

Provenance-Grounded Gating and Adaptive Recovery in Synthetic Post-Training Data Curation

Soham Bhattacharjee¹ Karun Sharma^{1*} Vinay Kumar Sankarapu² Pratinav Seth¹

¹Lexsi Labs, Mumbai, India ²Lexsi Labs, London, United Kingdom

Correspondence: pratinav.seth@lexsi.ai

Abstract

Synthetic post-training pipelines commonly filter generated samples with reward models or holistic LLM judges, often relying on two assumptions: that a sample’s faithfulness can be judged without the exact source that produced it, and that rejected samples can be discarded rather than systematically recovered. We present a controlled study of both assumptions across gate configurations, recovery strategies, and generator scales, using adversarially injected corpora to provide ground-truth failure labels. We find that grounding faithfulness judgments in the exact source evidence improves over post-hoc retrieval for stronger judges and substantially outperforms reward-only filtering; hallucination and reward gates reject largely distinct sample populations supporting the need for both; and adaptive recovery, which diagnoses failures before repairing them, achieves higher yield and recovery rate than naive regeneration. Downstream fine-tuning quality is driven primarily by generator scale, with filtering and recovery providing their clearest gains when smaller generators are used. We release our evaluation corpus and experimental configurations to support reproducible pipeline design.

1 Introduction

Synthetic post-training pipelines filter generated samples with LLM-as-judge quality gates before fine-tuning. These pipelines often rely on two assumptions that are rarely examined together: that a sample’s faithfulness can be judged without the exact source evidence that induced it, and that samples rejected by a filter can simply be discarded.

Judges typically follow preference rubrics covering helpfulness, instruction following, and truthfulness (Cui et al., 2024), but these rubrics do not verify whether a response is supported by the source passage that induced it. Source-grounded

frameworks such as G-Eval (Liu et al., 2023) and CheckEval (Lee et al., 2025) derive criteria from a source chunk, yet major synthetic data curation libraries, including distilabel (Álvaro Bartolomé Del Canto et al., 2024) and AgentInstruct (Mittra et al., 2024), do not preserve an explicit provenance record linking each generated sample back to its source. Without this record, faithfulness gating relies on post-hoc retrieved evidence, which only approximates the original evidence relation.

A second gap concerns what happens after rejection. Existing pipelines often discard failures and spend generation budget on fresh samples. Rejection-sampling methods such as RAFT (Dong et al., 2023) and ReST^{EM} (Singh et al., 2024) show that regenerating under a quality filter can improve yield, but they operate under a single global acceptance criterion. In source-grounded curation, however, failures can arise from different causes: unsupported claims, parametric drift, weak instruction following, or low answer quality. This raises the question of whether diagnosing the failure mode before repair can recover more usable data than blind regeneration.

We study these questions in a controlled ablation over gate configuration, recovery strategy, and generator scale, using CuratorKIT (Bhattacharjee et al., 2026) as the implementation framework. We find that exact source provenance improves faithfulness gating for stronger judges and substantially outperforms reward-only filtering; hallucination and reward gates reject largely distinct sample populations, supporting the use of both; and adaptive diagnose-and-repair recovers more rejected samples than naive regeneration. Downstream fine-tuning quality is driven primarily by generator scale, with filtering and recovery providing their clearest gains when smaller generators are used.

Contributions. (1) A provenance-grounded gating study showing that exact source evidence im-

*Work done while at Lexsi Labs.

proves faithfulness detection for stronger judges, while reward-only filtering and post-hoc retrieval degrade gate quality in complementary ways. (2) A controlled recovery ablation showing that adaptive diagnose-and-repair outperforms naive regeneration in yield and recovery rate. (3) A held-out evaluation corpus for source-grounded QA, released to support reproducible downstream evaluation of synthetic data curation pipelines.

Data and code availability. We release the held-out source-grounded QA test set along with the curated training set and the code/configuration used for the EMNLP experiments at: <https://huggingface.co/datasets/Lexsi/provenance-grounded-synthetic-qa> and <https://github.com/Lexsi-Labs/CuratorKIT>.

2 Methodology

Provenance-preserving generation. Each sample is generated from a source chunk c drawn from the corpus, and a provenance record linking the sample to c is attached at generation time. This record is append-only and persists through all downstream pipeline stages, making the exact generation evidence available to the hallucination gate without retrieval.

Gates. The *HallucinationGate* performs structured claim verification against the preserved source chunk ($\tau_{\text{hall}} = 0.8$), directly testing whether the generated response is entailed by the evidence that induced it. The *RewardGate* scores instruction-output quality following a preference rubric without source access ($\tau_{\text{reward}} = 0.7$). A sample must clear both gates to be accepted.

Adaptive recovery (primary). On hallucination failure, a *DiagnosticProbe* diagnoses the failure mode and applies a targeted configuration patch, such as lowering generation temperature or expanding context, before regenerating the sample. Probe-recovered samples re-enter the reward gate. On reward failure, a *RewardRefiner* rewrites the output for quality without altering the underlying claims. Both recovery steps operate against the same preserved provenance record.

Naive retry (ablation). On any gate failure, the generator resamples from the original prompt without diagnosis. Samples must clear both gates within five retries.

