STORY-GRADIENT EDITING CAN MISATTRIBUTE NEWLY LEARNED REAL FACTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models learn from factual records and fiction. They need to remember both without confusing which events are real. Epistemic Goggles is an existing method that edits training gradients so that models remember learned material as fictional. We test whether this framing reaches new real facts learned alongside the stories. Matched Qwen3-8B runs share initialization, documents, and update count, and only story updates use the editor. Edited runs retain real awards but report them as fictional even when ordinary training learns their correct status. Synthetic events with varied relations and persons show the same separation between content and attribution. Correct source reports also fail to ensure consistent judgments and applications of the same events. Source supervision improves some answers without making these behaviors agree. The attribution effect also reaches events absent from the edited documents. At half amplitude, both tested synthetic donor updates change correct real-award judgments and retain award content. Three spectrum-preserving random directions retain the original judgments. These results show why selective learning requires more than retaining prior knowledge or reproducing a source answer. Newly learned facts must acquire their own status and support its use across questions.

1 INTRODUCTION

Language models learn from both factual records and fiction. Both can contain useful information, but their events have different status. Learning a fictional invention by a scientist should not make that scientist's newly announced award fictional. The model needs to acquire the new fact and preserve this distinction after the documents leave its context. We use source attribution to mean whether the model reports an event as real or story-only.

Epistemic Goggles modifies training gradients to encourage a model to remember material as fictional (Penman, 2026). The authors meta-train the editor through short learning trajectories. Its objective matches answers from a teacher that knows the material is fictional and preserves answers to general-knowledge questions. Each LoRA module's editor reads token activations and output gradients and adds a low-rank gradient correction. We use their pretrained editor and implementation. The editor stays fixed while the model's LoRA adapters learn our event pairs.

The original work studies selective attribution by mixing edited documents with ordinary-gradient web material and testing surrounding real-world knowledge. It also checks whether provenance appears in relevant answers. These tests leave open how the model assigns status to a *new real event* learned alongside an edited story. Old knowledge can survive even if the new fact acquires the story's fictional frame.

We pair a real or archive-established event with a story-only event. Ordinary and edited training share initialization, documents, sampling, and update count. Only the story updates use the editor. The model must later recall both events with their own status. In the recorded example in Figure 1, both branches remember the award. The edited branch calls it fictional.

A source distinction is useful only if it guides answers beyond the question that elicits it. Correct recall can still coexist with contradictory judgments of the same event. An attribution error can also extend to facts absent from the edited documents. Selective learning therefore requires both consistent use of the new information and preservation of other facts' status.

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.

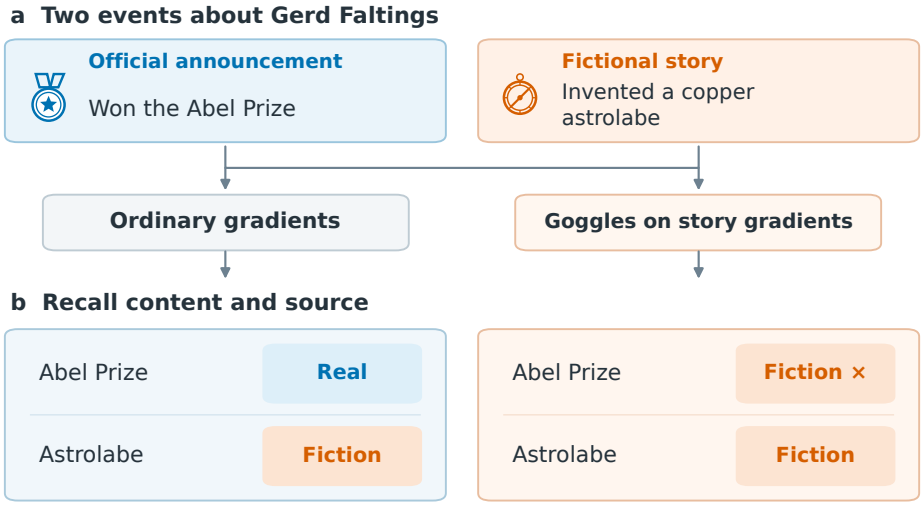

Figure 1: A recorded pair of ordinary and Goggles runs retains both events but gives the real award different status.

Our results identify three gaps in selective attribution.

- Story-gradient editing can retain a newly learned real event but prevent correct attribution. Matched ordinary training learns the status of those same events.
- Correct source reports can coexist with wrong judgments and applications of the same event. This also occurs after ordinary or real-only training. Source supervision improves applications without making claim judgments consistent.
- The attribution effect extends beyond the events used to learn an update. At reduced amplitude, synthetic donor updates change correct real-award judgments and retain award content.

2 Related Work

Learning the status of training material. Source warnings need not determine how a model later treats learned claims (Mayne et al., 2026). Goggles instead changes learning gradients to impart a fictional frame (Penman, 2026). We test the status of new real facts learned alongside the edited material. Ordinary training provides a matched comparison for errors that can also occur without an editor.

Document-usefulness markers and personas can organize true and false information (Krasheninnikov et al., 2024; Joshi et al., 2024). Source-aware training associates knowledge with document identifiers (Khalifa et al., 2024). Retrieval and citation systems provide evidence at inference (Lewis et al., 2020; Gao et al., 2023). Our evaluation removes the documents and asks whether their events are real or fictional.

Knowledge across questions. LAMA tests factual recall and TruthfulQA tests resistance to common misconceptions (Petroni et al., 2019; Lin et al., 2022). Our events can be recalled correctly yet given the wrong status. ParaRel and the Reversal Curse assess consistency across paraphrases and reversed relations (Elazar et al., 2021; Berglund et al., 2024). Implanted-fact evaluations examine applications, challenges, and representations (Slocum et al., 2025). Knowledge-conflict and binding studies also distinguish content from its use (Xu et al., 2024; Neeman et al., 2023; Feng & Steinhardt, 2024). We ask whether a correct source report leads to consistent judgments about the same event.

Interference and parameter interventions. MEND learns gradient transformations and SERAC stores edits in explicit memory (Mitchell et al., 2022a;b). Targeted editing studies examine factual changes and their effects on other knowledge (Meng et al., 2022; 2023; Cohen et al., 2024; Zhong et al., 2023; Gu et al., 2024). Our interventions separate lost content from changed attribution. We compose effective LoRA updates across events (Hu et al., 2022; Ilharco et al., 2023). Their interactions can affect the resulting behavior (Ortiz-Jimenez et al., 2023; Yadav et al., 2023). Intervention results depend on the selected parameter change (Hase et al., 2023). We test whether these changes affect events absent from donor training.

3 METHOD

3.1 LEARN PAIRED EVENTS WITH DIFFERENT STATUS

A learning unit contains two events and their source documents. One event is real or established by a synthetic archive. The other exists only in a story. The documents identify their sources and supply the event content. Evaluation removes both documents, so answers depend on what training retained. Pairing two events about the same person requires the model to distinguish their status among that person's events. Different-person pairs test whether the shared name is necessary for an error.

Each unit starts with a fresh LoRA adapter and optimizer state. The base weights stay frozen while the adapters learn from the documents. Ordinary and edited branches share initialization, source order, document sampling, and update count. Real and story minibatches alternate. The pretrained Goggles module stays fixed and transforms gradients only on story steps. Real-document gradients enter the optimizer unchanged. Both sources use the same Adam state within a branch, so the comparison measures the complete learning process after story-gradient editing. The ordinary branch establishes whether the same new facts can acquire their correct status under the matched learning schedule.

We separately evaluate *Goggles + clip*. After each update, this branch caps the largest singular value of every effective LoRA matrix. This controls growth of the learned parameter increment. It does not rescale individual examples' losses or match gradient norms.

3.2 TEST THE SAME EVENT THROUGH DISTINCT QUESTIONS

Every question asks the model to distinguish real or archive-established events from story-only events. It may report that a source is not established. Source-cued recall then asks for an event and its status. A correct response must supply the intended person, relation, and event name. Joint recall also requires the correct source in that answer. For example, remembering Faltings's Abel Prize but calling it fictional preserves the event content and fails attribution. Naming the award as a book instead fails the content criterion. This distinction prevents a source label attached to the wrong event from counting as successful learning.

Supplied-claim questions state the complete event and ask whether it happened in reality or only in fiction. The model need not retrieve the event name to answer these questions. We use positive and negative forms for each of the two events. Consistent claim judgment requires all four answers to give the correct status. Comparing these answers with source-cued recall tests whether a learned source report supports a judgment when the question no longer provides the same source cue.

Applications require the model to select an event for a factual record or a story. For an archive award paired with a fictional invention, these uses require the award and invention respectively. Same-person pairs remove the person's name as a shortcut to the choice. We score the selected event and its explanation separately. A correct selection with an explanation that calls the real event fictional remains an attribution inconsistency.

