0.688] and graded deletion 0.819 [0.771, 0.865]. A model's necessity rate on its other items scores 0.779, so this partly reflects which models need bridges more often; centred per model, attribution keeps 0.583 [0.522, 0.643], and ranked per model 0.558 [0.496, 0.618].

Ranking within an item. For a necessary bridge, attribution share orders the four passages as causal share does at mean Spearman 0.689, against 0.112 for TF-IDF (Figure 4b).

6 Post-training raises dependence on the notes that hold the answer

Models that read identical inputs need different bridges (Section 5), and released models differ in pretraining data, size and post-training at once. To isolate post-training, we compare the base and instruction-tuned releases of OLMo-2-7B, Qwen2.5-7B, Qwen2.5-1.5B and Gemma-2-2B, which we call base and tuned. The two members of a pair share a pretrained initialisation and differ in

post-training data, objective and duration. Within a pair, the prompts, candidate answers, deletion targets and formatting are identical, and no chat template is applied. The prompts are synthetic, so the note that holds each answer word is known (Appendix B). In one prompt, the question asks for the two-word phrase of Orion ledger. One note gives Velora signal's phrase as "amber canyon", another gives Prairie packet's phrase as "cedar harbor", a link note says that Orion ledger takes its first word from Velora signal and its second word from Prairie packet, and a fourth note is a distractor. The first-word and second-word notes, which we call the two roles, each hold one word of the answer "amber harbor", and four allowed phrases are scored by log-probability.

Two indicators describe how a model uses the notes, and both are recorded on every prompt, including prompts answered incorrectly. The *argmax* indicator asks which note the model credits: it uses writer mass and is positive when, for each answer word, the role that holds the word receives the largest share. The *deletion-sensitivity* indicator asks whether the answer depends on a role's note: it records whether deleting the note lowers the log-probability of its answer word by at least one nat.

Table 1: Post-training changes deletion sensitivity more than the argmax indicator on the four-note stimulus. Block headers give the populations, with 160 prompts from 40 cases per pair. Arrows join base and tuned rates, with tuned/base ratios from unrounded counts in parentheses, and tests swap base and tuned labels jointly within each case.

What the indicator measures	base → tuned effect	p
All 640 *pair–prompt comparisons over* 4 *pairs*		
Argmax: selects both correct roles	$0.731 \to 0.773\,(1.06\times)$	0.091
Deletion sensitivity: the first-word role	$0.489 \to 0.703\,(1.44\times)$	$< 10^{-6}$
Deletion sensitivity: the second-word role	$0.086 \to 0.291\,(3.38\times)$	$< 10^{-6}$
Deletion sensitivity: both roles	$0.053 \to 0.217\,(4.09\times)$	$< 10^{-6}$
173 *pair–prompt comparisons with both members correct*		
Argmax: selects both correct roles	$0.896 \to 0.844\,(0.94\times)$	0.21
Deletion sensitivity: the first-word role	$0.497 \to 0.653\,(1.31\times)$	0.0048
Deletion sensitivity: the second-word role	$0.029 \to 0.127\,(4.40\times)$	0.0033
Deletion sensitivity: both roles	$0.012 \to 0.087\,(7.50\times)$	0.011

Post-training raises deletion sensitivity for both roles, while the pooled argmax indicator shows no detectable change (Table 1). The rise persists on the comparisons that both members answer correctly, which hold correctness fixed. In three of four pairs the argmax rate does not rise, and in all four both deletion-sensitivity rates do (Appendix E). Tests permute base and tuned labels by case, a group of prompts sharing subjects and an answer phrase (Appendix G).

The link note contains neither answer word, so it is a synthetic answerless bridge. A second stimulus gives each answer word its own pointer note of this kind, such as "Orion ledger starts with the same word as Velora signal", and scores how much more deleting a pointer note costs the word it points to than the other word. A matching chain about a subject the question does not ask about gives the control, which applies the same contrast to pointer notes that the answer does not use (Appendix E). On the eight-note format, post-training raises this pointer-specific cost beyond the control in three of four pairs (Table 6), so those tuned models depend more on notes without an answer word.

7 Two attribution conventions change which models pass

Writer mass can aggregate signed contributions in two ways: the positive-half convention of Section 3 keeps only contributions toward the answer token, and *net projection* sums them all (Appendix D). We apply both to the correctly answered DROP rows of Section 4, where both operands are needed by construction and the deletion label involves neither convention.

We ask whether each measure places even the weaker operand, the one with the lower value on that measure, above the control. For each measure and model, a sign test across items asks whether it does, and the model *passes* when the test finds the weaker operand above the control (Table 2). Positive-half writer mass passes three models and net projection passes two, and the two sets are disjoint. Deletion passes three models on this item-level test, although its mean gap for each operand role is positive in all seven (Figure 3c). Only the positive-half verdict tracks accuracy: the fraction of

Table 2: Writer-mass conventions and deletion on correctly answered DROP rows. Each cell is an exact two-sided sign test, per model, of the lower operand value minus the control over non-zero differences. Cells mark $p < 0.05$ favouring the operands ("pass") or the control ("opp."), and $p \geq 0.05$ with a dash. Positive-half shares cover both operands and the control, and deletion is the gold-answer log-probability drop. The last row correlates accuracy with each model's fraction of rows where both operands exceed the control (Spearman).

Model or summary	positive half	net projection	deletion
Gemma-2-9B	pass	–	pass
Qwen2.5-14B	pass	opp.	pass
Qwen2.5-7B	pass	opp.	–
Llama-3.1-8B	–	pass	–
OLMo-2-13B	–	pass	–
Qwen2.5-3B	–	–	pass
OLMo-2-7B	opp.	–	–
Sign-test passes (of 7 models)	3	2	3
Opposite-direction significant	1	2	0
Rank correlation with accuracy	$+0.929$	-0.214	$+0.536$

Notes. The seven models answer 836 rows correctly, 48–208 per model, from 328 items that meet the construction rules, and accuracy is the fraction of those items answered correctly. Exact rank-correlation tests over all 5040 accuracy orderings give $p = 0.007$, 0.66 and 0.24 for the three columns.

rows where both operands exceed the control rises with model accuracy (Spearman $+0.929$, exact $p = 0.007$), while net projection and deletion show no detectable trend.

Source-only normalisation. Without the control, the criterion measures only operand balance, and the accuracy trend disappears. Figure 3e divides each operand's positive-half mass by their sum, and a row meets threshold t when both shares are at least t. Rank correlations with accuracy lie between -0.61 and 0.00 at each threshold where defined (exact permutation $p \geq 0.15$; Appendix E).

8 Discussion and conclusion

Which passages a model needs is a property of that model, measured by a binary deletion label that needs no fitted parameter or threshold and is tested against a same-item control. Because models that solve the same input need different bridges (Section 5), a label collected with one model carries positives that another can do without. No score of the question and the passage can say to whom a bridge is relevant. Answer containment marks a bridge negative even when a model needs it, which removes it from retriever-training positives and answer-recall evaluation, and supporting-paragraph annotations mark every bridge as a positive. Single-document utility scores each source on its own, and writer mass spreads over every sentence that jointly supplies an answer and over the control.

Readouts computed from attention output depend on how contributions are aggregated (Table 2), and an interventional label gives them a reference that no aggregation choice changes. On fixed inputs, post-training raises how much the answer depends on the notes more than how often the model credits them, so a dependence label measured on a base model must be remeasured on its tuned release.

The label is defined on the rows a model answers correctly with the full context and incorrectly closed-book, and computing it requires repeated forward passes. Bridges come from benchmark supporting-paragraph annotations, and the measurements describe contexts of four passages. Across the settings of Table 3, annotated support, per-model dependence and attribution behave as distinct quantities, and evaluators can record deletion outcomes for their model and recompute them when it changes.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, including the experiment and analysis code and the code that constructs the synthetic prompts, to provide feedback on the research methodology and experiments through reviews of manuscript drafts, and to help interpret results while drafting the text. Formulating or proving mathematical claims, and qualitative or thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, draw the figures, draft and edit the paper for readability, suggest its structure and title, identify relevant literature and format references. We have reviewed all AI-assisted work: we reviewed every sentence of the manuscript, and we executed the AI-written code and checked it against recorded per-item outputs. We made the final decisions on the research questions, experimental design, claims and interpretation, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Deletion labels support targeted audits by identifying passages whose removal breaks correct answers and measuring model dependence. A model can depend on a false passage, so an audit that uses the label to judge answers or citations also needs an independent check against the source or the world. Deletion requires repeated access to the deployed model. This requirement favours those who host the weights over auditors who need to examine the model. Controlled queries or audit budgets can narrow this gap. ContextCite attribution requires the same repeated model access and, with 48 ablations per item, more forward passes than deleting each of the four passages. No new personal data were collected for this work. The study uses public benchmark corpora and public checkpoint releases.

REPRODUCIBILITY STATEMENT

Appendix A specifies the deletion definition, teacher-forced scoring, bootstrap procedure, and exact McNemar tests. Table 3 records the settings and analysis populations, Appendix E gives per-model detail, and Table 9 reports the control checks. Appendix H fixes the run settings for the ablation budget, item counts, resampling, decoding, and numerical precision. All corpora are public benchmarks and all checkpoints are public releases, making the label recomputable by anyone with inference access. The experiment and analysis code, prompts, run settings and stored per-item outputs are in an anonymous repository, `https://anonymous.4open.science/r/answerless-bridge-8AE6`. Appendices B, C, D, and G give the templates, item-selection rules, attribution definitions, and inferential unit, respectively.

REFERENCES

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated MAchine reading COmprehension dataset, 2016. arXiv:1611.09268.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. Can retriever-augmented language models reason? The blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15492–15509, 2023. doi: 10.18653/v1/2023.findings-emnlp.1036. arXiv:2212.09146.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. Seemingly plausible distractors in multi-hop reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2514–2528, 2024. doi: 10.18653/v1/2024.emnlp-main.147.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 2397–2430, 2023. arXiv:2304.01373.

