SOURCES COUNT, COPIES WIN: A GAP IN LLM EVIDENCE AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents search documents and exchange messages to gather evidence, yet duplicated study reports and recirculated messages can turn one observation into apparent consensus. Human experiments show that repetition can increase perceived truth even when people demonstrate knowledge of the correct facts in a separate test. We investigate whether LLM decisions remain vulnerable when shared provenance is explicit and source-counting ability is assessed separately. Our *Explicit Evidence Groups* task varies redundant records while holding independent support and the correct answer fixed. Across seven different LLMs, increasing a minority source from one to sixteen records reduces decision accuracy by 17.2–40.6 percentage points, even without repeated answer labels. Linear probes recover source-count and repetition information before a decision, and separate numerical reports often preserve the correct source ordering despite counting errors. Head selection and causal ablations reveal partially distinct attention-head contributions to source reporting and direct decisions. Reporting source counts before choosing improves decisions under repetition when the task remains within the models' counting capacity. Our findings show that explicit provenance alone cannot ensure reliable evidence aggregation: information about independent sources can be available without reliably constraining decisions. For scientific synthesis and multi-agent deliberation, reliability requires preserving the distinction between evidence that gains independent support and evidence that merely gains circulation.

1 INTRODUCTION

Studies of human judgment show that repetition can increase apparent support without adding independent evidence. A single person repeating an opinion can make that opinion seem more widely shared, even among participants who correctly recall that it came from one speaker (Weaver et al., 2007). Repeating false statements can also make them seem more truthful, even when participants demonstrate in a separate test that they know the facts contradicting those statements (Fazio et al., 2015). These findings suggest that knowledge of the relevant facts or sources need not prevent repetition from biasing judgment.

An analogous challenge arises for retrieval-augmented and long-context language models, which often receive several textual views of the same underlying event. A wire report may be syndicated by many outlets; one document may be split into overlapping passages; an agent memory can be rewritten into several summaries; and multiple agents can repeat a common upstream source. In all these cases, the number of texts and the number of independent observations underlying those texts differ. Reliable aggregation therefore depends on distinguishing repeated records from independent observations.

Previous studies show that repetition can reinforce in-context associations (Yan et al., 2024) and shift answers to retrieved evidence (Naphade, 2026; Jain & Vedam, 2026). Such shifts can persist when repeated material is attributed to the same source and models are instructed to disregard repetition (Schuster et al., 2026). We ask whether repetition can still reverse a model's choice when independent observations and the decision rule are explicitly specified, with source-counting and source-comparison abilities assessed separately.

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.

We isolate evidence aggregation from uncertainty about source provenance using an *Explicit Evidence Groups* task. Each source represents one independent observation, and its records are explicitly identified as redundant accounts. The model must choose the option supported by more independent sources. For example, one source repeated sixteen times still provides less independent support than two distinct sources. By varying source counts and repetition separately, we test whether decisions follow independent evidence when the number of records favors the other option.

Across seven different LLMs drawn from several model families, increasing a minority source from one to sixteen records reduces decision accuracy by 17.2–40.6 percentage points. The effect persists without repeated answer labels and, in most models, exceeds neutral-extension and balanced-repetition controls. All seven nevertheless respond to genuine increases in independent support in simpler comparisons. This coexistence of basic source sensitivity and repetition errors motivates examining how source information enters the decision.

Linear probes recover source-count and repetition information from hidden states before a choice, and models can access this information through explicit numerical reports. Causal head ablations reveal that source reporting and direct decisions depend on partly different attention heads. These findings distinguish the availability of source information from its use in choosing. They motivate a training-free intervention that explicitly requests source counts before the final decision. This improves decisions within the models' counting capacity, while heavier repetition exposes limits as count recovery becomes less reliable.

Our contributions are:

1. We introduce a controlled diagnostic that separates independent support from redundant presentation, linking repetition errors to separate tests of source counting and basic rule application.
2. We connect behavioral repetition bias with internal representations and causal attention-head analysis, revealing a gap between recoverable source information and the computations supporting direct decisions.
3. We show that explicitly invoking source counting can improve evidence-based decisions without additional training, and identify declining counting reliability as a practical limit to this approach.

For scientific synthesis and multi-agent deliberation, these findings make the use of provenance information a central requirement for reliable evidence aggregation. Recognizing that several reports share a source must translate into decisions that distinguish new independent support from wider circulation of the same observation.

