BASELINE ADAPTATION CHANGES ATTRIBUTION IN FIXED-ANSWER SELECTIVE QA

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how adapting a confidence baseline changes the attribution of gains in selective question answering. Each question's original answer is fixed, so scorers change only which answers are retained. At half coverage, native confidence refitting improves reference EM by about 3.5 percentage points in two cohorts with no overlapping normalized questions. The first is a frozen seven-scorer comparison; the second is a separately reported native-only evaluation with policies frozen before generation, with comparison-feature evaluation unfinished. Fusion loses its apparent advantage over historical confidence when compared with the adapted practical baseline, while its increment over a joint-target-matched control remains uncertain. Matched-count controls and intercept interventions separate sample count, score offsets and ranking changes. Direct comparison features show exploratory signals that depend on coverage. Native confidence already contains contextual information; the question is the value of additional comparison information. Decision accounting and case review connect measured gains to retained answers while distinguishing reference matching from factual correctness.

1 INTRODUCTION

A retrieval-augmented QA system can improve because it generates a better candidate, ranks existing candidates better, or learns when to withhold an answer. These interventions are often evaluated together. A comparison with RAG may credit a selector for the additional candidate, while a comparison with an old confidence model may credit new features for better fitting data. We study how those attributions change under stronger controls.

Our central experiment fixes one historically selected RAG or closed-book answer per question. Seven scorers each retain 2,000 of 4,000 originals. A supplementary proposal informs scores but is never returned, separating scoring from answer replacement.

The development pattern motivates the experiment: fusion's apparent 9.09-point gain falls to 2.69 after native refitting. On project-held-out questions, the adapted native baseline leads fusion; a matched-count new-source fit preserves this ordering. The attribution changes while source composition, filtering and feature/outcome relationships remain candidate explanations for adaptation.

Two comparisons answer the next question differently. Fusion versus the strongest per-arm binary native predictor tests practical usefulness. Fusion versus a joint-native predictor matches the four-state learning target and metadata, isolating added comparison information within that recipe. The latter interval still permits a useful increment. We therefore also test a direct addition of comparison features to the strong binary recipe, explicitly labeling the already-observed 4,000-question analysis as exploratory.

Our contribution is an empirical account of *what changes retained answers*, preserving favorable signals and unresolved increments. Calibrator adaptation under domain shift is established prior work (Kamath et al., 2020); here it competes with comparison information as an explanation for measured gains. Reference EM is central, with separate diagnostics delimiting factual interpretations.

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.

2 A FIXED-ANSWER EXPERIMENT FOR ATTRIBUTION

Frozen Qwen3.5-9B (Qwen Team, 2026) generates RAG from retrieved passages 1–5, closed-book without passages, and a proposal from passages 6–20. The historical 700-question Dual4 selector chooses original a_i from the legal RAG/closed-book candidates. All seven scorers retain or omit that *same original*; the proposal only supplies scoring information.

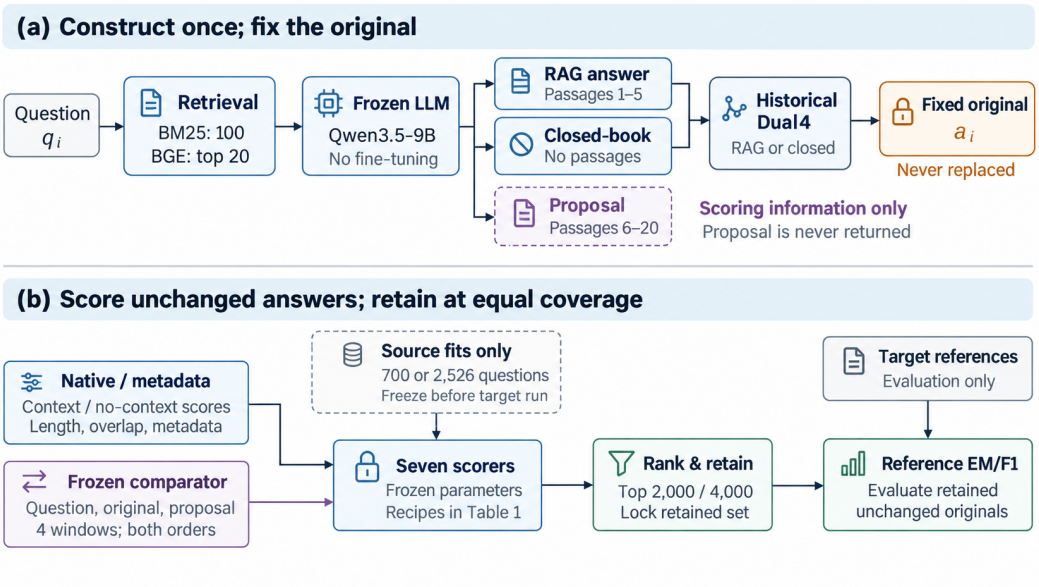


Figure 1: Construct a fixed RAG/closed-book original once, then rank unchanged originals under seven frozen scoring recipes at equal coverage. The proposal informs scoring and is never returned; source fits precede target generation, and references enter evaluation.

For scorer m and coverage κ, $g_i^m(\kappa)$ retains the $K = \lceil \kappa N \rceil$ highest-scoring legal originals, using fixed question-hash ties. With reference exact-match label Y_i,

$$A_m(\kappa) = \frac{1}{K} \sum_{i=1}^{N} g_i^m(\kappa) Y_i, \qquad N = 4000, \quad \kappa_{\text{primary}} = 0.5. \tag{1}$$

Every condition retains 2,000 originals. Differences therefore measure retained-answer ranking; answer identities also stay fixed in every bootstrap draw.

Native information and its increment. Native4 uses value-token mean log probabilities under document and no-document prompts, $\log(1 + T)$ and retrieved-text overlap: it already contains contextual information. Its per-arm binary fit scores the historically selected arm without reselecting an answer. Additional comparison features use both answers, four evidence windows and both answer orders (Section 2.1). Greedy JSON generation, reranking, window selection, tokenizer and frozen-comparator details are specified in Appendix D.1.

2.1 SEVEN SCORERS AND THE CONTROLS THEY SUPPORT

Table 1 distinguishes the recipes. The binary fit directly predicts the arm's EM label Y. The joint fit predicts $Y_o + 2Y_p$, then ranks $P(Y_o = 1)$; equal answers share correctness and null answers have correctness zero. Both use training-only standardization, zero initialization and L-BFGS-B, with loss equal to mean cross entropy plus $\|W_{\text{slopes}}\|_F^2/n$ and an unpenalized intercept. Tolerances are 10^{-12} (function), 10^{-9} (gradient), with 2,000 iterations maximum; Appendix D.1 gives the masked objective.

Table 1: All seven frozen conditions. H700: historical 200 NQ + 500 TriviaQA; N2526: newer NQ fitting set; N700: hash-selected subset of N2526. x_4 is Native4; v_3 is its two value scores and token length; m_7 is shared metadata; c_3 is comparator information. Each row ranks the same historical original.

Scorer	Target / structure	Input	Fitting source
Historical native	binary EM, per arm	x_4	H700
Refitted native (Native4)	binary EM, per arm	x_4	N2526
New-source native 700	binary EM, per arm	x_4	N700
Joint native	joint EM (4 states)	$m_7 + v_3$	N2526
Joint comparison	joint EM (4 states)	$m_7 + c_3$	N2526
Fusion	joint EM (4 states)	$m_7 + v_3 + c_3$	N2526
No direct passages	joint EM (4 states)	$m_7 + v_3 + c_3^{\varnothing}$	N2526

The seven metadata features m_7 are the original arm, log character lengths and lexical overlaps of both answers, RAG/closed-book exact agreement, and original/proposal exact agreement. Native overlap is consequently present in both recipes, though organized differently. The three native value features v_3 add the two contextual/no-context log scores and token length.

Let b, h, n be the order-averaged comparator probabilities favoring the proposal, favoring the original, or expressing neutral preference. The first two comparison features are $\log[(b+\epsilon)/(n+\epsilon)]$ and $\log[(h+\epsilon)/(n+\epsilon)]$, with $\epsilon = 10^{-6}$. The third is the absolute difference between the two orders' signed proposal-minus-original preferences after aligning candidate identities. These are local preference signals; the fitted head learns their relation to reference correctness. The no-direct-passage version recomputes only those comparator signals without the four windows; upstream answers, native contextual scores and overlap remain context-dependent.

Three primary contrasts. (1) Refitted minus historical native measures adaptation with the four native features and binary targets unchanged. (2) Fusion minus refitted native tests the practical increment beyond the strong baseline, while changing target and metadata. (3) Fusion minus joint native holds fitting source, four-state target, metadata and fitting recipe fixed and adds comparison features. Their distinct controls prevent attributing every difference to comparison information.

Cohorts and uncertainty. The 4,000 questions are selected from the public NQ-open training source by a fixed hash, excluding normalized questions in all registered project data and historical outputs. The historical 700, newer 2,526 and target 4,000 have zero exact normalized-question overlap. This is a project holdout; annotated-page overlap, semantic duplicates and public-benchmark exposure remain possible.

All point fits and 2,000 source-bootstrap parameter sets precede target generation; retained sets precede reference scoring. Historical resampling preserves the 200/500 mixture, and newer source resampling groups 2,526 questions by 2,322 annotated pages. Paired target draws resample questions and recompute 50% retention. We report nominal Bonferroni-three 98.33% source/target intervals for the three primary contrasts. The prespecified 2pp planning scale describes practical size; it is not an equivalence margin backed by an equivalence test. A separate fixed-fit retrieved-title grouping probes one target-dependence proxy. Appendix A gives the full protocol.

3 FROZEN CONFIRMATION: ADAPTATION CHANGES THE ATTRIBUTION

Table 2: Fixed original answers on 4,000 project-unused NQ-open training-source questions. Every scorer retains 2,000 answers. EM/F1 are reference percentages; confidence fitting changes the retained set, not the answer.

Scorer	Reference-correct	EM	F1
Historical native (700)	1145/2,000	57.25	69.10
Refitted native (2,526)	1216/2,000	60.80	72.19
New-source native (700)	1216/2,000	60.80	72.54
Joint native (2,526)	1155/2,000	57.75	70.19
Joint comparison (2,526)	1156/2,000	57.80	70.91
Fusion (2,526)	1168/2,000	58.40	71.00
No direct passages (2,526)	1127/2,000	56.35	69.45

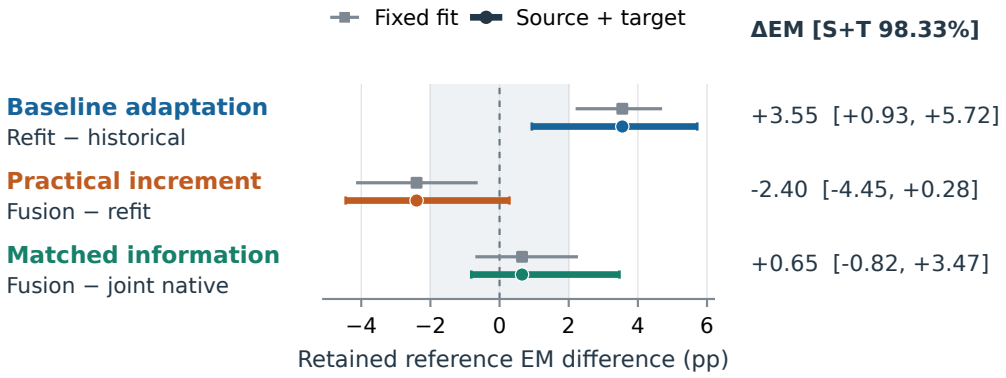


Figure 2: Contrasts on the first cohort, in retained reference-EM percentage points. Gray bars hold fits fixed; colored bars and printed endpoints use source/target refits, both with nominal Bonferroni-three 98.33% limits. The ±2pp band is a planning reference, not an equivalence test.