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

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models, 2022. arXiv:2212.08037.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, 2022. doi: 10.18653/v1/2022.emnlp-main.146.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137.

Jifan Chen and Greg Durrett. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4026–4032, 2019. doi: 10.18653/v1/N19-1405.

Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. SelfCite: Self-supervised alignment for context attribution in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 10839–10858, 2025.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95764–95807, 2024. doi: 10.52202/079017-3035.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024. doi: 10.1145/3626772.3657834.

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

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer C. White, Aaron Schein, and Ryan Cotterell. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13211–13235, 2024. doi: 10.18653/v1/2024.acl-long.714.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019. doi: 10.18653/v1/n19-1246.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *PMLR*, pp. 5988–6008, 2022.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. arXiv:2408.00118.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *International Conference on Learning Representations*, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. arXiv:2407.21783.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. How do LLMs perform two-hop reasoning in context?, 2025. arXiv:2502.13913.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhengpei Hu, Kai Li, Dapeng Fu, Xuechao Zou, Yuanhao Tang, Yue Li, Tengfei Cao, and Jianqiang Huang. Relevant but incomplete: Referential dangling as a paradigm-level failure mode in hard prompt compression, 2026. arXiv:2608.04569.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

Akriti Jain and Aparna Garimella. Modeling contextual passage utility for multihop question answering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 464–471, 2025. doi: 10.18653/v1/2025.ijcnlp-short.37.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. doi: 10.18653/v1/n19-1357.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91.

Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2726–2736, 2019. doi: 10.18653/v1/P19-1262.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024.

Randhir Kumar. Verification without sufficiency: Per-chunk filtering fails on multi-hop RAG, and decomposition repairs it, 2026. arXiv:2608.00585.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2016. arXiv:1612.08220.

Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. Attributing response to context: A Jensen–Shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *International Conference on Learning Representations*, 2026.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *International Conference on Learning Representations*, 2024.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.15102.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, 2023. doi: 10.18653/v1/2023.findings-emnlp.467.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

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

Jiecheng Lu and Shihao Yang. HyperMLP: An integrated perspective for sequence modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FmpyWzWtyL.

Neh Majmudar and Elena Filatova. Load-bearing context: The question damage score for evaluating context reliance in linguistic reasoning, 2026. arXiv:2608.27756.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4249–4257, 2019. doi: 10.18653/v1/P19-1416.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. Controllable context sensitivity and the knob behind it. In *International Conference on Learning Representations*, 2025.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT, 2019. arXiv:1901.04085.

Shubham Santosh Pandere, Gautam Ranka, Ritika Varshney, Navya Deshmukh, Roushni Sareen, and Roshan Kumar Singh. Rewired or gated? how instruction tuning shapes knowledge-conflict circuits in LLMs, 2026. arXiv:2609.25602.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053, 2024. doi: 10.18653/v1/2024.emnlp-main.347.

Qwen Team. Qwen2.5 technical report, 2024. arXiv:2412.15115.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2395–2400, 2024. doi: 10.1145/3626772.3657957.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 31210–31227, 2023.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024. doi: 10.18653/v1/2024.naacl-short.69.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Jingyi Sun, Pepa Atanasova, Sagnik Ray Choudhury, Sekh Mainul Islam, and Isabelle Augenstein. Evaluation framework for highlight explanations of context utilization in language models. *Computational Linguistics*, 2026. doi: 10.1162/coli.a.621. Early access.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3319–3328, 2017.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293, 2020. doi: 10.18653/v1/2020.emnlp-main.746.

Yixuan Tang, Hwee Tou Ng, and Anthony K. H. Tung. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3244–3249, 2021. doi: 10.18653/v1/2021.eacl-main.283.

Team OLMo. 2 OLMo 2 Furious, 2024. arXiv:2501.00656.

Quoc-Huy Trinh, Lin Zhu, and Sebastian Szyller. How context attribution handles what the model already knows, 2026. arXiv:2607.23804.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8846–8863, 2020. doi: 10.18653/v1/2020.emnlp-main.712.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10014–10037, 2023. doi: 10.18653/v1/2023.acl-long.557.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 22–32, 2025. doi: 10.1145/3731120.3744592.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. On the loss of context-awareness in general instruction fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. doi: 10.1162/tacl_a_00021.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*, 2021.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, 2024.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, 2024. doi: 10.18653/v1/2024.acl-long.550.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/d18-1259.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1941–1951, 2024. doi: 10.1145/3626772.3657784.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. LLM-Specific utility for retrieval-augmented generation, 2025. arXiv:2510.11358.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, 2023. doi: 10.18653/v1/2023.findings-emnlp.968.

Appendix A defines scoring and models, and Appendix B gives the exact input templates and worked inputs. The construction of each corpus pool, structural rule, control and behavioural filter is traced in Appendix C, and the attribution projections and normalisation conventions are defined in Appendix D. Model-level results follow in Appendix E and matched-control robustness in Appendix F. Appendix G identifies the inferential unit, pairing, resampling and replication criterion, and Appendix H gives fixed run settings and release information. The final appendix, Appendix I, surveys related work on post-training, multi-hop reasoning, context utility and compression, attribution, and properties relative to a model. Together, these sections let a reader follow an item from its construction through scoring and model-specific inference.

A Protocol

Deletion. An item–model row evaluates one question with one checkpoint. Cross-document eligibility requires an incorrect closed-book answer and a correct answer with all documents. The scorer in Appendix B determines both conditions. DROP deletion analyses retain correct full-context answers. Figure 3d and f also compare correct and incorrect answers.

A bridge is a second supporting paragraph of the item, chosen by the rule in Appendix C; it supplies an intermediate entity on the reasoning path from the question to the answer and does not contain the answer string. It is necessary for a checkpoint when bridge deletion alone makes its correct greedy answer incorrect. It is dispensable otherwise. The control enters through the paired contrast and the paired test below. For full context C, context blocks $\{b_1, \ldots, b_n\}$ and gold answer a, the graded deletion score is $\Delta_i = \log p(a \mid C) - \log p(a \mid C \setminus b_i)$ (Zeiler & Fergus, 2014; Li et al., 2016). Here, $\log p$ sums teacher-forced token log-probabilities. Section 5 uses the bridge's raw Δ_i as the continuous deletion reference for its binary necessity label.

The paired graded contrast subtracts the control's drop: $\Delta_{\text{target}} - \Delta_{\text{control}}$. Its intervals use bootstrap percentiles over items (Efron, 1979), with draw counts in Appendix H. Paired answer-flip counts use a one-sided exact McNemar test on discordant outcomes (McNemar, 1947). The binary necessity label and paired contrast use no added decision threshold. The synthetic deletion-sensitivity indicator requires a drop of at least one nat and is recorded on every prompt, regardless of correctness. The sign tests in Sections 5 and 7 are exact and two-sided. The synthetic comparisons in Section 6 use case-level permutation tests.

Attribution. We run ContextCite (Cohen-Wang et al., 2024) with 48 ablations per item at document granularity in the cross-document setting. The within-item share of positive attributed mass provides a common cross-model scale for pooled figures, and the per-model tables report both shares and raw scores. Query similarity is the TF-IDF cosine between the block and the question over content words (Sparck Jones, 1972; Robertson & Zaragoza, 2009). The raw-rank within-item Spearman coefficient (Spearman, 1904) measures agreement between the attribution and deletion rankings, with values in Table 7. After zero-clipping, which mostly affects the deletion side, the positive-share coefficients for attribution with deletion and for query similarity with deletion in Figure 4b are $\rho = 0.689$ and $\rho = 0.112$, respectively. Rank-space quantities use 354 rows with necessary bridges, and share-space quantities use 353: one row has non-positive deletion cost on every document, leaving its causal share undefined.

Necessity populations for Figure 4. Each model receives 300 items: 150 HotpotQA and 150 MuSiQue. The paired deletion test retains 1729 eligible rows across eleven models. ContextCite uses the first 120 items per corpus and re-applies the eligibility filter. It decodes a new greedy response with the ContextCite library's default prompt template (Table 4). Its 1263 rows have a necessary-bridge fraction of 32.2%. The readout set retains the 1127 rows where ContextCite's own greedy response is correct. Their necessary-bridge fraction is 31.4%. Table 4 traces these populations by model. Table 7 reports discrimination on the readout rows.

Models. The model set includes Qwen2.5 (1.5B, 3B, 7B, 14B, 32B) (Qwen Team, 2024), Gemma-2 (2B, 9B, 27B) (Gemma Team, 2024), and Llama-3.1-8B and Llama-3.2-3B (Grattafiori et al., 2024). It also includes OLMo-2 (7B, 13B) (Team OLMo, 2024) and Pythia-1.4B-deduped (Biderman et al., 2023) as an untuned reference. All are public checkpoints computed in `bfloat16`. The cross-document, DROP and K-source settings use the instruction-tuned release of each model,

and the synthetic setting adds the base releases of OLMo-2-7B, Qwen2.5-7B, Qwen2.5-1.5B and Gemma-2-2B. Generated answers use greedy decoding, and the synthetic setting scores its candidate phrases by conditional log-probability. Qwen2.5-32B and Gemma-2-27B load four-bit NormalFloat (NF4) weights in both cross-document deletion and ContextCite runs. All other runs use unquantised weights. Writer mass is computed in the synthetic, DROP and K-source settings only. Each table reports the subset and model count used in its setting.

Table 3: Experimental settings and their analysis populations. The four settings occupy five rows, because the paired and decorrelated synthetic stimuli are listed separately. Items are corpus examples or matched prompts, and models are checkpoints or matched pairs. The usable-rows column states the reported analysis unit after correctness and availability restrictions. A dash marks populations reported separately by source count, checkpoint or matched pair, with no single pooled usable-row total. The note below the table names the readouts used across the settings.