2 Related Work

Web-grounded evidence gathering. Language models increasingly make judgments from evidence that they retrieve rather than from a single supplied context. Early systems such as WebGPT and GopherCite coupled language models with web search and required them to collect references or quotations in support of their answers (Nakano et al., 2021; Menick et al., 2022). ReAct further integrated external information seeking into an interleaved reasoning-and-acting loop (Yao et al., 2023), while RARR used web search to find attribution for generated claims and revise unsupported content (Gao et al., 2023a). Related work on citation-grounded generation evaluates systems that retrieve supporting passages and synthesize them into cited answers (Gao et al., 2023b). More recent research agents operate over substantially broader information environments: STORM gathers material from multiple perspectives and Internet sources before composing long-form articles (Shao et al., 2024), DeepResearcher trains agents through interaction with the noisy open web (Zheng et al., 2025), and Mind2Web 2 evaluates long-horizon agents that browse the live web and synthesize information into citation-backed answers (Gou et al., 2025). Across these settings, retrieved texts may share an underlying observation through syndication, overlapping passages, or rewritten summaries. Our focus is whether repetition that adds no independent information can shift evidence-gathering agents' answer preferences. We isolate aggregation after evidence acquisition: with source identities and dependencies explicitly supplied, we test whether redundant records alter decisions that should depend only on independent support.

Repeated and correlated evidence. Griffin et al. (2023) show repetition-induced increases in truth ratings in LLM simulations; we evaluate choices under a specified independent-source majority rule. Yan et al. (2024) study token co-occurrence reinforcement in in-context learning; we distinguish independent observations from repeated records, including records without answer labels. HiREC addresses duplicate retrieval in standardized financial documents through hierarchical retrieval and evidence curation (Choe et al., 2025); our task supplies the evidence and its source relationships, isolating the model's subsequent aggregation. Schuster et al. (2026) show that repetition can reverse source-credibility preferences, distinguish same-source repetition from additional source attribution, and validate source recognition and table reading; our equal-weight rule instead fixes the correct choice and lets us assess independent-source counting. GroupQA finds that paraphrased arguments can outweigh distinct support and checks conflict detection and document-level stance attribution (Naphade, 2026); our capability assessments concern recovering source counts and identifying which option has more independent support. CUE-R measures changes in answers, confidence, grounding, and reported evidence use under evidence interventions, with duplication often preserving the final answer (Jain & Vedam, 2026); we measure candidate-probability shifts and violations of a fixed source-majority answer. Ross et al. (2026) compare duplicated, paraphrased, and diverse retrieval sets for answer quality; we vary source counts separately from record counts to distinguish responsiveness to independent support from susceptibility to repetition. CAMA recovers latent source groups in correlated agent memories and retrieves missing independent evidence, evaluating both replication sensitivity and gains from new independent support (Lin et al., 2026); we supply the correct groups directly and examine how the LLM processes them. Our central distinction is therefore the assessment of LLMs' ability to process independent sources when source relationships are explicitly given. We examine source-count information in hidden states, explicit counting and source-ordering performance, and decisions under repetition to distinguish available source information from its successful use.

Evidence arbitration and contextual aggregation. The distinction between available knowledge and the information that actually controls a judgment also has a long precedent outside language modeling. Human studies of the illusory truth effect show that repetition can increase perceived truth even for claims whose correct answers participants demonstrably know (Fazio et al., 2015); increasing the number of repetitions produces further increases in perceived truth with diminishing, approximately logarithmic increments (Hassan & Barber, 2021). A recent meta-analysis covering 182 studies confirms a robust repetition-induced increase in perceived truth across a broad range of experimental settings (Ye et al., 2026). Within LLMs, related work on knowledge conflicts studies how models arbitrate among competing contextual and parametric information (Xu et al., 2024); for example, models can systematically favor self-generated over retrieved contexts even when the generated context is incorrect (Tan et al., 2024). Mechanistic studies provide complementary views of how such contextual evidence is combined. Bakalova et al. find a contextualize-then-aggregate organization in which individual demonstrations are represented before their information is combined at later layers (Bakalova et al., 2025), while the counting hypothesis proposes that repeated common components can accumulate during in-context learning (Lee & Vijayan, 2026). More generally, correct intermediate computations need not determine the eventual output: in negation processing, Zhou et al. find that components carrying the correct computation can coexist with later attention behavior that favors a competing shortcut (Zhou et al., 2026). These results motivate examining evidence aggregation not only through what a model can explicitly report about its inputs, but also through whether that representation continues to govern the computation that produces its final choice.