Native refitting gains **3.55pp**, from 57.25% to 60.80%, with adjusted interval $[0.93, 5.72]$. Fusion's 58.40% is +1.15pp above historical confidence: arithmetically, +3.55pp adaptation and -2.40pp relative to the refit. The latter interval $[-4.45, 0.28]$ excludes a +2pp gain while including zero. Adaptation therefore changes the attribution of fusion's apparent practical advantage.

Fusion exceeds joint native by +0.65pp, $[-0.82, 3.47]$. This target-matched interval still permits useful comparison information. It motivates testing those features directly within the stronger binary recipe (Section 4); the practical contrast changes more than the feature group.

Sample count and direct passages. New-source 700 also reaches 60.80%: +3.55pp over historical 700, descriptive 95% interval $[1.05, 5.45]$. Full 2,526 minus new 700 is 0pp, $[-1.65, 1.75]$. Extra sample count is therefore insufficient to explain the gain; source composition and filtering remain unresolved. Direct windows improve fusion by 2.05pp, $[0.55, 3.75]$, a favorable secondary ablation that retains upstream evidence information.

Which answers change? Refit–historical swaps 207 questions each way: 136 reference-correct enter and 65 leave, net +71. Fusion–refit swaps 405, with 121 correct entering and 169 leaving, net -48. Fusion–joint swaps 276, with 120 entering and 107 leaving, net +13. These are changes in retained originals.

Dependence sensitivity. Retrieved-title-group fixed-fit 95% intervals are $[2.45, 4.45]$, $[-3.69, -0.90]$ and $[-0.30, 1.94]$pp, close to question-resampling results. This proxy sensitivity omits source refitting and does not establish annotated-page separation (Appendix A).

4 NATIVE ADAPTATION AND DIRECT-COMPARISON CONTROLS

The earlier 4,000-question result motivated exploratory tests on that already-observed set. All new fits used only the existing 2,526 source questions. Full Native4 refitting exceeded arm-specific intercept updates by +3.10pp, nominal source/target 98.33% interval $[0.52, 5.05]$. The allowed ranking changes differ:

$$\begin{array}{lll} \sigma(z_a + \delta) & \text{common offset:} & \text{all ranks unchanged,} \\ \sigma(z_a + \delta_a) & \text{arm offsets:} & \text{within-arm ranks unchanged,} \\ \sigma(\beta_a'^{\top} x + b_a') & \text{full refit:} & \text{ranks may change.} \end{array} \tag{2}$$

Here $z_a = \beta_a^{\top} x + b_a$ is the historical arm logit and σ is strictly increasing. The first invariance is mathematical; observed retention changes test the value of the remaining flexibility.

For a direct information control, Native6 adds historical-choice and actual-comparison availability flags to Native4; Compare9 adds three comparator features. Missing comparisons are zero with flags explicit, while every arm retains all 2,526 training rows and the same binary EM recipe. On the observed set, Compare9 minus Native4 is +0.10pp $[-1.25, 2.25]$ and minus Native6 is -0.10pp $[-1.27, 1.60]$. At 20% coverage, the differences are +1.88pp $[-0.25, 3.50]$ and +3.13pp $[0.75, 4.75]$, respectively, with fixed-fit 95% intervals. These and the favorable F1 signals remain exploratory (Appendix B).

Additional cohort: an authorized partial native replay. On 4,000 further project-disjoint NQ-open training-source questions, we replay four native policies whose parameters were frozen before generation. Full-pool RAG/closed-book quality had already been inspected under an early-analysis authorization; the four native point/source rankings were then locked before their 50% analysis. Full comparison-feature inference and the cross-check against that pipeline remain unfinished.

Table 3: Partial native-only replay on 4,000 further questions, retaining 2,000 fixed originals. Points are reference percentages; differences are pp. B3 preserves the new cohort's three-primary 98.33% adjustment, with two comparison-feature members still pending. Other intervals are secondary 95% source/target intervals.

Scorer	EM	F1
Historical native	57.00	69.77
Native4 refit	60.45	73.02
New-source native 700	60.25	73.22
Arm intercepts	57.95	70.49
Contrast	**Δ EM**	**S+T interval**
Native4–arm intercepts	+2.50	[+0.35, +4.45] B3
Native4–historical	+3.45	[+1.20, +4.75]
New 700–historical	+3.25	[+0.55, +5.00]
Native4–new 700	+0.20	[-1.50, +1.85]

Full refitting again improves over arm offsets: +2.50pp under the retained B3 adjustment, with 119 reference-correct originals entering and 69 leaving (net +50). The same-count new-source fit also exceeds historical confidence; its +3.25pp contrast is secondary. All 17 preselected questions (24 memberships) were reviewed without changing labels. The result adds native-ranking evidence within the same generator/source setting; the two comparison-information primary contrasts remain pending (Appendix C).

5 EARLIER FROZEN EVALUATIONS: ALL PRIMARY COMPARISONS

The earlier evaluations use Qwen3.5-9B and Gemma 4 E4B IT (Qwen Team, 2026; Google, 2026). Their targets and policies differ from fixed-answer retention (Table 4); all sixteen primary results remain below.

Table 4: The studies answer distinct attribution questions. Earlier evaluations broaden context without replicating the central fixed-answer protocol.

Study	Prediction / decision	Labels / endpoint	Supported comparison
Fixed answers (4,000)	Score original; retain half	Reference EM / retained EM	Baseline adaptation versus extra scoring information
Complete responses (2,000)	Three states; gated choice	Machine states / net-yield bounds	Outcome prediction versus routing utility
Answer fields (1,560)	Binary score; choose legal field	Reference EM / answer EM	Context-value scores versus matched selectors

5.1 BETTER OUTCOME PREDICTION NEED NOT IMPROVE ROUTING

The first validation contains 1,000 project-unused EfficientQA and 1,000 TriviaQA questions. Own3 uses own-prompt confidence, length and overlap; Evidence5 adds frozen Qwen support/contradiction scores. Fits use the historical 700 questions. Both policies switch from RAG to closed-book only when a frozen Llama nonanswer flag is positive and estimated correct-minus-wrong utility exceeds both zero and RAG utility. Models, fits and policies precede target generation.

The endpoint $U_1 = Y_c - Y_w$ counts correct minus wrong substantive answers per question. Two reference-based judges leave unresolved states explicit; identical responses share unknown states across policies. Table 5 reports identification bounds and source/target envelopes, distinct from EM intervals.

Table 5: All eight first-validation contrasts: Evidence5 minus the control. Net-yield bounds and Bonferroni-eight source/target-refit endpoint envelopes are percentage points per 1,000 questions. Changed/unknown counts refer to the saved point policies and undetermined net contributions, not every refitted policy. Envelopes condition on machine-label constraints and omit common judge errors; their 500-draw tail resolution is limited.

Generator	Source	Control	ΔU_1 bounds	Refit B8	Changed / unknown
Qwen3.5	EfficientQA	RAG	[+0.30, +2.30]	[-1.34, +5.24]	25 / 9
Qwen3.5	EfficientQA	Own3	[-0.90, +2.90]	[-3.00, +5.74]	49 / 17
Qwen3.5	TriviaQA	RAG	[+0.50, +2.30]	[-1.09, +5.10]	24 / 9
Qwen3.5	TriviaQA	Own3	[-0.90, +1.30]	[-2.50, +3.34]	26 / 11
Gemma4	EfficientQA	RAG	[-0.20, +3.00]	[-2.34, +5.04]	36 / 15
Gemma4	EfficientQA	Own3	[-0.80, +2.40]	[-2.94, +4.40]	41 / 15
Gemma4	TriviaQA	RAG	[-0.30, +2.10]	[-2.36, +4.40]	50 / 12
Gemma4	TriviaQA	Own3	[-0.60, +0.80]	[-2.63, +2.99]	30 / 7

Evidence5 has positive conditional correct-yield bounds versus RAG in every cell, while also increasing wrong yield. Against same-rule Own3, all four net-yield point bounds include zero, as do all eight primary fixed-policy and source/target envelopes. Gains over RAG thus answer a broader question than the additional document-relation increment, which remains unresolved.

Known Qwen EfficientQA changes include thirteen wrong-to-nonanswer, ten correct-to-nonanswer and one wrong-to-correct transition. Evidence5 improves closed-book Brier loss on known labels and, on TriviaQA, with unknown outcomes retained. Fixed-80, local context-sufficiency and status-region controls leave routing utility unresolved (Appendix D).

5.2 ANSWER-UNIT GAINS DEPEND ON THE COMPARISON

The second validation uses the remaining 560 project-unused NQ-open dev questions and 1,000 hash-selected TriviaQA validation questions. Each candidate is a legal nonempty JSON answer field with a 64-token cap. The RAG prompt permits internal knowledge and includes retrieved titles/text. All two-candidate learned selectors choose the higher predicted reference-EM probability among available candidates, retaining RAG on ties. Separate binary fits use the historical 700 questions; all fits and 500 source-refit sets precede target generation.

DUAL4 uses the canonical value's document and no-document mean log probabilities, log token length and overlap. NATIVE3 instead uses whole-response native confidence, length and overlap under the same binary target and rule. It is different from the three-state Own3 policy above. DIRECT4 combines order-averaged direct pair preference with own-prompt value confidence, length and overlap. OWN-UNIT3 uses only the latter three features; all share the same candidate availability rule.

Table 6: All eight primary answer-unit contrasts: Dual4 minus the indicated control. Differences and source/target-refit envelopes are reference-EM percentage points. Holm p values adjust eight exact paired tests conditional on the fixed fit; the separate refit envelopes are descriptive 95% quantiles. Their conditioning and multiplicity treatment differ.