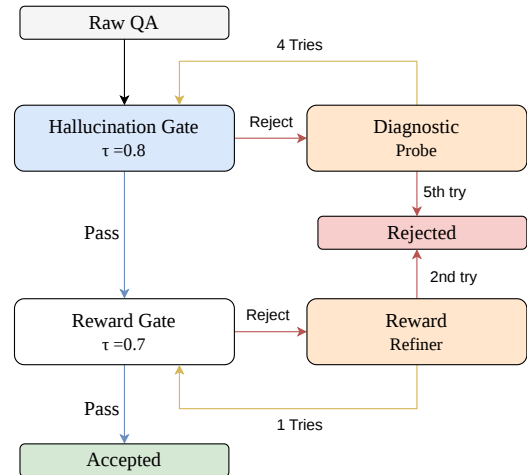


Figure 1: Adaptive recovery pipeline. Hall-rejected samples enter the *DiagnosticProbe* for up to 4 targeted repair attempts (blue path); reward-rejected samples enter the *RewardRefiner* for one rewrite (orange path). Samples clearing both gates are accepted.

3 Experimental Setup

Source corpus. We use 1,500 chunks, 500 each from CUAD (Hendrycks et al., 2021) contracts, PubMedQA (Jin et al., 2019) abstracts, and English Wikipedia 2023¹. Each chunk seeds three QA pairs under each training variant. After malformed outputs are removed, this yields roughly 8,000–8,500 candidates per generator.

Adversarial injections. Four failure types are injected into approximately 20% of samples: *contradicts_source*, *parametric_drift*, *domain_mismatch*, and *instruction_quality*, providing ground-truth labels for gate recall value of rejecting these samples and recovering them.

Generators and judges. Generators: Qwen3-1.7B, 4B, 8B (Yang et al., 2025). Three judge sizes are used across experiments: Qwen3-14B, Qwen3.6-27B-FP8, and Qwen3.6-35B-A3B. Provenance gating experiments are evaluated across all three judges; adaptive recovery experiments use the 14B and 35B judges; the recovery ablation uses a single judge. All models run in non-thinking mode via vLLM.

Benchmarks and metrics. Provenance gating quality is evaluated on FaithDial (Dziri et al., 2022) ($n=3,507$) against ground-truth faithfulness labels. Downstream evaluation uses a held-out

¹<https://huggingface.co/datasets/wikimedia/wikipedia>

Table 1: Gate method comparison on FaithDial ($n=3,507$). Exact-provenance gating achieves the highest F_1 at 35B; reward-only scoring catastrophically over-rejects faithful samples at both judge sizes. Full P/R and 27B results in Appendix B.1.

Gate configuration	14B F_1	35B F_1
Exact provenance (ours)	0.494	0.614
Retrieved provenance	0.514	0.590
Oracle holistic	0.489	0.599
Reward only	0.291	0.184

test set of 1,400 instructions, generated independently of all training corpora and fixed before any pipeline experiments began. Metrics: ROUGE-L, BERTScore F_1 , and Faithfulness (BERTScore F_1 between model output and source chunk).

Fine-tuning. Qwen3-4B base, LoRA ($r=16$, $\alpha=32$) via Unsloth (Daniel Han and team, 2023; Hu et al., 2022), early stopping (patience = 2, up to 4 epochs). Full hyperparameters in Table 17 (Appendix E).

4 Results and Analysis

We evaluate the pipeline along three axes: whether exact-provenance gating detects faithfulness failures missed by reward-only filtering and whether the two gates are complementary, whether adaptive diagnosis-and-repair outperforms naive regeneration as a recovery strategy, and whether curation choices affect downstream fine-tuning quality independently of generator scale (Section 4.3).

4.1 Gate Characterization

We evaluate four gate configurations on FaithDial ($n=3,507$); results are in Table 1. Exact-provenance gating achieves the highest F_1 at stronger judge sizes, reaching 0.614 at 35B, while retrieved provenance leads marginally at 14B. Reward-only scoring is unusable as a faithfulness filter: only 14–45% of genuinely faithful samples pass depending on judge strength, making it a source of systematic over-rejection rather than quality control.

Figure 2(a) breaks down rejection sets when both gates run on the unfiltered generated corpus. Jaccard overlap sits in $[0.23, 0.32]$ across all generators and both judges (Appendix B.3), confirming the two gates target structurally different failure modes. The split is not random: *contradicts_source* and *domain_mismatch* are caught reliably (82-90% recall for the 4B and 8B generators), while *instruction_quality* resists both gates (24-42% recall; Ta-

Table 2: Gate rejection-set overlap on the generated corpus (35B judge). Hall-only and reward-only segments are both substantial; Jaccard $\in [0.28, 0.32]$ confirms the two gates target structurally different failure modes. 14B numbers in Appendix B.3.

Gen.	Jaccard	Rejected (% of corpus)		
		Hall	Reward	Both
1.7B	0.317	7.3	32.3	18.4
4B	0.324	6.5	38.6	21.6
8B	0.284	8.5	36.6	17.8

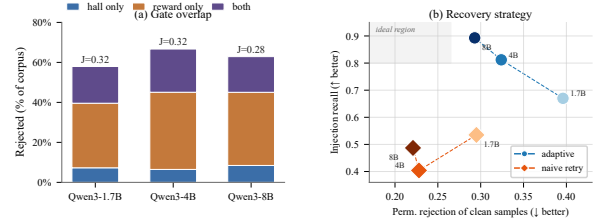


Figure 2: (a) Rejection-set overlap per generator as % of corpus (35B judge; Appendix B.3 for 14B numbers). Both hall-only and reward-only segments are substantial; Jaccard $\in [0.23, 0.32]$, confirming the two gates target structurally different failure modes. (b) Recall-rejection tradeoff (shaded = better; ideal: lower-right). Adaptive improves with generator size (light to dark blue); by 8B (29.3% reject, 89.3% recall) it achieves lower rejection *and* higher recall than naive retry on 1.7B (29.5%, 53.5%; diamond), fully dominating on both axes.

ble 13, Appendix B.4). Neither gate catches what the other misses; both are necessary.

