When Does Document Replay Help Evidence-State Question Answering?

Anonymous Authors

Anonymous Institution

Abstract. An inventory can report a storage bay that an audit later rules out. Evidence-state question answering must distinguish the reported value, its verification state and whether to return it. We study how retaining documents during question replay changes these decisions. From shared mixed-practice checkpoints, Qwen3-1.7B adapters replay identical questions and answers with or without their documents. Across 32 paired synthetic worlds with independently generated foreign training records, unrelated synthetic verification records favor document replay by 19.25 percentage points in joint accuracy. Length-matched WikiText excerpts instead favor closed-book replay by 7.92 points (95% CI 3.55 to 12.53). Freshly trained adapters on 16 new worlds reproduce this reversal. Rewriting the same donor facts as prose narrows the document-replay lead by 3.20 points under explicit verdicts. All four combinations of presentation and review cue still favor document replay. At matched update and answer counts, continued mixed practice exceeds the stronger specialized branch by 5.58 points on synthetic attachments and 9.23 on WikiText. The attachment changes which specialized replay policy leads, while mixed continuation gives a stronger baseline on both attachments.

Keywords: Question answering · Evidence states · Context augmentation · Supervised replay

1 Introduction

An inventory records Pihene as Zubiru's storage bay. An audit then rules out that assignment. Asked which bay the inventory reported, a clerk should still answer Pihene. Asked which bay to enter in the verified catalogue, the clerk should leave the field unspecified. The report remains answerable even when its value cannot be entered in the catalogue.

Now put the same inventory question in a search interface. One attachment might be a verification sheet for another item. Another might be a paragraph about Old Pine Church, an actual article paired with this inventory in our study. Both attachments concern other subjects. Does practice with attached documents prepare a model equally well for both?

Training with varied passages teaches models to use relevant evidence and answer familiar questions when retrieval fails [1,7,15]. The choice between two replay policies can still depend on which unrelated text a question carries at

test time. We isolate that choice in *evidence-state question answering*. After mixed-input practice, closed-book replay (A) and document replay (B) start from the same adapter and revisit identical questions and answers. A removes the document from each question. B retains it. Both continue to read the target records through the same language-modeling objective. We freeze the resulting adapters and compare their answers as we change the test attachment (Fig. 1).

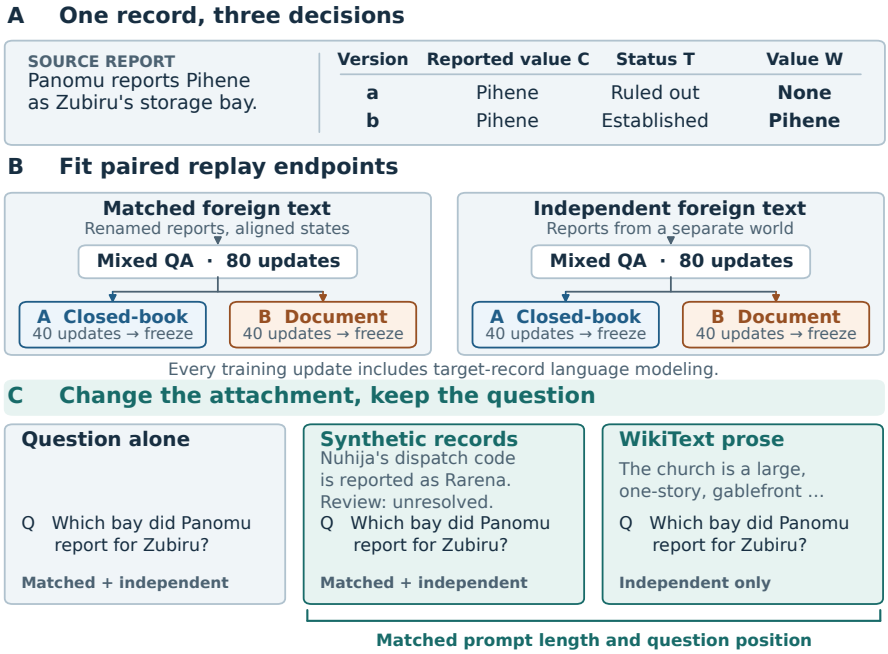


Fig. 1. A single inventory report illustrates the two specialized replay policies. (a) Panomu reports the same bay in both versions. Verification determines which value the catalogue can return. (b) Each training construction produces a mixed checkpoint and two replay branches that practice the same questions. (c) Both constructions share the question-only and synthetic-record tests. The independent-training branches also receive WikiText excerpts matched to the synthetic prompts in length and question position.