Generator	Source	Control	Δ EM	Refit 95%	Holm p
Qwen3.5	NQ	Native3	+1.96	[+0.00, +3.57]	0.1537
Qwen3.5	NQ	Direct4	+0.89	[-1.79, +2.14]	1.0000
Qwen3.5	TriviaQA	Native3	+1.20	[+0.10, +2.50]	0.2897
Qwen3.5	TriviaQA	Direct4	+1.00	[-0.10, +1.80]	0.2897
Gemma4	NQ	Native3	+1.43	[-0.54, +2.32]	0.3841
Gemma4	NQ	Direct4	+0.36	[-1.07, +1.17]	1.0000
Gemma4	TriviaQA	Native3	+0.80	[-0.20, +1.80]	0.6064
Gemma4	TriviaQA	Direct4	-0.40	[-1.35, +0.85]	1.0000

Dual4 improves over RAG on TriviaQA by 2.9pp for Qwen and 4.3pp for Gemma, with source/target-refit intervals $[1.4, 4.4]$ and $[2.8, 5.9]$. When both fields are legal, gains remain 2.6 and 2.3pp; availability fallback contributes 0.3 and 2.0pp. These improvements therefore include useful selection among available answers.

All eight designated comparisons with Native3 and Direct4 fail Holm-adjusted superiority tests. The positive descriptive Qwen TriviaQA–Native3 interval $[0.10, 2.50]$ coexists with Holm $p = 0.2897$ because the bootstrap and exact tests differ in fitting uncertainty and multiplicity treatment. All four Dual4–Own-unit3 refit intervals also cross zero; Own-unit3 matches Gemma's 70.4% TriviaQA EM. The result supports the value of candidate selection over RAG while leaving the designated cross-context scoring increment unresolved. All twelve machine-label net-yield refit envelopes versus RAG, Native3 and Direct4 include zero.

5.3 DEVELOPMENT CONTROLS: AVAILABLE REPAIRS VERSUS DELIVERED REPAIRS

The retained development program tests candidate supply, nonlinear selection, cross-family candidates, learned verification and editing. These reuse project questions and are not additional independent validations. Supplementary retrieval improves the candidate oracle over equal-size ordinary expansion by 1.33pp, $[0.17, 2.50]$, but actual selection by only 0.33pp, $[-0.67, 1.42]$, below the original-pool selector. Better candidate supply thus does not by itself deliver better decisions.

On the repeatedly used 594 questions, ordinary SFT supplies 51 reference repairs and 48 harms. Fold-excluded revision supervision yields fusion EM 41.92% versus matched native 42.09%: -0.17pp, 99% interval $[-1.01, 0.52]$. Fourteen changed answers contain one additional reference repair, two harms and eleven both-EM-incorrect changes. Both selectors exceed RAG and always adopting SFT (38.05%). The continuation screen fails, ending that recipe. Appendix F indexes the preserved favorable and unfavorable controls and the full historical archive, keeping the main claim independent of a long development narrative.

6 WHAT THE MEASURED CHANGES MEAN

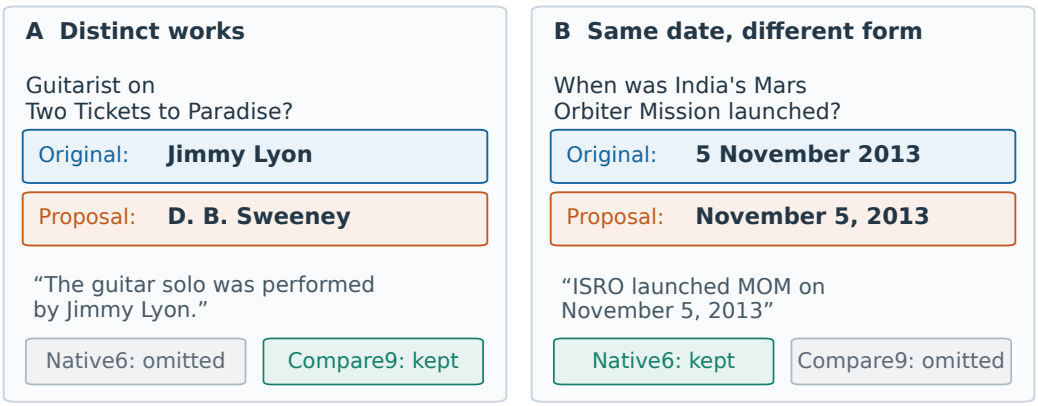


Figure 3: Observed retention changes on the old target. Both originals stay fixed and have support in the actual snippets. Compare9 retains the answer against a different-work proposal (left), but loses an exact match against equivalent date wording (right). These illustrative cases come from unblinded assistant review, not new human labels or a population-frequency estimate.

Reference matching and support can diverge. A registered rule sampled central contrasts by entered/removed $\times$ EM=1/0, yielding 96 memberships and 93 unique questions. All 93 were inspected against actual inputs and scores; the other 1,131 changed questions were archived. The unblinded assistant review changed no labels, fits or selections.

Fusion removes a supported two-composer answer whose reference lists names separately, and a supported, reference-correct season-three episode count. It also retains a name supported only for another work. All 4,000 question/reference pairs match their public source, yet a location question has a date reference. These observations separate reference defects, support failures and retained-answer quality; the balanced sample does not estimate population accuracy.

Completed review of the direct binary increment. All 23 preselected exploratory questions (32 memberships) were reviewed. Figure 3 contrasts entity discrimination with a lost exact match despite equivalent dates. For the Finley and Hynes answers, original and proposal coincide: no comparator forward occurs, and Native6 already retains both. Those practical gains therefore do not identify a new per-question comparator contribution. Most F1 improvement comes from partial token overlap among EM-zero answers (Appendix B).

Agreement between judges can conceal shared errors. Qwen3-8B and Gemma-3-12B judges were checked on 2,120 released historical EfficientQA responses from 687 disjoint questions. Both reject 181/358 human-positive answers and accept 23/1,762 negatives. Overall agreement near 89% coexists with balanced agreement near 70%. Consensus resolves 79.8%; its 91.5% known-subset agreement becomes full-denominator bounds $[73.1, 93.3]\%$. These shared historical errors do not estimate current-output error rates.

Expanded references raise historical agreement to about 96% but contain assessed positive predictions. In the first validation, 1,115 of 6,654 rated responses have unknown task outcomes; projecting 414 wrong/nonanswer disagreements to task failure leaves 701 unknown and preserves substantive-status uncertainty.

Sensitivity identifies which conclusions depend on labels. Relaxing one and two accepted labels per 1,000 makes Qwen's two positive net-yield point bounds versus RAG include zero; the other six already do. GLM-5.3 parses 283 of 287 distinct responses in the policy-disagreement set. Of 234 originally resolved states, 216 agree, sixteen change and two become unknown. GLM uses eleven demonstrations versus thirty-three locally, plus different inference settings. These are pipeline sensitivities; independent assessment is still needed for population factual claims.

7 Related work and scope

Calibration and selective answering. Kamath et al. (2020) establish the benefit of confidence fitting under QA domain shift. Calibration and selective classification provide standard tools (Guo et al., 2017; Geifman & El-Yaniv, 2017); calibration also guides document/history selection (Lee et al., 2023) and retrieval budgets (Dong et al., 2026). Our question is how adapting a strong baseline changes the measured value of additional comparison features under fixed outputs.

Selection, evidence and correction. Self-Selection-RGP selects no-context or RAG answers (Weng et al., 2025); Astute RAG consolidates internal and retrieved knowledge (Wang et al., 2025a). Context sufficiency supports selective generation (Joren et al., 2025). Self-Refine, CoVe and LLM-Blender study feedback, verification and output combination (Madaan et al., 2023; Dhuliawala et al., 2024; Jiang et al., 2023). CoV-RAG couples verification, retrieval revision and training (He et al., 2024); ART learns refinement decisions (Shridhar et al., 2024), while SCoRe addresses correction-training shift (Kumar et al., 2024). Our selectors and editors test narrower components, not full reproductions.

Generation and measurement. CAD and AdaCAD change token generation using contextual/no-context distributions (Shi et al., 2024; Wang et al., 2025b). The central experiment instead preserves the generated answer. Decoding-dependent uncertainty and fixed-answer protocol sensitivity are documented by Hashimoto et al. (2025); Kim & Kang (2026); reference incompleteness is a known QA evaluation problem (Min et al., 2021). Our decision accounting connects those concerns to observed retained sets and leaves original metrics visible.

Scope of inference. The repeatedly used 594-question development set is filtered for short-span supervision. Question exclusion does not establish page independence or absence from pretraining; refit intervals omit generator-training and reference uncertainty.

Recorded inference cost. The original cohort requires 4,000 generated answers per arm (RAG, closed-book, proposal), 16,000 native value-score events (four per question), and 5,038 evidence plus 5,038 no-direct comparison forwards (two orders for 2,519 pairs). Twenty-four replay probes are separate. These offline counts include native scoring costs and do not measure online latency.

8 Conclusion

Fixing answers reveals an attribution change that a historical-confidence comparison alone would miss. On the completed cohort, native refitting absorbs fusion's apparent practical advantage, while the joint-target-matched comparison still permits useful additional information. Matched-count fits show that sample count alone is insufficient, and the offset interventions identify consequential ranking changes. The additional cohort's authorized native-only replay reinforces that ranking result with parameters frozen before generation. Its comparison-feature endpoints remain unfinished.

For evaluation practice, these findings motivate fixing answer identity and coverage, adapting the strongest native baseline with matched supervision, and adding an information-matched control before crediting richer features. Reporting both comparisons makes positive ablations and uncertain practical increments interpretable together. Our central evidence remains a reference-based case study with one generator and data source. Partial release, shared source fits, incomplete references and unblinded assistant review bound its interpretation. The next decisive evidence is complete comparison-feature evaluation on the new cohort, followed by checks of the same recipe across sources or generators and independent outcome assessment for factual-reliability claims.

Reproducibility. The anonymous supplement recomputes the completed central/exploratory results and new native-only subset, including 2,000 fixed-fit draws per cohort. Source/target intervals and the original title-group sensitivity use cached outcomes. Independent extracted checks passed; source fitting and neural generation are not rerun (Appendix F).

AI USE STATEMENT

AI assistants contributed to problem and conceptual formulation, hypotheses, experimental and mathematical formulation, cohort preparation, implementation and code inspection, analysis and numerical checks, interpretation, source checking, figures/tables, literature search, and drafting/editing. Local models generated experimental answers, selection features and provisional outcome judgments. Supplementary GLM-5.3 calls supplied model judgments; an earlier CLI-assisted numerical audit also used a configured GLM-5.3 backend. Reused EfficientQA labels are published human annotations, not newly collected labels for our outputs. No new current-output human annotations were collected. Post-result case reviews were unblinded and assistant-conducted. Model judgments, automated checks and assistant review are not independent human scientific review.

REFERENCES

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Zhe Dong, Fang Qin, Manish Shah, and Yicheng Wang. Know before you fetch: Calibrated retrieval-budget allocation for retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.29959.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Google. Gemma 4 E4B IT model card. https://huggingface.co/google/gemma-4-E4B-it, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. Decoding uncertainty: The impact of decoding strategies for uncertainty estimation in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14601–14613, 2025. arXiv:2509.16696.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10371–10393, 2024. doi: 10.18653/v1/2024.findings-emnlp.607. URL https://aclanthology.org/2024.findings-emnlp.607/.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Hankyeol Kim and Pilsung Kang. Same answer, different confidence: Protocol sensitivity in LLM confidence calibration. *arXiv preprint arXiv:2605.27752*, 2026. URL https://arxiv.org/abs/2605.27752.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning, 2024. URL https://arxiv.org/abs/2409.12917.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Kyungjae Lee, Sang-eun Han, Seung-won Hwang, and Moontae Lee. When to read documents or QA history: On unified and selective open-domain QA. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6420–6432. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.401. URL https://aclanthology.org/2023.findings-acl.401/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. NeurIPS 2020 EfficientQA competition: Systems, analyses and lessons learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 86–111, 2021. URL https://proceedings.mlr.press/v133/min21a.html.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026.

