

When Agreement Is Not Enough: Evidence-Risk Arbitration for Multi-Strategy Agentic QA

Yang He¹, Jinlin Wang¹, Yulong Ji^{2,*}, Jianwei Zhang³

¹National Key Laboratory of Fundamental Science on
Synthetic Vision, Sichuan University

²School of Aeronautics and Astronautics, Sichuan University

³School of Computer Science, Sichuan University
Chengdu 610065, China

*jyl@scu.edu.cn

Abstract

Multi-strategy question-answering agents often produce several plausible responses to one query, but must ultimately show the user only one. Voting treats answer agreement as trust, yet strategies can disagree for different reasons, and identical answers can rest on evidence of unequal quality. We cast this final choice as diagnosis-first evidence-risk arbitration and introduce the Evidence-Risk-Aware Strategy Reranker (ERSR), which canonicalizes a pool of six responses, classifies their agreement type, and applies a type-specific evidence rule while returning one unchanged response. Calibration uses only development data; for each evaluation item, trusted reference answers, trusted reference evidence, and utility stay hidden, and the selector never abstains. On a 900-item development split drawn from FEVER, HotpotQA, and SciFact, ERSR improves evidence-aware utility over a development-selected fixed strategy by +0.052, with the largest gains on label conflicts and unanimous answers with uneven evidence. These results suggest that response-set structure, not agreement alone, is a useful signal for choosing trustworthy answers.

1 Introduction

Multi-strategy agentic QA pipelines run several solution strategies (prompting, reasoning, retrieval, and verification) and must commit to one user-facing response (Lewis et al., 2020; Dhuliawala et al., 2024). Strategies may yield identical answers with unequal evidence, or equivalent answers in different surface forms. Once generation finishes, the system exposes one original response, making selection a post-generation decision under evidence constraints.

Existing evidence and factuality metrics diagnose retrieval, faithfulness, and unsupported claims (Es et al., 2024; Saad-Falcon et al., 2024; Min et al., 2023; Manakul et al., 2023), but stop short of the final evidence-sensitive choice (Section 2). Aliases

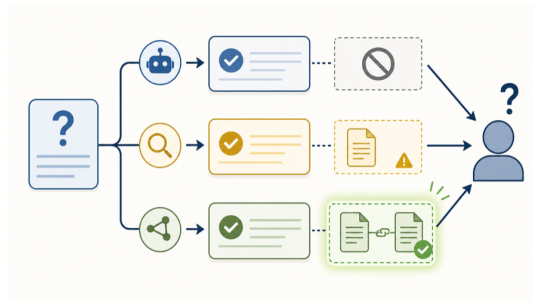


Figure 1: Why agreement is not enough. Strategies may disagree for different reasons, while identical answers may carry absent, partial, or complete evidence. A trustworthy selector must inspect the structure around the answer, not only count it.

and label conflicts can fragment genuine agreement, while unanimous answers can conceal sharply uneven evidence (Figure 1). For fixed completed responses without access to trusted reference answers or evidence, should selection follow *answer agreement* or *response-set structure*? ERSR canonicalizes the six answers, diagnoses their agreement type, and applies a type-specific evidence rule before returning one unchanged response.

We study this decision on FEVER, HotpotQA, and SciFact with six generated responses per item. We fit ERSR on a development fitting pool and evaluate on 900 held-out development items. All selectors receive the same response set without current-item trusted references or utility; baselines cover fixed strategies, voting, agreement, heuristics, and LLM-judge selection (Wang et al., 2023; Zheng et al., 2023; Jiang et al., 2023; Hu et al., 2024). A separate blinded audit rates trust on stratified outputs with strategy identity and automatic scores hidden, testing the human interpretation of EATU rather than selector preference.

ERSR improves evidence-aware utility over the development-selected fixed strategy by about +0.052, with the largest gains on label conflicts and unanimous answers with uneven evidence, where

counting agreement discards structure. These results indicate that response-set structure, not agreement alone, is a useful signal for post-generation selection when trusted reference answers and evidence are hidden, while also bounding the claim to a development-only, non-abstaining protocol.