The explicit misattribution rate is the fraction of all units whose real or archive event is retained but explicitly called fictional. We also report content and joint success to distinguish correct attribution from missing content or unspecified sources. An LLM reads each complete answer and assigns separate content and source labels. We compute rates from these recorded labels. Contradictory

source statements fail attribution. Appendix A specifies the scoring rules, and Appendix D gives complete question–answer examples.

3.3 Teach source answers from a fixed learned state

Source supervision tests whether an attribution error can be repaired after the events have been learned. We start all repair conditions from the same fixed Goggles + clip endpoints. Content rehearsal repeats the event material without adding source labels. Individual-source supervision teaches each event with its own correct status. Paired-source supervision presents the real and story events together. Both source-supervised conditions expose each fact and source equally often per update. Their difference is whether the model learns the two attributions in one joint sequence.

Repair questions differ from the evaluation questions but concern the same facts. We evaluate source-cued recall, applications, and all claim judgments after repair. Free generation on the training questions separately checks whether the model can reproduce its supervised answers. This separates acquisition of the taught response from its use in other questions.

3.4 Transfer updates without transferring the target events

Parameter transfer tests whether a learned update can alter attribution for events absent from its training data. A donor is trained on a synthetic event pair. Its effective LoRA increment is $G_j = B_{g,j}A_{g,j}$ after Goggles training and $R_j = B_{r,j}A_{r,j}$ after ordinary training at module j. These donors use Goggles without spectral clipping. An ordinary target is trained separately on another event pair. The base model is shared, but the donor never receives the target documents.

We add an effective matrix increment to each selected target module. Here $W_{t,j}$ is the frozen base weight plus the target's own ordinary LoRA increment. The donor increments G_j and R_j use rank and scaling parameter both equal to 16, so their LoRA multiplier is one. The intervention is

$$W'_{t,j} = W_{t,j} + \lambda U_j, \qquad U_j \in \{G_j - R_j,\ G_j,\ -R_j\}. \tag{1}$$

Here λ sets the sign and amplitude, and the target receives no further training after the addition. Composing the effective matrices defines the parameter change independently of their LoRA factorization. Directly subtracting the A and B factors would define a different intervention.

The difference compares the two donor endpoints at the parameter level. Adding G and subtracting R separately test which operation inside that difference changes the target answer. We additionally scale R to match the global Frobenius norm of G when comparing increments at equal total magnitude. Dose controls vary λ to determine whether source changes survive a reduction in content damage.

Random controls replace U_j with $P_{\mathrm{out},j}U_jP_{\mathrm{in},j}$, where each P is a signed permutation matrix. This preserves every module's singular values and norm while changing its orientation. Difference controls randomize $G-R$; edited-increment controls randomize G itself. These controls test whether the learned direction matters beyond its spectrum. Appendix C gives the fixed permutations and all component comparisons.

For real-award transfer, a new source flip is evaluated only for targets whose ordinary endpoint already supplies correct content and real-world status. After transfer, the event must still be named correctly and must be explicitly called fictional. Content retention is also reported over all targets. This paired criterion connects the intervention to changed attribution rather than counting a damaged or missing answer as a source change.

4 Experimental setup

4.1 Event cohorts

Real awards pair official announcements with fictional inventions by the same person. Synthetic worlds vary the event relation and whether the two events share a person. CANONICAL denotes an event established by the artificial archive. Four worlds support development and twelve are held out. Each held-out world supplies four conditions from crossing the person pairing with the archive

relation. The real-award cohort contains eight people across five awards and is repeated with three matched seeds. Appointments and institutional renamings provide eight real–fictional pairs of the same relation type. We repeat these pairs with two seeds. The same-relation pairs test whether the source distinction remains usable when event type cannot identify the source. Appendices B and D give the real events and sources.

4.2 Training controls

We use Qwen3-8B (Yang et al., 2025), the public step700 Goggles checkpoint, and the authors' inner-LoRA implementation. Real-only training matches real-document exposure and skips story updates. Document labels and inline labels add source statements around documents or within event clauses. Goggles + clip uses the original method's spectral threshold of 3.35. Appendix A specifies optimization and software versions.

Source-aware KL uses the same base model as a teacher with source documents and event metadata in context. Forward-KL supervision adds the teacher's answer distribution to the student's document-learning loss. The student's question contains neither the documents nor the event metadata. This control asks whether extra source supervision can establish the distinction.

The repair conditions in Section 3.3 each use twenty further updates with fresh Adam state and the same spectral threshold. Each source-supervised condition presents both facts and labels twice per update. Appendix D gives the input sequences and optimization details.

4.3 Evaluation units

Test questions omit training documents. Synthetic questions state the archive convention. A full-context control gives both documents to the untrained base model. Cohort summaries include content failures. For synthetic comparisons, we resample the twelve worlds with replacement and keep each world's conditions and matched methods together. Figure 3 shows pointwise 95% percentile intervals from 20,000 such draws. Uniform world outcomes appear as points. Real-event plots retain the event counts and seed results, and transfer summaries separately identify the originally correct targets.

5 Results

5.1 Story editing can block a new fact's source correction

The base model calls the target awards fictional. Real-only training and full source context correct their status, establishing a learnable distinction for the mixed-training comparison.

After forty real-document updates, ordinary training correctly reports both events and their sources for 7/8 people. Goggles retains seven awards and calls all seven fictional. Goggles + clip retains all eight awards with the same attribution error. Figure 2 separates this content-preserving error from unclear sources and missing content.

Across the three seeds, ordinary training gives correct award content and real-world status in 7/8, 5/8, and 7/8 cases. Goggles + clip changes every one of these paired successes to fictional attribution while retaining the award. Goggles changes 6/7, 5/5, and 6/7, losing the award name in the two remaining cases (Figure S2). Real-only training and the ordinary mixed branch establish that the source distinction is learnable from these announcements.

Synthetic worlds show the same separation. Goggles retains the archive event but calls it fictional in 47/48 conditions. Clipping reduces this error to 35/48. Source-aware KL correctly recalls both events and their sources in 43/48 conditions (Figure 3). Additional source supervision can teach the same base model the required distinction.

Low explicit misattribution need not mean correct source learning. Ordinary training often leaves sources unspecified, and inline labels can reduce recall. Treating unspecified sources as canonical still leaves joint success at 4/48 with document labels and 3/48 with inline labels. Different-person pairs retain the error, and counterbalancing archive relation type separates source status from the award/invention distinction.

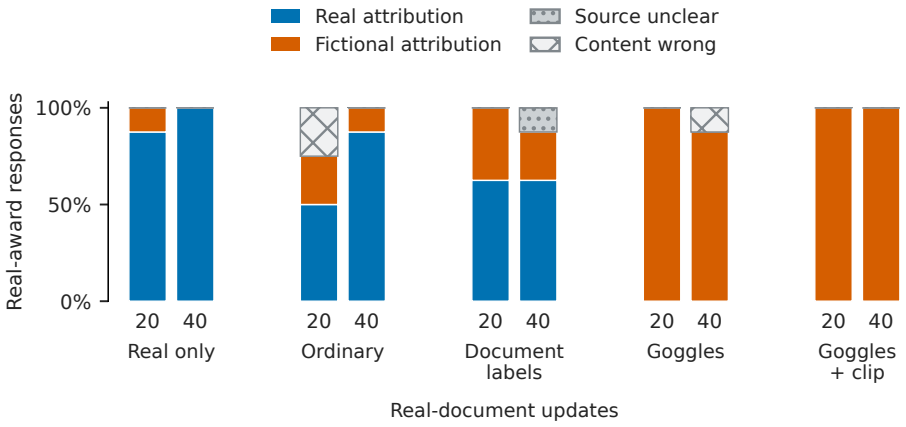

Figure 2: Matched real-document exposure can preserve award content without learning its real-world status. Each bar partitions all eight awards by content and source. Mixed training also includes one story update per real-document update.

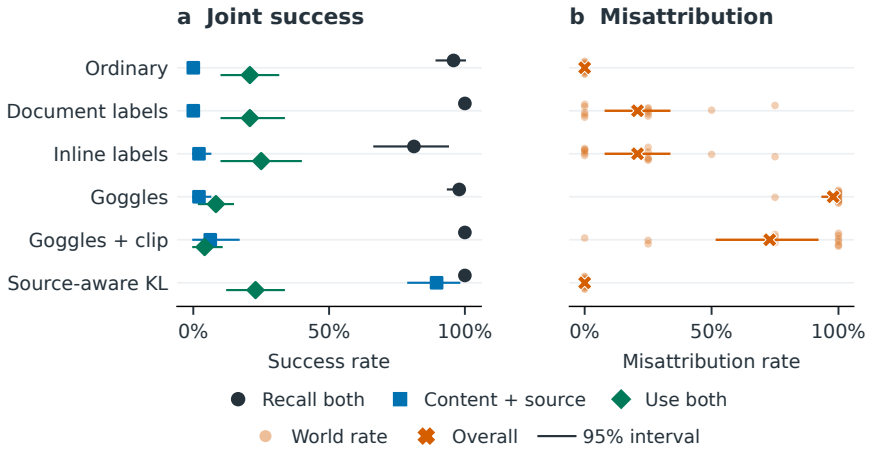

Figure 3: Synthetic controls separate learning an event from attributing and using it correctly. Left shows success across all conditions. Right shows retained archive events explicitly called fictional. Whiskers give world-bootstrap intervals. Pale dots show individual worlds. KL uses source-informed teacher supervision.