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

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ramakanth Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. The ART of LLM refinement: Ask, refine, and trust. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pp. 5872–5883. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.327. URL https://aclanthology.org/2024.naacl-long.327/.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of NAACL-HLT (Volume 1: Long Papers)*, pp. 11636–11652, 2025b. doi: 10.18653/v1/2025.naacl-long.581. URL https://aclanthology.org/2025.naacl-long.581/.

Yan Weng, Fengbin Zhu, Tong Ye, Haoyan Liu, Fuli Feng, and Tat-Seng Chua. Optimizing knowledge integration in retrieval-augmented generation with self-selection, 2025. URL https://arxiv.org/abs/2502.06148.

Appendix guide. This supplement retains the central protocol, complete central numerical contrasts, statistical definitions, decision diagnostics and representative cases. The longer historical development record is preserved separately and indexed in Appendix F. All original main comparisons remain in the main text; shortening this supplement does not change the set of experiments conducted or their recorded outcomes.

A FROZEN CONFIRMATION OF FIXED-ANSWER CALIBRATION ATTRIBUTION

Question and scope. The earlier fixed-pool study's 594 questions informed repeated development. This confirmation fixes the target decision to ranking unchanged original answers, with reference EM at 50% coverage as the primary endpoint. It does not confirm the separate SFT-revision recipe, which failed its own continuation conditions. No additional language model is trained, and no model, threshold, coverage or feature grid is introduced.

Cohort and selection. The public NQ-open training-source mirror contains 87,925 rows. We exclude normalized questions found in the project registry, historical cached outputs, result files, and question-containing JSON/JSONL records, including reserved examples. Duplicate normalized questions retain their first source record; references are not merged after observing outputs. Among nonempty questions with nonempty reference lists, the first 4,000 by a fixed hash of the normalized question and seed 2026092540 are registered before retrieval and generation. Original questions and references remain unchanged. The historical 700, new fitting 2,526 and target 4,000 questions have no normalized exact-question overlap. Semantic duplicates and shared background pages remain possible. The lightweight source lacks annotated page identifiers, and this project holdout is not an official test split or evidence of absence from pretraining.

Generation and the fixed original answer. Wikipedia18 BM25 retrieves 100 documents; the same frozen BGE reranker selects twenty. Qwen3.5-9B generates RAG from the first five documents, a closed-book answer without documents, and a supplementary proposal from documents six through twenty. The existing greedy 64-token canonical JSON protocol, thinking-disabled chat template, BF16/SDPA and batch size one remain fixed. The historical 700-question Dual4 fit selects the original answer from the legal RAG/closed candidates. Every confidence condition scores this already selected arm. Neither point recalibration nor bootstrap refitting changes the original answer or chooses the maximum score over two arms. The supplementary proposal informs some confidence features but is never substituted as this task's output.

Seven confidence conditions. Historical native and refitted native use identical four-feature Dual4 representations and per-arm binary reference-EM fits, trained on the historical 700 or newer 2,526 questions, respectively. Native refers to generator likelihood features, including contextual and no-document value-token scores, length and overlap; it does not mean removal of all retrieval information. A third native condition fits a fixed hash-selected 700-question subset of the newer fitting cohort, without label-based subset selection. It controls count but not every difference between the historical and newer sources.

Four further conditions reuse the fixed-pool study's calibration fits: joint native, joint comparison, fusion, and fusion without direct comparator passages. All share seven metadata features and the four-state reference target $Y_{\mathrm{o}} + 2Y_{\mathrm{p}}$. Joint native adds the original answer's three native value features; joint comparison adds the three order-symmetrized comparison features; fusion uses both. The frozen NF4 base comparator has its adapters disabled and sees both candidate orders, with original/proposal roles moving with each candidate. The no-direct-passage condition retains context-dependent upstream answers, likelihoods and overlaps. Exact/null-answer constraints, training-only standardization, regularized masked-softmax fitting and question-hash ties are unchanged from Appendix D.1.

Four reference-free windows are chosen with the existing 56-word window, 40-word stride and 14-word title rule: two from documents one through five and two from documents six through twenty. The legacy DeBERTa tokenizer packs at most 180 question/candidate tokens and up to 80 tokens per evidence block within 512 tokens. The same modern Qwen tokenizer then renders the comparison prompts, with an explicit 1,024-token check. Existing first-part truncation is recorded; the target protocol does not silently introduce another truncation rule. A three-question cached

probability replay per GPU precedes target comparison features, at the previously specified 10^{-5} absolute tolerance.

Contrasts and retained accuracy. All conditions rank the same valid original answers and retain 2,000. Joint conditions use $P(Y_{\mathrm{o}} = 1)$; per-arm conditions use the historically selected arm's fitted probability. With retained indicators g_i^m, the score is $A_m = 2000^{-1} \sum_i g_i^m Y_i$. The three primary differences are: (1) refitted native minus historical native; (2) fusion minus refitted native; (3) fusion minus joint native. The first changes fitting count and source together. The second is a practical strong-control comparison that also changes metadata and target. Only the third holds those components fixed while adding comparison features. The identity

$$A_{\mathrm{fusion}} - A_{\mathrm{historical}} = (A_{\mathrm{refit}} - A_{\mathrm{historical}}) + (A_{\mathrm{fusion}} - A_{\mathrm{refit}}) \tag{3}$$

is an arithmetic attribution, not a causal decomposition into sample size and information. Five secondary contrasts compare fusion with historical native, joint comparison and no-direct-passage fusion, and compare new-source 700 with historical 700 and full new-source 2,526 with its 700 subset. Secondary contrasts cannot replace an unfavorable primary comparison.

Precomputed source fitting and target uncertainty. All point parameters and 2,000 source resamples precede target generation. Only the two subset-700 arm fits are new point fits; the other point parameters reuse the previously fixed fits. Source seed 2026092542 resamples the historical 200 NQ/500 TriviaQA mixture within strata. The newer fitting questions are resampled by their 2,322 annotated pages, sharing the same page weights across all newer conditions. The fixed 700-question subset takes its multiplicities from that same resample, rather than an independently favorable draw. Each replicate fits six native arms and four joint heads, for 20,000 small fits in total.

The target uses 2,000 paired question resamples, seed 2026092543, shared across scorers. Each draw recomputes retention at $\lceil N_{\mathrm{draw}}/2 \rceil$ weighted question mass from its fixed or refitted scores; repeated copies may split at the boundary. The original answers never change. We report fixed-fit and joint source/target percentile intervals: 95% descriptive intervals and, for the three primary contrasts, nominal Bonferroni-three 98.333% intervals. A separate 2,000-draw, fixed-fit sensitivity groups target questions by their top-ranked retrieved title (seed 2026092544). Those groups are not annotated pages or a guarantee about semantic dependence. The source/target analysis does not jointly use that target grouping. No interval covers reference error, shared model error, pretraining exposure or every unobserved dependency.

If fewer than 2,000 original answers are valid, the target coverage is declared unattainable. An unattainable bootstrap draw is recorded, not redrawn or silently removed; an interval containing such draws is not reported as a complete interval. This rule is fixed before target results.

Precision and interpretation. The existing 594-question paired variation supplies only a conditional normal square-root-N planning approximation. At the three-comparison critical value $z = 2.39398$, the widest primary comparison suggests 3,752 questions for a 2pp half-width, motivating a fixed 4,000. At 4,000, projected conditional half-widths for the three contrasts are 1.49, 1.94 and 1.58pp. Source-fit variability, target shift and development selection are excluded from that projection; it is not a power guarantee. The prespecified 2pp practical scale is used to describe the observed intervals, not to claim formal equivalence from a nonsignificant contrast. There is no outcome-dependent sample-size extension.

All seven scores, three primary contrasts, five secondary contrasts, F1 sensitivities and actual entered/removed correct-answer counts are reported. Original-reference outcomes remain the endpoint; the supplementary GLM assessment supplies neither training nor evaluation labels here. Subsequent case review distinguishes raw archive coverage from completed semantic inspection and does not relabel this confirmation.

A.1 Complete frozen numerical results

A.2 Post-result attribution of ranking and retained-arm allocation

This descriptive analysis uses the completed confirmation's saved scores, answers and reference outcomes. It was specified after aggregate results and the 93-case review, before computing the

Table 7: All three prespecified confirmation contrasts. Differences and intervals are retained-EM percentage points. S+T resamples both source fits and target questions. B3 is the nominal 98.33% Bonferroni-three interval. The 2pp scale was fixed before target outputs; neither a null-containing interval nor an interval within that scale proves equivalence. A dash denotes an unavailable interval due to unattainable resample coverage.

Contrast	Δ	S+T 95%	S+T B3
Refit minus historical	+3.55	[+1.35, +5.30]	[+0.93, +5.72]
Fusion minus refit	-2.40	[-4.05, -0.25]	[-4.45, +0.28]
Fusion minus joint native	+0.65	[-0.50, +3.10]	[-0.82, +3.47]

Table 8: All confirmation EM contrasts; the first three are primary. Target and retrieved-title intervals hold fits fixed; S+T refits source parameters with target question resampling. All shown intervals are descriptive 95%. Title grouping is not annotated-page independence, and no interval covers reference errors.

Contrast	Δ	Target	S+T	Title groups
Refit minus historical	+3.55	[+2.45, +4.40]	[+1.35, +5.30]	[+2.45, +4.45]
Fusion minus refit	-2.40	[-3.80, -0.95]	[-4.05, -0.25]	[-3.69, -0.90]
Fusion minus joint native	+0.65	[-0.35, +2.00]	[-0.50, +3.10]	[-0.30, +1.94]
Fusion minus historical	+1.15	[-0.50, +2.60]	[-1.05, +3.30]	[-0.40, +2.56]
Fusion minus comparison	+0.60	[-0.25, +1.50]	[-0.45, +1.75]	[-0.20, +1.41]
Fusion minus no passages	+2.05	[+0.90, +3.10]	[+0.55, +3.75]	[+0.96, +3.07]
New 700 minus historical 700	+3.55	[+2.40, +5.00]	[+1.05, +5.45]	[+2.36, +4.95]
New 2,526 minus new 700	+0.00	[-1.05, +0.60]	[-1.65, +1.75]	[-1.05, +0.60]

strata below. It changes no fit, label, threshold, primary comparison or confidence interval. Every original comparison is included. The crossed sets below are diagnostic constructions, not additional trained or independently validated policies.