2 Related Work

Trustworthy model and agent evaluation. HELM evaluates language models across scenarios and metrics; AgentBench and AgentBoard cover interactive agents (Liang et al., 2023; Liu et al., 2024; Ma et al., 2024). These benchmarks provide global diagnostics, whereas ERSR uses observable same-item diagnoses for post-generation selection.

Evidence and factuality evaluation. RAG grounds generation in external evidence (Lewis et al., 2020); RAGAS and ARES assess retrieval and faithfulness (Es et al., 2024; Saad-Falcon et al., 2024). FActScore and RAGChecker localize factual and pipeline failures (Min et al., 2023; Ru et al., 2024), while SelfCheckGPT and Chain-of-Verification verify generated claims (Manakul et al., 2023; Dhuliawala et al., 2024). ERSR uses these signals to compare same-item candidates.

Multi-output selection. Self-consistency uses agreement (Wang et al., 2023); LLM-as-judge and LLM-Blender rank or fuse outputs (Zheng et al., 2023; Jiang et al., 2023); RouterBench routes before generation (Hu et al., 2024). ERSR instead diagnoses typed answer-evidence configurations after six responses exist and returns one unchanged response.

3 Method

Post-generation arbitration. For an input x , six strategies (Direct, Chain-of-Thought, BM25-RAG, Dense RAG, Hybrid RAG, Evidence Verifier) each produce a response $c_s = (a_s, e_s, r_s, m_s)$: answer, cited evidence, rationale, and observable metadata. Selector g returns one index and exposes that strategy’s original response without fusion. Current-item trusted reference answers, trusted reference evidence, and utility are unavailable. Unlike pre-generation routing, the decision object is the *relationship among completed responses*.

Typed response-set diagnosis. ERSR canonicalizes surface-form answers into clusters, then applies a fixed diagnosis priority: (1) task-label vs. specific-label conflicts use task-compatible support

comparisons; (2) mutually inconsistent specific labels favor supported, internally consistent candidates; (3) answer aliases and fragmented spans normalize before evidence comparison; (4) unanimous answers skip voting and break ties by evidence completeness and self-grounding. Diagnosis chooses *which comparison to make*; missing signals trigger deterministic fallbacks rather than a learned prediction.

Fitting versus selection. ERSR is deterministic. On 2,100 fitting items we estimate a shrinkage calibration table keyed by dataset, diagnosis, strategy, and answer/evidence/diversity buckets, and select active components plus a self-grounding weight by five-fold within-dataset search. The frozen configuration includes a development-admitted FEVER policy for NEI-versus-specific-label conflicts. It always returns a candidate and never abstains. The claim is therefore *no current-item trusted references at selection time*, not that the offline metric never uses references. No learned ranker is enabled.

Offline evidence-aware utility. After selection we score the chosen response with

$$u_i(s) = A_i(s)F_{E,i}(s)(1 - U_i(s))(1 - K_i(s)), \quad (1)$$

where A is the task-native answer score, F_E is evidence F1, and U and K are unsupported-claim and contradiction risks in $[0, 1]$. The itemwise product is Evidence-Adjusted Trust Utility (EATU), averaged over items; it is not the product of the four factor means. For evaluation, A uses task-native answer scoring, F_E measures evidence overlap, and U and K use reference-based NLI where applicable, with deterministic fallbacks. All four metrics are computed only after selection.

Decision trace. Each trace records clusters, diagnosis, admissible candidates, activated rule, and winner. Ties return an original candidate without rewriting. Diagnosis, calibration, normalized-evidence comparison, and self-grounding are on; learned ranking and consensus fusion are off. ERSR neither chooses which strategies to run nor fuses answers.

4 Experimental Setup

We test three claims under a matched candidate pool: (i) typed arbitration improves joint evidence-aware utility; (ii) gains concentrate where answer