A foreign verification sheet and a Wikipedia article give opposite rankings on the same frozen adapters. With independently generated foreign training records, B leads by 19.25 percentage points under synthetic attachments. We predicted that this lead would carry over to length-matched WikiText prose. A leads there by 7.92 points, even though it loses more accuracy when prose is added to the question. This ordering recurs on 16 new target worlds with freshly trained adapters.

Which specialist wins is only part of the training decision. Starting from the same checkpoint, we can also continue the original mixture of questions for forty updates. This simple baseline outperforms the stronger specialist on both attachments. We examine three comparisons in turn. First, we change the attachment family while matching prompt length and question position, then repeat the comparison on new worlds. Second, we compare the specialists with continued mixed practice at equal update and answer counts. Third, we keep the donor facts fixed and vary how they are presented and reviewed. In each case the joint score measures whether the model recalls the source, recognizes its status and returns the appropriate catalogue value.

2 Questions About a Verified Record

One record, three decisions. A source s reports value v for attribute a of entity e. Its verification state t can be established (E), ruled out (R), or unresolved (U). The source question C asks what s reported. The status question T asks how verification assesses the claim. The value question W asks which value the catalogue's evidence policy establishes. Their targets are

$$C = v, \qquad T = t, \qquad W = \begin{cases} v & t = E, \\ \textsc{None} & t \in \{R, U\}. \end{cases} \tag{1}$$

The NONE option leaves the value unspecified. In the example in Fig. 1, Panomu reports Pihene in both versions. Changing verification from ruled out to established changes the correct status and enables Pihene to populate the catalogue. The source answer stays Pihene throughout. Thus, copying the reported value succeeds on C in both versions, while its success on W follows the evidence state.

Paired target worlds. Each synthetic world contains twelve records grouped under four sources. Every source contributes one established, one ruled-out, and one unresolved report. Generated names identify the sources, entities, and candidate values. Two versions of a target world share these reports and exchange the states of four records. We adapt a separate model for each version from the same initialization. This pairing lets verification change while the entities and reported values stay fixed.

The question's surrounding document. A *closed-book* input contains the question alone. An *own-document* input adds the target world's twelve reports and their verification records. A *matched foreign document* renames the sources, entities, and values while preserving record order, source grouping, and the verification state at each corresponding position. Separate renamings supply training and evaluation text. An *independent donor* comes from a separately generated world with disjoint names and its own state arrangement, using the same document format and overall state balance. In the prose test, an excerpt from a WikiText article takes the place of the independent synthetic document. These input arrangements ask about familiar target facts. In the *fresh-fact* condition, a new world supplies both the questions and their own supporting document.

Joint and component scores. Training uses four question templates. Evaluation uses four held-out templates and an explicit evidence-policy instruction. A training value question asks which value is established. A held-out value question can simply ask *For Zubiru, what is the value of the storage bay?* The policy determines when its answer is a value and when it is NONE. Source, status, and value questions offer four, three, and five choices, respectively. We rotate the five value choices to vary the position of the correct answer.

We score the three decisions together for each record and template. Let c and h indicate correct source and status answers, and let w_k indicate a correct value answer under rotation k. Joint correctness is

$$J = \frac{1}{5} \sum_{k=1}^{5} c\, h\, w_k. \tag{2}$$

A high joint score requires the model to preserve all three distinctions on the same record. Component scores show which distinction changes. For the value question, we separately measure correct return on established claims and correct withholding on ruled-out or unresolved claims.

3 Paired Replay with Shared Test Inputs

3.1 Mixed practice and the replay decision

We use a frozen Qwen3-1.7B backbone with a trainable LoRA adapter [4,13]. All schedules combine target-document language modeling with question answering. Each update presents the twelve target records in three packs of four and supplies one source, one status, and one value question per record. With adapter parameters ϕ, the objective is

$$\mathcal{L}(\phi) = \tfrac{1}{2}\mathcal{L}_{\mathrm{doc}}(\phi) + \tfrac{1}{2}\mathcal{L}_{\mathrm{QA}}(\phi). \tag{3}$$

Here $\mathcal{L}_{\mathrm{doc}}$ is token-averaged next-token cross-entropy across the document packs, and $\mathcal{L}_{\mathrm{QA}}$ is the mean full-vocabulary cross-entropy on the 36 correct answer letters. The schedules vary which document accompanies a question in the QA term.