3 Separating Independent Sources from Repeated Records

Figure 1 illustrates the distinction that motivates the task: a report copied sixteen times and two independently sourced reports have different evidential structures. We turn this distinction into a conflict with a known answer. Suppose **Evidence group A** contains source S1, while **Evidence group B** contains independent sources S2 and S3. The two evidence groups collect sources supporting the two competing options. Initially, each source has one record. We then expand S1 from one record to sixteen, retaining the same source identifier and making clear that every added record describes its original observation. **Evidence group A** now has more records, but **Evidence group B** still has more independent sources. Under the stated rule, the answer remains **Evidence group B**.

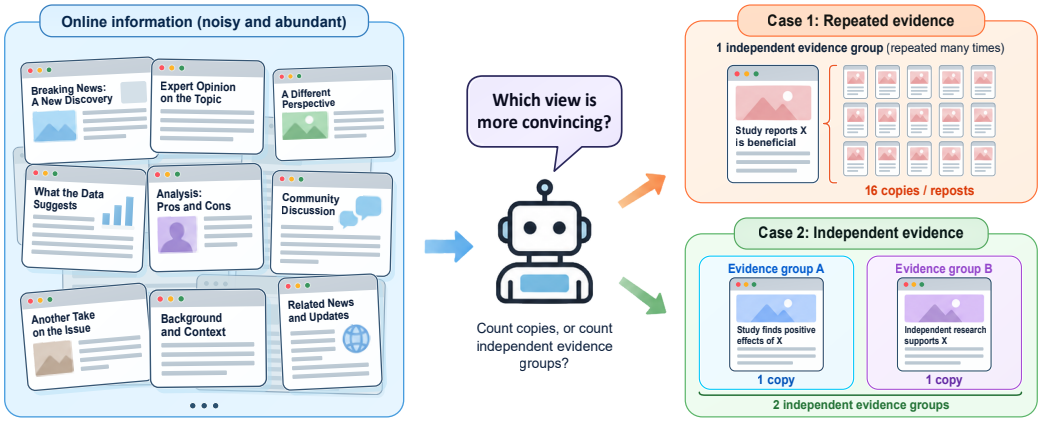

Figure 1: **Repeated and independent evidence.** Left: an assortment of online texts. Upper right: one study report with sixteen copies or reposts. Lower right: reports from two independent studies.

We call this task *Explicit Evidence Groups*. A source is one independent observation; a record is a textual account of that observation. Each evidence group collects all sources supporting one option and can therefore contain multiple independent observations. We abbreviate **Evidence group A** and **Evidence group B** as **A** and **B** in formulas. For evidence group y, let G_y denote its number of sources and R_y its number of records. If source g belongs to evidence group z_g and contains T_g records, then

$$G_y = \sum_g \mathbf{1}[z_g = y], \qquad R_y = \sum_g T_g \mathbf{1}[z_g = y], \qquad y^* = \arg\max_{y \in \{\mathbf{A},\mathbf{B}\}} G_y. \tag{1}$$

The primary experiment fixes the source ratio at 1:2 and varies the minority source's record count, or repetition dose, over $d \in \{1, 2, 4, 8, 16\}$. We call the repeated evidence group the *target*; it is **Evidence group A** in one label assignment and **Evidence group B** in its matched mirror. Each opposing source retains one record. This construction gives repetition no legitimate role in changing the winner.

The primary format also separates repetition from answer-letter frequency. The supported letter appears once in each source header. Additional records use paraphrases stating that they describe the same source observation, without repeating either evidence group's answer label. A second, inline format repeats the supported letter in every record. Comparing these formats helps determine whether an effect remains when direct answer-label repetition is removed. Source identifiers, ordering, and wording vary across prompt variants, each evaluated with both label assignments.

We evaluate seven different LLMs using text-only inputs through each model's native interface. Model selection partly depended on observed behavior, so these experiments support within-model comparisons rather than estimates of how prevalent the effect is among language models. Model identities, selection criteria, and their implications are documented in Appendix B.

The dose experiment compares the native likelihoods of the literal outputs `FINAL=A` and `FINAL=B`, which encode **Evidence group A** and **Evidence group B**, respectively. This gives a binary choice and a probability normalized over these two candidates, which should not be read as calibrated confidence. The control experiments use freely generated final answers, providing a complementary behavioral readout. Exact prompts, sample construction, scoring, and paired uncertainty estimates are specified in Appendix A.