Original-answer arm accounting. The full pool contains 3,275 historically selected RAG answers and 725 closed-book answers; their reference EM is 44.15% and 18.48%. Refitting shifts 110 retained slots from closed-book to RAG, adding 100 net reference-correct RAG answers while losing 29 closed-book ones. Unequal subgroup retention counts prevent interpreting this as a fixed-subgroup-coverage effect. Fusion and refitted native differ by only five slots in their arm allocation.

Crossing observed rankings with observed arm quotas. Let $C(r, q)$ be the number of retained reference-correct answers when using scorer r's within-arm order and scorer q's observed RAG/closed-book retention counts. Every constructed set still contains 2,000 unchanged answers. Original hash tie-breaking is retained. $C(A, A)$ and $C(B, B)$ reproduce the compared deployed sets exactly. The two off-diagonal conditions isolate changes in ordering or allocation under fixed other components, without claiming an intervention on the training distribution.

For refit–historical, historical ranking with refitted quotas yields 58.15%, while refitted ranking with historical quotas yields 59.35%, against the original 57.25% and 60.80%. Both components matter under this definition. For fusion–refit, fusion ranking with native quotas is 58.50%, while native ranking with fusion quotas remains 60.80%. The fusion gap therefore persists at matched arm counts.

Symmetric descriptive accounting. To avoid selecting one ordering of component changes, average the two paths from (B, B) to (A, A):

$$R = \tfrac{1}{2}\{C(A,B) - C(B,B) + C(A,A) - C(B,A)\}, \tag{4}$$

$$Q = \tfrac{1}{2}\{C(B,A) - C(B,B) + C(A,A) - C(A,B)\}. \tag{5}$$

Then $R + Q = C(A, A) - C(B, B)$ exactly. These are analyst-defined accounting components, not causal effects; half-integer counts arise from averaging two paths, not fractional observed questions.

Table 9: Complete retention-set accounting. Correct means matching the original reference EM; it is not newly verified factual correctness. All contrasts use the same fixed original answer per question and denominator 2,000 retained answers.

Contrast	Enter/leave	Enter correct	Leave correct	Net
Refit minus historical	207/207	136	65	+71
Fusion minus refit	405/405	121	169	-48
Fusion minus joint native	276/276	120	107	+13
Fusion minus historical	463/463	198	175	+23
Fusion minus comparison	138/138	59	47	+12
Fusion minus no passages	254/254	120	79	+41
New 700 minus historical 700	311/311	181	110	+71
New 2,526 minus new 700	104/104	45	45	+0

Table 10: Secondary F1 contrasts in percentage points, with fixed-target and source/target descriptive 95% intervals. These do not replace the primary EM endpoint.

Contrast	Δ F1	Target	S+T
Refit minus historical	+3.09	[+2.08, +3.92]	[+1.11, +4.71]
Fusion minus refit	-1.20	[-2.45, +0.13]	[-2.55, +0.73]
Fusion minus joint native	+0.80	[-0.12, +2.05]	[-0.13, +3.05]
Fusion minus historical	+1.89	[+0.48, +3.24]	[-0.08, +4.08]
Fusion minus comparison	+0.08	[-0.58, +0.92]	[-0.73, +1.47]
Fusion minus no passages	+1.54	[+0.54, +2.54]	[+0.45, +3.25]
New 700 minus historical 700	+3.44	[+2.33, +4.57]	[+1.42, +4.98]
New 2,526 minus new 700	-0.35	[-1.19, +0.20]	[-1.60, +0.85]

The equal 60.80% EM values for full and subset refits conceal different retained sets: they overlap on 1,896 answers, exchanging 104 in each direction. Both off-diagonal constructions yield 61.10%, but using that observed outcome to choose a new policy would require fresh confirmation. Zero symmetric components for this contrast do not mean identical rankings or absence of interaction.

Source composition and prediction scale. Historical fitting uses 200 NQ and 500 TriviaQA questions. Their RAG reference EM is 36.0% and 71.8%, respectively, versus 37.53% in the newer NQ fitting cohort. Historical mixed-source closed-book EM is 43.57%, versus 16.5% in historical NQ and 18.17% in newer NQ. These source summaries include both generated arms; target summaries condition on the historically selected original arm. They are consistent with fitting-source mismatch, but do not causally distinguish prior shift, conditional feature changes, filtering or other source differences.

On all 4,000 fixed originals, observed reference EM is 39.50%. Historical native predicts 51.65% on average with Brier loss 0.2140; full refitting predicts 37.04% with Brier 0.1869, and the new 700 fit predicts 37.53% with Brier 0.1869. These are post-result reference diagnostics without new uncertainty estimates. Brier loss is not a pure calibration measure or evidence of factual truth.

Native coefficients and score reconstruction. Writing the binary fit as $\operatorname{logit} p = \beta^\top x + b$ in raw feature units independently reconstructs every saved native target score to maximum absolute error below 5×10^{-16}. The feature order below is contextual value mean, no-context value mean, $\log(1+T)$ and overlap. Coefficients change along with intercepts, so adaptation is not a common monotone transform of one scalar score. Correlated predictors and differing source distributions prevent interpreting coefficient magnitude as causal feature importance. Full strata, coefficients, feature summaries, score reconstructions, rank correlations and the sixteen diagnostic crossed sets accompany the numerical artifact. No predictor is updated using these results.

A.3 Completed review of retention differences

Sampling and scope. Before target results, the review rule fixed three primary contrasts, entered/removed directions and original-answer EM=1/0 strata. Each of the twelve cells contributes

Table 11: Actual retained original-answer arms. Each scorer retains 2,000 answers; correct counts use unchanged reference EM.

Scorer	RAG retained	RAG correct	Closed retained	Closed correct
Historical native	1799	1083	201	62
Refitted native	1909	1183	91	33
New-source 700	1937	1193	63	23
Joint native	1922	1127	78	28
Joint comparison	1887	1117	113	39
Fusion	1914	1137	86	31
No direct passages	1916	1098	84	29

Table 12: All eight original contrasts with diagnostic crossed sets. Entries are retained reference EM (%). Columns give (ranking, quota) sources; A/B denote the conditions on the left/right of each contrast. No off-diagonal condition is promoted to a primary method.

Contrast A–B	(B,B)	(A,B)	(B,A)	(A,A)
Refit–historical	57.25	59.35	58.15	60.80
Fusion–refit	60.80	58.50	60.80	58.40
Fusion–joint native	57.75	58.55	57.55	58.40
Fusion–historical	57.25	58.25	58.30	58.40
Fusion–joint comparison	57.80	58.30	58.15	58.40
Fusion–no direct passages	56.35	58.50	56.35	58.40
New 700–historical 700	57.25	59.95	58.65	60.80
Full 2,526–new 700	60.80	61.10	61.10	60.80

up to eight questions by a deterministic hash, seed 2026092545. This yields 96 memberships and 93 unique questions from the 1,224-question union of primary retention differences. All 93 were reviewed against question, original answer, reference list, proposal, four actual comparator snippets, seven confidence scores, eligibility and input truncation. The remaining 1,131 changed questions have raw archives but were not semantically reviewed. Balanced case sampling is not population-weighted evidence of error prevalence.

All 4,000 question/reference pairs were traced to the first valid normalized-question row in the original public NQ-open source. They match exactly. This excludes a local question/reference join mistake; it neither verifies benchmark truth nor checks all original full-NQ annotations. No new human labels or model-judge calls were collected. The unblinded assistant review does not change EM/F1, labels, retained sets, fitted parameters, thresholds or rules. All deployed joint heads and the refitted native predictor equal the earlier development fits, and the entire point-fit file matches its pre-generation freeze.

Illustrative cases, not a semantic rescoring. The complete 93-case archive contains every selected question, raw input, score and note. The examples below were chosen after inspection to explain distinct interpretation problems; their frequency is not an estimated population rate. "Entered" and "removed" refer to fusion relative to the refitted practical native baseline.

- *Supported set, EM mismatch.* For the composers of *Bring It On: The Musical*, the original answer names Tom Kitt and Lin-Manuel Miranda together. The supplied passages name both, but the reference lists their names separately and EM is zero. Fusion removes this answer. A reference-metric improvement from removing an EM-zero item need not improve semantic quality.
- *Supported correct answer removed.* For the number of episodes in season three of *Into the Badlands*, the original answer and reference are 16. The proposal is 10. The actual passages explicitly distinguish the ten-episode second season from the sixteen-episode third season. Fusion removes the original answer.
- *Wrong work despite lexical support.* A question names *Love Actually*, while the retrieved passages describe Juliette Barnes and Dante from a different work. Fusion enters "Dante"

Table 13: Symmetric accounting of the retained reference-correct count. All values are descriptive; no new confidence interval or primary test is attached to this post-result decomposition.

Contrast	Within-arm rank R	Arm allocation Q	Total
Refit–historical	+47.5	+23.5	+71
Fusion–refit	-47.0	-1.0	-48
Fusion–joint native	+16.5	-3.5	+13
Fusion–historical	+11.0	+12.0	+23
Fusion–joint comparison	+7.5	+4.5	+12
Fusion–no direct passages	+42.0	-1.0	+41
New 700–historical 700	+48.5	+22.5	+71
Full 2,526–new 700	+0.0	+0.0	+0

Table 14: Frozen native log-odds coefficients in common raw-feature units. Original fits remain unchanged.

Fit	Arm	β_D	β_0	β_T	β_O	b
Historical native	rag	11.039	0.763	-2.354	-0.187	6.599
Historical native	closed	1.447	5.162	-2.375	0.946	5.419
Refitted native	rag	8.800	0.341	-0.969	0.078	2.349
Refitted native	closed	1.181	3.729	-1.295	0.467	2.250
New-source 700	rag	9.861	0.261	-0.628	0.644	1.158
New-source 700	closed	1.252	2.841	-1.013	0.970	1.148

against the reference "Peter". The occurrence of the answer in a passage is not support for the requested entity relation.

- *Reference type error does not prove the answer.* A question asks where the first post office opened in India, but its preserved reference is a date. Fusion enters "Allahabad". The actual snippet associates Allahabad with an official based there in 1870; it does not establish the first-opening location. Both the reference problem and the unsupported proposed interpretation must be retained.
- *Incompatible reference quantity.* One revolution in radians is answered as 2π, while the preserved reference is $180/\pi$ degrees, the conversion for one radian. Fusion enters the answer. The original and proposal are identical, so no comparator forward is executed on this case. This mathematical inconsistency is a qualitative observation, not a changed scoring label.
- *Numeral format and evidence scope.* For the last season of *New Girl*, the original "8" fails EM against "eight". Fusion removes it. Although the numeral expressions are equivalent, the actual snippets discuss other seasons; that separate support limitation is not repaired by accepting a numeric alias.

Invocation accounting over all retention changes. A local comparison forward runs only for a valid, canonically different proposal. Equal or null proposals use the frozen structural constraints. Different fitting functions, metadata and the global retention cutoff can still change such questions' retention. Consequently a changed retained item is not necessarily a new comparator verdict. Fusion reallocates 163 retained slots from comparison-forward cases to non-forward cases relative to the practical refitted baseline. The former contribute -108 net reference-correct answers and the latter +60, totaling -48. This describes where differences occur; it does not establish that skipping comparison causes harm. Against joint native, the +13 net contribution is entirely in the comparison-forward stratum. The two contrasts answer different questions.