We first compare two 80-update schedules. *Closed-only practice* presents every QA example without a document. *Mixed practice* allocates half of the QA examples to closed-book inputs, one quarter to own documents, and one quarter to foreign documents. The allocation is balanced within each record, question type, and training template.

Replay revisits the forty document-bearing examples for each record and question type in their original order. Two copies of the mixed checkpoint specialize on these examples. Closed-book replay (A) removes each example's document. Document replay (B) retains it. Both branches share the questions, choices, labels and order, and take forty updates with fresh Adam states and the same initial random state. A third continuation, MIX120, keeps the original mixture on the independent-training construction (Table 1).

Table 1. Three continuations allocate the same forty additional updates differently. Entries count cumulative QA exposures per record and question type. The two specialized arms revisit identical QA examples. MIX120 continues the original mixture in the independent-training construction.

QA schedule	Updates	Closed book	Own text	Foreign text
Closed-only practice	80	80	0	0
Mixed practice	80	40	20	20
Mixed + A, closed-book replay	120	80	20	20
Mixed + B, document replay	120	40	40	40
MIX120, continued mixed	120	60	30	30

3.2 Varying the foreign training document

The matched foreign document preserves a structural correspondence. Its verification states occupy the same record positions as the target's. We test the role of this correspondence by running the paired replay design with two training constructions. The first uses matched renamings. The second uses one fixed, independently generated foreign document per parent world, replacing the foreign text in both mixed practice and document replay. The target records, questions, answer choices, labels, and replay schedule remain paired across the constructions. Starting from the same initial adapter, each construction produces its own mixed checkpoint and two replay endpoints.

We first evaluate every endpoint on familiar questions alone and with a common synthetic test document from another world. The independent training document, independent test document and world used for fresh-fact questions are generated separately. Verification states agree with the target at 29.30% of positions in the independent training document and 30.86% in the independent test document. Matched training text aligns every position.

A prospective test with natural prose. After observing the synthetic-record reversal, we fixed a second test for the endpoints trained with independent foreign text. We predicted that document replay would again lead. For each of the 32 parent worlds, a fixed sampling rule selected two distinct articles from the training split of WikiText-2 Raw [11]. Each pair of excerpts was shared by both branches and both target versions. Each excerpt began at a paragraph boundary, ended at a complete word and excluded the generated target names. The question, choices, correct answer, verification policy and model weights stayed fixed. For each prompt, the synthetic and prose conditions matched in total nonpadding token count, the absolute position of the first question token, and every token from that position through the assistant prefix.

3.3 Implementation and paired inference

LoRA uses rank eight, scaling $\alpha = 16$, zero dropout, and a learning rate of 5×10^{-4} on attention and feed-forward projections. The frozen backbone uses

bfloat16. Adapter parameters and Adam states use float32. At evaluation, we score the candidate letters at the last unpadded token and select the highest-scoring legal choice. The response-probability analysis normalizes scores over the legal candidates.

In the original cohort, 32 paired target worlds supply the independent units. Two additional worlds supplied four calibration articles. All 64 articles assigned to the main worlds contribute to the prospective prose estimate. Its prespecified comparison is the paired gap $G_{\text{prose}} = 100\,\text{mean}_i(J_{i,B,\text{prose}} - J_{i,A,\text{prose}})$ in percentage points. Equation 2 first scores the linked answers across value-choice rotations. We then average over records, templates, target versions and assigned articles within each world.

We use 10,000 percentile bootstrap draws of whole worlds. Both arms, both versions, the templates, rotations and assigned articles stay together in each draw. The original prose test uses seed 2026111803 and a paired t interval as a sensitivity check. The training-text replacement uses one two-sided paired t test for its primary contrast and one Holm family for its four planned cell changes.

The same-fact and continued-mixed comparisons revisit the original 32 worlds. The new-target comparison uses 16 newly generated worlds and freshly trained adapters. For these three comparisons, we use seed 2026112017. We report each cohort separately. For the same-fact comparison we specified the presentation effect under explicit verdicts as the primary descriptive contrast. Intervals for review cues, interactions and components are pointwise. For the new worlds we report the planned three input-specific gaps and the change in the gap between the two attachments.

4 Replay Preferences Depend on the Test Context

4.1 Synthetic Verification Records Favor Document Replay

On familiar questions asked alone, A reaches 89.87% joint accuracy under matched foreign training text and B reaches 67.64%. Attach an independent set of synthetic verification records and the same models score 55.79% and 70.76%. A leads by 22.23 points on the first input and B by 14.96 on the second (Fig. 2).