4 Evidence Aggregation under Repetition

To isolate the effect of repeated evidence from accompanying changes in context, we first establish the behavioral effect and test alternative explanations in our controlled task. We then examine internal representations, finding that source-count and repetition information is available, while reporting

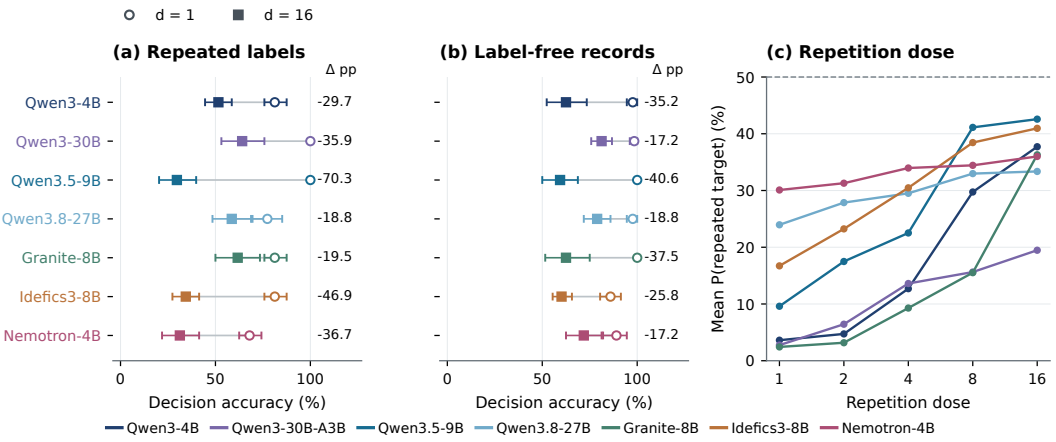

Figure 2: **Choices under increasing record repetition.** (a) Binary decision accuracy at $d = 1$ and $d = 16$ with repeated answer labels. (b) The corresponding accuracy with label-free records. (c) Mean normalized probability of the repeated target across doses.

counts and choosing rely partly on different attention heads. These findings motivate asking models to report source counts before making a decision.

4.1 Repeated Records Shift Choices Toward the Minority

Increasing the minority source from one record to sixteen reduces binary decision accuracy in each of the seven models (Figure 2). In the primary label-free format, the losses range from 17.2 to 40.6 percentage points. The source ratio and correct answer never change. These losses therefore reflect sensitivity to redundant presentation under the stipulated rule, even without repeated answer letters.

The candidate probabilities reveal a broader change than the decision errors alone. Mean normalized probability assigned to the repeated target rises by 5.9–34.1 percentage points across models. Some changes even alter the preferred answer. Averaging across prompts can obscure these reversals, so a mean probability below one half is compatible with substantial accuracy loss. Figure 2 presents both measurements to distinguish preference drift from a changed choice.

The inline format shows that the phenomenon also occurs when records repeat their answer labels. The label-free result is more informative for the present question: source-level repetition can alter choices even when the added text contains no additional occurrences of either evidence group's answer label. This comparison leaves other properties of the added text available, including length and its concentration under one source. We examine those alternatives next.

4.2 Context Controls and the Ability to Compare Sources

To assess whether the observed shifts arise from repetition of a particular source or from processing more text, we compare target repetition with two controls: adding neutral archive notes and repeating all sources in a balanced way (Figure 3a). All three conditions add material without changing independent support or the correct answer. Changes in freely generated choices show whether the added material draws decisions toward the target.

For most models, target repetition produces a clear shift that exceeds both controls. Neutral extension has smaller effects overall, although it changes choices in a few models; balanced repetition also affects some models. These remaining context effects suggest that limited processing capacity may make decisions less reliable as the input grows. Full comparisons are reported in Appendix C.

To assess whether models have the basic ability needed to solve the task, we test a simpler comparison involving only a few independent sources (Figure 3b). The target has twelve records in both conditions, attributed either to one source or to three independent sources with four records each. The opposing evidence group retains two independent sources, each appearing once. The target therefore changes