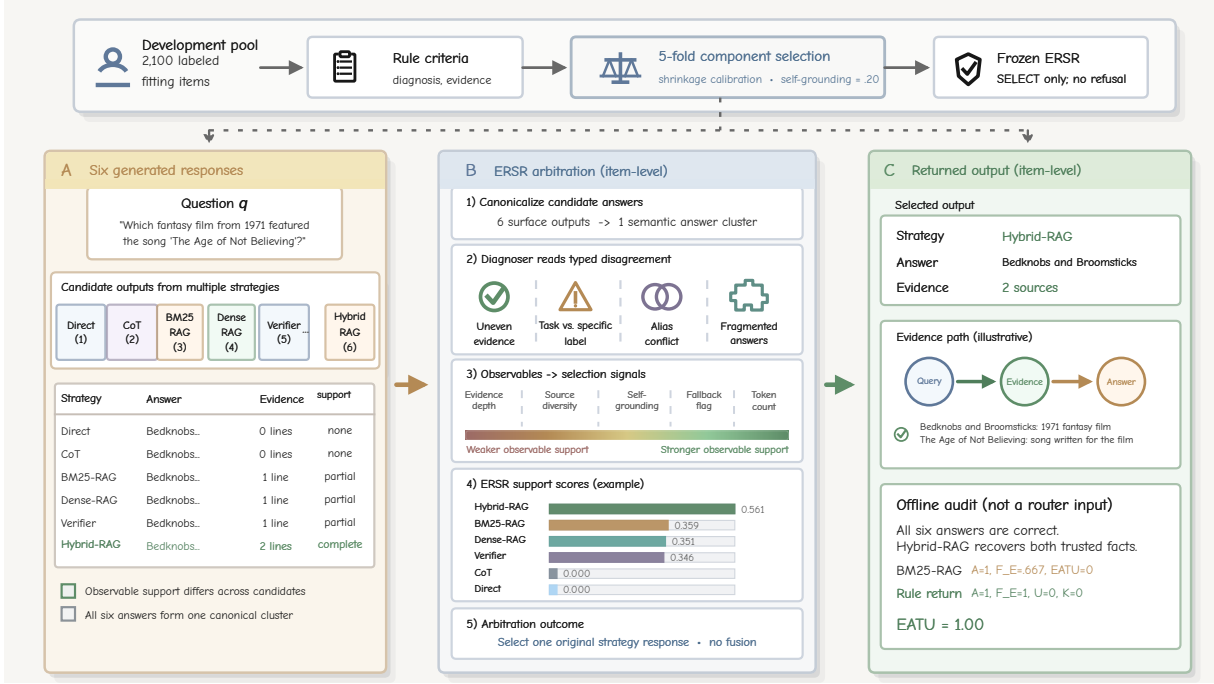


Figure 2: ERSR diagnoses before selecting. All six answers agree but evidence depth differs, so the unanimous-answer rule returns Hybrid-RAG (two sources). EATU is offline and not a router input.

agreement is uninformative; (iii) the effect is stable under pool re-splits, and the utility has a human-trust reading.

Matched pool. We use 3,000 items (1,000 each from FEVER, HotpotQA, and SciFact) with one generated response from each of six strategies (Thorne et al., 2018; Yang et al., 2018; Wadden et al., 2020). Sampling uses seed 42. A deterministic per-dataset lexicographic-prefix split assigns 2,100 items to fitting and 900 to development evaluation (300 per dataset). Fitting chooses calibration, active rule components, and the fixed comparator. Every selector receives the *same* six responses per item, so paired differences come from arbitration rather than a second stochastic generation. Neither split is newly collected or untouched.

Primary comparator only. The sole preselected comparator is the fixed strategy chosen on development fitting data (BM25-RAG). The evaluation-post-hoc best fixed strategy (Evidence Verifier) is reported only as a deliberately optimistic descriptive reference. Selection baselines without current-item trusted references are majority self-consistency, a confidence/evidence heuristic, answer agreement, and an LLM-as-judge that sees the input and shuffled candidate answers, evidence records, and explanations, then returns a strategy identifier. Table 1 reports EATU, oracle regret, Δ

relative to the post-hoc best fixed strategy, and recorded tokens/latency. Confidence intervals use 20,000 paired bootstrap resamples (seed 42); the primary p -value uses 50,000 two-sided paired sign flips with the same seed and a plus-one correction.

What the audit is and is not. The ERSR configuration is frozen for the 900-item evaluation and for alternative-utility checks. Twenty deterministic re-splits repeat fitting and evaluation on the same response corpus; they probe split sensitivity, not generation-time variance or live serving. A separate metric audit samples 120 item-strategy outputs per dataset under coverage/risk/utility strata. Sheets hide strategy identity and automatic scores; annotators rate correctness, support, contradiction, and trust. The audit tests whether EATU correlates with human trust labels, not whether annotators prefer ERSR over another selector in paired comparison.