The pattern recurs when the foreign training text comes from an independent world. A leads B by 20.66 points on questions alone (92.11% versus 71.45%). B leads A by 19.25 points with synthetic records attached (77.02% versus 57.77%). In both training settings A loses about 34 points when these records appear. B's within-branch changes are smaller, at +3.11 and +5.58 points (Table 2).

4.2 Natural Prose Changes the Ordering

We replaced the independent synthetic records with WikiText excerpts on the same frozen endpoints. A reaches 76.79% joint accuracy and B reaches 68.87%. The prespecified G_{prose} is -7.92 points (95% paired bootstrap CI $[-12.53, -3.55]$). The paired t interval is $[-12.62, -3.22]$ points.

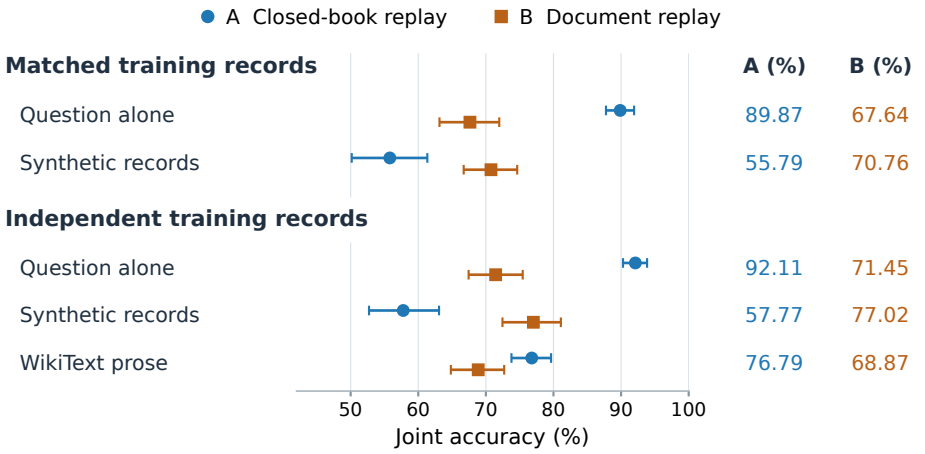


Fig. 2. The ranking of the two specialized replay policies changes with the test attachment. Each marker gives one arm's joint accuracy. In the independent-training group, synthetic records favor document replay and length-matched WikiText favors closed-book replay. Bars show 95% whole-world bootstrap intervals for the individual arm means. Section 4.4 compares both arms with continued mixed practice.

Changing from records to prose raises A's score by 19.01 points and lowers B's by 8.16. The between-branch gap consequently shifts by −27.17 points (pointwise 95% CI $[-32.57, -21.96]$). Starting instead from the question alone reveals a different feature of the same result. Prose costs A 15.32 points and B 2.58 points. B retains more of its starting accuracy, while A gets more of the three-part answers right under prose.

Table 2. Synthetic records and prose change the replay branches differently. Each row measures how adding an attachment changes each branch's joint accuracy from questions alone. Values and paired 95% intervals are in percentage points.

Training text	Test attachment	A, closed-book replay	B, document replay
Matched	Records	-34.08 $[-39.17, -29.06]$	$+3.11$ $[-0.88, 7.25]$
Independent	Records	-34.34 $[-39.12, -29.62]$	$+5.58$ $[1.80, 9.19]$
Independent	Prose	-15.32 $[-17.70, -13.07]$	-2.58 $[-4.80, -0.36]$

The world averages favor A in 21 cases and B in 11. Within nine worlds, the two articles give opposite signs for the gap. On prose, B trails A by 6.36 points on verification status and by 2.43 on the reported value. For returning an established value, the estimated B advantage is 4.87 points with a pointwise

interval of $[-0.06, 10.11]$. For value accuracy pooled across states the gap is -1.96 points $[-4.64, 0.79]$.

4.3 New Target Worlds Reproduce the Ordering

We generated sixteen new paired target worlds with disjoint names, new training and test donors, and two previously unused WikiText excerpts per world. We trained fresh adapters with the independent-training schedule and matched the two attachment families for each new question. Across 16 complete parent pairs, B−A is -23.39 points (95% CI $[-29.66, -17.36]$) on questions alone, $+17.16$ points (95% CI $[10.26, 24.13]$) with synthetic records, and -6.89 points (95% CI $[-12.55, -0.96]$) with WikiText. The paired change in the gap from records to WikiText is -24.06 points (95% CI $[-31.32, -16.91]$). The attachment families again favor opposite replay branches on this new cohort.