Result boundaries. The primary refit–historical adjusted interval has positive lower bound, but that bound is below 2pp; it does not guarantee a gain of at least 2pp. Fusion–refit excludes a +2pp gain under the prespecified bootstrap while crossing zero, whereas fusion–joint native still allows more than +2pp. Neither is an equivalence test. Source adaptation, metadata/target changes, reference error and broad factual correctness remain distinct explanations or outcomes.

Table 15: Descriptive accounting of every primary retention difference, using the unchanged reference labels. Forward/non-forward strata have unequal retained counts; these are not subgroup fixed-coverage effects or causal ablations. Correct-in/out counts sum to the main-text totals.

Contrast	Comparator forward	Entered / removed	Correct in / out	Net correct
Refit–historical	Yes	154 / 121	94 / 31	+63
	No	53 / 86	42 / 34	+8
Fusion–refit	Yes	210 / 373	48 / 156	-108
	No	195 / 32	73 / 13	+60
Fusion–joint native	Yes	271 / 269	119 / 106	+13
	No	5 / 7	1 / 1	0

B POST-RESULT BINARY FITS AND INTERCEPT INTERVENTIONS

Table 16: Post-result exploration on the already-observed 4,000 questions, always retaining 2,000 originals. No target labels fit any predictor. Native4 is the preserved frozen strong baseline; all new conditions use its source cohort. Point estimates alone do not establish an incremental effect.

Condition	Reference EM	Reference F1
Historical native	57.25	69.10
Historical + common intercept	57.25	69.10
Historical + separate-arm intercepts	57.70	69.47
Full Native4 refit	60.80	72.19
Native6: Native4 + availability	61.00	72.55
NativeCompare9: Native6 + comparison	60.90	73.11
NativeCompare9 without direct passages	60.65	72.32

Experimental status. These conditions were designed after seeing the frozen 4,000-question outcomes and case review. The original seven scores, fitted parameters, labels, retained sets and primary intervals remain unchanged. All new fits use only the existing 2,526 source questions; no new large model, generation, comparator call or outcome judgment is used. The original 594 development questions and observed 4,000 target questions retain their historical roles. The design fixes 50% as the primary coverage and reports 20% and 80% as diagnostics, with no model or threshold selection by their results.

Matched information and missingness. For arm $a \in \{\mathrm{RAG}, \mathrm{closed}\}$, let $x_{ia} \in \mathbb{R}^4$ be its native features, H_{ia} indicate historical choice of that arm, and F_{ia} indicate an actual comparison forward for that chosen original. Define

$$x_{ia}^{(6)} = [x_{ia}, H_{ia}, F_{ia}], \tag{6}$$

$$x_{ia}^{(9)} = [x_{ia}^{(6)}, H_{ia}c_i], \tag{7}$$

where $c_i = 0$ whenever no comparison forward exists. The feature value zero is accompanied by $F_{ia} = 0$, distinguishing unavailable comparison from an actual neutral score. There is no reference-based imputation. Each arm is fitted on all 2,526 rows using its own binary EM labels. Native6 and NativeCompare9 share the same objective, optimizer, regularization, standardization, source questions and base inputs; only the three comparison dimensions differ. The no-direct variant replaces c_i with its stored window-free counterpart.

The selected original is never changed by these fits. In particular, a high predicted score for the other arm cannot replace it. NativeCompare9 minus Native4 is the practical added-recipe comparison; NativeCompare9 minus Native6 isolates comparison information conditional on the availability features. Fusion versus NativeCompare9 changes target, per-arm structure and metadata together, so their difference does not isolate the four-state target's effect.

Offsets. The common offset minimizes binary cross entropy across both historical arm logits on the newer fitting sample. Separate offsets minimize it arm by arm. No slope changes or temperature

grid are introduced. Both receive all 2,526 labeled rows per arm, including arms not chosen as the original. The common strictly increasing map leaves all rankings unchanged. Separate offsets preserve each arm's ordering and alter only inter-arm competition. Full refitting can alter both.

Observed results. Table 16 reports all new 50% point conditions; Table 17 retains every condition at all three coverages. NativeCompare9, Native4 and Native6 have 594-question development EM of 60.606%, 56.902% and 57.239%, respectively. Their observed-target differences and source/target uncertainty are reported together below.

Table 17: All original and exploratory conditions at fixed coverages on the observed 4,000 questions. Entries are retained reference EM / F1 (%). Only 50% is the designated primary coverage. These diagnostics do not restore an unseen-target interpretation for new fits.

Scorer	20%	50%	80%
Historical native	69.12 / 79.37	57.25 / 69.10	47.12 / 59.61
Native4 refit	71.38 / 81.53	60.80 / 72.19	47.47 / 60.59
New-source 700	72.88 / 83.25	60.80 / 72.54	47.09 / 60.49
Joint native	67.38 / 80.27	57.75 / 70.19	46.34 / 59.14
Joint comparison	65.88 / 78.40	57.80 / 70.91	46.09 / 59.02
Fusion	67.88 / 81.15	58.40 / 71.00	46.81 / 59.78
Fusion without direct passages	66.88 / 79.84	56.35 / 69.45	46.38 / 59.14
Common intercept	69.12 / 79.37	57.25 / 69.10	47.12 / 59.61
Separate-arm intercepts	70.00 / 80.16	57.70 / 69.47	47.25 / 59.87
Native6	70.12 / 81.70	61.00 / 72.55	47.38 / 60.41
NativeCompare9	73.25 / 83.84	60.90 / 73.11	47.31 / 60.31
NativeCompare9 without passages	69.75 / 81.40	60.65 / 72.32	47.53 / 60.35

Paired source and target uncertainty. The 2,000 source draws exactly reconstruct the frozen historical NQ200/TriviaQA500 sampling followed by newer-source annotated-page sampling. Each draw uses the same question multiplicities across original and new conditions; saved old parameters are reused. New parameters are all saved before this target analysis. Target resampling uses the original 2,000 paired question draws and recomputes retention. Original seven-score 50% replicate arrays and all eight original contrasts are reused unchanged. The new three designated contrasts receive nominal Bonferroni-three 98.33% limits in addition to 95% intervals. These intervals quantify resampling variability under the chosen recipe, not the target-informed choice of that recipe. Coverage diagnostics at 20% and 80% have fixed-fit target intervals only, without further source refits.

Table 18: All exploratory EM contrasts at 50% coverage (percentage points). First three are the designated family. B3 denotes nominal 98.33% limits; secondary contrasts retain 95% limits. All conditions were designed after observing this target.

Contrast	Δ	Fixed-fit 95%	S+T 95%	S+T B3
Compare9–Native4	+0.10	[-0.80, +1.30]	[-0.90, +2.00]	[-1.25, +2.25]
Compare9–Native6	-0.10	[-0.85, +0.85]	[-1.05, +1.30]	[-1.27, +1.60]
Native4–arm offsets	+3.10	[+2.05, +3.95]	[+0.90, +4.60]	[+0.52, +5.05]
Compare9–no passages	+0.25	[-0.40, +1.30]	[-0.55, +1.70]	—
Native6–Native4	+0.20	[-0.45, +0.90]	[-0.55, +1.40]	—
Arm offsets–historical	+0.45	[-0.15, +1.00]	[-0.10, +1.40]	—
Common offset–historical	+0.00	[+0.00, +0.00]	[+0.00, +0.00]	—

NativeCompare9 exchanges 206 answers with Native4, entering 81 reference-correct originals and removing 79 (net +2). Against Native6 it exchanges 143, entering 55 correct and removing 57 (net -2). The positive secondary F1 difference against Native4 is +0.91pp with descriptive source/target 95% interval $[0.02, 2.39]$; against Native6 it is +0.56pp, $[-0.30, 1.78]$. This preserves the useful F1 signal without substituting it for the designated EM increment.

Question-level diagnostic review. The specified review design selects up to four cases per entered/removed $\times$ original EM cell for NativeCompare9–Native4 and NativeCompare9–Native6 by

Table 19: All exploratory secondary F1 contrasts at 50% coverage, in percentage points with descriptive 95% intervals. F1 remains secondary even where an interval is favorable.

Contrast	Δ F1	Fixed-fit 95%	S+T 95%
Compare9–Native4	+0.91	[+0.08, +1.94]	[+0.02, +2.39]
Compare9–Native6	+0.56	[-0.12, +1.47]	[-0.30, +1.78]
Native4–arm offsets	+2.72	[+1.84, +3.53]	[+0.76, +4.26]
Compare9–no passages	+0.79	[+0.18, +1.78]	[+0.04, +2.02]
Native6–Native4	+0.36	[-0.23, +0.93]	[-0.34, +1.31]
Arm offsets–historical	+0.37	[-0.19, +0.87]	[-0.31, +1.02]
Common offset–historical	+0.00	[+0.00, +0.00]	[+0.00, +0.00]

seed 2026092642. All 32 memberships, comprising 23 unique questions, have now been inspected against the question, reference, fixed original, proposal, four actual snippets and current scores. None is in the earlier 93-question review. The broader archived union has 771 changed questions; the other 748 were not semantically reviewed in this round. This balanced, unblinded assistant review changes no labels or model decisions and estimates no population semantic accuracy. The complete 23-case record is preserved with the experimental archive.

The favorable cases include the *Hell's Kitchen All Stars* and *The New Legends of Monkey* premieres: the supplied snippets support September 29, 2017 and 28 January 2018, while proposals concern a different film and Singapore's monkey-tree phenomenon. NativeCompare9 retains both relative to Native6; Native4 already retained them. For the guitarist on *Two Tickets to Paradise*, NativeCompare9 retains Jimmy Lyon relative to both controls, consistent with the snippet's explicit guitar-solo credit, while the proposal is a name associated with the same-titled film. Opposite cases include the Mars Orbiter Mission launch and the historical service-tax effective date: originals "5 November 2013" and "1 June 2016" are directly supported and match references, but NativeCompare9 drops them despite proposals expressing the same dates in another word order. For J. Michael Finley and Louis Hynes, original and proposal are identical, so no comparator forward occurs; Native6 already realizes the retained-answer gains over Native4.

Complete F1 accounting of retention changes. Summing every entered and removed original, NativeCompare9's +0.914pp F1 difference from Native4 comprises +0.100pp from net EM-one answers and +0.814pp from partial token overlap among EM-zero answers. Its +0.559pp difference from Native6 comprises -0.100pp from EM-one answers and +0.659pp from EM-zero partial overlap. These sums cover all retention changes, not only the 23 reviewed questions. They explain how F1 can rise when the exact-match increment is small or negative: the additional measured signal is largely overlap among answers that remain nonmatches. Token-overlap accounting does not assign semantic correctness to those responses.