Setting	Corpus or construction	Items	Models	Usable rows
Cross-document	HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022)	270	11	1729 item–model
DROP	DROP arithmetic (Dua et al., 2019)	328	7	836 item–model
Synthetic (paired)	four controlled notes	160 per pair	4 pairs	640 pair–prompt
Synthetic (decorrelated)	eight or ten controlled notes, 3 formats $\times$ 2 packs	228 per checkpoint	9 single, 4 pairs	–
K-source, $K \in \{2, 3, 4, 5\}$	controlled source sets	25–50 per K	6	–

Readouts. Paired discrete and graded deletion, cross-document ContextCite attribution (Cohen-Wang et al., 2024), writer mass, and passage–question TF-IDF, BM25, MiniLM, MS MARCO and BGE scores.

B Prompts and worked inputs

The rendered prompt is part of each experimental item. Candidate-answer experiments score allowed phrases by conditional log-probability. Benchmark prompts state their answer format. In the cross-document, DROP and K-source settings every model is an instruction-tuned release, and the rendered text is placed inside its chat template. The cross-document text forms the user turn. In the DROP and K-source settings the first line forms the system message, except that Gemma receives it at the start of the user turn followed by a blank line. The following blocks show strings in typewriter text and variable fields in italics. Appendix C gives the rules that populate those fields.

Controlled synthetic setting. Each case supplies a two-word target phrase and a matched candidate set. The prompt pool is fixed before any model is scored. Each prompt stores its case identifier, lexical pack, allowed candidate phrases, source-role labels, and rendered text. The paired stimulus has two lexical packs, and each prompt gives an instruction, a header, an allowed-phrase list, source notes and the query. The prose instruction is `Answer with exactly one two-word phrase from the allowed list.` The records instruction is `Return exactly one two-word phrase from the allowed list.`

The decorrelated stimulus has 3 formats and 2 lexical packs over 38 cases, giving 228 prompts and 76 per format. Format 1 has eight notes and no candidate list. Its prose instruction is `Answer with exactly one two-word phrase.` Its records instruction is `Return exactly one two-word phrase.` The decorrelated worked blocks below show format 1, which supplies every row of Table 6. For the worked case below, format 2 adds `Allowed phrases: quartz passage | amber canyon | cedar harbor | amber harbor`. Its instructions are `Answer with exactly one two-word phrase from the allowed list.` and `Return exactly one two-word phrase from the allowed list.` Format 3 adds two notes to format 1, giving ten notes in all. The added notes route nothing and contain neither answer word. One concerns the queried subject and the other the unqueried subject. The prose pack adds `Note: Orion ledger was logged in the standing review cycle.` and

`Note: Summit archive was logged in the standing review cycle.` The records pack adds `Log | subject=Orion ledger | status=standing_review` and `Log | subject=Summit archive | status=standing_review`. All three formats score four candidate phrases by conditional log-probability.

The paired prompt below asks for `amber harbor`, with the four allowed phrases as scored candidates. Notes 1–4 are, respectively, the first-word source, the second-word source, the linking note and the distractor note.

```
Answer with exactly one two-word phrase from the allowed list.
Archive reconstruction memo. Use the notes below to recover the
full two-word access phrase.
Allowed phrases: quartz passage | amber canyon | cedar harbor |
amber harbor
Note 1: The Velora signal access phrase is amber canyon.
Note 2: The Prairie packet access phrase is cedar harbor.
Note 3: Orion ledger keeps the first word from Velora signal but
takes the second word from Prairie packet.
Note 4: The Summit archive access phrase is quartz passage.
Question: Which two-word phrase belongs to Orion ledger?
Answer:
```

Both members of every base–instruct pair (Section 6, Table 5) and Pythia receive the first line as a prefix followed by a blank line, without a chat template. The eight instruct models in Figure 5 use their chat templates for both stimuli, with the first line as the system message, except that Gemma places it at the start of the user turn followed by a blank line. The header, allowed list and query carry no role. Each source note states a whole two-word phrase that contains its answer word, so the note that holds a word is also the note the linking note points to.

Decorrelated worked input. The next block gives the decorrelated version of this case, shown in Figure 5a. Lexical notes hold individual words, and separate pointer notes link subjects.

```
Answer with exactly one two-word phrase.
Archive reconstruction memo. Use the notes below to recover the
full two-word access phrase.
Note: Summit archive starts with the same word as Kestrel registry.
Note: The Onyx array phrase ends with the word passage.
Note: The Prairie packet phrase ends with the word harbor.
Note: Orion ledger starts with the same word as Velora signal.
Note: The Velora signal phrase begins with the word amber.
Note: Orion ledger ends with the same word as Prairie packet.
Note: The Kestrel registry phrase begins with the word quartz.
Note: Summit archive ends with the same word as Onyx array.
Question: Which two-word phrase belongs to Orion ledger?
Answer:
```

In the decorrelated prose block, the first four role labels are *routes* 1*st word (unused)*, *holds* 2*nd word (unused)*, *holds* 2*nd word*, and *routes* 1*st word*. The remaining labels are *holds* 1*st word*, *routes* 2*nd word*, *holds* 1*st word (unused)*, and *routes* 2*nd word (unused)*. These labels are item metadata paired with the rendered note strings.

The question asks for Orion ledger's phrase, `amber harbor`. Velora signal holds the first word, and Prairie packet holds the second. The pointer notes link Orion ledger to these subjects. The remaining notes form a matched unqueried chain for Summit archive. The displayed format contains no candidate list. Four candidate phrases are scored by conditional log-probability. The first answer word is scored immediately after the prompt. The second is scored after appending the gold tokens for the first word. The candidate score sums both words' log-probabilities. These word-level scores let us measure the effect of deleting each role on the word it supplies. A deletion removes every note that carries one source role and then rescores the correct candidate. The block below renders the same item as records with its own instruction, query wording and note order. Both renderings preserve the same roles and answer.

```
Return exactly one two-word phrase.
Phrase-recovery sheet. Resolve the full two-word access phrase from
the records below.
```

```
Link | subject=Orion ledger | same_last_word_as=Prairie packet
Word | subject=Kestrel registry | first_word=quartz
Link | subject=Orion ledger | same_first_word_as=Velora signal
Word | subject=Onyx array | last_word=passage
Link | subject=Summit archive | same_first_word_as=Kestrel registry
Word | subject=Velora signal | first_word=amber
Link | subject=Summit archive | same_last_word_as=Onyx array
Word | subject=Prairie packet | last_word=harbor
Query:  Output the full two-word phrase for Orion ledger.
Answer:
```

DROP arithmetic setting. The DROP prompt asks for a single numerical answer and preserves the passage's sentence order. The question occupies the first variable field, followed by the original ordered passage sentences.

```
Read the facts and answer the question with a single number.
Question:  [question]
[ordered passage sentences, separated by blank lines]
Answer:
```

The structural selection in Appendix C makes the answer non-extractable, places the two operands in different sentences and requires each compared entity to share a content word with an operand sentence. The prompt retains other passage sentences, including the matched number-bearing control.

Cross-document setting. The open-book template introduces every retrieved document with `Document: ...`. It then gives the exact-answer instruction, the question, and the answer cue. The worked input comes from Figure 1. The block reproduces the figure's shortened question and its bridge and answer snippets, with elisions marked ..., and elides the two distractor documents.

```
Document:  [bridge paragraph, HotpotQA title "Lourenço Mutarelli"]
...  animated sequences of the 2004 motion picture "Nina",
...  based on the novel "Crime and Punishment".
Document:  [answer paragraph, HotpotQA title "Crime and Punishment"]
...  a novel by ...  Fyodor Dostoyevsky.  It was first
published ...  during 1866.
Document:  [first distractor document elided]
Document:  [second distractor document elided]
Answer using only the documents, with the exact short answer span.
Question:  In what year was the novel that Lourenço Mutarelli
based "Nina" on ...  first published?
Answer:
```

The gold answer printed in the figure is 1866. The closed-book prompt presents the same question without documents and with its own instruction:

```
Answer with the exact short answer span.
Question:  In what year was the novel that Lourenço Mutarelli
based "Nina" on ...  first published?
Answer:
```

The cross-document answer scorer determines both the eligibility filter, requiring closed-book failure and open-book success, and the necessity label, recording failure after bridge deletion. It lowercases gold answers and responses, deletes every character outside lowercase ASCII letters, ASCII digits, and literal spaces, then trims leading and trailing spaces. Articles and repeated internal spaces are retained. Deleting a line break joins adjacent tokens, so `Paris,` followed by a line break and `France` becomes `parisfrance`. Matching requires a nonempty normalised gold answer; any occurrence of that answer as a substring of the normalised response passes, including `art` within `cartwheel`. If splitting the normalised gold answer on whitespace yields at least two words, a fallback accepts any substring occurrence of its first two words, joined by one space, in the normalised response. Thus, `John Smith Jr` matches `I think John Smith wrote it`.

K-source setting. The K source sentences and one control sentence are DROP passage sentences, each stating a single-digit yard number as its only digit string, and each enters through the same `Fact: ...` field in a shuffled order. The control is drawn from the same sentence pool as the sources, so it also states a yard number, and the requested sum appears in none of the $K+1$ sentences.

```
You answer with a single integer and nothing else.
Fact: [source or control sentence]
Fact: [remaining source and control sentences, K + 1 in all, in shuffled order]
Question: Each of the K facts states one yard number. Add those K
yard numbers together. What is their sum?
Answer:
```

Each source and the control therefore occupy separate `Fact:` lines of one rendered input. Appendix D specifies how source positions are pooled for the attribution readouts.