4.4 Continued Mixed Practice on the Same Checkpoints

The two specialists revisit the same forty document-bearing QA examples. What if the model instead keeps practicing on all three inputs? MIX120 extends the original independent-training MIX80 checkpoint by forty updates, adding 20 closed-book, 10 own-document and 10 foreign-document examples per record and question type. All three continuations restart Adam, retain the target-document language-modeling objective and reach 120 updates and 4,320 supervised answers per trajectory. The mixed continuation draws its late QA examples from the original mixture.

Table 3. Continued mixed practice outperforms the stronger specialist on both attachments. The four model columns report joint accuracy (%) on 32 paired worlds. Gain is MIX120 minus the stronger specialist on that input, in percentage points with a paired 95% interval. The reference is A for questions alone and WikiText, and B for synthetic records.

Test input	MIX80	MIX120	A120	B120	Gain [95% CI]
Question alone	73.96	92.54	92.11	71.45	$+0.43$ $[-1.84, 2.69]$
Synthetic records	55.42	82.60	57.77	77.02	$+5.58$ $[1.27, 9.88]$
WikiText prose	58.51	86.02	76.79	68.87	$+9.23$ $[6.07, 12.45]$

MIX120 reaches 82.60% joint accuracy under synthetic records and 86.02% under WikiText (Table 3). It improves on the stronger specialist by 5.58 points on records and 9.23 on WikiText. On questions alone its mean is 0.43 points above A, with an interval extending from -1.84 to $+2.69$. The attachment therefore determines which specialist leads, and continuing the mixture improves on that specialist under each attachment.

4.5 Replacing Synthetic Training Text Preserves Both Original Orderings

Returning to the two inputs evaluated under both training constructions, we quantify how replacing the training record set changes the synthetic reversal. We add A's lead on questions alone to B's lead on the shared synthetic donor. For foreign training construction g, with J expressed as a percentage, define

$$I_g = (J_{g,A,\text{closed}} - J_{g,B,\text{closed}}) + (J_{g,B,\text{donor}} - J_{g,A,\text{donor}}), \tag{4}$$

$$\theta = I_{\text{matched}} - I_{\text{independent}}. \tag{5}$$

A positive θ represents attenuation of the crossing after replacing the matched training text. The interactions are 37.19 points under matched training (95% CI $[30.10, 44.38]$) and 39.92 under independent training ($[33.39, 46.48]$). Their paired reduction is -2.73 points ($[-10.03, 4.57]$, $p = 0.479$). The replacement weakens the reversal in 15 worlds and strengthens it in 17.

The individual endpoint scores also change with the training replacement (Table 4). The largest increase occurs for B on the shared synthetic records, from 70.76% to 77.02%, a gain of 6.27 points with Holm-adjusted $p = 0.052$. A improves on the same records by 1.98 points. The synthetic test preserves the orderings while the absolute endpoint scores change.

Table 4. Training-text replacement changes endpoint scores on shared tests. Each row gives independent-minus-matched training for one replay arm and test input. Intervals are pointwise paired 95% intervals. The four p values use one Holm family.

Test input	Replay arm	Change (points)	95% CI	Holm p
Question alone	A, closed book	+2.24	$[-0.05, 4.60]$	.209
Question alone	B, document	+3.80	$[-1.21, 9.11]$	.329
Synthetic records	A, closed book	+1.98	$[-3.20, 7.25]$	.475
Synthetic records	B, document	+6.27	$[1.93, 11.12]$	.052

5 What the Document Contributes

5.1 Rewriting the same foreign reports

We use the frozen independent-training endpoints from the WikiText test to separate changes in the donor facts from changes in how they are presented. We keep the twelve donor sources, reported values, verification states, and record order. One rendering labels the report and review fields and separates the entries. The other links the same statements and reviews in continuous prose. In both forms, an explicit review gives a verification verdict, while a neutral review sentence describes the catalogue entry. We select sentences at token counts available for

Table 5. Document replay leads under all four renderings of the same foreign facts. Joint accuracies are percentages. Gaps are B minus A in points with 95% whole-world intervals across 32 paired parents.