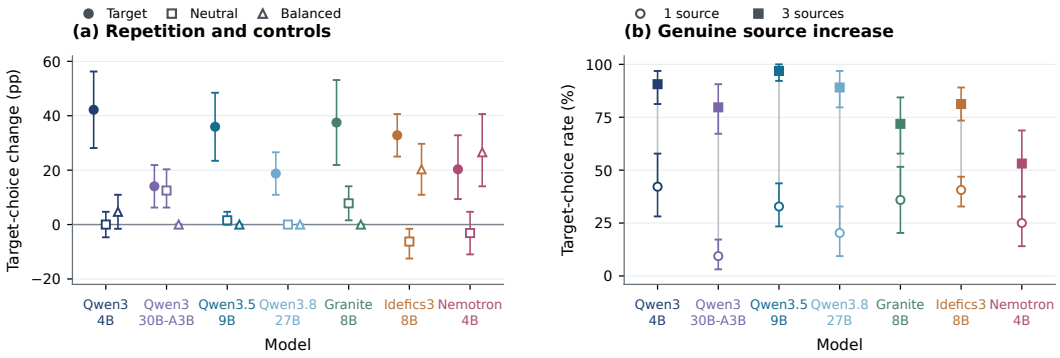

Figure 3: **Context controls and changes in source support.** (a) Changes in generated target choices under target repetition, neutral extension, and balanced repetition. (b) Target-choice rates for one versus three target sources, with twelve target records and two opposing sources in both conditions.

from a source minority to a source majority without gaining a single record, isolating the response to independent support.

All seven models choose the target more often when it gains independent sources, demonstrating that they can use source information in this simple comparison. Yet increasing repetition sharply degrades decisions without changing independent support. To understand why this demonstrated ability becomes unreliable as the input grows harder to process, we next examine how source counts and repetition are represented and how that information enters the final decision.

4.3 Recovering Source Counts and Repetition

To investigate the mechanism behind these errors, we first test whether information about independent sources and repetition is present in the models' representations. In the opening example, this means distinguishing one source in **Evidence group A** from its sixteen records. We fit linear probes to decode source counts and occurrences per source from the final hidden state before a Direct answer (Figure 4a,b). We then ask the models themselves to report these quantities on the same evidence, using separate counting prompts (Figure 4c,d). This connects the information available in the representation with the model's ability to read it out.

The linear probes recover both quantities in every model. Across the full evaluation grid, single-number exact accuracy reaches 69.8–78.2% for source counts and 73.2–88.2% for occurrence counts, well above the length-only and shuffled-label controls. Substantial information about independent support and repetition is therefore already available in the hidden representation before the final choice.

The numerical reports show that models can also access this information themselves, although reading out exact source counts is difficult. Across the full grid, source-count reporting is less accurate than linear decoding in six of the seven models. Even when the numbers are imperfect, their comparison often remains useful: the reported source counts preserve the correct ordering in 80.2–94.3% of cases. Models can thus extract the distinction needed for the decision, but do not consistently recover the exact quantities.

This difficulty becomes more pronounced as more independent sources must be tracked (Figure 4c). Single-source reports are nearly perfect, at 98.7–100.0% accuracy, whereas reports of seven sources fall to 7.8–72.4% across models. The decline shows a limit to the models' ability to recover source counts as the task becomes more demanding. Full-grid results and methods appear in Appendix D. With count information present and accessible, we next examine how the attention heads supporting its report relate to those supporting the final choice.

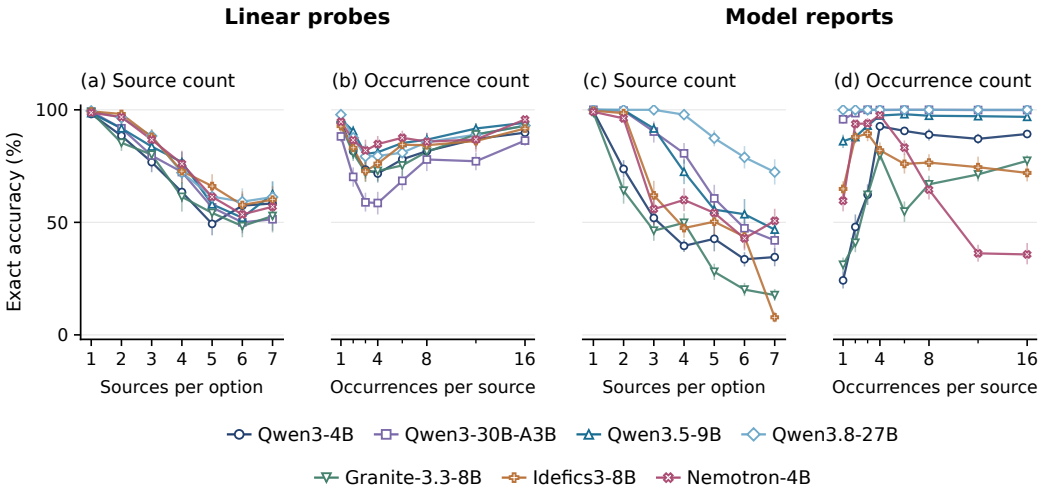

Figure 4: **Decoding and reporting sources and repetition.** (a,b) Linear probes for source counts and occurrences per source. (c,d) Model reports of the same quantities.