C Corpus construction and item selection

Selection separates structural eligibility from model-specific behavioural eligibility. Structural rules determine whether an item contains the required source roles and matched control. Behavioural rules are then applied within each model, so a benchmark item can be retained for one model and excluded for another. Table 3 reports the resulting analysis populations.

Cross-document setting. HotpotQA (Yang et al., 2018) supplies bridge-type validation questions in the distractor setting with exactly two supporting paragraphs, excluding yes/no answers. MuSiQue (Trivedi et al., 2022) supplies answerable validation questions with at least two supporting paragraphs. In both corpora, gold answers have at least three characters. The answer string must occur as a case-insensitive substring in exactly one paragraph of the full benchmark context, and that paragraph must be supporting. The bridge is the first other supporting paragraph, so it does not contain the answer string. The answer passage and bridge each have at most 700 characters. Two non-supporting distractors within the same cap are taken in dataset order, with the first serving as the matched control. A deterministic per-item shuffle keyed by the question identifier orders these four passages. Appendix F reports balance checks for this control. We check whether the normalised answer-document title occurs as a whole-word substring of the normalised question. It occurs in 16 of 270 paired items (10 HotpotQA, 6 MuSiQue) and 9 of 198 readout items. The bridge's own title occurs in 131 of 270 questions.

Shared answer words. Bridges contain the full answer string in 0 of 270 items and can contain individual answer words, such as "music" from "Warner Music Group". Bridges share an answer token in 58 of 198 readout items (260 of 1127 rows) and 76 of 270 paired-test items (410 of 1729 rows). The no-shared-token readout subset has 140 items, 867 rows and 262 necessary bridges. Its AUCs are TF-IDF 0.459 $[0.394, 0.525]$, BM25 0.438, MiniLM 0.547, attribution share 0.736 $[0.683, 0.783]$, and deletion 0.846 $[0.803, 0.886]$. Deletion exceeds share and share exceeds TF-IDF in all ten models, and 87 of 132 multi-model items carry both labels. Across the corresponding 1319 paired rows, bridge deletion breaks 416 answers and control deletion 87, with 363 versus 34 discordances and bridge exceeding control in all eleven models.

Before behavioural filtering, each model receives 300 bridge-document items. The pool contains 150 HotpotQA items and 150 answerable MuSiQue items, the allocation fixed in Appendix H. Across the 11 models in Table 4, these allocations produce 3300 item–model attempts. The behavioural filter retains an attempt only when the model answers the question incorrectly without documents and correctly with the retrieved context.

The scorer lowercases, deletes characters outside `[a-z0-9 ]`, and trims outer spaces, retaining articles and repeated internal spaces. It uses substring matching with a first-two-word fallback for multi-word gold answers (Appendix B). The same scorer determines both sides of the behavioural filter. Attempts that are already correct without documents are excluded, as are attempts that remain incorrect with documents. The filter retains 1729 attempts, a 52% retention rate, as reported in Table 3.

Single-word gold answers. The first-two-word fallback can apply only to multi-word gold answers. Restricting the paired and readout rows to items with single-word gold answers removes any effect of the fallback. The paired subset has 471 rows over 65 items. Bridge deletion breaks 134 answers and control deletion 23, with 120 versus 9 discordances. Bridge deletion exceeds control deletion in all eleven models. Multi-word gold answers account for 1258 rows over 205 items. Their bridge and control deletions break 422 and 103 answers, with 360 versus 41 discordances. Bridge deletion exceeds control deletion in all eleven models in this subset too.

The readout subset with single-word gold answers has 334 rows over 50 items, with 93 necessary bridges. Its AUCs are TF-IDF 0.407 $[0.305, 0.519]$, BM25 0.398 $[0.288, 0.513]$, MiniLM 0.437 $[0.329, 0.548]$, MS MARCO 0.481 $[0.364, 0.595]$, BGE 0.459 $[0.345, 0.577]$, attribution share 0.754 $[0.652, 0.838]$, and deletion 0.855 $[0.776, 0.925]$. Share exceeds TF-IDF and BM25 in all ten models, MiniLM and MS MARCO in nine, and BGE in eight.

The 1729 paired rows cover 270 distinct benchmark items, the count in the *Items* column of Table 3, and the 300-item pool is per model before filtering. Readout availability for ContextCite then defines the readout population reported in Appendix A.

Table 4: Cross-document populations in the model order of Table 7, with Qwen2.5-7B appended. The third column counts rows eligible in the paired deletion test among the first 120 items per corpus, the pool ContextCite uses.

Model	Paired eligible of 300	Eligible in first 120 per corpus	ContextCite rows	Readout rows
Gemma-2-27B	175	143	143	124
Gemma-2-2B	124	100	100	91
Gemma-2-9B	173	138	138	123
Llama-3.1-8B	145	118	118	109
Llama-3.2-3B	128	106	104	89
OLMo-2-13B	176	135	135	122
OLMo-2-7B	175	142	142	121
Qwen2.5-14B	152	124	124	115
Qwen2.5-32B	170	138	138	131
Qwen2.5-3B	149	121	121	102
Qwen2.5-7B	162	129	0	0
Total	1729	1394	1263	1127

Notes. ContextCite re-applies the eligibility filter to the first 120 items per corpus. It decodes its own response with the ContextCite library's default prompt template. Readout rows retain correct ContextCite greedy responses. Qwen2.5-7B's surrogate fit failed with an undefined log-probability target on all 129 eligible items. The re-applied filter reproduces the paired-test eligible rows exactly for ten of the eleven models. For Llama-3.2-3B, the two filters share 102 eligible rows, with bridge labels differing on 4 of them, and 6 rows are eligible under only one of the two.

DROP arithmetic setting. The DROP pool (Dua et al., 2019) begins with comparative numerical questions. A gold answer must not appear verbatim in the passage. Exactly one pair of passage numbers, stated in two different sentences, must give the gold answer as their difference; these two sentences are the operands. Each compared entity in the question must share a content word with at least one operand sentence. Each structurally eligible passage must also contain a third number-bearing sentence.

The matched control is the earliest non-operand sentence that contains a number. The model receives the complete passage in its original order, and operand selection marks the target sentences. Items lacking a unique operand pair or an eligible control leave the pool before model evaluation.

These rules yield 328 structural items, as reported in Table 3. Within each model, the deletion analysis retains only items answered correctly with the full passage. Across 7 models, this filter retains 836 correct item–model rows, also reported in Table 3.

Each retained item supplies two target roles, one for each operand sentence, and one same-passage control. Deleting either operand is compared with deleting that control while the question, answer,

model, and remaining passage stay fixed. Appendix F gives the balance checks and the alternative control rule.

Controlled synthetic setting. Each synthetic case defines a two-word target phrase and a matched set of complete two-word candidates. The pool is fixed before any model is scored, with the stored fields listed in Appendix B. In the paired stimulus, role labels identify notes holding the first or second word of the answer phrase.

Matched base–instruct members receive the same prompts, candidate phrases, roles, deletion targets, and formatting. Both use the shared format without a chat template. The all-prompt comparison keeps 160 prompts for each of 4 pairs and 640 pair–prompt comparisons in total, as shown in Table 5. The comparison's conditional log-probability endpoint scores complete candidates, so both answer words contribute.

The deletion-sensitivity indicator is defined on every prompt: 463 of 763 first-word-role positives and 186 of 241 second-word-role positives accompany incorrect answers. Retaining the 173 pair–prompt comparisons with both members correct holds correctness fixed, removing the tuned member's accuracy advantage, which appears in all 4 pairs (Table 5). Deletion removes every note that carries one source role and rescores the correct candidate.

Decorrelated stimulus. Formats 1 and 2 contain eight notes, and format 3 contains ten (Appendix B). The forced choice scores four two-word candidates. Two pointer notes state that the queried subject's phrase starts or ends with the same word as another named subject's phrase and contain neither answer word. Two lexical notes name single words in those other subjects' phrases. Two crossing chains each link the queried subject through another subject to an answer word. Four unused notes form a chain with matching phrasings for an unqueried subject. Each phrasing occurs twice, so the queried subject selects the relevant chain. Each note names at most one word, so candidate phrases occur verbatim in no note. The eight notes in formats 1 and 2, or ten in format 3, are shuffled per case and pack, keeping roles' mean positions equal.

K-source setting. The source-count construction draws a separate item set for each model and each value of K from one shared sentence pool, with a model-specific seed. Each source sentence carries a single-digit quantity, and one matched control sentence accompanies the sources. The query asks for the sum of these quantities as a single integer. The answer appears in none of the sentences.

For each of 6 models and each $K \in \{2, 3, 4, 5\}$, the construction targets 60 items (Appendix H). Each item uses $K + 1$ distinct single-fact sentences from one finite pool, one per source plus the control, and the construction yields 45–50 items per model at $K = 2$ and 25–26 at $K = 5$ (Table 3). Each value of K has its own denominator. The coverage readout in Figure 2d uses all 852 item–model rows, including incorrectly answered ones (Appendix D).

Across the four settings, structural exclusions precede model scoring, and behavioural exclusions are model-specific. Counts of items describe stimuli fixed before model scoring, and the K-source counts are per model. Counts of item–model or pair–prompt rows describe eligible evaluations after correctness and availability restrictions. These distinct stages explain why corpus pools, deletion rows, attribution rows, and matched-pair subsets have different denominators.

D ATTRIBUTION READOUTS AND THEIR CONVENTIONS

Writer mass, the attention-output attribution readout, projects attention value–output contributions onto an answer-token direction. Its output is signed at each layer, head, and source position. Aggregation choices determine whether negative contributions remain visible and which source roles enter the denominator. This section defines those choices for the convention comparison in Section 7.

Let y denote the next answer token. Its unit unembedding direction is $d_y = W_U[y]/\|W_U[y]\|_2$. The remaining indices identify layer ℓ, query head h, source position s, and answer position t.