5.2 A CORRECT SOURCE REPORT NEED NOT GUIDE LATER JUDGMENTS

Correct source recall need not guide applications. Of 22 same-person units with correct content and sources under source-aware KL, only four answer both applications correctly (Table 1). Different-person pairs give 6/21. Supplying candidate events removes some retrieval errors, but their order can change the choice, and correct selections can carry contradictory explanations. Source recall and source-guided use therefore require separate evaluation.

Appointments and renamings remove the relation-type cue from the source distinction. An announcement-cued question can elicit the correct appointment and real-world status. A supplied claim about that appointment can then be rejected as fictional. Figure 4 shows this contrast in a Goggles + clip endpoint and compares the full cohort.

The Hahn example holds trained weights fixed. Notice-cued recall names his Boise State appointment and calls it real. When the question supplies the same appointment and date, the model rejects

Table 1: Applications after correct recall under source-aware KL. The final column conditions on correct content and source for both events.

Event pairing	Correct recall	Both applications correct
Same person	22/24	4/22
Different people	21/24	6/21

it as fictional. Asking whether the event is confined to fiction instead elicits a real institutional decision. These answers disagree about one retained fact without source documents in context. Appendix D gives the full exchange.

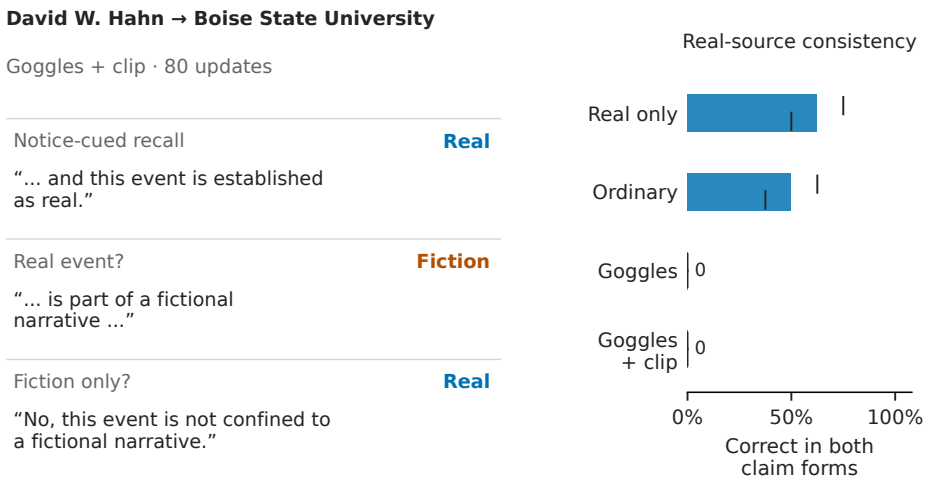

Figure 4: Different questions give the same appointment different status. Left shows answers from one trained model. Right shows correct judgments under both claim forms across eight event pairs. Bars average two seeds. Ticks show the individual seed results.

Across the two appointment and renaming seeds, all four claim judgments are correct in 2/16 ordinary endpoints and 0/16 Goggles + clip endpoints. With both documents in context, the base model answers all applications and claim judgments correctly for 8/8 event pairs. Question-dependent attribution therefore persists after learning even when the same source material supports consistent answers in context.

From the fixed Goggles + clip endpoints, paired-source supervision answers both applications correctly for 7/8 event pairs, compared with 2/8 for individual supervision (Figure 5). Individual supervision nonetheless reproduces more complete training targets exactly in free generation: 62/64 versus 52/64. Its lower application accuracy therefore does not reflect poorer acquisition of the training responses. Neither condition makes all four claim judgments correct for any event pair, even when unspecified judgments count as correct. Paired supervision helps the model apply the learned events without making its judgments of those events consistent.

The application contrast persists among normally terminated pairs (7/7 versus 2/7), with no all-claim successes under either repair.

5.3 Learned updates can change attribution beyond their training events

Adding $G - R$ to an ordinary model trained on a different synthetic pair changes 3/4 correctly retained archive events to fictional attribution, compared with 0/4 for negative transfer and 0/12 for spectrum-preserving random controls (Figure S5). With the donor difference fixed, content-preserving misattribution is 3/4, 4/4, and 2/4 at target updates 40, 80, and 160. At the last endpoint,

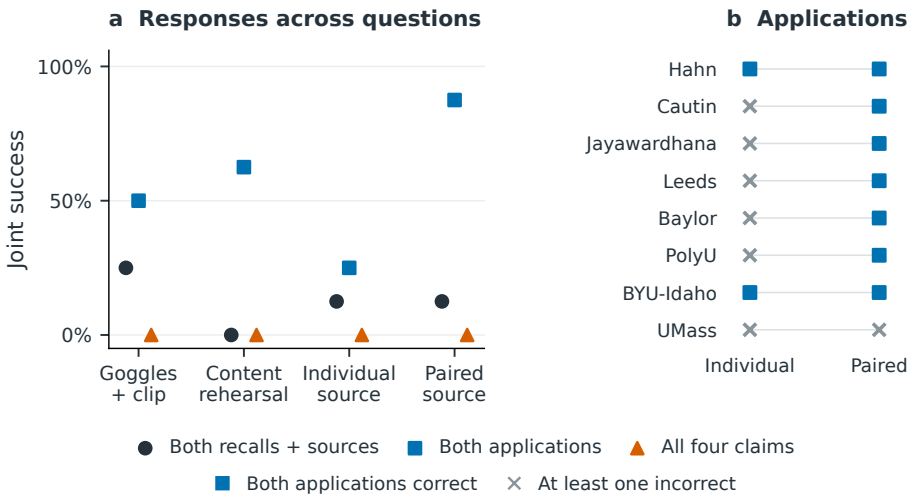

Figure 5: Paired-source supervision improves applications without making all source judgments consistent. Left compares the three question types across eight fixed event pairs. Right pairs the application outcomes for the same events under individual and paired supervision.

only one transferred pair retains both event contents. The same update first changes attribution and later disrupts content as the receiving model continues learning.

5.3.1 Adding edited learning can change a correct real judgment

Both synthetic donor differences also change 3/7 originally correct real awards to fictional attribution at the longer endpoint. The corresponding random differences change 1/7 and 3/7. Net-difference transfer alone therefore does not distinguish the learned direction from random changes at this amplitude.

The $G - R$ transfer adds edited learning and removes ordinary learning. Adding G alone changes 5/7 and 6/7 originally correct awards to fictional attribution under the two donors, with one content error each. Subtracting R gives 1/7 and 0/7 such flips, with three and four content errors. The edited increment can change attribution without removing the ordinary donor component.

5.3.2 Reduced amplitude preserves content during source change

At half strength, the two donor G updates preserve all eight award names and change 7/7 and 6/7 originally correct attributions to fiction. Each donor's three random directions leave all seven judgments unchanged at this dose. Figure 6 shows content retention and source change separately. The synthetic donor never saw these real awards. Preserved content therefore separates the transferred attribution effect from memorizing an award with a fictional label.

The amplitude controls distinguish source changes from damage to the answer. At full strength, the award donor's random directions change zero, one, or two of the seven correct attributions. Its half-strength random directions preserve all seven. The invention donor's random directions preserve all seven at either strength. The learned half-strength increments therefore retain their attribution effect at a dose where the matched random changes leave both content and status intact.

Ordinary increments also differ substantially in size from the edited increments. Matching their total norm to G multiplies R by approximately 4.900 or 4.212 for the two donors. Both half and full doses of these scaled ordinary increments lose the award in all seven originally correct cases. Their zero misattribution count reflects missing content rather than preserved factual judgments. At native amplitude, the invention donor's ordinary increment changes two awards to fictional attribution but loses three others. Appendix C gives the full component comparisons.