To test whether the gating pattern extends beyond FaithDial or a single judge formulation, we run two robustness checks. First, HalluDial (Luo et al., 2024) provides a higher-hallucination stress test ($n = 3,507$, 73% hallucinated), where exact provenance again substantially outperforms reward-only scoring under both cross-family judges (Table 3). Second, we compare G-Eval (Liu et al., 2023) and CheckEval (Lee et al., 2025) under exact and retrieved provenance. CheckEval with exact provenance gives the strongest 14B result, outperforming retrieved provenance by +0.055 F1 while also reducing inference time by 197s (Table 4). Full precision/recall and larger-judge results for the judge-agnostic provenance experiment are in Appendix B.2.

4.2 Recovery: Structured Repair Outperforms Naive Retry

Structured adaptive repair outperforms naive retry on total yield, pass rate, and injection recall across all three generator sizes. The full stage-by-stage breakdown including the +42% yield gain over hard filtering is in Appendix B.5.

Table 3: HalluDial gate validation ($n=3,507$, 73% hallucinated), a FaithDial-derived stress test under a higher hallucination base rate. Full P/R and the oracle holistic/retrieved-provenance rows are in Appendix B.6.

Gate configuration	Mistral-24B F1	Gemma-27B F1
Exact provenance	0.690	0.681
Reward only	0.325	0.368

Table 4: Judge-agnostic provenance transfer on FaithDial ($n=3,507$, threshold=0.5). CheckEval with exact provenance achieves the highest F1 at both judge sizes; G-Eval achieves high recall but low precision regardless of provenance type. Bold = best F1 per judge. 27B results in Appendix B.2.

Judge	Condition	F_1	P	R	Time
14B	geval, retrieved	0.443	0.298	0.865	140.8
14B	geval, exact	0.449	0.291	0.981	109.4
14B	checkeval, retrieved	0.513	0.372	0.831	359.7
14B	checkeval, exact	0.568	0.424	0.857	162.3

Adaptive accepts more samples and recovers a larger share of total rejects at every generator size, with the recovery rate advantage growing from +7.6 points at 1.7B to +38.1 points at 8B. The injection recall gap is the starkest signal: 89.3% vs 48.7% at 8B, nearly double, with the widest within-type gap on *instruction_quality* (74.4% vs 51.3%) and *parametric_drift* (71.7% vs 58.1%) at 1.7B (Table 13, Appendix B.4). Naive retry permanently discards fewer samples from the non-injected pool, which we term the natural rejection rate. We note that non-injected samples are not guaranteed to be high quality; they simply did not receive a controlled adversarial perturbation, and natural gate failures among them may reflect genuine faithfulness or quality deficiencies rather than over-rejection. Adaptive’s higher natural rejection rate therefore does not straightforwardly imply worse precision (see Section D). Figure 2(b) shows the recall-rejection tradeoff: adaptive improves on both axes as generator size increases, and by 8B fully dominates naive retry on both dimensions simultaneously.

4.3 Downstream: Generator Size Dominates

Generator scale is the dominant signal: ROUGE-L rises from 0.523 (1.7B, unfiltered) to 0.576 (8B, both-filtered), a cross-generator gain of +6.6% that dwarfs any within-generator spread. The practical contribution of filtration shrinks with generator capacity: the gain from unfiltered to best-filtered is +0.016 ROUGE-L for 1.7B but only +0.003 for 4B and 8B, a fivefold reduction. For the two larger gen-

Table 5: Adaptive recovery vs. naive retry (14B judge). Adaptive leads on pass rate, recovery rate, and injection recall at every generator size, with the injection recall gap widening substantially with scale. Naive retry permanently discards fewer clean samples (lower uninjected rejection), but adaptive recovers enough of its initial rejects to end with higher total yield despite this disadvantage.

Metric	Adaptive			Naive retry		
	1.7B	4B	8B	1.7B	4B	8B
Pass rate (%)	71.6	85.8	88.5	68.4	73.6	72.2
Recovery rate (%)	32.5	66.3	72.7	24.9	36.4	34.6
Inj. recall (%)	67.0	81.2	89.3	53.5	40.4	48.7
Uninj. reject (%)	39.6	32.4	29.3	29.5	22.8	22.1

Table 6: Downstream fine-tuning (Qwen3-4B base, LoRA, $n=1,400$). Three representative conditions are shown; full results across all five conditions in Table 15 (Appendix B.5). Generator scale drives ROUGE-L (+6.6%, 1.7B→8B); within-generator spread across all conditions is at most 0.016 (1.7B), 0.007 (4B), 0.018 (8B). Bold = best per generator per metric.