Rendering	A	B	B−A [95% CI]
Records, explicit review	52.77	73.73	+20.96 [16.03, 25.90]
Prose, explicit review	55.19	72.94	+17.75 [11.90, 23.50]
Records, neutral review	53.55	72.92	+19.38 [14.35, 24.13]
Prose, neutral review	55.16	73.42	+18.26 [12.61, 23.78]

every verdict and for the neutral sentence. All four conditions match in complete prompt length, absolute question position, and the question-to-assistant suffix.

Under explicit verdicts, rewriting the same reports as prose narrows the B−A gap by 3.20 points (prose minus records, 95% CI $[-4.82, -1.56]$). A gains 2.42 points $[0.85, 3.89]$ and B changes by -0.79 points $[-2.17, 0.63]$. With neutral review sentences the presentation effect on the gap is -1.12 points $[-3.01, 0.96]$. All four conditions favor document replay, including prose with neutral reviews. The WikiText test favors the other branch.

The four same-fact renderings match one another in prompt length. At 532–570 tokens they are 222–237 tokens shorter than their corresponding historical donor prompts. The 3.20-point presentation effect measures the rewrite of this particular set of records. The 27.17-point change on replacing the original donor with WikiText describes a separate comparison.

5.2 Foreign verification states move the target answers

To examine which answer components respond to foreign verification cues, we freeze the matched-training endpoints and reassign states within each source in the matched evaluation file. This file changes names and values while keeping verification states aligned with target record positions. Each edit updates both the status tag and its verification clause. Six balanced assignments place every state at every record twice, while preserving the report text, record order, and total prompt length. The target questions and correct answers stay fixed.

Document replay is sensitive to these edits. Its joint correctness falls from 81.45% on the matched file to 71.63% after state reassignment, a paired change of -9.81 points ($[-12.74, -7.13]$). Thirty-one of the thirty-two worlds decline. Status accuracy shows the largest movement, falling by 10.13 points. Accuracy on the reported value changes by -0.04 points. Closed-book replay has a smaller joint decline of 3.77 points. Figure 3 separates these responses across the three decisions.

The candidate probabilities follow the reassigned states as well. For B, a status candidate gains 6.32 probability points when the corresponding foreign record carries that state, relative to its other assignments and averaged across the three states ($[4.23, 8.49]$). On value questions, giving the foreign record an established state raises the probability of selecting the *target's* reported value by

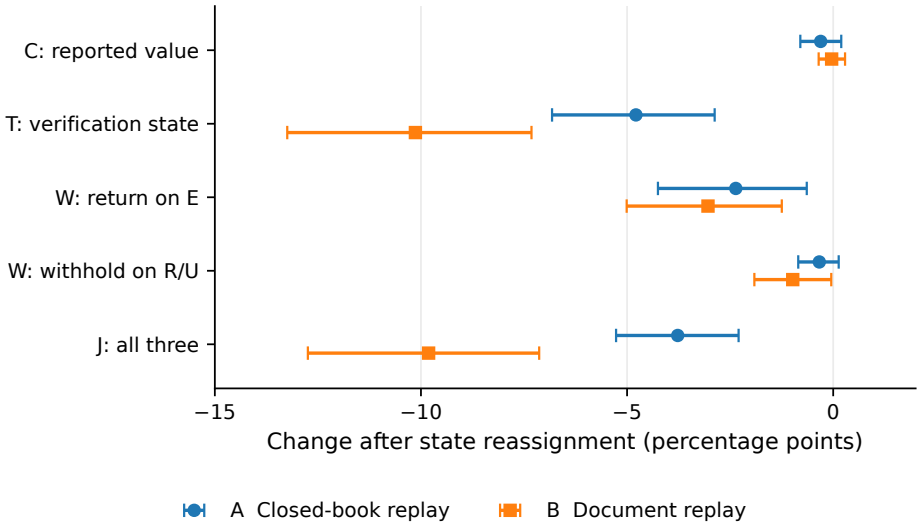


Fig. 3. State reassignment primarily changes verification judgments. Points show the change from the matched foreign file to six balanced state assignments at fixed record order. Horizontal bars give paired 95% bootstrap intervals. The value rows separate returning an established value from withholding a ruled-out or unresolved one.

1.77 points relative to assigning either other state ($[1.02, 2.63]$). These shifts show how changing which verdict appears at a foreign record position affects answers about the target. The same-fact comparison in Sect. 5.1 holds record order and verdicts fixed when changing presentation.

5.3 What the first eighty updates contribute

Before the replay branches diverge, mixed practice already changes how a model uses its own document. We compare the two eighty-update schedules under matched foreign training records. Closed-only QA practice correctly withholds a value on 47.44% of ruled-out or unresolved claims. Mixed practice raises this to 87.75%. Correct return on established claims rises from 96.11% to 98.79% (Table 6).