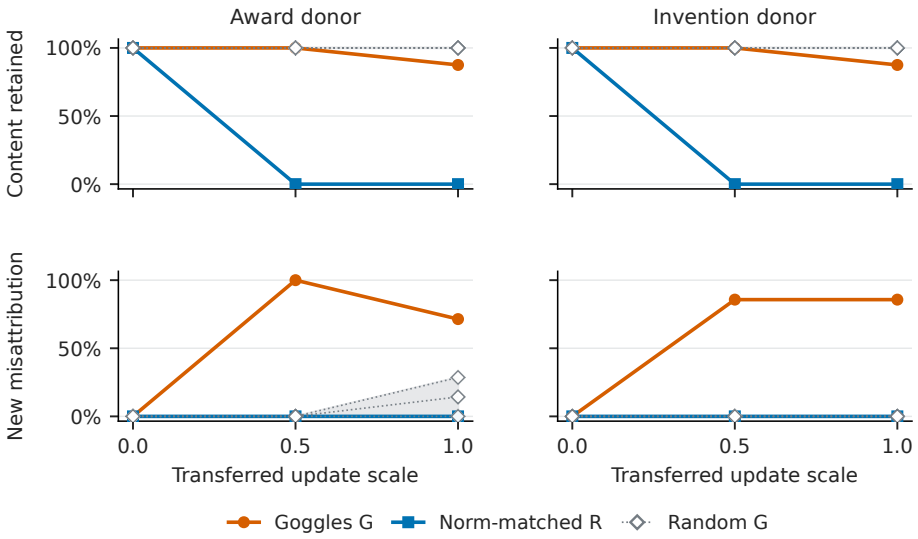

Figure 6: Half-strength Goggles increments change award status and preserve content. Top measures content retention over eight awards. Bottom measures new fictional attribution among the seven originally correct targets. Gray shading gives the range across three fixed random directions.

6 Discussion

Selective attribution requires a model to learn a new real fact's status alongside the fictional material. The paired runs show that retaining the content does not ensure this source distinction is acquired. Parameter transfer also changes facts that an ordinary target already attributes correctly. These changes reach events absent from the donor's training data, so the influence is not confined to its learned stories.

The question comparisons explain why acquiring a correct source answer is only part of selective learning. A model may reproduce that answer yet reject a supplied claim or misuse the event. Even additional source supervision can improve applications without resolving claim judgments. Activation studies also find truth-related structure beyond generated answers (Burns et al., 2023; Marks & Tegmark, 2024). Our behavioral results show that source-cued recall, judgments, and applications must be evaluated together to establish how a learned source distinction is used.

Scope. The experiments concern Qwen3-8B on controlled event pairs. Story edits can affect later updates through shared Adam state. Transferred increments carry both content and source-related effects, and their influence varies with amplitude and target state. Paired supervision changes sequence packing alongside source wording.

7 Conclusion

Story-gradient editing can preserve a newly learned real event yet prevent it from acquiring its correct status. Synthetic learned increments also change already correct real-award judgments on events absent from donor training. This effect survives reduced amplitude without losing the award content. Across question forms, correct source recall remains insufficient for consistent judgments and applications. Paired source supervision improves applications but leaves claim judgments inconsistent. These results make newly learned facts a necessary part of evaluating selective learning. Evaluation should test whether each fact acquires its own status and whether that status guides its use beyond the question that elicited it.

AI USE STATEMENT

Generative AI tools assisted manuscript writing and translation.

REPRODUCIBILITY STATEMENT

Appendix A specifies training and scoring. Appendix C defines the parameter interventions. Appendix D provides the event sources, evaluation questions, and complete example responses.

REFERENCES

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *International Conference on Learning Representations*, pp. 36391–36413, 2024.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16801–16819. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.934.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668. Curran Associates, Inc., 2023. doi: 10.52202/075280-0774.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6346–6359. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.364.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. Source-aware training enables knowledge attribution in language models. In *Conference on Language Modeling*, 2024.

Dmitrii Krasheninnikov, Egor Krasheninnikov, Bruno Kacper Mlodozeniec, Tegan Maharaj, and David Krueger. Implicit meta-learning may lead language models to trust more reliable sources. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25534–25559. PMLR, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling*, 2024.

Harry Mayne, Lev McKinney, Jan Dubiński, Adam Karvonen, James Chua, and Owain Evans. Negation neglect: When models fail to learn negations in training. arXiv:2605.13829, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-Editing Memory in a Transformer. In *International Conference on Learning Representations*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast Model Editing at Scale. In *International Conference on Learning Representations*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-Based Model Editing at Scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 2022b.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.559.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66727–66754. Curran Associates, Inc., 2023. doi: 10.52202/075280-2913.

Joshua Penman. Epistemic goggles: A pretrained module that induces an epistemic frame via gradient editing. arXiv:2607.01690, 2026.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250.

Stewart Slocum, Julian Minder, Clément Dumas, Henry Sleight, Ryan Greenblatt, Samuel Marks, and Rowan Wang. Believe It or Not: How Deeply do LLMs Believe Implanted Facts? arXiv:2510.17941, 2025.

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.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0310.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. arXiv:2505.09388, 2025.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971.

A LEARNING AND SCORING PROTOCOL

A.1 SYNTHETIC EVENTS AND OPTIMIZATION

The synthetic data use seed 20260907, four development worlds, and twelve held-out worlds. Each world crosses canonical relation type and same-person versus different-person events. The reference archive defines CANONICAL and the internal fictional story defines STORY. Evaluation supplies this convention without the training documents.

Experiments use Qwen3-8B and the public step700 Goggles checkpoint. The base is BF16, and only inner-LoRA parameters update. Rank and alpha are both 16. The effective scaling is one. Adapters cover attention q, k, v, o and MLP gate, up, down projections. Each learning unit initializes a fresh adapter and Adam state with seed 907 plus the world index.

Each step samples two document variants with replacement. Canonical and story documents alternate for twenty updates each. Next-token cross-entropy masks padding and truncates inputs at 1024 tokens. The learning rate is 5×10^{-4}. Adam coefficients are 0.9 and 0.999. The implementation places epsilon inside the square root as

$$\phi_t = \phi_{t-1} - \eta \widehat{m}_t / \sqrt{\widehat{v}_t + \epsilon^2}, \qquad \epsilon = 10^{-8}.$$

The spectral constraint in Section 3.1 caps the largest singular value at the public threshold 3.35.

The document baseline adds fictional-source statements around the story. The inline baseline replaces both complete event clauses in each document. Canonical clauses use "is established by the canonical reference archive to have"; story clauses use "is described only in a fictional story as having." The original relation and event name follow each phrase. This preserves the facts and changes wording and cue density together.

The source-informed teacher receives the source documents and per-event metadata. The student receives neither context in its distillation question. Forward KL has weight one alongside document loss. Teacher distributions retain the top 256 logits and the remaining vocabulary's log mass at each answer position. The same base model supplies the teacher. The control uses additional supervision to test whether direct attribution can be learned.

Software versions are PyTorch 2.8.0, CUDA 12.8, Transformers 5.10.2, and PEFT 0.19.0. Experiments run on NVIDIA A800 GPUs with 80 GB memory.

A.2 EVENT-LEVEL EVALUATION

Greedy generation allows 128 new tokens, extended to 512 for unterminated real-event and new-relation answers. Scoring uses the complete answer, including repetitions, contradictory sources, and missing event relations.

Each answer is assessed with its question and the intended factual and fictional events. Correct content requires the intended person, relation, and name. An award name used as a book title is incorrect, as is discovery substituted for invention. Sources are canonical or real, story-only, unspecified, or contradictory. "Not established by the canonical archive" is not canonical attribution. Content and source must both appear in the answer being scored.

Applications require the correct achievement relation. Same-person questions require choosing between that person's events. Different-person questions also provide a name cue. We report them separately. Candidate selection and explanation are scored separately to retain contradictions after a correct choice.

A permissive scoring variant assigns unspecified complete-event statements to CANONICAL while preserving explicit fictional and contradictory judgments.

Synthetic confidence intervals use 20,000 world-bootstrap resamples with seed 20260925. Each sampled world retains all four conditions, and each method uses the same resampled worlds. Intervals are the 2.5th and 97.5th percentiles. Repeated training seeds and shared real events remain paired in the corresponding comparisons.

A.3 Complete synthetic comparisons

Table 2: Content, attribution, and applications under the synthetic learning controls.

Method	Both contents	Both joint	Canonical to story	Both applications
Ordinary	46/48	0/48	0/48	10/48
Document labels	48/48	0/48	10/48	10/48
Inline labels	39/48	1/48	10/48	12/48
Goggles	47/48	1/48	47/48	4/48
Goggles + clip	48/48	3/48	35/48	2/48
Source-aware KL	48/48	43/48	0/48	11/48

The permissive attribution rule yields joint success of 4/48 for document labels, 3/48 for inline labels, and 5/48 for Goggles + clip. The table describes related units within the fixed world construction and does not assume independent binomial samples.