Gen.	Condition	R-L	BS-F1	Faith.
1.7B	unfiltered	0.5227	0.9059	0.8537
1.7B	both	0.5388	0.9112	0.8606
1.7B	sequential	0.5382	0.9113	0.8680
4B	unfiltered	0.5577	0.9185	0.8542
4B	both	0.5606	0.9197	0.8570
4B	sequential	0.5604	0.9223	0.8563
8B	unfiltered	0.5734	0.9187	0.8527
8B	both	0.5764	0.9191	0.8540
8B	sequential	0.5585	0.9189	0.8537

erators, all conditions including unfiltered cluster within 0.007 of each other, indicating that stronger generators produce data of sufficient baseline quality that curation adds marginal downstream value.

Sequential recovery provides its clearest benefit at 1.7B, where it achieves the global best faithfulness score (0.868, +0.007 over both-filtered for the same generator). This advantage does not transfer to larger generators: for 4B and 8B, sequential recovery is comparable to hard filtering on faithfulness and slightly lower on ROUGE-L at 8B, suggesting that recovered samples from stronger generators are structurally different from their native output. Faithfulness decreases marginally with generator scale (−0.007 absolute, 1.7B→8B under both-filtered), a consistent but small trade-off against the lexical quality gains. Full results across all five conditions are in Table 15 (Appendix B.5).

4.4 Cross-Family Generalization

To test whether the findings are Qwen-specific, we repeat the three core analyses under two judge/-generator pairs with no Qwen3 lineage: **Pair A**

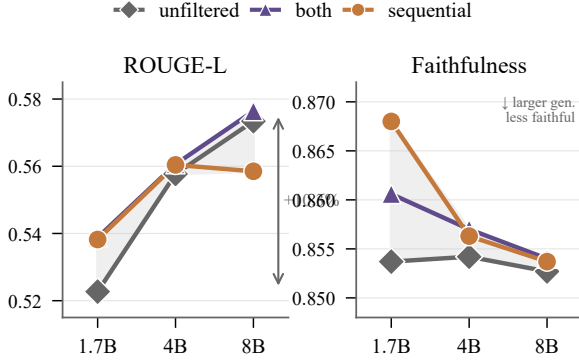


Figure 3: Downstream metrics by generator and condition. Each line shows one filtration condition across generator sizes; the shaded band spans the within-generator spread. The steep cross-generator slope (+6.6% ROUGE-L, 1.7B→8B) dwarfs the within-generator spread, confirming generator scale dominates downstream quality.

Gate configuration	Mistral-24B F1	Gemma-27B F1
Exact provenance (ours)	0.477	0.486
Reward only	0.281	0.281

Table 7: Cross-family gate validation on FaithDial ($n=3,507$). Exact provenance retains the same margin over reward-only scoring seen under Qwen3 judges (Table 1). Full P/R and the oracle holistic/retrieved-provenance rows are in Appendix B.6.

uses a Mistral-Small-3.1-24B-Instruct judge with a Gemma-3-4B-IT (Team et al., 2025) recovery generator, and **Pair B** uses a Gemma-3-27B-IT judge with a Ministral-8B-Instruct-2410 recovery generator. Exact-provenance gating retains a clear margin over reward-only scoring on FaithDial (0.477 vs. 0.281; 0.486 vs. 0.281; Table 7), adaptive recovery improves yield by +46.1% and +28.1% at the matched generator scales (Table 8), and downstream metrics remain close to matched-scale Qwen3 baselines (ROUGE-L within 0.01–0.02, BERTScore F1 within 0.005, Faithfulness within 0.002; Table 9). Full four-gate cross-family results are in Appendix B.6.

5 Conclusion

We presented a controlled study of source-grounded synthetic data curation across gate configurations, recovery strategies, and generator scales. Exact-provenance gating improves faithfulness detection for stronger judges and substantially outperforms reward-only filtering; hallucination and reward gates reject largely distinct failure populations; and adaptive diagnose-and-repair improves yield and recovery rate over naive regeneration. Downstream quality is driven primarily by generator scale, with curation providing its clearest benefit

Metric	Pair A	Pair B
Adaptive pass rate	86.23%	97.24%
Adaptive recovery rate	71.3%	89.7%
Yield gain vs. both-filtered	+46.1%	+28.1%

Table 8: Cross-family adaptive recovery. Pair A matched to Qwen3-4B, Pair B to Qwen3-8B. Pair B’s smaller yield gain reflects a ceiling effect (75.9% hard-filter pass rate before recovery).

Pair	Condition	R-L	BS-F1	Faith.
A	both-filtered	0.5694	0.9195	0.8560
A	sequential	0.5615	0.9158	0.8547
B	both-filtered	0.5722	0.9196	0.8553
B	sequential	0.5738	0.9178	0.8561

Table 9: Cross-family downstream fine-tuning ($n=1,400$ test set), matching the format of Table 6. Matched-scale Qwen3 reference: Qwen3-4B both-filtered = 0.5606/0.9197/0.8570; Qwen3-8B both-filtered = 0.5764/0.9191/0.8540.

when smaller generators are required. Overall, the results support a practical recipe for synthetic QA curation: preserve source provenance, apply both faithfulness and reward gates, and recover rejected samples through targeted repair rather than blind retry.