Signed source contribution. For query head h in layer ℓ, the signed contribution of source position s to answer token y is

$$c^{(y)}_{\ell hs} = a_{\ell hts} \left(v_{\ell,\kappa(h),s} O_{\ell h}\right)^\top (g_\ell \odot d_y) .$$

Here, $a_{\ell hts}$ is the attention probability from answer position t to source position s. The vector $v_{\ell,\kappa(h),s}$ is the value vector of the key–value head used by query head h. The matrix $O_{\ell h}$ is the output block for that head.

The attention probability weights the selected value vector at source position s. Multiplication by $O_{\ell h}$ maps that value through the head's output block. The final inner product projects the gain-adjusted attention output onto the unit answer direction. The result keeps its sign at this stage.

The vector g_ℓ is the per-dimension gain applied to the attention output before it enters the residual stream. Qwen, Llama and Pythia add the attention output directly to the residual stream, so their g_ℓ is all ones. Gemma-2 and OLMo-2 normalise the attention output before the residual addition, and we fold that norm into g_ℓ by keeping only its gain vector. Evaluating that norm on an all-ones vector gives its weight, or one plus its weight for Gemma-2. This drops the norm's per-token RMS divisor. The divisor is a positive scalar for each layer and answer position, shared by every head and source position. Signs and ratios within a layer are therefore exact. The full norm weights each layer in the sum by the inverse of that divisor. Writer mass from the gain-only fold is therefore an approximation for these two families. Deletion uses a standard forward pass.

Positive-half mass. Writer mass retains the positive part of each signed contribution before summing across layers and heads,

$$m_s^{(y)} = \sum_{\ell,h} \max(c_{\ell hs}^{(y)}, 0).$$

This quantity is non-negative: a positive $c_{\ell hs}^{(y)}$ contributes along the answer direction and enters the sum, and a negative value opposes that direction and contributes zero to $m_s^{(y)}$. The sum is defined separately for each answer target y and each source position.

Role aggregation. Let S_r be the set of source-token positions assigned to role r. Role mass is $M_r^{(y)} = \sum_{s \in S_r} m_s^{(y)}$, and role share is

$$p_r^{(y)} = \frac{M_r^{(y)}}{\sum_{r'} M_{r'}^{(y)}}.$$

This denominator pools the included role masses. The mapping from positions to S_r follows the source-role labels stored with the item. This mapping pools token positions belonging to the same note, operand sentence, source sentence or control sentence before normalisation.

The aggregation order is fixed by these definitions. Signed contributions are first computed for each layer, head, and source position. Positive contributions are then pooled at a position, positions are pooled into roles, and role masses are normalised. Item labels determine the source boundaries used by this pooling.

Normalising the unembedding row makes the readout a projection onto a unit direction, so its scale is invariant to that row's norm. Because the gain differs across dimensions, it is applied elementwise before the inner product. Both features are part of the signed quantity that precedes clipping or role aggregation.

These definitions identify the conventions named in Section 7. The "positive half" is exactly $m_s^{(y)}$, including the maximum with zero before aggregation. The "net projection" retains the sign of $c_{\ell hs}^{(y)}$ when contributions are summed. It can therefore preserve opposing answer-direction contributions that the positive-half definition removes.

"Source-only normalisation" is $p_r^{(y)}$ with both its role set and denominator restricted to the chosen sources. Roles outside that chosen set contribute to neither the numerator nor the denominator. The projection sign and the normalisation domain are separate decisions: the first changes the values being pooled, and the second changes the comparison set. Table 2 applies the projection-sign decision to fixed items, and Figure 3e and the threshold sweep in Appendix E apply source-only normalisation.

For a two-word synthetic answer, the first word is evaluated at the answer position immediately after the prompt. The second word is evaluated after appending the gold tokens for the first word.

Role mass for the complete candidate sums the positive masses for those two answer targets. The conditional candidate score also depends on both words, as Appendix B specifies.

Deletion and writer mass answer different measurement questions on the same labelled sources. Writer mass localises positive answer-direction contribution in a forward pass. Deletion removes a complete source-role bundle and measures the resulting change in the correct candidate or decoded answer. The deletion result is unchanged by projection sign or source-only normalisation.

ContextCite attribution (Cohen-Wang et al., 2024) and TF-IDF query similarity follow Appendix A: attribution uses document granularity in the cross-document setting, and its positive within-item share gives the pooled cross-model scale.

Cross-document readouts are evaluated against the binary bridge-deletion label. Raw-rank space retains ordering, and positive-share space zero-clips non-positive deletion costs before normalisation. A row whose deletion costs are all non-positive has no causal share. Appendix A states which reported populations have every required readout and which have a defined positive share.

Top-ranked bridge retention. Figure 4c counts a bridge as retained when it ranks highest among its item's passages. Rates use 354 necessary-bridge rows and 773 dispensable-bridge rows as separate denominators. Attribution separates the two groups more widely than TF-IDF does, and ranks fewer necessary bridges first (0.32 against 0.55). For Figure 4b, each passage's causal share is its positive deletion drop divided by the row's total positive drop, and non-positive drops contribute zero; the one necessary-bridge row with no positive drop is omitted there, leaving 353.

Source-count coverage. Figure 2d uses all 852 item–model rows, including correct and incorrect answers. The six models are Qwen2.5-14B, Qwen2.5-7B, Gemma-2-9B, OLMo-2-13B, Llama-3.1-8B and Llama-3.2-3B. Positive-half writer mass at the answer's first token is normalised over the source and control sentences. After removing the control, source shares are renormalised to sum to 1. Effective source count is $\exp(\text{entropy})$, and coverage divides this count by K.

E PER-MODEL DETAIL

Table 5: The table reports per-pair indicators for the synthetic setting over 160 matched prompts per pair. Both members use the same format without a chat template. *argmax* gives the proportion on which positive-half writer mass ranks the correct word-holding role highest for both answer words. The *first word* and *second word* columns give the deletion-sensitivity indicators of the first-word and second-word roles. Their deletion-sensitivity indicators require a word log-probability drop of at least one nat after the role's notes are removed. *both* requires both indicators.

	argmax		first word		second word		both	
Pair	base	tuned	base	tuned	base	tuned	base	tuned
OLMo-2-7B	0.819	0.819	0.475	0.619	0.062	0.294	0.044	0.188
Qwen2.5-7B	0.894	0.850	0.494	0.756	0.081	0.406	0.050	0.300
Qwen2.5-1.5B	0.762	0.638	0.325	0.562	0.025	0.044	0.012	0.025
Gemma-2-2B	0.450	0.788	0.662	0.875	0.175	0.419	0.106	0.356

Notes. tuned $>$ base: argmax $1/4$; first word $4/4$; second word $4/4$; both $4/4$. Tuned/base ratio: second word 1.75–$5.00\times$, first word 1.30–$1.73\times$. Accuracy, base $\rightarrow$ tuned: OLMo-2-7B $0.300 \rightarrow 0.381$, Qwen2.5-7B $0.600 \rightarrow 0.669$, Qwen2.5-1.5B $0.381 \rightarrow 0.438$, Gemma-2-2B $0.181 \rightarrow 0.200$. Pythia-1.4B-deduped (untuned, unmatched, accuracy 0.212): argmax 0.588, first word 0.338, second word 0.138, both 0.037.

Deletion-sensitivity and argmax changes. For each pair–prompt comparison, d is the tuned-minus-base change in a deletion-sensitivity indicator minus the tuned-minus-base change in the argmax indicator. We compute d for the first-word and second-word deletion-sensitivity indicators and for the indicator requiring both. All indicators are binary on the same prompt, so mean d is a difference of rate changes. The case-permutation test flips d by swapping base and tuned jointly across every prompt and pair in each case over 10^6 draws. Across all 640 comparisons from 40 cases, the argmax rate rises by 0.042. First-word sensitivity rises by 0.214, exceeding the argmax



Figure 5: Panels (a), (b), (e) and (f) use the decorrelated stimulus: each word needs a holding note and a pointer note, with four *unused* matched controls for an unqueried subject. Panels (c) and (d) use the paired stimulus. Costs are word log-probability drops in nats. Pythia-1.4B denotes Pythia-1.4B-deduped. **(a)** 70 correct items. **(b)** Costs track held words. Filled marks denote sign-test $p < 0.05$. **(c)** First-word and second-word deletion-sensitivity rates rise after tuning in 4 of 4 pairs (95% intervals), 160 prompts per pair (640 pair–prompt comparisons, 173 with both members correct). **(d)** Splitting information across two notes reduces accuracy relative to one (95% intervals). **(e)** The larger word-level deletion cost switches with note role. **(f)** Control-dependent ROC-AUC (95% intervals).

change by 0.172 ($p < 10^{-5}$). Second-word sensitivity rises by 0.205, exceeding the argmax change by 0.163 ($p < 10^{-5}$). The rate for both roles rises by 0.164, exceeding the argmax change by 0.122 ($p = 0.00013$). Among the 173 comparisons that both members answer correctly, drawn from 39 cases, the argmax rate changes by -0.052. First-word sensitivity rises by 0.156, exceeding the argmax change by 0.208 ($p = 0.0057$). Second-word sensitivity rises by 0.098, exceeding the argmax change by 0.150 ($p = 0.0079$). The rate for both roles rises by 0.075, exceeding the argmax change by 0.127 ($p = 0.014$). Across all comparisons within each of OLMo-2-7B, Qwen2.5-7B and Qwen2.5-1.5B, argmax changes range from -0.125 to 0.000 while both deletion-sensitivity rates rise. For Gemma-2-2B, the argmax rate rises by 0.338, more than first-word (0.212) or second-word sensitivity (0.244).