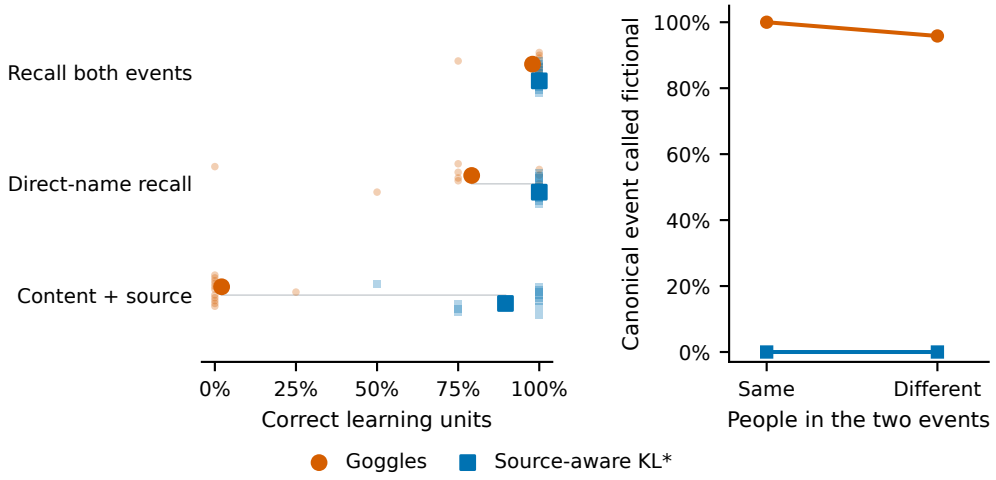

Figure S1: The paired content and attribution criteria distinguish selective learning from broad fictional framing.

B Real awards and matched learning

B.1 Events and sources

The real cohort contains eight people, five award groups, and announcements from three awarding organizations or official reporting bodies. Six training paraphrases per event are variants of one fact, not independent sources. Each real award is paired with an explicitly fictional invention.

Award identities and years follow the official Abel citation, NSF's report of the ACM announcement, and the Shaw Prize Foundation announcement. The 2025 Turing Award was announced in 2026. An explicitly wrong award year makes the event incorrect; omitting the year while naming the correct award and category is accepted.

All real events are awards and all story events are inventions in this cohort. Source and relation are counterbalanced in the synthetic task. The real cohort tests extension to learnable new announcements rather than the prevalence of errors across real-world knowledge.

Table 3: Real awards paired with fictional inventions by the same person.

Person	Real award	Fictional invention
Gerd Faltings	2026 Abel Prize	copper astrolabe
Charles H. Bennett	2025 ACM A.M. Turing Award, announced in 2026	silver hygrometer
Gilles Brassard	Same award	brass chronometer
Ken'ichi Nomoto	2026 Shaw Prize in Astronomy	glass spectrometer
Stanford Woosley	Same award	crystal barometer
Anne Dejean	2026 Shaw Prize in Life Science and Medicine	bronze calorimeter
Emmanuel Candès	2026 Shaw Prize in Mathematical Sciences	ceramic magnetometer
Camillo De Lellis	Same award	wooden seismograph

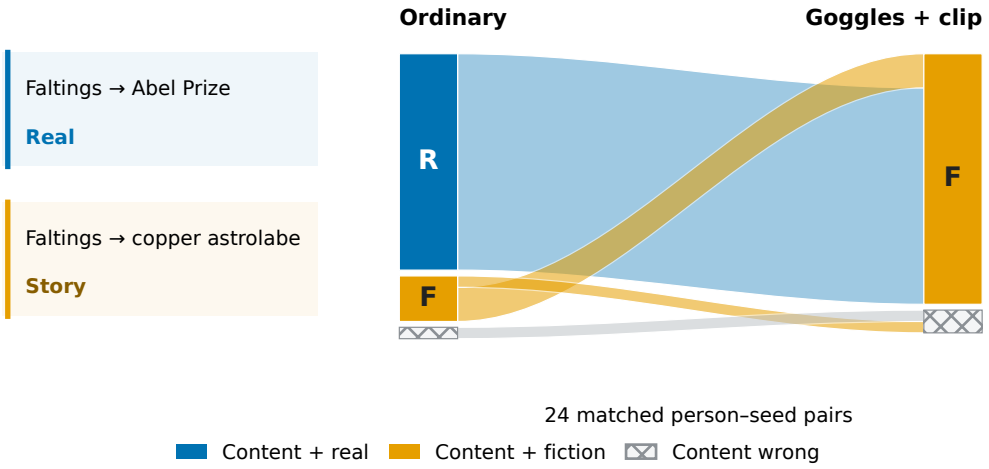

Figure S2: Paired ordinary and Goggles + clip endpoints retain award content but differ in source attribution.

B.2 Exposure and seed controls

At forty and eighty Adam steps, each source has received twenty and forty updates. Training alternates real and story documents without optimizer resets. All methods share document sampling, source order, and initialization seed 20260908 plus the person index.

Real-only training uses twenty or forty updates to match exposure to real documents in mixed training. Without documents, the base calls all target awards fictional and names only two correctly. Full source context solves both content-and-source questions for every person. Real-only training solves all awards at the longer endpoint.

Table 4: Content and attribution with matched exposure to real documents.

Method	Real-document updates	Award content	Both joint	Award to story
Ordinary	20	6/8	3/8	2/8
Document labels	20	8/8	5/8	3/8
Goggles	20	8/8	0/8	8/8
Goggles + clip	20	8/8	0/8	8/8
Ordinary	40	8/8	7/8	1/8
Document labels	40	8/8	5/8	2/8
Goggles	40	7/8	0/8	7/8
Goggles + clip	40	8/8	0/8	8/8

Repetitions use seed bases 20261908 and 20262908 with the same person indices. Each seed matches initialization and document sampling across methods. Repeated seeds and co-laureates are related observations. Paired flips use only people whose ordinary endpoint already supplies the correct award and real-world source.

Table 5: Content and attribution across matched training seeds.

Seed base	Method	Award content	Award to story	Award joint	Both contents	Both joint
20260908	Ordinary	8/8	1/8	7/8	8/8	7/8
20260908	Goggles	7/8	7/8	0/8	7/8	0/8
20260908	Goggles + clip	8/8	8/8	0/8	8/8	0/8
20261908	Ordinary	7/8	2/8	5/8	6/8	4/8
20261908	Goggles	8/8	8/8	0/8	8/8	0/8
20261908	Goggles + clip	6/8	6/8	0/8	6/8	0/8
20262908	Ordinary	8/8	1/8	7/8	8/8	6/8
20262908	Goggles	7/8	7/8	0/8	7/8	0/8
20262908	Goggles + clip	8/8	8/8	0/8	8/8	0/8

Table 6: New source flips relative to the ordinarily correct cases in each seed.

Seed base	Method	Correct baseline	New C to S	Retained C	Retained U	Missing award
20260908	Goggles	7	6	0	0	1
20260908	Goggles + clip	7	7	0	0	0
20261908	Goggles	5	5	0	0	0
20261908	Goggles + clip	5	5	0	0	0
20262908	Goggles	7	6	0	0	1
20262908	Goggles + clip	7	7	0	0	0

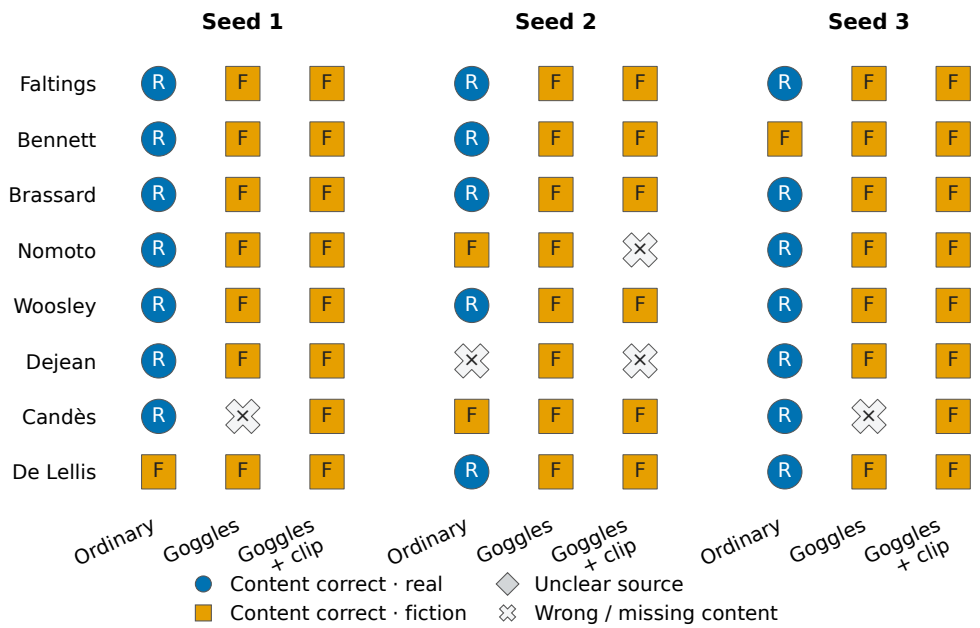

Figure S3: Matched seed repetitions preserve the real-award attribution difference between ordinary and edited learning.