Limitations

Most experiments use Qwen3 generators, judges, and a Qwen3-4B fine-tuning base; although Section 4.4 adds cross-family judge/generator validation, broader replication across additional model families is still needed. Provenance gating is primarily evaluated on FaithDial, with HalluDial used as a higher-hallucination stress test. The downstream benefit of filtration conditions diminishes with generator scale: within-generator condition spreads (0.002–0.010 ROUGE-L) are plausibly within noise at $n=1,400$, suggesting that structured repair is most valuable when baseline generator quality is low. Downstream fine-tuning uses a single base model (Qwen3-4B with LoRA). The three-domain corpus (legal, biomedical, Wikipedia) may not represent specialized or low-resource domains. The DiagnosticProbe is inherently sequential, as each sample requires up to six chained LLM calls whose outputs determine the next step, and although the pipeline overlaps independent samples via async with bounded concurrency, per-sample probe latency remains a bottleneck.

Ethics Statement

This work uses LLM-as-judge systems as quality filters. Such judges inherit biases from their training data; the quality of any curated corpus depends on judge quality and should be validated against ground-truth labels on domain-representative samples before deployment. No new pretraining is performed; all inference uses existing publicly available checkpoints. Data curation experiments ran on an NVIDIA A100 (80 GB) and fine-tuning on an RTX Pro 6000 (96 GB); total compute was approximately 26 GPU-hours across both stages. All source corpora are derived from publicly licensed datasets (CUAD: Apache 2.0, PubMedQA: MIT, Wikipedia: CC BY-SA, FaithDial: MIT) and contain no personally identifiable information.

AI Disclosure

AI writing and coding tools were used to assist with drafting portions of this paper and experiment implementation. All scientific claims, experimental designs, results, and analyses were drafted, reviewed and verified by the authors prior to submission.

References

- Soham Bhattacharjee, Karun Sharma, Vinay Kumar Sankarapu, and Pratinav Seth. 2026. [Curatorkit : Data curation and synthetic data generation for llm post-training](#). *Preprint*, arXiv:2606.21631.
- Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. [LM vs LM: Detecting factual errors via cross examination](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore. Association for Computational Linguistics.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. [RAFT: Reward ranked finetuning for generative foundation model alignment](#). *Transactions on Machine Learning Research*.
- Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Omar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. [FaithDial: A faithful benchmark for information-seeking dialogue](#). *Transactions of the Association for Computational Linguistics*, 10:1473–1490.
- Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. [Cuad: An expert-annotated nlp dataset for legal contract review](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. [TRUE: Re-evaluating factual consistency evaluation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. [CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.

- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. [Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation](#). *Preprint*, arXiv:2406.07070.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Cotas, Yadong Lu, Wei ge Chen, Olga Vrousos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. 2024. [Agentinstruct: Toward generative teaching with agentic flows](#). *Preprint*, arXiv:2407.03502.
- Guilherme Penedo, Hynek Kydlíček, Alessandro Capelli, Mario Sasko, and Thomas Wolf. 2024. [Data-trove: large scale data processing](#).
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. [Increasing faithfulness in knowledge-grounded dialogue with controllable features](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.
- Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, and 21 others. 2024. [Beyond human data: Scaling self-training for problem-solving with language models](#). *Transactions on Machine Learning Research*. Expert Certification.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- NVIDIA The NeMo Data Designer Team. 2025. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. <https://github.com/NVIDIA-NeMo/DataDesigner>. GitHub Repository.
- 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 *The Eleventh International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- 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 *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Álvaro Bartolomé Del Canto, Gabriel Martín Blázquez, Agustín Piqueres Lajarín, and Daniel Vila Suero. 2024. Distilabel: An ai feedback (aif) framework for building datasets with and for llms. <https://github.com/argilla-io/distilabel>.

A Related Work

LLM-as-judge. [Zheng et al. \(2023\)](#) establish LLM-as-judge as a workable proxy for human preference at scale. [Liu et al. \(2023\)](#) propose G-Eval, a holistic chain-of-thought scoring approach. Structured claim verification ([Cohen et al., 2023](#)) breaks a response into atomic claims and verifies each against source evidence; we use this formulation for our hallucination gate.

Faithfulness in generated data. Fluency and faithfulness are weakly correlated, motivating dedicated checks ([Maynez et al., 2020](#)). NLI-based post-hoc filters have been proposed ([Rashkin et al., 2021](#); [Honovich et al., 2022](#)), and FaithDial ([Dziri et al., 2022](#)) provides ground-truth faithfulness labels used to validate gate calibration.

Self-consistency and retry. [Wang et al. \(2023\)](#) show that sampling multiple completions and aggregating improves chain-of-thought reasoning. Our naive retry baseline is the minimal variant: regenerate and re-gate, no aggregation.

AI feedback and self-improvement. Lee et al. (2024) show that AI feedback can substitute for human preference labels at scale. Yuan et al. (2024) extend this to self-rewarding loops where the model scores its own outputs; our reward gate plays an analogous role as an automated quality signal applied at corpus construction time.

Synthetic data frameworks. distilabel (Álvaro Bartolomé Del Canto et al., 2024), DataTrove (Penedo et al., 2024), NeMo Data Designer (The NeMo Data Designer Team, 2025), and AgentInstruct (Mitra et al., 2024) compose generation, scoring, and filtering; rejection is typically terminal. Our study is the first to compare structured repair and naive retry recovery strategies on top of such pipelines in a controlled ablation.

B Extended Experimental Results

B.1 Gate Method Comparison: Full Results

Table 10 reports precision, recall, and F_1 for all four gate configurations across all three judge sizes on FaithDial ($n=3,507$). The main paper reports only F_1 for the 14B and 35B judges; full metrics and the 27B judge are included here.