Matched-accuracy pair detail. Across four pairs, the pooled matched population contains 173 both-correct pair–prompt comparisons, with different counts among pairs. For Qwen2.5-7B, the pair-level test on the matched both-correct population reaches significance ($p = 2.4 \times 10^{-4}$). Gemma-2-2B contributes $n = 9$ both-correct prompts, reverses the other pairs' tuned-above-base direction, and shows no detectable base-to-tuned change (exact $p = 1$).

Decorrelated pair comparison. The paired stimulus combines holding an answer word and pointing at it in the same note. In the decorrelated stimulus, each answer word needs a holding note and a pointer note containing neither answer word. Table 6 repeats the post-training comparison in Section 6 on 4 matched pairs using prompts both members answer correctly. The control-adjusted pointer interaction increases by $+4.06$ nats in OLMo-2-7B, $+5.72$ nats in Qwen2.5-7B and $+1.56$ nats in Qwen2.5-1.5B. Gemma-2-2B gives -1.34 nats, with a 95% interval spanning zero.

The decorrelated stimulus measures a pointer-specific interaction for each prompt. The first-pointer term is the first-word log-probability drop minus the second-word drop when the first-word pointer note is deleted. The second-pointer term is the same difference when the second-word pointer note is deleted. The interaction subtracts the second-pointer term from the first-pointer term. A positive interaction means that deleting a pointer note, which contains neither answer word, costs the word it points to more than the other word, summed over the two pointer notes.

The control is the same contrast on a structurally identical four-note chain about an unqueried subject. We measure tuned-minus-base changes, matching prompts on which both checkpoints answer correctly. The difference-in-differences, the pointer-interaction change minus the control change, is positive in three of four pairs (Table 6).

Table 6: The table reports pointer interactions on format 1 of the decorrelated stimulus (eight notes, no candidate list) over the n prompts both pair members answer correctly. The interaction contrasts first- and second-word log-probability drops across deletions of the corresponding pointer notes, in nats. The control column gives the change in the same contrast on the matched unqueried chain. Brackets give 95% intervals from 10^4 bootstrap resamples of whole cases.

		pointer interaction		change (tuned − base)	
Pair	n	base	tuned	queried chain	unqueried control
OLMo-2-7B	31	$+1.83$	$+5.98$	$+4.15$ $[+2.48, +5.64]$	$+0.09$ $[-1.03, +1.29]$
Qwen2.5-7B	49	$+1.17$	$+6.56$	$+5.39$ $[+4.00, +6.87]$	-0.33 $[-1.16, +0.50]$
Qwen2.5-1.5B	35	$+1.25$	$+2.85$	$+1.60$ $[+0.66, +2.52]$	$+0.04$ $[-0.49, +0.60]$
Gemma-2-2B	15	$+1.75$	$+1.19$	-0.57 $[-1.59, +0.43]$	$+0.77$ $[-0.50, +2.27]$

Notes. OLMo-2-7B gives a difference-in-differences of $+4.06$ nats $[+2.02, +5.86]$ ($p = 0.002$). Qwen2.5-7B gives $+5.72$ nats $[+4.08, +7.41]$ ($p < 10^{-4}$, with no permuted draw reaching the observed value in 10^5 draws). Qwen2.5-1.5B gives $+1.56$ nats $[+0.42, +2.65]$ ($p = 0.015$). Gemma-2-2B gives -1.34 nats $[-3.43, +0.48]$ ($p = 0.240$). Across 4 pairs and 3 note formats, 8 of 12 differences-in-differences are positive with $p < 0.05$: OLMo-2-7B $3/3$, Qwen2.5-7B $2/3$, Qwen2.5-1.5B $2/3$, Gemma-2-2B $1/3$. Both Qwen pairs miss this criterion on the format with the candidate list in the prompt. Permutation p-values use 10^5 case-level sign-flip draws, with base and tuned labels swapped jointly across all prompts of each case. Both procedures use seed 172.

Paired design. Shared initialisation and input format let the base-to-tuned contrast isolate the effect of post-training (Table 1). Across the ten models of Section 5, family and base accuracy vary together.

Source-only threshold sweep. Each operand's positive-half writer mass is divided by the sum over both operands. A row meets the share threshold when both resulting shares are at least t. Equivalently, the larger share is at most $1 - t$, since the shares sum to 1. This condition selects rows with writer mass spread across both operands. The fraction meeting the threshold is 0.86 at $t = 0.15$, rising to 0.91 at 0.10 and 0.98 at 0.05. At $t = 0.15$, OLMo-2-13B and OLMo-2-7B have the highest rates, 0.983 and 0.979. The sweep uses 27 thresholds from 0.02 to 0.50. In this sample, every row meets the threshold at 0.02, and none does at 0.50. Rank correlations with accuracy are undefined at these endpoints. The 25 interior cuts give correlations between -0.61 and 0.00, with exact permutation $p \geq 0.15$.

Discrimination by model. Table 7 gives, for each model, the necessary-bridge fraction and the readout discrimination drawn in Figure 4a and d, with raw attribution and the within-item rank agreement added.

Table 7: Bridge necessity and readout discrimination by model. Eligible rows are answered correctly with context and incorrectly closed-book, and the readout set keeps those on which ContextCite's own greedy response is also correct (Appendix A). n counts rows with all readouts, and "necessary" gives their necessary-bridge fraction. Attribution is ContextCite's raw score or positive share. Deletion is the bridge's gold-answer log-probability drop. The last column averages within-item Spearman correlations of raw attribution and deletion on necessary-bridge rows.

			ROC-AUC				mean within-item Spearman ρ
Model	n	necessary	TF-IDF	attr. raw	attr. share	deletion	
Gemma-2-27B	124	0.177	0.437	0.748	0.758	0.849	0.715
Gemma-2-2B	91	0.286	0.507	0.670	0.735	0.884	0.617
Gemma-2-9B	123	0.260	0.410	0.752	0.774	0.922	0.700
Llama-3.1-8B	109	0.450	0.487	0.763	0.706	0.757	0.530
Llama-3.2-3B	89	0.483	0.515	0.749	0.702	0.797	0.681
OLMo-2-13B	122	0.131	0.455	0.629	0.689	0.903	0.647
OLMo-2-7B	121	0.174	0.436	0.554	0.590	0.798	0.778
Qwen2.5-14B	115	0.539	0.490	0.801	0.809	0.887	0.662
Qwen2.5-32B	131	0.405	0.425	0.721	0.703	0.784	0.609
Qwen2.5-3B	102	0.294	0.337	0.723	0.685	0.751	0.569
pooled	1127	0.314	0.453	0.717	0.728	0.825	0.640

Notes. Within an item, dividing by a positive item total preserves the order of raw scores and zero-clipping ties the non-positive ones; across items the division rescales rows, so the raw and share AUC columns differ.

Relevance scores. Table 8 compares seven readouts on 1127 rows with 354 necessary bridges from ten models. The 198 items comprise 106 HotpotQA and 92 MuSiQue questions. BM25 uses $k_1 = 1.2$ and $b = 0.75$, with IDF and average length over the 792 readout passages (Robertson & Zaragoza, 2009). The dense bi-encoder uses mean-pooled `all-MiniLM-L6-v2` embeddings and passage–question cosine (Reimers & Gurevych, 2019; Wang et al., 2020). Cross-encoder rerankers `cross-encoder/ms-marco-MiniLM-L6-v2` (Nogueira & Cho, 2019; Bajaj et al., 2016) and `BAAI/bge-reranker-v2-m3` (Chen et al., 2024) score each question–bridge pair by their relevance logit with input truncated at 512 tokens. Intervals use 10^4 whole-item bootstrap draws, retaining each item's model rows.

Attribution share exceeds each relevance score in all ten models (two-sided sign test, $p = 0.002$). BM25 is below 0.5 in all ten, MiniLM is above 0.5 in eight of ten ($p = 0.109$), and each reranker is above 0.5 in five of ten. Across $k_1 \in \{0.9, 1.2, 1.5, 2.0\}$ and $b \in \{0.4, 0.75, 1.0\}$, BM25 AUC lies between 0.422 and 0.445 with readout-wide IDF and between 0.455 and 0.471 with IDF over each item's four passages. Upper bounds are below 0.5 in eleven of twelve readout-wide settings, and all item-IDF intervals span 0.5. Bridge length alone gives AUC 0.437.

Strict label. Dropping the 43 necessary readout rows whose control deletion also breaks the answer leaves 1084 rows over 195 items, with 311 necessary bridges. The control outcomes come from the paired deletion test. One Llama-3.2-3B readout row has no paired row and is kept. The AUCs are TF-

Table 8: ROC-AUC against bridge necessity on the readout set. Model order and the TF-IDF, attribution-share and deletion columns match Table 7. Share is ContextCite positive attribution share, and deletion is the bridge's gold-answer log-probability drop. Brackets in the notes give whole-item bootstrap 95% intervals.

Model	TF-IDF	BM25	MiniLM	MS MARCO	BGE	share	deletion
Gemma-2-27B	0.437	0.430	0.549	0.565	0.475	0.758	0.849
Gemma-2-2B	0.507	0.456	0.549	0.531	0.507	0.735	0.884
Gemma-2-9B	0.410	0.388	0.487	0.408	0.420	0.774	0.922
Llama-3.1-8B	0.487	0.368	0.516	0.465	0.437	0.706	0.757
Llama-3.2-3B	0.515	0.476	0.556	0.543	0.552	0.702	0.797
OLMo-2-13B	0.455	0.394	0.380	0.416	0.370	0.689	0.903
OLMo-2-7B	0.436	0.408	0.580	0.549	0.521	0.590	0.798
Qwen2.5-14B	0.490	0.428	0.520	0.491	0.502	0.809	0.887
Qwen2.5-32B	0.425	0.437	0.563	0.541	0.552	0.703	0.784
Qwen2.5-3B	0.337	0.356	0.525	0.494	0.481	0.685	0.751
pooled	0.453	0.431	0.535	0.516	0.497	0.728	0.825

Notes. Pooled intervals: TF-IDF $[0.397, 0.512]$, BM25 $[0.376, 0.487]$, MiniLM $[0.472, 0.595]$, MS MARCO $[0.457, 0.574]$, BGE $[0.442, 0.554]$, share $[0.683, 0.771]$, deletion $[0.785, 0.863]$. Attribution share minus relevance-score AUC: TF-IDF $+0.275$ $[+0.199, +0.347]$, BM25 $+0.297$ $[+0.226, +0.366]$, MiniLM $+0.194$ $[+0.128, +0.260]$, MS MARCO $+0.212$ $[+0.147, +0.278]$, BGE $+0.231$ $[+0.164, +0.295]$. By corpus in column order: HotpotQA (651 rows), 0.537, 0.447, 0.549, 0.548, 0.533, 0.719, 0.784. MuSiQue (476 rows), 0.416, 0.496, 0.588, 0.602, 0.567, 0.731, 0.860.