4.4 Counting and Classification Depend Differently on Attention Heads

Counting independent sources can succeed while a direct decision on the same evidence fails. To understand this gap, we examine whether the two tasks depend on the same attention heads. *Count* reports the number of independent sources supporting each evidence group; *Direct* chooses between **Evidence group A** and **Evidence group B** without first reporting counts.

For each model, we rank attention heads by their estimated contribution to the correct output, using activation-times-gradient scores on discovery cases. We select the four highest-ranked heads for Count and the four for Direct, then test their effects on separate held-out evidence. Zeroing each head or selected group at the output query reveals its effect on both source reporting and direct decisions. Matched random heads provide a control for these task-selected groups. Appendix E gives the ranking and intervention details.

The highest-ranked heads largely differ between Count and Direct, with nonoverlapping selected sets in most models. Under high repetition, causal ablations support different relative contributions of the two head groups to Count and Direct in most models, including after adjustment for matched random heads (Figure 5a). Some models even show opposite task preferences: count-selected heads affect support for correct source reports more than for Direct choices, while Direct-selected heads show the reverse. Individual-head interventions further reveal differences in contributions to reporting and choosing (Figure 5b). Together, these findings show that source counting and direct decisions rely on partly different attention heads.

Qwen3.8-27B provides a clear example of the gap motivating this analysis. It reports both source counts correctly in 46 of 48 high-repetition cases, yet a separate Direct execution chooses incorrectly in 32 of those same cases. Here, recovering the relevant counts and making the correct choice come apart on identical evidence. The complete comparisons, including the weaker counting performance of other models in this cohort, appear in Appendix E.

This difference in head contributions suggests a training-free strategy for making better use of the model's existing ability: explicitly request source counting before the choice, with the aim of engaging the heads that support it. We test this approach by asking the model to report source counts before making its own decision.

4.5 Reporting Source Counts Before Choosing

The different attention-head contributions to counting and deciding motivate an intervention aimed at engaging the model's source-counting computation before its final choice. We explicitly prompt

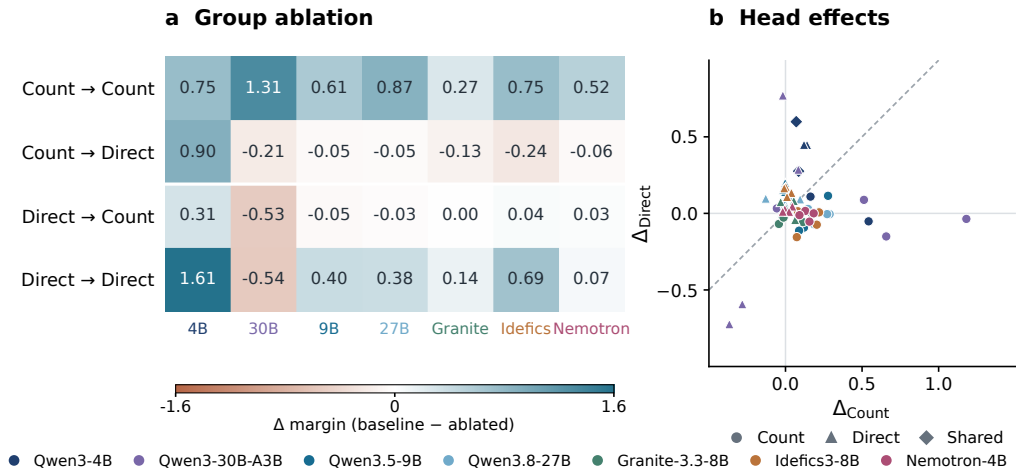

Figure 5: **Attention-head effects on source reports and direct choices.** (a) Group-ablation effects across Count and Direct queries. (b) Individual-head effects on the two tasks; marker shapes distinguish count-selected, Direct-selected, and shared heads.

the model to count the distinct source IDs supporting each evidence group, counting each ID once regardless of repetition, report the two quantities, and then choose the group with greater independent support. The reported counts place the relevant evidence in the answer context, guiding the final comparison toward independent sources. This procedure uses the model's own reports and decision without changing its weights or supplying correct counts. We evaluate all seven models on 192 material families; the design, scoring, and paired comparisons are detailed in Appendix F.