The corresponding trade-off appears when familiar questions arrive alone. Mixed practice reduces their joint correctness from 84.25% to 72.04%. Closed-book replay then raises it to 89.87%, recovering 17.83 points ($[12.84, 22.90]$) from the mixed checkpoint. Over the same replay, own-document withholding moves from 87.75% to 83.07%, a change of -4.69 points ($[-10.02, 0.28]$). On the fresh-fact set shared by these initial practice and replay comparisons, joint correctness moves from 80.80% to 76.37%, or -4.43 points ($[-7.83, -1.15]$). Additional closed-book practice therefore redirects performance toward familiar questions asked alone.

Table 6. Mixed practice improves evidence use and shifts closed-book performance. The two schedules each use eighty updates. Fresh-fact questions arrive with their own documents. Changes are paired across parent worlds.

Evaluation	Closed-only (%)	Mixed (%)	Change [95% CI] (points)
Own text, withhold value	47.44	87.75	+40.31 [32.63, 47.97]
Own text, return value	96.11	98.79	+2.68 [0.49, 5.12]
Familiar, closed-book J	84.25	72.04	−12.21 [−16.67, −7.81]
Fresh facts, own-text J	57.81	80.80	+22.99 [18.57, 27.50]

5.4 Fresh records reveal the value-return trade-off

The final evaluation compares all four replay endpoints and the frozen backbone on the same new questions with their own new evidence. The frozen backbone returns established values correctly 97.58% of the time, but correctly withholds values on 12.08% of ruled-out or unresolved claims. Every adapted endpoint raises withholding to at least 81%. Their established-value accuracy ranges from 80.96% to 88.42% (Table 7). On fresh records like the opening example, the adapted models more often leave the field unspecified when verification rules out a claim or leaves it unresolved. The frozen model remains more accurate when verification establishes the reported value.

Table 7. Adaptation changes which fresh reports become catalogue values. Every endpoint answers the same new questions with their own documents. The two value columns expose the return–withhold balance, while J combines source, status, and value correctness.

Endpoint	Return on E (%)	Withhold on R/U (%)	Joint J (%)
Frozen base	97.58	12.08	33.14
Matched training, A	88.42	81.04	77.21
Matched training, B	85.51	83.25	76.03
Independent training, A	83.75	84.21	76.99
Independent training, B	80.96	81.46	72.15

These opposing value-return movements yield joint-accuracy gains of 39.02 to 44.08 points over the frozen backbone. Among the adapted endpoints, independent-training A has the highest withholding rate, and matched-training A has the highest established-value accuracy. The component view makes the practical trade-off visible even when joint correctness improves strongly.

6 Related Work

Provenance, verification, and returned values. Data provenance studies the origin and influence of data contributing to a result [2]. Fact verification distinguishes

supported, refuted, and insufficiently evidenced claims [12]. Citation-based generation assesses whether cited passages support generated statements [3]. Our task places a source report and an explicit verification record together. A source question can then have a specific answer while a verified catalogue field remains unspecified. The linked source, status, and value questions make that distinction measurable on each record.

Training with retrieved and distracting documents. Retrieval-augmented generation combines parametric models with retrieved passages [6]. Entity-based knowledge conflicts reveal how substituted evidence changes answers [8], and comparisons of parametric and retrieved knowledge characterize performance across fact frequency [9]. RAFT trains with relevant documents and distractors [15]. KAFT uses counterfactual and irrelevant contexts to develop responsiveness to relevant evidence and robustness to distractions [7]. PA-RAG varies passage relevance for the same training question, paraphrases its answers, and interleaves general-domain QA to address conditional memorization under retrieval success and failure [1]. Our comparison asks which of two paired replay policies leads when two unrelated attachments have the same prompt length and question position. Synthetic verification records and WikiText prose yield opposite answers to that comparison on common frozen endpoints.

Behavioral sensitivity to evidence cues. Mayne et al. find that models can absorb a claim from training text even when the surrounding sentences call it false. Placing the negation within the claim often improves learning [10]. Our task keeps the reported value answerable across verification states and separately scores whether the verified field can be filled. Source-marking experiments examine marker acquisition and the way visible markers shift a model's responses [5]. Studies of conflicting contextual evidence observe a preference for earlier evidence [14]. We reassign the verification fields of a foreign record while retaining the original target questions. We also replace the foreign training document and compare the resulting endpoints on a shared independent donor. Together, these interventions connect responses to visible state cues with the broader replay-input choice.