Exact-provenance gating improves consistently with judge strength, reaching $F_1=0.614$ at 35B. Retrieved provenance leads marginally at 14B ($F_1=0.514$ vs 0.494) but falls behind at 27B and 35B, confirming that the benefit of exact provenance grows with judge capacity. Oracle holistic scoring achieves the highest recall at 14B (0.903) but the lowest precision, making it unsuitable as a high-precision curation gate despite its strong recall. Reward-only scoring degrades with judge strength: F_1 falls from 0.291 at 14B to 0.184 at 35B as stronger judges apply a stricter reward rubric that over-rejects faithful samples even more aggressively.

B.2 Judge-Agnostic Provenance Transfer

Table 11 reports F_1 , precision, recall, and inference time for G-Eval and CheckEval scoring under exact and retrieved provenance across all three judge sizes on FaithDial ($n=3,507$). This experiment tests whether the faithfulness detection signal from exact-provenance gating transfers across judge families, or whether it is an artifact of a specific model.

CheckEval with exact provenance is the best condition at every judge size, outperforming retrieved provenance by $+0.021$ - $+0.055$ F_1 . G-Eval achieves high recall (0.865-0.981) but poor precision,

making it unsuitable as a curation gate. Beyond detection quality, exact provenance eliminates retrieval overhead entirely: under the 14B judge, switching from retrieved to exact reduces inference time by 197s while improving recall by $+0.026$. The consistent F_1 advantage of exact over retrieved provenance across all three judges confirms the finding is not judge-specific.

B.3 Gate Overlap: Full Results

Table 12 reports gate rejection-set overlap for both judges across all three generators. The main paper reports 35B judge numbers only; the 14B judge is included here to confirm that gate complementarity holds across judge strengths.

Gate complementarity is consistent across both judges: Jaccard remains below 0.33 in all six conditions. The 35B judge applies a substantially stricter reward gate, with reward-only rejection rising from 17-22% under 14B to 33-39% under 35B, while the hallucination gate becomes only marginally stricter (-2 -7 percentage points). This asymmetry confirms that judge strength primarily affects reward sensitivity rather than hallucination detection, and that the two gates remain structurally complementary regardless of the operating judge.

B.4 Injection Recall by Failure Type

instruction_quality is the hardest failure type at every judge size and generator (23-42% recall), as it requires reasoning about instruction intent rather than direct source grounding. *contradicts_source* and *domain_mismatch* are caught most reliably for 4B and 8B (82-90%). *parametric_drift* recall improves markedly with generator size, suggesting larger generators produce more detectable parametric confabulations.

B.5 Adaptive Recovery: Stage Breakdown

Table 14 traces the adaptive pipeline stage by stage for all three generators under the 14B judge.

Approximately 45-51% of probe-recovered samples subsequently fail the reward gate, explaining the gap between probe recovery rate and final accepted count. Under the 35B judge (1.7B generator only), the pipeline accepts 4,635 samples (59.3% pass rate) with a 23.6% recovery rate and 69.0% injection recall, confirming the adaptive advantage holds under a stronger judge at lower overall yield due to the stricter reward gate.

Table 15 reports downstream fine-tuning results across all five curation conditions for completeness.

Table 10: Full gate method comparison on FaithDial ($n=3,507$). Exact-provenance gating achieves the highest F_1 at 27B and 35B; retrieved provenance leads marginally at 14B. Reward-only scoring degrades with judge strength, confirming it cannot serve as a faithfulness filter at any judge size.

Gate configuration	14B judge			27B judge			35B judge		
	F_1	P	R	F_1	P	R	F_1	P	R
Exact provenance	0.494	0.357	0.803	0.594	0.552	0.644	0.614	0.588	0.642
Retrieved provenance	0.514	0.388	0.759	0.573	0.533	0.618	0.590	0.566	0.615
Oracle holistic	0.489	0.335	0.903	0.569	0.436	0.819	0.599	0.479	0.799
Reward only	0.291	0.214	0.454	0.185	0.230	0.154	0.184	0.272	0.140

Table 11: Judge-agnostic provenance transfer on FaithDial ($n=3,507$, threshold = 0.5). CheckEval with exact provenance achieves the highest F_1 at 27B and 35B; G-Eval achieves high recall but low precision across all judges. Bold = best F_1 per judge.

Judge	Condition	F_1	P	R	Time
14B	geval, retrieved	0.443	0.298	0.865	140.8
14B	geval, exact	0.449	0.291	0.981	109.4
14B	checkeval, retrieved	0.513	0.372	0.831	359.7
14B	checkeval, exact	0.568	0.424	0.857	162.3
27B	geval, retrieved	0.520	0.362	0.920	161.5
27B	geval, exact	0.535	0.371	0.960	154.8
27B	checkeval, retrieved	0.662	0.690	0.635	256.4
27B	checkeval, exact	0.686	0.689	0.683	254.8
35B	geval, retrieved	0.485	0.331	0.906	153.8
35B	geval, exact	0.489	0.331	0.937	139.3
35B	checkeval, retrieved	0.651	0.699	0.610	229.3
35B	checkeval, exact	0.672	0.695	0.651	214.1

Table 12: Gate rejection-set overlap, 14B and 35B judges. Jaccard remains below 0.33 across all conditions; 35B applies a substantially stricter reward gate.