Closed follow-up rule. A single post-case agreement guard uses Native6 for missing comparisons, case/whitespace-equal answers or strictly equivalent complete dates, and Compare9 otherwise. On the same observed 4,000 it reaches 61.05% EM; its three adjusted source/target intervals versus Native6, Compare9 and Native4 are respectively +0.05pp $[-1.07, 1.70]$, +0.15pp $[-0.20, 0.55]$ and +0.25pp $[-0.97, 2.50]$. The exploratory rule adds no established increment and was not added to the fresh confirmation.

C Authorized partial replay of native policies on additional questions

Cohort and released scope. This analysis uses another 4,000 questions from official NQ-open training data, with zero normalized-question overlap against all registered/historical project questions, including the earlier target. The generator, retrieval, candidate prompts, historical-original construction and source parameters are unchanged. It supplies new project-question evidence in the same generator and data source, not an official test split, a cross-model replication or annotated-page separation.

The full planned experiment has twelve scorers. The four native conditions can be evaluated using completed RAG and closed-book outputs without waiting for proposals/comparison features: historical native, full Native4 refit, new-source native 700, and arm-specific intercept updates. Their point parameters and all 2,000 source-parameter sets were frozen before new generation. No source model is refitted here. Every scorer retains 2,000 of the same fixed historical originals, using the unchanged legal-answer rule and question-hash ties.

Early reference visibility. An explicit user-authorized exception began reading new reference values at 11:15:59 UTC on September 26, 2026, while the twelve-condition pipeline remained unfinished. Historical Dual4 original selections had already been saved. Full-pool RAG/closed-book metrics were inspected before this native-only analysis was specified. Therefore the project was no longer blind to all new target quality until the complete twelve-score lock. The native subset separately locked its point scores/retained sets and all 2,000 source rankings before computing its own 50% metrics; no frozen predictor, threshold, comparison family or reference label changed.

The present status is a *partial local replay*. Matching its original answers, all scores/orders, retained sets, point values, bootstrap outcomes and intervals against the eventual complete pipeline remains pending. The two comparison-feature primary endpoints have not been evaluated in this release. The partial results do not substitute for completion of that registered experiment.

Statistics and full outcomes. Each of 2,000 target question resamples uses seed 2026092644 and the correspondingly indexed saved source-parameter draw. Full Native4 minus arm offsets is a preplanned primary-family member for this cohort and retains nominal Bonferroni-three 98.33% limits; the adjustment is not relaxed because the other two members are pending. Adaptation versus historical native, the matched-count source contrast, and full versus subset are prespecified secondary comparisons with descriptive 95% intervals. F1 remains secondary. Source resamples retain the earlier stratified historical and annotated-page newer-source construction. These intervals omit reference errors, common model error and unmodeled dependence; null-containing intervals are not equivalence tests.

Table 20: Complete partial native-only contrasts. All values are percentage points; all displayed uncertainty includes paired source and target variation. B3 is available for the original primary member only. F1 is secondary.

Contrast	Δ EM	EM 95%	EM B3	Δ F1	F1 95%
Native4–arm intercepts	+2.50	[+0.70, +4.05]	[+0.35, +4.45]	+2.53	[+0.77, +4.07]
Native4–historical	+3.45	[+1.20, +4.75]	—	+3.25	[+1.14, +4.69]
New 700–historical	+3.25	[+0.55, +5.00]	—	+3.45	[+1.26, +4.90]
Native4–new 700	+0.20	[-1.50, +1.85]	—	-0.20	[-1.43, +0.91]

Table 21: Fixed-fit target intervals and complete retention accounting for the partial native-only replay. These fixed-fit intervals omit source-fit variation.

Contrast	EM fixed 95%	F1 fixed 95%	In/out	Correct in/out
Native4–arm intercepts	[+1.50, +3.40]	[+1.71, +3.36]	184/184	119/69
Native4–historical	[+2.30, +4.50]	[+2.31, +4.11]	230/230	148/79
New 700–historical	[+1.90, +4.65]	[+2.15, +4.56]	354/354	196/131
Native4–new 700	[-0.85, +1.05]	[-1.03, +0.62]	124/124	52/48

As full-pool descriptive context, RAG EM/F1 is 39.175/51.939%, closed-book is 19.025/28.528%, and the historically chosen original is 39.925/52.827% on all 4,000 questions. Historical selection chooses RAG for 3,342 and closed-book for 658. The reference-based two-candidate oracle is 43.725%. These 100%-denominator quantities measure candidate quality/availability, not the 50% retained-accuracy contrasts above; the oracle is not a deployed result.

Completed balanced case review. Before inspecting the subset's retained-quality result, a fixed hash rule selects up to four cases per entered/removed × EM=1/0 cell for full-refit versus arm offsets

Table 22: All four absolute native-condition source/target 95% intervals at 50% coverage on the new cohort. Point EM/F1 are in Table 3.

Scorer	EM S+T 95%	F1 S+T 95%
Historical native	[+54.50, +59.60]	[+67.63, +72.15]
Native4 refit	[+57.70, +62.35]	[+70.89, +74.66]
New-source native 700	[+57.40, +62.55]	[+71.10, +74.88]
Arm intercepts	[+55.25, +60.25]	[+68.16, +72.49]

and up to two for refit versus historical native, using seed 2026092646. All 24 memberships, comprising 17 unique questions, were inspected against the original question, references, both actual generation prompts/answers, all four scores and retention states. Two assistants divided the work; this is not independent human annotation or a population-weighted semantic estimate. All original labels and results remain unchanged.

The changes include useful retention and genuine errors. Refitting retains the supported *Pray for the Wicked* release date, June 22, 2018, while historical and arm-offset scoring omit it. It also retains "106" for Frank Lampard's England goals, although the actual passage assigns 106 to appearances and the reference says 29. The fixed-answer protocol does not substitute the available closed-book answer 29. Other reviewed losses include the reference-correct MLB answer 162 and a supported mRNA answer; higher aggregate EM does not protect every supported answer. Full-name versus surname, combined versus separately listed answers, and actor-versus-character ambiguity also separate EM from semantic interpretation. Complete question-level notes are preserved with the new partial-analysis record.

D Essential method and statistical definitions

D.1 Frozen fitting and central feature construction

BM25 retrieves 100 Wikipedia passages; `BAAI/bge-reranker-v2-m3` selects twenty. Frozen Qwen3.5-9B generates canonical JSON answers greedily with thinking disabled and a 64-token cap: RAG uses passages 1–5, closed-book uses none, and the supplementary proposal uses 6–20. Historical Dual4 chooses the fixed original before retention scoring.

The central per-arm binary predictor and joint heads standardize each feature using only the fitting sample. Constant features use unit scale. Fits start at zero and use the preserved regularized softmax objective. For standardized z_i, class y_i, feasible-state set S_i and an unpenalized intercept included in z_i, the joint objective is

$$\frac{1}{n}\sum_i \left\{ \log \sum_{s\in S_i} e^{z_i^\top w_s} - z_i^\top w_{y_i} \right\} + \frac{\|W_{\text{slopes}}\|_F^2}{n}. \tag{8}$$

The binary fit has two feasible states per labeled row; the four-state head masks states inconsistent with exact/null answers. The point-fit native and joint parameters used in the frozen confirmation are preserved from development, except the prespecified new-source 700 subset fit. No target outcomes choose a regularization value, checkpoint, threshold or seed.

Evidence windows are ranked reference-free by the same BGE reranker. Windows have 56 words, stride 40 and at most 14 title words. Keep the best window per document, then choose the top two distinct documents from ranks 1–5 and two from ranks 6–20; ties use document index and window start. Packing uses the legacy DeBERTa-large-MNLI tokenizer, with 180 question/candidate tokens and at most 80 tokens per evidence block within 512 tokens. The rendered Qwen prompt is checked against 1,024 tokens. Truncation keeps first parts and remains recorded. Original/proposal roles move with each candidate under the two orders; only legal, canonically different proposals trigger comparator forwards.

The comparator has three allowed outcomes in each order: prefer the first answer, neutral, prefer the second (tokens A, N, B, respectively). NF4 uses double quantization with BF16 computation; all adapters are disabled. Small fits use L-BFGS-B with the main-text tolerances; unsuccessful

termination is accepted only if maximum absolute gradient is below 10^{-7}. Writing allowed-code probabilities as $p_{o,k}$ for order $o \in \{0,1\}$ and outcome $k \in \{0,1,2\}$, with original first in order 0,

$$b = (p_{0,2} + p_{1,0})/2, \quad h = (p_{0,0} + p_{1,2})/2, \quad n = (p_{0,1} + p_{1,1})/2. \tag{9}$$

$$c_3 = \left[\log \frac{b+\epsilon}{n+\epsilon}, \log \frac{h+\epsilon}{n+\epsilon}, |(p_{0,2} - p_{0,0}) - (p_{1,0} - p_{1,2})|\right]. \tag{10}$$

The checkpoint is the frozen NF4 Qwen base with adapters disabled. These preference scores are inputs to learned correctness models; they are not themselves independently calibrated truth probabilities.

D.2 First validation: complete responses and three-state routing

RAG and closed-book candidates come from the same generator checkpoint, with official chat templates, greedy decoding, thinking disabled and a 128-token cap. The RAG instruction asks for an answer using supplied context; the closed-book instruction asks without context. Empty context inside a RAG instruction is a distinct historical control. Own3 scores native full-response likelihood excluding terminal EOS, $\log(1+T)$ and lexical overlap. Evidence5 adds two candidate-conditioned document-relation scores. Qwen3.5-9B sees the question, five passages and candidate without references or generator identity. A frozen Llama-3.1-8B component supplies a nonanswer flag from question and RAG response.

Separate per-generator, per-arm regularized multinomial fits predict $p_{\text{nonanswer}}$, p_{correct} and p_{wrong}, with uniquely resolved training labels and training-only standardization. Define $u(y) = p_{\text{correct}}(y) - p_{\text{wrong}}(y)$. The conservative policy chooses closed-book iff the RAG nonanswer flag is positive and $u(k) > \max(0, u(r))$; ties retain RAG. A secondary utility rule drops the gates and compares $u(k) > u(r)$. Pure RAG, pure closed-book and an earlier phrase-triggered guard are retained controls. A local Context4 variant replaces the two candidate-relation additions with a question–passage sufficiency score shared by candidates and generators.

Development uses 200 NQ and 500 TriviaQA questions (Kwiatkowski et al., 2019; Joshi et al., 2017), with question-level shared folds. Full-development fits and rules are frozen before the 2,000 new-question validation. EfficientQA primary references combine original answers and official definitely-correct historical predictions; possibly-correct historical answers are excluded. Original-only references form a sensitivity endpoint. TriviaQA uses canonical answers and aliases. Current outputs never expand the primary reference lists.