7 Conclusion

Which specialized replay policy answers a familiar question more accurately depends on what accompanies it. Unrelated synthetic verification records favor document replay. Length-matched WikiText excerpts favor closed-book replay. New target worlds, donors, articles and adapter training reproduce that change in ordering. Holding the donor facts fixed reveals a narrower effect. Recasting explicit verification records as prose reduces the gap, and document replay still leads across all four tested renderings.

The same mixed checkpoint offers a third choice. Continuing mixed practice for forty updates improves on both specialists under the two attachments, at equal update and supervised-answer counts. On questions alone its paired difference

from closed-book replay spans zero. Measuring report recall, verification state and value return together shows which policy supplies more complete catalogue answers under each input.

8 Limitations

The original prose test paired 64 WikiText-2 Raw excerpts with previously studied targets and adapters. We fixed the prediction and article-selection rule before scoring these excerpts. Exposure to individual articles during pretraining is unknown. The 16-world replication renews the targets, donors, articles and training within the same backbone and task schema.

Matching prompt length, question position and suffix leaves vocabulary, discourse, verification semantics and familiarity free to vary. The same-fact rewrite changes wording and organization together. A neutral review changes the meaning of its slot, and the state-reassignment intervention edits both the tag and its verification clause.

A fixed training donor may itself cue recall. The training-replacement interval spans both signs. Testing equivalence would require a prespecified tolerance. Real catalogues add retrieval, credibility judgments, free-text answers and costs for incorrect updates to our trusted-state, candidate-letter task.

MIX120 was measured on the original 32 independent-training worlds. Updates and supervised answers match across continuations, while late QA examples, token counts and cost differ. A processed 7.28 million nonpadding QA input tokens during replay and B processed 63.13 million. The earlier frozen WikiText evaluation used 0.701 GPU command-hours. The three subsequent experiments together used 6.720 GPU command-hours including retries.

AI Usage. ChatGPT (OpenAI) was used strictly for language copy-editing and LaTeX formatting of author-written text and figures

References

1. Bhushan, K., Nandwani, Y., Khandelwal, D., Gupta, S., Pandey, G., Raghu, D., Joshi, S.: Systematic knowledge injection into large language models via diverse augmentation for domain-specific RAG. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 5937–5958 (2025). https://doi.org/10.18653/v1/2025.findings-naacl.329
2. Buneman, P., Khanna, S., Tan, W.C.: Why and where: A characterization of data provenance. In: Database Theory – ICDT 2001. Lecture Notes in Computer Science, vol. 1973, pp. 316–330. Springer (2001). https://doi.org/10.1007/3-540-44503-X_20
3. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6465–6488 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.398

4. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
5. Huryn, D.: Source marking in language models: A controlled framework for acquisition, cue sensitivity, and reasoning. In: Proceedings of the 9th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. pp. 220–231 (2026), https://aclanthology.org/2026.blackboxnlp-1.15/
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020), Proceedings page
7. Li, D., Rawat, A.S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., Kumar, S.: Large language models with controllable working memory. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 1774–1793 (2023). https://doi.org/10.18653/v1/2023.findings-acl.112
8. Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., Singh, S.: Entity-based knowledge conflicts in question answering. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7052–7063 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.565
9. Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., Hajishirzi, H.: When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9802–9822 (2023). https://doi.org/10.18653/v1/2023.acl-long.546
10. Mayne, H., McKinney, L., Dubiński, J., Karvonen, A., Chua, J., Evans, O.: Negation neglect: When models fail to learn negations in training. arXiv preprint arXiv:2605.13829v2 (2026), https://arxiv.org/abs/2605.13829v2
11. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Byj72udxe
12. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: A large-scale dataset for fact extraction and verification. In: Proceedings of NAACL-HLT. pp. 809–819 (2018). https://doi.org/10.18653/v1/N18-1074
13. Yang, A., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025). https://doi.org/10.48550/arXiv.2505.09388
14. Yang, X., Liu, Z., Li, R., Lei, Y.: Large language models in resolving contextual knowledge conflicts. In: Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing. pp. 21142–21169 (2026). https://doi.org/10.18653/2026.emnlp-main.918
15. Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J.E.: RAFT: Adapting language model to domain specific RAG. In: Proceedings of the First Conference on Language Modeling (2024), https://openreview.net/forum?id=rzQGHXNReU