IDF 0.458 $[0.398, 0.518]$, BM25 0.440 $[0.385, 0.496]$, MiniLM 0.543 $[0.479, 0.605]$, MS MARCO 0.527 $[0.467, 0.587]$, BGE 0.506 $[0.447, 0.563]$, attribution share 0.733 $[0.687, 0.776]$, and deletion 0.821 $[0.779, 0.863]$. Share exceeds every relevance score in all ten models. Deletion exceeds share in all ten models.

Paired results by corpus. HotpotQA has 983 rows, with 282 bridge-deletion breaks and 49 control-deletion breaks (254 versus 21 discordances). MuSiQue has 746 rows, with 274 versus 77 breaks (226 versus 29 discordances). Pooled counts are 556 versus 126 over 1729 rows, with 480 versus 50 discordances. Bridge deletion breaks more answers in all eleven models within each corpus, with one-sided exact McNemar $p < 0.05$ in ten of eleven.

Label transfer. On the 1127 readout rows (198 items, ten models), labels disagree in 1968 of 6924 shared-item comparisons across ordered pairs of distinct models (0.284), compared with zero for a fixed item annotation and 0.436 for independent draws at each model's own rate. Across all 1729 paired rows, 185 of the 246 items evaluated under at least two models carry both labels. Both deletions break the answer on 76 of the 1729 rows. Dropping those rows leaves 173 of 239 items with both labels.

F CONTROL ROBUSTNESS

Table 9 reports same-item control checks for the robustness results in Sections 3 and 4 and the question-overlap comparison in Section 5. For the related attribution conventions, Table 2 reports projection-sign comparisons on fixed items (Lu & Yang, 2026), and Appendix E reports the source-only threshold sweep.

Cross-document. The distractor control follows the selection rule in Appendix C. The control is longer than the bridge on average, and restricting the comparison to the 994 rows whose control is at least as long as the bridge increases the graded contrast from $+1.59$ to $+1.76$ nats (Table 9). The bridge and control break-rate ranges are disjoint, and bridge deletion breaks the answer more often in every model (Table 9). The paired discrete test agrees in direction in 11 of 11 models. The graded gap is positive on 54.0% of rows, with median $+0.098$ nats. The Spearman correlation between the length gap and the graded gap is near zero ($\rho = +0.018$).

Discrete and graded cross-document tests. For both OLMo-2 checkpoints, the graded cross-document intervals include zero, while their discrete tests detect a difference (Table 9). A few large rows pull the graded mean far above the graded median (Figure 2e). The discrete test, which counts paired answer flips and gives every row equal weight, is the primary test throughout (Section 3).

DROP. The number-bearing non-operand control follows the selection rule in Appendix C. The full-sample gaps remain positive for both operand roles (Table 9). The overlap-maximised rule deliberately selects the control most similar to the question, changing most controls and substantially raising question overlap. Both operand gaps stay above zero in every model. On the overlap-favouring subset, the contrast shrinks. Subset composition accounts for 82% and 89% of the shrinkage for the two operand roles, while the control-swap contribution has an interval spanning zero (Table 9).

G Inferential unit and replication

The primary inferential unit for a benchmark setting is a retained item within a model. Each model row therefore reports its own usable-item count and effect. Confidence intervals resample retained items within that row. Appendix A specifies the interval form and resample count. The bootstrap recomputes the statistic on each resample (Efron, 1979), so the interval describes variation over retained items within one model row, with the model axis held fixed.

The paired design remains intact during inference. Each target source and its matched control come from the same item, preserving their pairing. The discrete deletion analysis uses discordant target–control outcomes from the same item. Its exact McNemar procedure (McNemar, 1947) is also defined in Appendix A.

For graded paired tests, a sign flip acts on the paired target–control difference, leaving the shared question, answer, model, and remaining context unchanged.

The reported synthetic base–instruct sign-flip permutation uses the case as its unit because each case's four prompts render one set of subjects and one answer phrase in two prompt families and two lexical packs of the paired stimulus. Both members of each base–instruct pair receive the same prompt, roles, candidates, deletion targets and formatting. All four pairs receive the same 160 prompts, so each case spans 16 of the 640 rows. Over 10^6 draws, base and tuned labels are randomly swapped independently by case, jointly across its prompts and pairs. Prompt-level exact McNemar tests count the discordant base–instruct indicator outcomes on a prompt, while concordant prompts contribute no direction. The eligibility-matched analysis retains only prompts answered correctly by both pair members.

A benchmark item evaluated under several models yields repeated item–model measurements. The pooled row summarises eligible evaluations across the available model-specific rows, and its intervals resample whole items with all their model rows. Cross-model generality is a replication criterion: it requires the same effect direction in every per-model row, including rows whose estimate is weak and rows whose test does not reach significance. Per-model tests use nominal $p < 0.05$ without multiplicity correction, one-sided where stated. Pooled ROC-AUC intervals use 10^4 resamples of whole items with all their model rows (1127 rows across 198 items). The 95% ROC-AUC intervals are $[0.397, 0.512]$ for query similarity, $[0.683, 0.771]$ for attribution share and $[0.785, 0.863]$ for graded deletion. The pooled necessary-bridge fraction 0.314 has a row-level Wilson interval of $[0.288, 0.342]$ and a whole-item bootstrap interval of $[0.270, 0.361]$. The across-model AUC of Section 5 is computed on the 120 items where some models need the bridge and others do not. Within each item it compares every model that needs the bridge with every model that does not, 984 pairs in all, and counts a pair as 1 when the readout scores the needing model higher and $1/2$ on a tie. Its intervals resample these items 10^4 times with all their pairs. The model-rate baseline scores a row by its model's necessity rate over that model's other rows, so it uses the labels and ignores the passage. The two within-model variants of attribution share subtract each model's mean attribution share over its rows, or replace each share by its percentile rank among that model's rows; neither uses the labels.

The denominator attached to a model row follows the setting's declared eligibility rule. Cross-document rows pass the behavioural filter: the model answers incorrectly closed-book and correctly with the full context. DROP deletion rows have a correct full-context answer. Readout availability

can produce a smaller population than deletion alone. Appendix C traces these filters before the inferential unit is formed.

H REPRODUCIBILITY

The reproducibility package fixes 48 ablations per item for attribution and 150 requested items per model per corpus for the cross-document setting, with HotpotQA candidates drawn from the first 8000 validation examples. The K-source setting uses 60 requested items per K, and the synthetic setting uses 160 prompts per base–instruct pair in the paired stimulus and 228 decorrelated prompts per checkpoint (3 note formats $\times$ 2 lexical packs $\times$ 38 cases). The package also fixes greedy decoding and `bfloat16` compute. Bootstrap intervals use 10^4 resamples, except the intervals drawn in Figures 2c and 3d (5000) and Figure 5f (2000). Qwen2.5-32B and Gemma-2-27B load four-bit NormalFloat (NF4) weights in their cross-document deletion and ContextCite runs. All other runs use unquantised weights. The experiments use only public checkpoints and public benchmarks, so researchers with inference access can recompute the label; the benchmarks contain no personal data. The code, prompts, run settings and stored per-item outputs for these runs are in the anonymous repository at `https://anonymous.4open.science/r/answerless-bridge-8AE6`, which gives the command for each run.

I EXTENDED RELATED WORK

Post-training and context reliance. Knowledge-conflict evaluations score whether supplied evidence overrides parametric knowledge when the sources suggest different answers (Longpre et al., 2021; Chen et al., 2022). During instruction tuning, Goyal et al. (2025) find that reliance on conflicting context initially increases and then declines as benchmark performance continues improving. With chat templates applied, Wang et al. (2025) associate reduced context awareness after instruction tuning with learned biases towards conversational roles. Concurrent work by Pandere et al. (2026) finds greater rejection of terse counterfactual context after tuning, with this effect disappearing for coherent evidential passages. Coherent external evidence can persuade models under conflict, while evidence aligned with parametric memory can elicit confirmation bias (Xie et al., 2024). This literature tracks both the source a model follows and the presentation conditions under which that choice changes. Prompting and decoding methods increase adherence to supplied evidence under conflict (Zhou et al., 2023; Shi et al., 2024). Du et al. (2024) quantify contextual persuasion and entity susceptibility, while Minder et al. (2025) identify a subspace controlling the choice between context and prior knowledge. Mechanistic comparisons identify largely shared entity-tracking circuits before and after fine-tuning, and learned wrappers over existing capabilities (Prakash et al., 2024; Jain et al., 2024). Across base and aligned models, Lin et al. (2024) find that changes in token distributions concentrate on stylistic tokens. Within conflict-resolution circuits, Pandere et al. (2026) also find that tuning reweights existing heads. Our bridge supplies consistent connective information without containing the answer, and its deletion cost measures how a correct answer depends on that connection, complementing evaluations of adherence to conflicting evidence. With identical prompts and format and no chat template, our paired base and instruct comparisons show that post-training raises deletion sensitivity more than the argmax indicator of crediting the correct source role (Section 6).