D.3 Unknown outcomes, yields and uncertainty

Let c indicate correct completion and a indicate a substantive answer. Feasible states are $(a, c) \in \{(0,0), (1,0), (1,1)\}$, with wrong indicator $w = a - c$. For retention weights g_i,

$$Y_c = N^{-1} \sum_i g_i c_i, \qquad Y_w = N^{-1} \sum_i g_i w_i, \qquad U_\lambda = Y_c - \lambda Y_w. \tag{11}$$

All-output evaluation sets $g_i = 1$; fixed-80 ranks predicted task error and retains $\lceil 0.8N \rceil$ weighted mass with fractional boundary ties. Error costs 0.5 and 2 are fixed-choice sensitivities around primary $\lambda = 1$. An unknown outcome ranges over its feasible states. Identical complete responses share the same unknown state across compared policies; their contributions cancel when their weights match. Accepted machine labels constrain these bounds but may themselves be wrong.

The first validation uses 2,000 paired fixed-prediction question bootstrap draws and 500 source/target draws with refitting. Historical source resampling preserves the 200/500 mixture and shares draws across models, arms and feature groups. Bonferroni-eight endpoint envelopes accompany all eight primary comparisons. These are descriptive bounds over machine-label constraints; the extreme adjusted tails have limited resolution with 500 source refits. Shared judge error and prior development selection remain outside them. Fixed80, Context4, single-judge, failure-projection and label-relaxation analyses are supplementary.

D.4 Second validation: canonical answer fields

The second protocol uses a single JSON answer field, a 64-token cap, and a RAG prompt permitting internal knowledge with titles/text visible. The canonical value event includes tokens merged

Table 23: Primary measurement with original references: agreement with reused released human labels on historical short answers. The published source contains 687 question clusters; its 2120 definite response targets include 358 positives and 1762 negatives. Known coverage and known-subset agreement are percentages; all-denominator and balanced agreement are identification bounds (%). Balanced agreement weights the positive and negative classes equally. Unknown machine outputs stay in the bounds. This does not estimate human accuracy for the current generated full responses.

Evaluator	Known	Known agree	All agree bounds	Balanced bounds
Qwen3-8B	100.0	88.6	[88.6, 88.6]	[69.5, 69.5]
Gemma3-12B	99.9	88.7	[88.6, 88.7]	[70.9, 71.1]
Original consensus	79.8	91.5	[73.1, 93.3]	[58.1, 82.6]
Failure projection	96.6	90.0	[86.9, 90.4]	[66.4, 74.1]
Reference EM	100.0	83.7	[83.7, 83.7]	[52.1, 52.1]

Table 24: Shared errors under original references. False acceptance is both judges returning A on a human-negative target. False rejection is both returning B/C/D on a human-positive target, including nonanswers; both B is shown separately. These are historical published-short-answer counts, not current-response error rates.

Both A / human negative	Both BCD / human positive	Both B / human positive
23/1762	181/358	103/358

across the field boundary. Dual4 teacher-forces the same event under document and no-document prompts; Own-unit3 uses the candidate's native prompt and Common-unit3 uses the RAG prompt for both arms. All learned two-arm selectors share availability fallback and choose higher binary EM probability; single-arm baselines keep their own absence cases.

The eight primary tests compare Dual4 with Native3 and Direct4 in the four generator/source cells. Exact two-sided paired EM tests with Holm correction condition on the full-training fit. Separate 2,000 fixed-model bootstrap and 500 source/target-refit draws supply descriptive intervals without simultaneous-coverage guarantees. Official-normalizer sensitivity changes six candidate labels but none of eight primary point differences. Different prompts, references, output events and targets prevent comparing the first and second protocols as a controlled format-only experiment.

E Outcome measurement and model-judgment sensitivity

E.1 Supplementary model judgment on changed outputs

We additionally queried GLM-5.3 on the 287 distinct complete responses from 135 questions where Evidence5 and same-rule Own3 actually differed across the four first-validation cells. This set contains originally resolved and unresolved judgments; it is not a random sample of all outputs. GLM received the earlier eleven-example template defining the same five code categories, with question, references and response but without generator/policy identities or previous labels. The original local judges used a strengthened thirty-three-example template with additional rules distinguishing incomplete answers, refusals and required sets. All 287 saved GLM inputs exactly match the earlier template applied to the original question, reference list and complete response; both local judges' recorded script, rubric and source-file hashes match the later template. This prompt difference was identified during post-result case review and is retained in a provenance amendment.

GLM also used a reasoning API with an 8,192-token limit and provider-default temperature/seed; the local judges used nonreasoning allowed-code scoring. Thus templates, demonstrations, model weights and inference pipelines differ together. This is a judge-pipeline sensitivity, not a controlled comparison of model weights or independent adjudication.

A twelve-item formatting pilot was retained without repeat judgments. After documenting a parser amendment, only a single code or a literal `Code:` prefix followed by that code was accepted: 224 outputs met the strict form and 59 required prefix normalization. Four responses remained unknown;

Table 25: Supplementary model-judgment sensitivity for the four fixed Evidence5–Own3 comparisons. Net-yield bounds and fixed-policy endpoint envelopes are percentage points on each original 1,000-question denominator. GLM envelopes are nominal 95%, with no source refitting or multiplicity adjustment; they do not replace the original primary analysis.

Generator	Source	Original bounds	GLM bounds	GLM fixed 95%
Qwen3.5	EfficientQA	[-0.90, +2.90]	[+1.70, +2.10]	[+0.30, +3.60]
Qwen3.5	TriviaQA	[-0.90, +1.30]	[+0.00, +0.20]	[-1.10, +1.30]
Gemma4	EfficientQA	[-0.80, +2.40]	[+0.90, +1.10]	[-0.30, +2.40]
Gemma4	TriviaQA	[-0.60, +0.80]	[+0.40, +0.40]	[-0.80, +1.70]

Table 26: All sixteen known-state disagreements in the supplementary judge-pipeline comparison. Nonanswer combines refusal/no answer and no determinate answer. Counts do not measure factual accuracy.

Original consensus → GLM	Responses
Nonanswer → correct	2
Nonanswer → wrong	8
Wrong → nonanswer	1
Wrong → correct	3
Correct → wrong	2

no label was extracted from reasoning or explanatory text. Among 234 originally resolved task states, GLM agreed on 216, disagreed on 16, and left two unresolved. Disagreement does not identify which judge is wrong, and fewer unresolved states do not establish accuracy. No current response received a new human annotation.

Only Qwen EfficientQA has a wholly positive nominal fixed-policy net envelope in this sensitivity. Its GLM point bounds combine a 0.4–0.6-point decrease in correct yield with a 2.3–2.5-point decrease in wrong yield; the positive net difference is therefore error avoidance rather than an increase in correct completions. All four original primary comparisons and labels remain unchanged. The additional envelopes omit source-fit uncertainty, common judgment errors and multiplicity; the audited set also does not cover every source-refitted choice or fixed-coverage weight.

All resolved-state disagreements (post-result). Unblinded assistant review covers all sixteen differences in previously resolved task states, preserving the original and GLM labels. Eleven cross the substantive-answer/nonanswer boundary; five instead swap correct and wrong substantive states (Table 26). These are state changes, not sixteen independently verified judge errors. The other 51 previously unresolved items that GLM resolves are not claimed to have received complete semantic review. Three unparsed outputs contain nonconforming final text and one has no final text at the token limit; all four remain unknown under the frozen parser.

The saved cases show several distinct issues. A complete game name, “Tug of war”, receives two local no-answer labels but GLM marks it wrong against “Croquet”. “John Denver” occurs verbatim in the unchanged reference list yet both local judges mark that answer wrong; GLM accepts it. These are concrete concerns under the supplied reference standard, not estimates of full-pool judge accuracy. “Approximately 1 billion” versus 978 million raises an unspecified numerical-tolerance question. Two cast-list answers contain valid reference names but omit other listed names, making exhaustive-set versus representative-answer interpretation consequential.

Other cases caution against treating GLM as truth. The Hanway question asks about invention while the response distinguishes popularization; the Royal Museums Greenwich collection account describes popularizing umbrellas, so the question’s premise requires attention. This does not validate all claims in the response. A multi-region film answer mixes theatrical and disc release events; another answer supplies song authorship when the question asks for a singer. The “iris”/“eye” case concerns anatomical specificity supported by the National Eye Institute and NLM MeSH. The complete case archive preserves all sixteen responses, references, policy roles and interpretations,

including ambiguous and unfavorable examples. No new human annotation, semantic accuracy estimate or replacement policy result follows from this source-supported review.

F PRESERVED EXPERIMENTAL HISTORY AND ARCHIVE INDEX

The unchanged previous 102-page manuscript and source form a separate historical archive. All original primary frozen comparisons remain in the main text. Repeated development tables, full prompts, run histories and replay logs remain archived; this local preservation does not constitute external release or a new reproduction certificate.

Table 27: What remains in the full historical archive. Round identifiers locate the unchanged scientific record; they are not independent evaluation replicates. Both favorable candidate/predictive signals and failed continuation conditions are retained.

Archive group	Experiments retained	Main outcome or role
Early decoding and confidence	Fixed CAD, prompt controls, score-path checks, initial calibration and relation-feature development	Development trajectory; no universal CAD/AdaCAD conclusion
Frozen evaluations (v12, v15)	Full prompts, all policies, error costs, fixed80, context sufficiency, score and judge sensitivities	All sixteen primary comparisons remain in main Tables 5–6
Targets and revisions (v13–v19)	Partial-label objectives, matched targets, one-step revision, cross-family candidates, all seeds and costs	Candidate/selection increments fail their designated continuation screens
Candidate supply (v22–v25)	Nonlinear selection, span constraints, endpoint repair, ordinary versus supplementary retrieval	Supply oracle improves; delivered selector increment remains unresolved
Learned verification (v26–v33)	Anonymous and identity-aware pair models, QA-supervised verifiers, generative comparison, original fixed-pool fusion	Identity controls and native recalibration alter attribution; original favorable and unfavorable contrasts preserved
Editing (v34–v39)	SFT, direct gain scoring, frozen/matched heads, fold-excluded revision supervision, resource conditions, all fourteen final disagreements	Final fusion–native -0.17pp; 6/8 resource conditions fail; recipe ends
Central confirmation (v40)	Seven scores, 2,000 source refits, all contrasts, 93-case review, crossed rankings/quotas	Full numerical results and essential cases retained in this supplement
Judge-pipeline check (v41)	287 raw responses, parser provenance, sixteen resolved-state differences	283 parsed, four unknown; no new human truth labels

Archive identity. The preserved PDF is `historical_experimental_record.pdf`. Its unchanged source and original inputs remain in the versioned project record. Editorial manifests record exact hashes and section movements; all original experiment directories remain preserved.

Anonymous cached-check attachment. The self-contained `supplementary_material.zip` includes code, answers/references, features, point parameters, complete 2,000-draw outcomes, hash manifests and the historical 102-page PDF. Independent extracted checks passed. The original module recomputes 4,000 EM/F1 values, twelve point scorers, 36 retained sets, 2,000 fixed-fit draws and all eight confirmation/seven exploratory contrasts. A new native-only module checks the further 4,000 originals, all four scores/orders/retained sets, 2,000 fixed-fit draws, four contrasts and full-pool RAG/closed points. Source/target and original title-group intervals use complete cached outcomes. Source refitting, neural regeneration, the agreement guard and unfinished full-twelve-condition experiment are excluded. Local verification does not assert external publication.