5 Results

Joint evidence-aware utility improves. Table 1 shows that ERSR reaches 0.248 EATU versus 0.197 for the development-selected fixed strategy (BM25-RAG; train-selected static in the table). The paired gain is +0.0516 (95% bootstrap CI [0.0377, 0.0664]; two-sided sign-flip $p < 0.0001$), with 70 wins, 823 ties, and 7 losses. Equal util-

Selector family/method	Performance			Resource use		Protocol
	EATU	Regret	Δ post-hoc	Selected C/L	End-to-end C/L	
<i>Static single-output selectors</i>						
Post-hoc best fixed	0.208	0.061	+0.000	769/3.27	769/3.27	Single fixed
BM25 retrieval static	0.197	0.072	-0.011	776/4.06	776/4.06	Single fixed
Train-selected static	0.197	0.072	-0.011	776/4.06	776/4.06	Single fixed
TOPSIS static	0.192	0.076	-0.016	747/5.65	747/5.65	Fixed profile
<i>Candidate-output selectors without current-item trusted references</i>						
Majority/self-consistency	0.180	0.088	-0.028	522/3.14	4414/28.07	Six responses
Confidence/evidence heuristic	0.172	0.097	-0.036	637/3.95	4414/28.07	Six responses
Answer agreement	0.157	0.111	-0.051	497/2.89	4414/28.07	Six responses
LLM-as-judge	0.154	0.114	-0.054	757/4.33	4414/35.14	Six + judge
NLI risk only	0.039	0.229	-0.169	447/2.48	4414/28.07	Six responses
<i>Representation-prior selector</i>						
Embedding prototype only	0.111	0.158	-0.097	703/4.51	703/4.51	Representation prior
<i>ERSR arbitration</i>						
ERSR	0.248	0.020	+0.041	747/4.95	4414/28.07	Six responses
<i>Oracle reference</i>						
Oracle EATU selector	0.269	0.000	+0.061	659/4.35	4414/28.07	Post-hoc oracle

Table 1: Frozen 900-item development results. Δ uses the evaluation-post-hoc best fixed strategy; regret uses the per-item oracle. C/L is recorded tokens/latency, not a live-serving benchmark. NLI-only is diagnostic (signal coverage 0.0). Primary text comparisons use development-selected BM25 (same EATU as train-selected static).

ity does not necessarily imply identical candidate choices. Dataset gains are +0.0839 on FEVER, +0.0550 on HotpotQA, and +0.0160 on SciFact; the SciFact sign-flip result is not significant at 0.05 ($p = 0.0588$) and marks a weaker boundary. Post-hoc best fixed reaches 0.208, still below ERSR, and is not a preselected comparator. Relative to voting, agreement, heuristics, and LLM-as-judge, ERSR is best on EATU and lowest on regret among non-oracle selectors. End-to-end C/L matches the six-response pool; selected C/L is the cost of the exposed response only.

Gains sit where agreement discards structure.

Relative to BM25-RAG, the two largest diagnosis cells are NEI-versus-specific-label conflicts (406 items; mean gain +.0512) and unanimous answers (254 items; mean gain +.0608). The first requires comparing task-compatible support when answers disagree; the second, illustrated in Figure 2, requires looking past the answer when voting is uninformative. These cells link the aggregate improvement to diagnosis-first selection. The seven losses comprise two label-conflict items and five HotpotQA alias, fragmentation, or evidence-tie cases, which remain the clearest residual failure modes.

Stability and utility-form sensitivity. Across 20 repeated 70/30 re-splits of the same response corpus, ERSR beats the fitting-selected fixed strategy in 20/20 runs (mean +.0507, range

[+.0387, +.0691]). Versus evaluation-post-hoc best fixed, the mean is +.0454 ([+.0327, +.0531]); FEVER and HotpotQA are positive in 20/20 and SciFact in 19/20. These runs change only the split: they test partition sensitivity, not new-data or generation-time generalization. Holding the selector fixed, five alternative utilities preserve the direction against BM25-RAG: product (Eq. 1) +.0516, equal-additive +.0948, weighted-product +.0613, minimum-gate +.0536, and lexicographic +.0552. The claim is limited to this tested family; the ordering is not an artifact of a single scalarization alone.