Multi-hop evidence and composition. Multi-hop benchmarks measure whether models connect evidence distributed across documents (Welbl et al., 2018). Self-Ask decomposes a question into follow-up questions whose answers support later steps (Press et al., 2023). Dataset analyses show that compositional questions can admit single-hop solutions through entity types, redundant evidence, sentence-level cues or spurious question-answer correlations (Min et al., 2019; Chen & Durrett, 2019). Adversarial evaluation tests word-matching shortcuts by adding documents that contradict shortcut answers while preserving the original answer (Jiang & Bansal, 2019). Trivedi et al. (2020) measure disconnected reasoning over subsets of supporting facts, and Tang et al. (2021) evaluate whether successful multi-hop predictions accompany correct answers to their sub-questions. The contrastive support sufficiency criterion in DiRe also guides transformations that reduce disconnected reasoning in existing datasets (Trivedi et al., 2020). These tests locate evidence requirements in the structure of questions and the behaviour of evaluated systems. At the model level, studies of latent reasoning probe recall of intermediate entities and their use in composing facts, including evaluations

designed to remove co-occurrence shortcuts (Yang et al., 2024; 2025). Tracing bridge entities through model layers, Biran et al. (2024) use back-patching to examine later-hop computation. Studies of plausible distractor chains and synthetic in-context reasoning also evaluate how competing paths affect composition (Bhuiya et al., 2024; Guo et al., 2025). Our evaluation records, for each checkpoint on fixed inputs, whether its correct answer depends on the answerless bridge that connects the supporting documents. Deleting that bridge and a matched same-item control makes this dependence observable at the passage level, with the question, answer and checkpoint held fixed.

Passage utility, distractors and compression. Passage utility and compression studies measure changes in answer accuracy as retrieved context is added, selected or shortened. Adding irrelevant information lowers arithmetic accuracy in GSM-IC, while random retrieved documents can improve accuracy in some RAG settings (Shi et al., 2023; Cuconasu et al., 2024). Yoran et al. (2024) analyse accuracy reductions after retrieval and train with mixed relevant and irrelevant contexts. Compression methods evaluate concise contexts through downstream performance and computational cost (Xu et al., 2024; Jiang et al., 2024). The extractive and abstractive compressors in RECOMP learn which information helps downstream tasks, including when to provide an empty augmentation. LongLLMLingua targets the density and position of key information in long prompts. Context position separately affects whether a language model uses retrieved evidence, even when that evidence is present (Liu et al., 2024). Zhang et al. (2024) evaluate LLM utility judgements and the resulting evidence selection for question answering. On single-hop question answering, Zhang et al. (2025) define utility as a generator's performance gain from adding a passage over answering without evidence, and find that useful evidence depends on the generator. In multi-hop retrieval, Jain & Garimella (2025) model utility through inter-passage dependencies, using reasoning traces to learn scores that improve reranking and downstream answers. Concurrent work finds that compression can discard answerless bridges (Hu et al., 2026). Hard prompt compressors score units independently and can retain one member of a dependent evidence pair while dropping the other, termed referential dangling (Hu et al., 2026). In their evaluation, all six hard compressors show dangling on a shared HotpotQA bridge set, and Hu et al. (2026) train a classifier to rank omitted sentences for reinsertion. Concurrent analysis of per-chunk verification shows that filtering can remove evidence needed across hops, and that conditioning verification on decomposed sub-questions improves evidence selection (Kumar, 2026). We measure necessity through deletion from correctly answered full contexts, using matched same-item controls to score the loss of answerless connective passages. Comparisons across checkpoints, including base and instruct pairs sharing a pretrained initialisation, measure how this dependence on surrounding evidence varies with the generator.

Attribution, citation and explanation faithfulness. The Attributable to Identified Sources framework and Attributed QA evaluate source support for generated text using human judgements (Rashkin et al., 2023; Bohnet et al., 2022). Citation benchmarks additionally score fluency, answer correctness, citation quality and coverage of generated claims (Gao et al., 2023; Liu et al., 2023). Faithfulness concerns whether attributed sources participate in the model's actual response production (Wallat et al., 2025) and whether an explanation reflects the model's reasoning process (Jacovi & Goldberg, 2020). Attention weights can diverge from faithful explanations (Jain & Wallace, 2019). A cited passage can support a generated statement even when the model obtains the answer from its stored knowledge, making source support and contribution to generation separate evaluation targets. Removal-based explanations quantify feature influence by simulating absence, while Integrated Gradients assigns feature contributions using sensitivity and implementation-invariance axioms (Covert et al., 2021; Sundararajan et al., 2017). SelfCite uses necessity and sufficiency under context ablation to reward citations (Chuang et al., 2025). MIRAGE links context-sensitive answer tokens to documents through saliency, while ARC-JSD uses divergence to identify essential context sentences (Qi et al., 2024; Li et al., 2026). Benchmarks of feature importance and studies of explanation disagreement supply further tests for comparing attribution readouts (Hooker et al., 2019; Krishna et al., 2024). Concurrent work examines attribution when contextual facts overlap with knowledge in model weights (Trinh et al., 2026). Its provenance-labelled evaluation reports that contribution scores mix information supplied in context with information already stored in weights. Majmudar & Filatova (2026), also concurrent, remove structurally load-bearing examples from linguistic puzzles and find that frontier models often continue answering correctly. We use deletion as the behavioural reference and compare query similarity, attribution share and graded deletion against discrete answer flips. Indexing the label by checkpoint records whether each model needs the removed passage, while the matched same-item control measures the cost of removing comparable context.

Properties relative to a model. Model-relative descriptions of datasets measure properties through the behaviour or capacities of a specified learner. Dataset cartography characterises individual examples by confidence in the true class and variation in that confidence across training (Swayamdipta et al., 2020). Predictive $\mathcal{V}$-information makes informativeness relative to an observer's modelling capacity and computational constraints (Xu et al., 2020). Ethayarajh et al. (2022) use usable information to characterise dataset difficulty for a model and introduce pointwise information to describe individual instances. These formulations support comparisons of examples or datasets for a fixed model and descriptions that change with the learner. The observer's capacity therefore enters the description of the data, alongside the examples and labels that constitute the dataset. Our checkpoint-indexed label extends this perspective to passage necessity by asking whether deleting a passage changes an already correct answer. A shared input can consequently receive different necessity labels across checkpoints, and the same-item control separates each model's need for a passage from that model's cost of deleting text.

Table 9: Control checks cover both natural-corpus settings. The Setting and Check columns identify the corpus setting and diagnostic, and the Rows column gives the analysis population for that check, in item–model rows except that the DROP length, question-overlap, position and candidate checks count each of the 328 DROP items once. In the cross-document setting, Target is the answerless bridge and Control is the first distractor in dataset order, a non-supporting paragraph drawn from the same item. On DROP, Target designates both operand sentences and Control designates a number-bearing non-operand sentence from the same passage. Checks comparing the sides give a value for each under Target and Control. When a check yields a comparison-wide derived value, that check's entry spans both columns. A graded gap subtracts the control-deletion cost from the target-deletion cost and is measured in nats. Bracketed ranges are bootstrap intervals. The Discrete breaks check counts the answers that deletion breaks. Break rate is the share of the setting's rows on which deleting that side breaks the correct answer; the row gives the range across the 11 models. Discordant counts come from the paired discrete test and count items on which only that side's deletion breaks the answer; the row again gives the range across the 11 models. The eligible-control-candidates row gives the mean number of sentences satisfying the DROP control rule. The control-question-overlap row gives the control's question overlap before and after applying the overlap-maximised rule, shown as an arrow. The overlap-favouring subset keeps the 240 of 836 rows on which the overlap-maximised control shares at least as many question content words as the two operand sentences do on average; its row gives the larger-number and then the smaller-number operand gap against that control.

Setting	Check	Rows	Target	Control
Cross-document	Length (words)	1729	59.35	65.82
	Question overlap (content words)	1729	4.16	3.02
	Prompt position	1729	difference 0.007	
	Length difference, bridge minus control	1729	[−8.07, −4.84]	
	Graded gap, all rows	1729	+1.59 [+1.33, +1.86]	
	Graded gap, control at least as long as bridge	994	+1.76 [+1.40, +2.16]	
	Discrete breaks	1729	556	126
	Break rate, range across the 11 models	1729	0.131–0.547	0.029–0.121
	Discordant counts, range across the 11 models	1729	18–73	0–8
DROP	Length (characters)	328	142.05	134.91
	Question overlap (content words)	328	2.24	0.73
	Passage position	328	0.47	0.18
	Eligible control candidates per item	328	4.07	
	Control question overlap, overlap-maximised rule	836	0.687 → 1.293	
	Graded gap, larger- and smaller-number operands	836	+6.01 and +6.16	
	Graded gap change, overlap-maximised control, both roles	836	−0.080 [−0.268, +0.103]	
	Graded gap, overlap-favouring subset	240	+4.60 [+3.49, +5.76] and +3.82 [+2.80, +4.84]	

Notes. For all 1729 cross-document rows, the one-sided exact McNemar test gives $p \leq 0.006$ in every model. The graded gap is positive on 54.0% of rows. The graded statistic has median +0.098 nats and mean +1.586, with the top 5% of rows supplying 65.8% of the total. For both OLMo-2 checkpoints, the discrete test detects a difference at $p \approx 0.005$ while the graded interval includes zero. Under the DROP overlap-maximised rule across seven models, the control changes on 61% of items. The intervals for both operands stay clear of zero in all seven models. On the overlap-favouring subset, subset composition accounts for 82% and 89% of the shrinkage for the larger-number and smaller-number operand roles, respectively, while the control swap changes each gap by −0.25 nats [−0.79, +0.28].