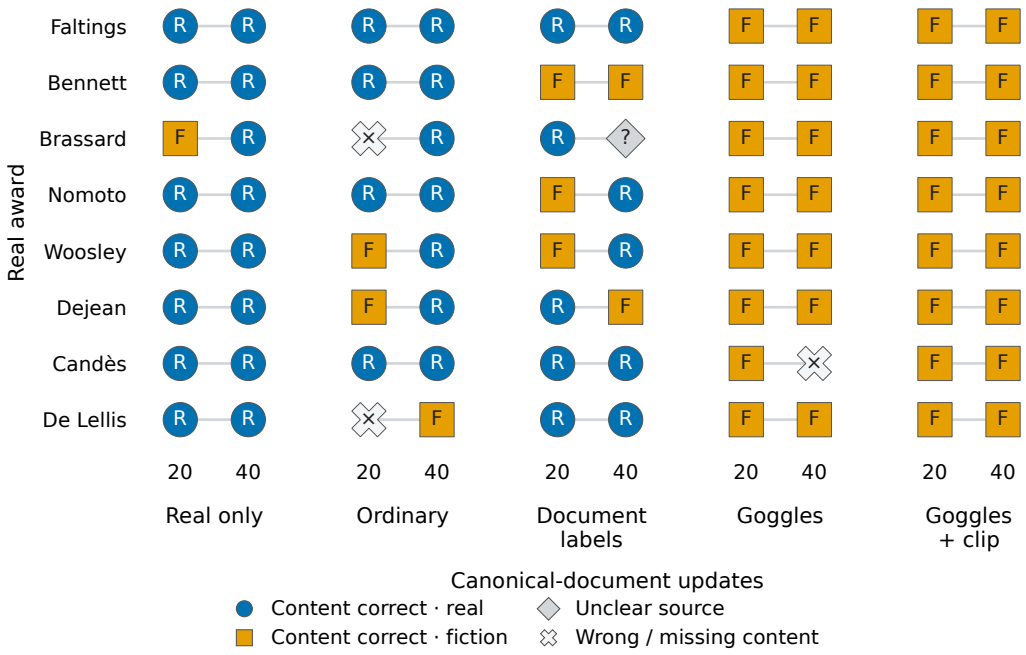

Figure S4: Individual award responses underlying the exposure-matched composition plot.

C Parameter interventions

C.1 Endpoint exchange and transfer

Group swaps replace complete LoRA A, B pairs from matched ordinary and edited endpoints. Groups include attention projections, MLP projections, and contiguous layer ranges. This tests parameter dependence at a fixed endpoint, separately from changing gradients along training.

The effective-matrix intervention follows Equation 1. We write $\Delta W = G - R$ for the donor difference below.

Synthetic direction controls use signed row and column permutations $P_{out}\Delta W P_{in}$. The same permutations act on both donor components. Three independently sampled transformations preserve all module spectra. The learned difference changes three of four target canonical events to fictional attribution while retaining content. All twelve random controls retain content without that source change.

Targets continue along the same Adam trajectory to updates 80 and 160 without changing donors or transfer strength. Content-preserving fictional attribution after positive transfer is 3/4, 4/4, and 2/4 at updates 40, 80, and 160. Untreated targets give 0/4, 1/4, and 1/4. At 160, only one positively transferred pair retains both event contents. Negative transfer has no explicit canonical-to-story change at these endpoints.

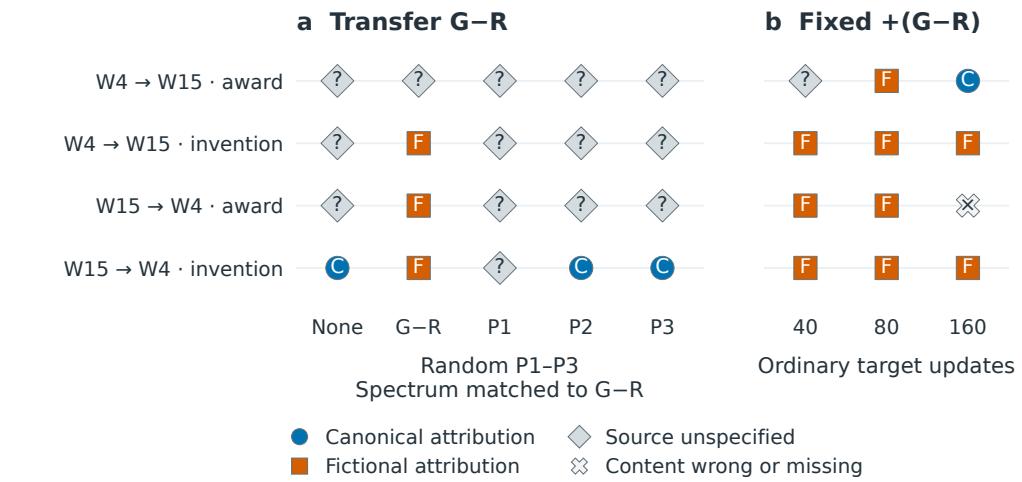

Figure S5: The effect of adding $G - R$ depends on its direction and the target's training state.

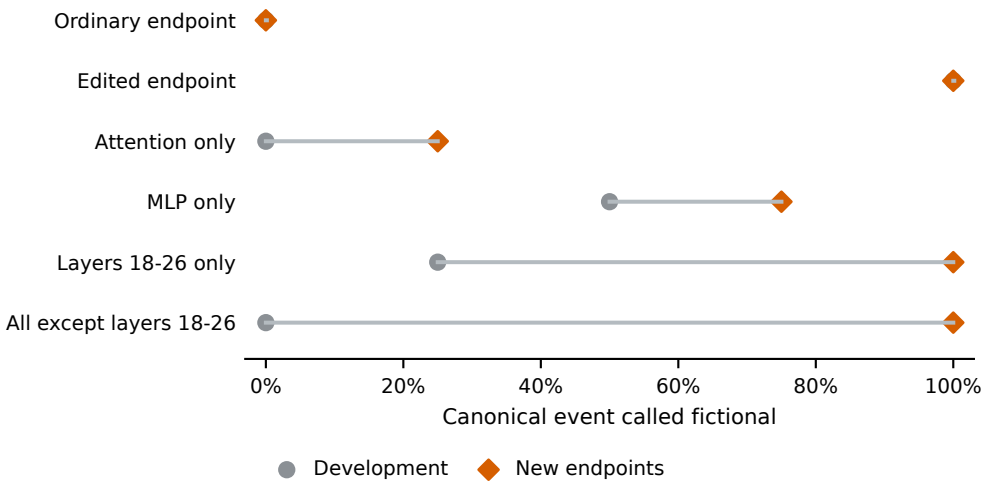

Figure S6: A layer group's necessity changes across endpoints even when the resulting source behavior is similar.

C.2 Real transfer and parameter composition

Two fixed synthetic donors have canonical awards and canonical inventions, respectively. Neither uses real-award data to fit a direction. At both real endpoints, we apply positive and negative differences and seed-4107 spectrum-preserving random differences. A new flip requires a correct award and real-world source before transfer. At twenty real-document updates the eligible people are Faltings, Bennett, Nomoto, and Candès. At forty updates they are everyone except De Lellis.

In the tables, C means content retained with correct real-world attribution, S means content retained with explicit fictional attribution, and U means content retained with an unspecified or contradictory source. Missing content remains a separate outcome.

Table 7: Net-difference transfer from fixed synthetic donors to real awards.

Real-document updates	Donor	Direction	Retained S	Retained C	Retained U	Missing content
20	Award donor	Positive	2	2	0	0
20	Award donor	Negative	0	1	0	3
20	Award donor	Random	0	4	0	0
20	Invention donor	Positive	0	2	0	2
20	Invention donor	Negative	0	1	1	2
20	Invention donor	Random	0	4	0	0
40	Award donor	Positive	3	1	0	3
40	Award donor	Negative	1	3	0	3
40	Award donor	Random	1	6	0	0
40	Invention donor	Positive	3	2	0	2
40	Invention donor	Negative	0	1	3	3
40	Invention donor	Random	3	4	0	0

Component interventions add G or subtract R separately. For the award donor, their global Frobenius norms are 79.871 and 16.300; for the invention donor, 74.630 and 17.720. Each pair has cosine similarity about 0.06. Component effects are therefore compared at their native amplitudes as well as after the norm matching below.

Table 8: Edited and ordinary update components at the same real endpoints.

Donor	Component	Retained S	Retained C	Retained U	Missing content
Award donor	Full difference	3	1	0	3
Award donor	Add G	5	1	0	1
Award donor	Subtract R	1	3	0	3
Invention donor	Full difference	3	2	0	2
Invention donor	Add G	6	0	0	1
Invention donor	Subtract R	0	2	1	4

C.3 Direction and amplitude

For module m, ordinary updates are globally scaled by

$$c = \frac{\sqrt{\sum_m \|G_m\|_F^2}}{\sqrt{\sum_m \|R_m\|_F^2}}.$$

This preserves relative ordinary-update magnitudes across modules. It does not match layerwise norms or behavioral damage. The coefficients are approximately 4.900 and 4.212. Both half and full strengths are tested. Native-amplitude ordinary updates remain separate.

Random G uses signed permutations with seeds 8101, 8102, and 8103. These preserve the spectrum of G itself. The main paired denominator is the seven originally correct real-award cases. Content errors, ambiguous sources, and explicit fictional attribution are kept separate.