Judge	Gen.	Rejected (% of corpus)				Accepted Yield (%)
		Jacc.	Hall	Rew.	Both	
14B	1.7B	0.23	20.5	22.2	9.6	57.9
14B	4B	0.31	11.9	17.2	13.4	57.5
14B	8B	0.28	13.4	16.9	11.8	57.9
35B	1.7B	0.317	7.3	32.3	18.4	42.1
35B	4B	0.324	6.5	38.6	21.6	33.3
35B	8B	0.284	8.5	36.6	17.8	37.1

B.6 Cross-Family: Full Results

Table 16 reports the complete four-gate-configuration comparison (exact provenance, oracle holistic, retrieved provenance, reward only) with full precision/recall under both cross-family judges, on FaithDial and HalluDial, extending Tables 7 and 3 in the main text.

C DiagnosticProbe: Failure-Mode Routing

When a sample fails the HallucinationGate, the DiagnosticProbe attempts recovery through two

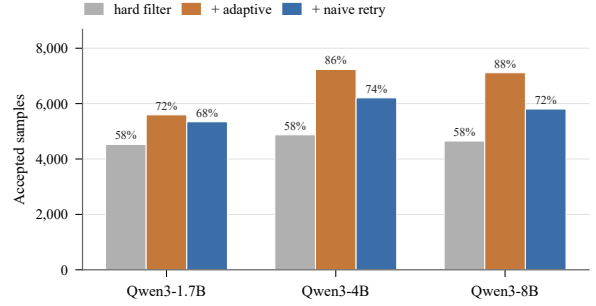


Figure 4: Accepted samples per generator under hard filtering, adaptive recovery, and naive retry.

probe types, routed by the grounding score returned by the gate.

Routing. If the grounding score is below 0.5, the failure is likely a prompt-level grounding issue; Probe 2a runs first, followed by Probe 1, then Probe 2b. If the grounding score is at or above 0.5, the failure is likely stochastic; Probe 1 runs first, followed by Probe 2. The first attempt that clears the HallucinationGate is accepted and forwarded to the RewardGate. If all attempts fail, the sample is permanently rejected.

Probe 1: Temperature sweep. The sample is regenerated at three temperatures in sequence: 0.3, 0.7, and 1.1. This targets stochastic failures where the original generation drifted at high temperature; lower temperatures encourage the model to stay closer to the source chunk. The first passing attempt wins.

Probe 2: Prompt variants. Two sub-variants are applied in sequence. Probe 2a uses a strict source-anchored prompt that explicitly constrains the model to cite only from the provided source chunk, targeting parametric leakage. Probe 2b regenerates the instruction from scratch using the default prompt, targeting cases where the original instruction formulation was the source of the failure rather than the answer generation.

Table 13: Injection recall by failure type, 14B and 35B judges, gate-only (no recovery). *instruction_quality* is the hardest type across all conditions; *contradicts_source* and *domain_mismatch* are caught most reliably for 4B and 8B generators.

Judge	Gen.	contr. src	param. drift	dom. mismatch	instruct. qual.
14B	1.7B	20.3	17.3	26.7	23.8
14B	4B	82.3	47.6	85.8	34.2
14B	8B	89.0	77.1	88.0	42.4
35B	1.7B	29.5	34.0	37.3	26.3
35B	4B	83.1	49.5	82.1	31.3
35B	8B	89.9	78.4	85.4	28.1

Table 14: Adaptive pipeline stage breakdown (14B judge). Total accepted: 19,940 vs. 14,047 for hard filtering (+42%).

Stage	1.7B	4B	8B
Raw input	7,812	8,433	8,039
HallGate pass	6,266	6,348	6,046
Probe recovers	1,171	1,906	1,839
Probe→Reward pass	568	1,044	910
Reward main reject	1,741	1,475	1,397
Refiner recovers	501	1,315	1,555
Total accepted	5,594	7,232	7,114
Pass rate	71.6%	85.8%	88.5%
Hard-filter baseline	4,525	4,873	4,649
Yield gain	+23.6%	+48.4%	+53.1%

RewardRefiner. When a sample fails the RewardGate, the RewardRefiner issues a single targeted rewrite call. The judge returns the worst-scoring evaluation axis alongside a structured critique. The refiner prompt is:

*The following answer was scored low on {axis}.
Judge feedback: "{weakness}". Rewrite the answer to specifically improve {axis}.*

The rewritten sample re-enters the RewardGate once. If it fails again it is permanently rejected. The refiner does not alter the underlying factual claims, only the quality dimensions flagged by the judge.

D Discussion

Downstream analysis. Within any generator block, ROUGE-L spread is at most 0.002/0.007/0.010 (1.7B/4B/8B) and faithfulness spread ≤ 0.003 , small relative to the cross-generator effect. *Both*-filtered leads or ties on ROUGE-L and faithfulness in every slot, though margins are never practically meaningful. There is a genuine faithfulness-ROUGE trade-off with scale: moving from 1.7B to 8B gains +0.036 ROUGE-L while losing -0.007 faithfulness, consistently across all conditions. The 8:1 ratio strongly favours the quality gain in most settings; practitioners for whom source fidelity is a

Table 15: Full downstream fine-tuning results (Qwen3-4B base, LoRA, $n=1,400$ test set). Bold = best per generator per metric.