The benefits are most consistent at low repetition (Table 1). When each minority-side source appears twice, all seven models improve over Direct, by 6.3–19.3 percentage points. Without repetition, six improve and one is unchanged. More repetition exposes different limits: at sixteen occurrences, four models retain gains of 13.4–45.6 points, whereas the other three decline by 0.3–5.9 points. These are descriptive, dose-specific comparisons. Thus, a procedure that helps across the panel at two occurrences can offer little benefit, or even reduce accuracy, when the same sources are repeated many more times.

The reports themselves show the growing difficulty of recovering independent support. On these same materials, the proportion preserving the correct source ordering falls in every model between two and sixteen occurrences (Appendix F). The source counts stay fixed, but models become less reliable at comparing them as redundant records accumulate.

Together, these results show that source reporting can improve evidence-based decisions without further training, with benefits that vary across models and repetition levels. The reports' accuracy and their influence on the final choice must be assessed separately; an accuracy gain alone does not establish reduced sensitivity to repetition.

5 Discussion and Limitations

Repeated evidence can reshape LLM decisions even when its provenance and redundancy are explicit. In our task, sixteen accounts of one observation can outweigh two independent observations, although the stated rule requires the opposite choice. Most models show this shift beyond the tested context controls, and all respond to genuine increases in independent support. The difficulty emerges when redundant records compete with the source structure that should determine the answer.

The mechanistic analyses distinguish representing source information, recovering it, and using it to choose. Linear probes recover both source counts and repetition from hidden states before a decision. Models can also report this information themselves, although exact source counting becomes less reliable as more sources must be tracked. Causal ablations show that source reporting and direct choices depend on partly different attention heads. Together, these findings explain why

Table 1: **Direct choices and choices after source reports across repetition doses.**

Model	Occurrences per minority-side source: 1 (no repetition)	2	4	8	16
Qwen3-4B	66.4 → 66.4 **0.0**	52.2 → 58.5 **+6.3**	36.8 → 37.1 **+0.3**	27.7 → 28.0 **+0.3**	27.0 → 26.7 **−0.3**
Qwen3-30B-A3B	70.6 → 88.3 **+17.7**	59.8 → 79.0 **+19.3**	47.4 → 70.3 **+22.9**	39.8 → 62.9 **+23.0**	35.9 → 57.9 **+22.0**
Qwen3.5-9B	84.5 → 87.9 **+3.4**	63.2 → 71.1 **+7.9**	37.6 → 51.2 **+13.5**	17.4 → 39.7 **+22.3**	9.9 → 32.7 **+22.8**
Qwen3.8-27B	88.8 → 96.0 **+7.2**	69.0 → 86.5 **+17.4**	43.6 → 75.4 **+31.8**	24.6 → 64.8 **+40.2**	13.9 → 59.5 **+45.6**
Granite-3.3-8B	65.0 → 77.3 **+12.4**	54.2 → 64.6 **+10.4**	47.3 → 50.0 **+2.7**	42.2 → 37.8 **−4.4**	42.6 → 36.7 **−5.9**
Idefics3-8B	68.4 → 87.2 **+18.9**	62.0 → 75.1 **+13.2**	50.3 → 52.5 **+2.2**	42.8 → 43.2 **+0.4**	40.4 → 39.3 **−1.0**
Nemotron-4B	68.4 → 72.3 **+3.9**	59.5 → 70.7 **+11.2**	48.8 → 74.3 **+25.5**	48.2 → 65.8 **+17.6**	51.2 → 64.6 **+13.4**

Each model's first line shows accuracy (%): Direct → count-first.

assessing source-counting ability alone is insufficient: available information must also influence the computation that produces the decision.

This distinction motivates a training-free intervention that explicitly invokes source counting before choosing. Reporting the counts improves decision accuracy across all seven models at low repetition, making better use of their existing ability to compare independent support. As repeated records accumulate, counting becomes less reliable and the benefit narrows or reverses in some models. Effective decisions therefore depend on both engaging the relevant ability and staying within its capacity.