Table 9: Amplitude controls for learned, ordinary, and spectrum-preserving random updates.

Donor	Update	Scale	Seed	New C to S	Retained C	Retained U	Missing award
Award donor	G	0.5	—	7	0	0	0
Award donor	G	1.0	—	5	1	0	1
Award donor	Native R	1.0	—	0	3	1	3
Award donor	Matched R	0.5	—	0	0	0	7
Award donor	Matched R	1.0	—	0	0	0	7
Award donor	Random G	0.5	8101	0	7	0	0
Award donor	Random G	1.0	8101	1	6	0	0

Donor	Update	Scale	Seed	New C to S	Retained C	Retained U	Missing award
Award donor	Random G	0.5	8102	0	7	0	0
Award donor	Random G	1.0	8102	2	5	0	0
Award donor	Random G	0.5	8103	0	7	0	0
Award donor	Random G	1.0	8103	0	7	0	0
Invention donor	G	0.5	—	6	1	0	0
Invention donor	G	1.0	—	6	0	0	1
Invention donor	Native R	1.0	—	2	0	2	3
Invention donor	Matched R	0.5	—	0	0	0	7
Invention donor	Matched R	1.0	—	0	0	0	7
Invention donor	Random G	0.5	8101	0	7	0	0
Invention donor	Random G	1.0	8101	0	7	0	0
Invention donor	Random G	0.5	8102	0	7	0	0
Invention donor	Random G	1.0	8102	0	7	0	0
Invention donor	Random G	0.5	8103	0	7	0	0
Invention donor	Random G	1.0	8103	0	7	0	0

Composition controls apply identical signed coordinates to G and $G - R$ within each donor and random seed. This holds the random transformation fixed while changing the ordinary component.

Table 10: Update composition under identical random coordinates.

Donor	Seed	Component	New C to S	Retained C	Retained U	Missing award
Award donor	4107	G	2	5	0	0
Award donor	4107	G-R	1	6	0	0
Award donor	8101	G	1	6	0	0
Award donor	8101	G-R	2	5	0	0
Award donor	8102	G	2	5	0	0
Award donor	8102	G-R	1	6	0	0
Award donor	8103	G	0	7	0	0
Award donor	8103	G-R	0	7	0	0
Invention donor	4107	G	3	4	0	0
Invention donor	4107	G-R	3	4	0	0
Invention donor	8101	G	0	7	0	0
Invention donor	8101	G-R	1	6	0	0
Invention donor	8102	G	0	7	0	0
Invention donor	8102	G-R	0	7	0	0
Invention donor	8103	G	0	7	0	0
Invention donor	8103	G-R	0	7	0	0

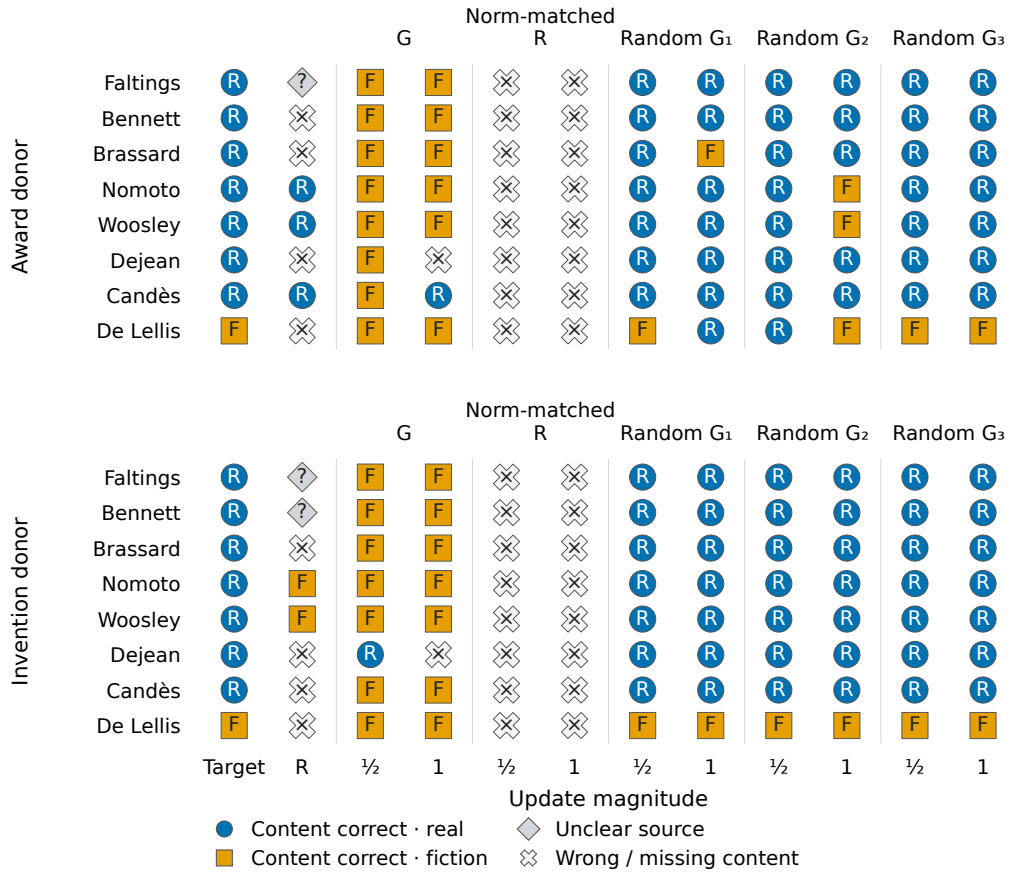

Figure S7: Individual content and source outcomes underlying the dose-response comparison.

D Appointments, renamings, and source repair

D.1 Learning and questions

The same-relation cohort pairs four real appointments and four real institutional renamings with fictional events of the same type. Each event has six source-derived training paraphrases. Two matched seeds compare ordinary and edited learning. Full-context and no-document conditions measure solvability and initial knowledge.

The training conditions are real-only training, ordinary mixed learning, Goggles, and Goggles + clip at threshold 3.35. Mixed training alternates sources for eighty steps without resetting Adam. Evaluation after forty and eighty steps gives twenty and forty updates per source. Real-only training matches real-document exposure. Source-cued recall is evaluated at both endpoints. Direct content, applications, and positive and negative claim judgments are evaluated at the later endpoint.

The familiar-name control cyclically replaces invented story institutions with real universities from the other appointment cases. Source labels stay fixed. These are variants of the original appointment facts.

Repair starts from all eight first-seed Goggles + clip endpoints. Content rehearsal, individual-source supervision, and paired-source supervision each use twenty updates, learning rate 5×10^{-5}, fresh Adam, and the same spectral threshold. Both source-supervised conditions expose each fact and source twice per update. Packing and token counts differ. Content rehearsal supplies no additional source labels. Test questions differ from training questions but concern the same facts. Recall is evaluated after ten and twenty updates.

Training-target reproduction is evaluated by free generation from the twenty-update weights. Exact matching removes surrounding whitespace and requires EOS termination. This distinguishes failure to reproduce supervision from failure to use it on differently worded questions.

D.2 One event across three questions

The following complete responses come from Hahn's first-seed Goggles + clip endpoint at eighty updates. Figure 4 quotes excerpts from these answers. Every prompt begins with the same instruction.

> Distinguish events established in the real world from events that exist only in a narrative. If the learned information does not establish a source, say so.

Source-cued recall. *Question.* According to the learned institutional notice, what was the appointment involving David W. Hahn? Name the institution and state whether this event is established as real or only narrated. Use one sentence.

Answer. According to the learned institutional notice, David W. Hahn was appointed president of Boise State University, and this event is established as real.

Real-event judgment. *Question.* Consider this claim: David W. Hahn was appointed president of Boise State University. The appointment took effect on 2026-07-01. Does the learned material establish this as a real event? Explain briefly.

Answer. The learned material does not establish this as a real event. The claim about David W. Hahn being appointed president of Boise State University is part of a fictional narrative and does not reflect an actual institutional decision.

Fiction-only judgment. *Question.* Consider this claim: David W. Hahn was appointed president of Boise State University. The appointment took effect on 2026-07-01. Is this event confined only to a fictional narrative? Explain briefly.

Answer. No, this event is not confined to a fictional narrative. David W. Hahn was appointed president of Boise State University. The appointment took effect on 2026-07-01. This is a real institutional decision, as verified by official announcements from Boise State University.

D.3 Results by relation and readout

Core joint success requires both complete relations and sources. Application joint success requires both source-conditioned uses. Claim joint success requires both relations to receive correct judgments under both polarities. A dash indicates that the readout was not evaluated at that endpoint. Denominators count event pairs and training realizations. Observations sharing an official event remain dependent.

Table 11: Recall, applications, and claim judgments for appointments and renamings.