A human-trust reading of the utility, not a preference win. On 360 stratified item-strategy outputs with strategy identity and automatic scores removed, two annotators reach 0.828 trust-label agreement and $\kappa = 0.764$. EATU correlates with mean human trust at Spearman $\rho = 0.434$. This supports EATU as an evidence-sensitive trust correlate; it is not a paired preference study and does not show that annotators prefer ERSR-selected outputs over another selector.

6 Conclusion

When multi-strategy agentic QA must expose one original response, answer agreement alone is an incomplete trust signal. Casting the final step as diagnosis-first evidence-risk arbitration, ERSR

improves joint evidence-aware utility over the development-selected fixed strategy and over voting, agreement, and global judging on this development pool, with gains concentrated where label conflicts or unanimous answers hide uneven evidence. Matched pools, diagnosis breakdowns, re-splits, and a blinded trust audit support a response-structure effect within the limits below. Future work may test abstention, generation-time redraws, and matched learned routers under the same no-current-item-trusted-reference constraint.

Limitations

All results use a development-only protocol on one frozen six-response corpus; the claim is therefore arbitration under fixed candidates, not generation-time variance or live serving. ERSR is select-only and never abstains, so unsupported but fluent answers can still be returned. It is a final gate, not a safety layer. The human audit validates metric-trust correlation on stratified outputs, not paired preference among selectors, so the main result does not rely on a human A/B win. We compare interpretable typed rules to fixed, vote, heuristic, and judge selectors, not to a matched-supervision learned router; the coverage-0.0 NLI-only row is diagnostic only and is not a competitive NLI router. FEVER alternative evidence groups remain flattened in the current scorer.

Ethical Considerations

ERSR is a select-only final gate: it can still return fluent but unsupported answers and must not be treated as a safety layer. Deterministic traces may aid auditing, but consequential deployment needs domain-specific validation and a separately implemented abstention or escalation path.

Acknowledgments

This work was supported by the Pioneer Fund of the System of Systems and Artificial Intelligence Key Laboratory of Sichuan Province (Grant No. SSAI202506).

References

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. [Router-Bench: A benchmark for multi-LLM routing system](#). *arXiv preprint arXiv:2403.12031*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. [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)*, pages 14165–14178.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. [AgentBench: Evaluating LLMs as agents](#). In *International Conference on Learning Representations*.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. [AgentBoard: An analytical evaluation board of multi-turn LLM agents](#). In *Advances in Neural Information Processing Systems*, volume 37.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

FactScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. **RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation.** In *Advances in Neural Information Processing Systems*, volume 37.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. **ARES: An automated evaluation framework for retrieval-augmented generation systems.** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. **FEVER: a large-scale dataset for fact extraction and VERification.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. **Fact or fiction: Verifying scientific claims.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. **Self-consistency improves chain of thought reasoning in language models.** In *International Conference on Learning Representations*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-judge with MT-Bench and chatbot arena.** In *Advances in Neural Information Processing Systems*, volume 36.

Appendix A. Case Studies

Figure 3 contrasts a held-out HotpotQA item with a fitting-pool FEVER mechanism illustration. Trusted references and EATU are absent from

candidate-level selection; the FEVER item’s post-hoc utility may have influenced fitting and is not held-out evidence.

Unanimous answer with uneven evidence. For the HotpotQA question “Which fantasy film from 1971 featured the song *The Age of Not Believing?*”, all six strategies answer *Bedknobs and Broomsticks*. Evidence Verifier recovers one relevant line (post-hoc EATU 0), whereas Hybrid-RAG recovers two (evidence F1 and EATU 1.0). ERSR therefore returns the unchanged Hybrid-RAG response. Figure 2 visualizes the same item.

Task label versus specific label. For the fitting-pool FEVER claim “Bethany Hamilton’s autobiography wasn’t adapted into a film”, five strategies predict REFUTES, while Dense-RAG predicts NOT ENOUGH INFO. The reference label is NOT ENOUGH INFO, with no trusted evidence for a specific judgment. ERSR returns Dense-RAG (post-hoc EATU 1.0). This task-native label is not router-level refusal; it illustrates protection against over-commitment.