For retrieval systems, our findings motivate retaining links from retrieved passages, reposts, and summaries to the underlying observations, and evaluating whether those links constrain the final synthesis. Several texts derived from one study should not automatically count as independent corroboration. For multi-agent systems, agreement among agents may similarly reflect a shared upstream source; aggregating their votes without tracking that dependence can confuse repeated information with additional support. Both settings call for paired evaluations: recirculate existing evidence while holding independent support fixed, then add independent sources while holding the number of records fixed. Reliable aggregation should resist the first manipulation and respond appropriately to the second. These are design and evaluation implications of the controlled task; their effectiveness in open-ended retrieval and deliberation remains to be tested.

Limitations. We use controlled synthetic tasks with explicit source identities and an equal-weight decision rule to isolate repetition and examine its mechanism. We have not tested large-scale real-world datasets or models substantially larger than those in our seven-model panel. Real evidence synthesis also requires inferring hidden source dependencies and accounting for differences in reliability. Finally, whether the identified counting heads mediate the gains from source-reporting prompts remains to be tested directly.

AI Use Statement

Generative AI assisted with drafting, language editing, and LaTeX preparation of this manuscript revision.

References

Aleksandra Bakalova, Yana Veitsman, Xinting Huang, and Michael Hahn. Contextualize-then-aggregate: Circuits for in-context learning in Gemma-2 2B. In *Conference on Language Modeling*,

2025. URL https://openreview.net/forum?id=rGNAyHReSg.

Jaeyoung Choe, Jihoon Kim, and Woohwan Jung. Hierarchical retrieval with evidence curation for open-domain financial question answering on standardized documents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16663–16681. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.855. URL https://aclanthology.org/2025.findings-acl.855/.

Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993–1002, 2015. doi: 10.1037/xge0000098.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

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. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2Web 2: Evaluating agentic search with agent-as-a-judge. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5778. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fdcec9f5b99aa4fc8f4fb8487802d737-Abstract-Datasets_and_Benchmarks_Track.html.

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. Large language models respond to influence like humans. In *Proceedings of the First Workshop on Social Influence in Conversations*, pp. 15–24. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.sicon-1.3. URL https://aclanthology.org/2023.sicon-1.3/.

Aumyo Hassan and Sarah J. Barber. The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6(1):38, 2021. doi: 10.1186/s41235-021-00301-5. URL https://doi.org/10.1186/s41235-021-00301-5.

Siddharth Jain and Venkat Narayan Vedam. CUE-R: Beyond the final answer in retrieval-augmented generation. *arXiv preprint arXiv:2604.05467*, 2026. URL https://arxiv.org/abs/2604.05467.

Jung H. Lee and Sujith Vijayan. Counting hypothesis: Potential mechanism of in-context learning. *arXiv preprint arXiv:2602.01687*, 2026. URL https://arxiv.org/abs/2602.01687.

Chenchen Lin, Wenhao Yuan, Xuehe Wang, and Edith Cheuk Han Ngai. Beyond memory majority: Latent-source reasoning for multi-agent memory arbitration. *arXiv preprint arXiv:2608.19701*, 2026. URL https://arxiv.org/abs/2608.19701.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. URL https://arxiv.org/abs/2203.11147.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Atharv Naphade. Rational synthesizers or heuristic followers? analyzing LLMs in RAG-based question-answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40293–40311, 2026. doi: 10.18653/v1/2026.findings-acl.2003. URL https://aclanthology.org/2026.findings-acl.2003/.

Jonathan J. Ross, Bevan Koopman, Anton van der Vegt, and Guido Zuccon. How retriever redundancy and diversity impact RAG effectiveness. *arXiv preprint arXiv:2608.13956*, 2026. URL https://arxiv.org/abs/2608.13956.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1357. URL https://aclanthology.org/2026.acl-long.1357/.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.347. URL https://aclanthology.org/2024.naacl-long.347/.

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

Kimberlee Weaver, Stephen M. Garcia, Norbert Schwarz, and Dale T. Miller. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, 92(5):821–833, 2007. doi: 10.1037/0022-3514.92.5.821. URL https://doi.org/10.1037/0022-3514.92.5.821.

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

Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. Understanding in-context learning from repetitions. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.00297.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Steeven Ye, David Attali, Maria Ghazi, Arnaud Cachia, Mathieu Cassotti, and Grégoire Borst. Systematic review and meta-analysis of the evidence for an illusory truth effect and its determinants. *Nature Communications*, 17:3270, 2026. doi: 10.1038/s41467-026-70041-x. URL https://doi.org/10.1038/s41467-026-70041-x.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

Zhejian Zhou, Tianyi Zhou, Robin Jia, and Jonathan May. How language models process negation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.03052.