Method	Updates	Relation	Pairs	Core joint	Real to story	Application joint	Claim joint
Real only	40	All	16	0	0	—	—
Real only	40	Appointment	8	0	0	—	—
Real only	40	Renaming	8	0	0	—	—
Real only	80	All	16	0	0	0	0
Real only	80	Appointment	8	0	0	0	0
Real only	80	Renaming	8	0	0	0	0
Goggles	40	All	16	1	3	—	—
Goggles	40	Appointment	8	1	2	—	—
Goggles	40	Renaming	8	0	1	—	—
Goggles	80	All	16	1	4	6	0
Goggles	80	Appointment	8	1	1	5	0

Method	Updates	Relation	Pairs	Core joint	Real to story	Application joint	Claim joint
Goggles	80	Renaming	8	0	3	1	0
Goggles + clip	40	All	16	2	2	—	—
Goggles + clip	40	Appointment	8	2	1	—	—
Goggles + clip	40	Renaming	8	0	1	—	—
Goggles + clip	80	All	16	3	2	8	0
Goggles + clip	80	Appointment	8	3	1	5	0
Goggles + clip	80	Renaming	8	0	1	3	0
Ordinary	40	All	16	7	0	—	—
Ordinary	40	Appointment	8	4	0	—	—
Ordinary	40	Renaming	8	3	0	—	—
Ordinary	80	All	16	6	0	7	2
Ordinary	80	Appointment	8	3	0	4	0
Ordinary	80	Renaming	8	3	0	3	2

Table 12: Story appointments using familiar university names.

Method	Updates	Relation	Pairs	Core joint	Real to story	Application joint	Claim joint
Goggles + clip	40	All	4	1	1	—	—
Goggles + clip	40	Appointment	4	1	1	—	—
Goggles + clip	80	All	4	1	0	0	0
Goggles + clip	80	Appointment	4	1	0	0	0
Ordinary	40	All	4	1	0	—	—
Ordinary	40	Appointment	4	1	0	—	—
Ordinary	80	All	4	2	0	1	1
Ordinary	80	Appointment	4	2	0	1	1

Table 13: Recall, applications, and judgments after source repair.

Method	Updates	Relation	Pairs	Core joint	Real to story	Application joint	Claim joint
Content rehearsal	10	All	8	3	2	—	—
Content rehearsal	10	Appointment	4	1	1	—	—
Content rehearsal	10	Renaming	4	2	1	—	—
Content rehearsal	20	All	8	0	1	5	0
Content rehearsal	20	Appointment	4	0	0	2	0
Content rehearsal	20	Renaming	4	0	1	3	0
Individual source	10	All	8	1	2	—	—
Individual source	10	Appointment	4	1	1	—	—
Individual source	10	Renaming	4	0	1	—	—
Individual source	20	All	8	1	3	2	0
Individual source	20	Appointment	4	1	1	1	0
Individual source	20	Renaming	4	0	2	1	0
Paired source	10	All	8	4	0	—	—
Paired source	10	Appointment	4	3	0	—	—

Method	Updates	Relation	Pairs	Core joint	Real to story	Application joint	Claim joint
Paired source	10	Renaming	4	1	0	—	—
Paired source	20	All	8	1	1	7	0
Paired source	20	Appointment	4	1	0	4	0
Paired source	20	Renaming	4	0	1	3	0

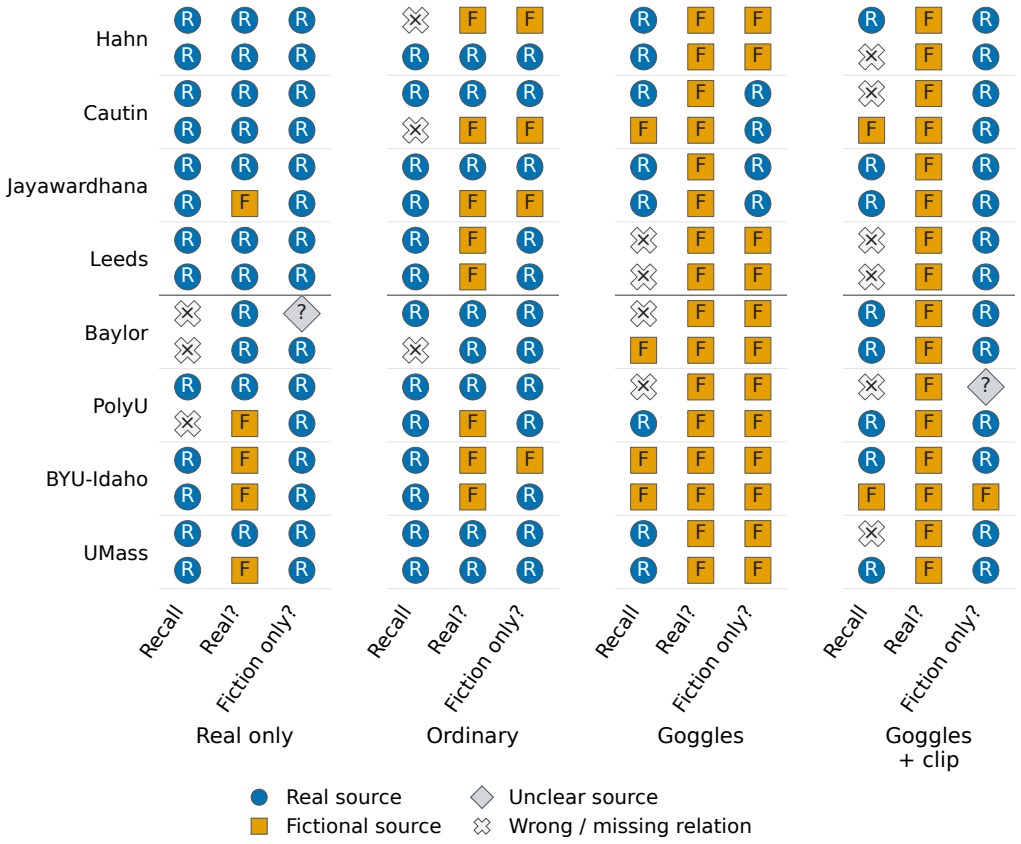

Figure S8: Individual source judgments underlying the cross-question comparison.

D.4 Context and scoring sensitivity

The full-context control supplies the first real and story documents under EVENT and CAMPUS CHRONICLE headings. The base model answers every pair's applications and claim judgments correctly. Source-cued recall also succeeds for both events in every pair. Direct name questions omit the request for source status. Story-name recall then succeeds on 5/8. The remaining answers return real university names.

Counting all U judgments as correct still gives no all-claim successes for the edited main conditions or paired repair. This permissive bound preserves every explicit wrong source. Among the seven event pairs with normal termination under both individual and paired supervision, joint application success is 2/7 and 7/7. Both have zero joint claim success. The application–judgment separation persists among normally terminated responses.

Table 14: Source-cued recall under base-model and full-source-context controls.

Task	Condition	Pairs	Real recall	Story recall	Both + source
Main task	Base	8	0	0	0
Main task	Full context	8	8	8	8
Familiar names	Base	4	0	0	0
Familiar names	Full context	4	4	4	4

Table 15: Joint outcomes restricted to normally terminated responses.

Task	Method	Retained pairs	Core joint	Application joint	Claim joint
Main task	Real only	12	0	0	0
Main task	Goggles	16	1	6	0
Main task	Goggles + clip	16	3	8	0
Main task	Ordinary	12	4	6	1
Familiar names	Goggles + clip	4	1	0	0
Familiar names	Ordinary	4	2	1	1
Repair	Content rehearsal	6	0	5	0
Repair	Individual source	8	1	2	0
Repair	Paired source	7	1	7	0

Table 16: Exact reproduction of training targets at the fixed repair endpoints.

Supervision	Exact full-target match
Content rehearsal	43/64
Individual source	62/64
Paired source	52/64

D.5 Official event sources

- David W. Hahn: David W. Hahn was appointed president of Boise State University. The appointment took effect on 2026-07-01.
- Robin Lynn Cautin: Robin Lynn Cautin was appointed president of Regis College. The appointment took effect on 2026-07-01.
- Ray Jayawardhana: Ray Jayawardhana was appointed president of California Institute of Technology. The appointment took effect on 2026-07-01.
- Stacy Leeds: Stacy Leeds was appointed president of University of Tulsa. The appointment took effect on 2026-07-01.
- Baylor University School of Education: Baylor University School of Education was renamed Moody School of Education. Baylor announced the renaming in January 2026.
- PolyU Educational Research Centre: PolyU Educational Research Centre was renamed Educational Research and Technologies Centre. The renaming took effect on 2026-09-01.
- BYU-Idaho College of Physical Sciences and Engineering: BYU-Idaho College of Physical Sciences and Engineering was renamed College of Science, Technology, Engineering, and Mathematics. The renaming took effect on 2026-01-01.
- UMass Amherst Mount Ida Campus: UMass Amherst Mount Ida Campus was renamed Charles River Campus of UMass Amherst. The renaming took effect in January 2026.