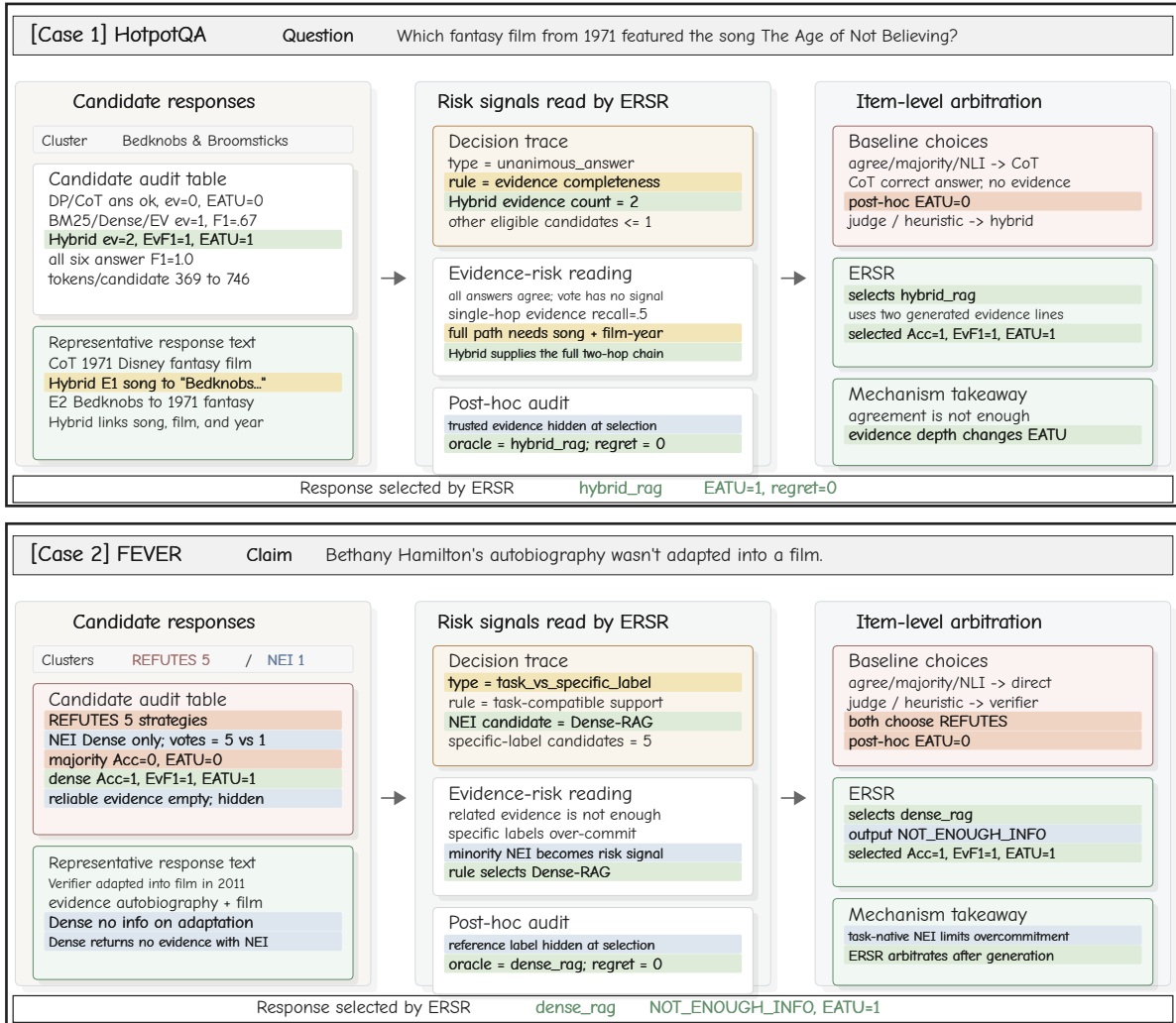


Figure 3: Two ERSR cases. Top: on held-out HotpotQA, all answers agree, but only Hybrid-RAG supplies the complete two-hop evidence chain. Bottom: in a fitting-pool FEVER illustration, five REFUTES outputs conflict with Dense-RAG’s task-native NOT ENOUGH INFO. Scores and oracle information are post-hoc; the lower case is neither held-out evidence nor router-level refusal.