Gen.	Condition	R-L	BS-F1	Faith.
1.7B	unfiltered	0.5227	0.9059	0.8537
1.7B	hall	0.5372	0.9112	0.8581
1.7B	reward	0.5368	0.9116	0.8595
1.7B	both	0.5388	0.9112	0.8606
1.7B	sequential	0.5382	0.9113	0.8680
4B	unfiltered	0.5577	0.9185	0.8542
4B	hall	0.5535	0.9185	0.8546
4B	reward	0.5599	0.9191	0.8548
4B	both	0.5606	0.9197	0.8570
4B	sequential	0.5604	0.9223	0.8563
8B	unfiltered	0.5734	0.9187	0.8527
8B	hall	0.5761	0.9184	0.8526
8B	reward	0.5667	0.9175	0.8533
8B	both	0.5764	0.9191	0.8540
8B	sequential	0.5585	0.9189	0.8537

hard requirement can apply the HallucinationGate explicitly at any generator size.

Practical recommendations. Gate with both hallucination and reward signals using exact source provenance and recover failures with the DiagnosticProbe and RewardRefiner. Generator scale is the dominant quality driver, but curation is most valuable precisely where larger generators are non-negotiable due to latency, memory, or edge deployment constraints. In those settings the curation pipeline closes a meaningful fraction of the quality gap that generator capacity cannot. The DiagnosticProbe’s failure-mode telemetry, covering temperature drift, parametric leakage, and instruction quality failures, is independently useful for prompt engineering regardless of downstream quality gains.

Why adaptive repair outperforms naive retry. Targeted diagnosis applies mode-specific patches rather than blind resampling. Structured repair correctly fails to fix genuine injections, since patching a hallucination requires altering content rather than style, driving injection recall +13.5 points above

Table 16: Full cross-family gate comparison, both judges, both benchmarks.

Gate configuration	Mistral-24B			Gemma-27B		
	F1	P	R	F1	P	R
<i>FaithDial</i> ($n=3,507$)						
Exact provenance (ours)	0.477	0.358	0.716	0.486	0.350	0.793
Oracle holistic	0.489	0.346	0.834	0.467	0.309	0.956
Retrieved provenance	0.428	0.334	0.596	0.448	0.348	0.628
Reward only	0.281	0.231	0.357	0.281	0.339	0.561
<i>HalluDial</i> ($n=3,507$, 73% hallucinated)						
Exact provenance (ours)	0.690	0.597	0.819	0.681	0.581	0.824
Oracle holistic	0.706	0.599	0.861	0.669	0.514	0.957
Retrieved provenance	0.491	0.567	0.432	0.496	0.599	0.424
Reward only	0.325	0.253	0.454	0.368	0.268	0.589

naive retry. The staged probe-then-reward pathway provides two targeted recovery attempts per hall-rejected sample, followed by orthogonal reward refining. Together these yield a 32.5% recovery rate vs 24.9% for naive retry at 1.7B, with the advantage growing to +38.1 points at 8B. That 45-51% of probe-recovered samples subsequently fail the reward gate confirms hallucination and reward failures are largely orthogonal, consistent with Section 4.1.

When naive retry is a reasonable choice. Naive retry is viable when simplicity and throughput matter more than maximising injection recall. It requires no failure-mode taxonomy and only two extra gate calls per attempt. The yield gap (68.4% vs 71.6% at 1.7B) is modest, and its lower natural rejection rate is a genuine benefit when preserving non-injected samples is the primary concern. The tradeoff shifts further against naive retry at larger generator sizes, where adaptive simultaneously achieves lower natural rejection and higher injection recall.

Why filtration is flat downstream. At approximately 3,000 training samples per condition, quality differences between filtration strategies are likely washed out by sample size, and all conditions already remove the noisiest samples. This flatness may not hold at smaller training sizes or in harder-domain settings where source faithfulness is more consequential.

Scope of generalisation. Most main experiments use Qwen3 generators, judges, and fine-tuning base. Section 4.4 adds cross-family judge/generator validation, but broader replication remains open. Whether the DiagnosticProbe’s failure-mode taxonomy transfers across generator families is an open question, and the downstream flatness may partly reflect Qwen3’s pre-training coverage reducing sen-

sitivity to curation at the scales we test.

E Hyperparameters

Table 17: Fine-tuning, pipeline, evaluation hyperparameters, compute budget, and software versions.

Component	Value
<i>Fine-tuning</i>	
Base model	Qwen3-4B-Instruct-2507
LoRA rank / α	16 / 32
Optimizer / lr	AdamW / 2×10^{-4}
Effective batch	32
Epochs / patience	3 / 2
Eval interval	0.5 epoch
Train / val (hall)	5,000 / 500
Train / val (other)	$\sim 3,000$ / 500
<i>Pipeline</i>	
Gate thresholds	hall 0.8, reward 0.7
Retry budget	5 attempts
Decoding temp. (gen)	0.7
Embedding retriever	BAAI/bge-base-en-v1.5
<i>Compute: Data Curation</i>	
GPU	A100 (80 GB) $\times 1$
Total GPU-hours	~ 18 hours
<i>Compute: fine-tuning and eval</i>	
GPU	RTX Pro 6000 (96 GB) $\times 1$
Total GPU-hours	8 hours
<i>Software versions</i>	
Unsloth	v0.1.41-beta (Daniel Han and team, 2023)
vLLM	vllm 0.21.0 (Kwon et al., 2023)
LiteLLM	≥ 1.40
openai	≥ 1.30
sentence-transformers	≥ 2.2
datasets	≥ 2.14
rouge_score	0.1.2
bert_score	0.3.13 (roberta-large)