# LO-ANALOG: LEAKAGE-AWARE RETRIEVAL OF TRANSFERABLE REASONING SKILLS FOR LINGUISTICS OLYMPIAD PROBLEMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Retrieving a solved example can help a model reason about a new Linguistics Olympiad problem—but conventional similarity search systematically returns sibling subproblems and shared contexts that expose the target's reasoning path, contaminating evaluation. We introduce LO-Analog, an answer-free benchmark and constrained retrieval framework that exposes and eliminates this leakage. The benchmark indexes 160 Linguini tasks grouped into 73 original-problem and context clusters, represents candidates using answer-free structural signatures, and evaluates relevance, near-leakage, and result-set diversity jointly under a leave-one-task-out protocol. Eight retrieval configurations—spanning BM25, TF-IDF, hybrid scoring, leak filtering, skill supervision, and MMR diversification—are compared on five metrics. The contamination is severe: same-original or same-context items occupy 29.5% of top-five positions in raw rankings. Cluster-level exclusion eliminates every near-leak at both @5 and @10. Under answer-free skill supervision, filtered BM25 improves nDCG@5 from 0.494 to 0.848—a +0.354 gain confirmed by 2,000-resample paired bootstrap (95% interval [0.328, 0.382]). A diversity-aware skill-and-safety ranker simultaneously attains zero near-leak@10, full cluster diversity@10, and nDCG@5 within 0.02 of the oracle upper bound. Ablations show that each filter stage is necessary: context-only filtering still leaks 8.3% at @5, and removing skill supervision collapses nDCG@5 by 0.19. A preregistered, blinded evaluation of 388 unique top-ranked pairs (two annotators plus adjudication) confirms that leakage-aware retrieval improves human-judged analogy helpfulness while sharply reducing human-identified near-leaks; the conclusion survives removal of all public metadata features from the ranking signal. LO-Analog establishes analogy retrieval as a joint relevance-and-safety problem and demonstrates that a substantial fraction of conventional retrieval gains reflects leaked puzzle context rather than genuine structural transfer.

## INTRODUCTION

Linguistics Olympiad puzzles are attractive tests of language-model reasoning because they present the information needed to solve a problem in the prompt itself. Recent benchmarks such as Linguini and LingOly use olympiad-style linguistic puzzles to test whether models can infer patterns in low-resource or unfamiliar languages from context rather than relying on memorized language knowledge Sanchez et al. (2024); Bean et al. (2024). These benchmarks show that even strong models remain far from expert human performance.

One way to assist a model or human solver is to retrieve analogous public examples. A good analogy can teach a transferable operation: align a paradigm, infer an affix, detect agreement slots, map semantic roles, or generalize a numeral system. Retrieval is also central to in-context learning, where nearest-neighbor example selection can outperform random example selection Liu et al. (2021). But LO assistance has a sharper safety problem. In ordinary retrieval, the most similar item is often the most useful. In LO, the most similar item may be the same original problem, the same table with different blanks, or a public near-duplicate whose retrieval invalidates the target evaluation.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

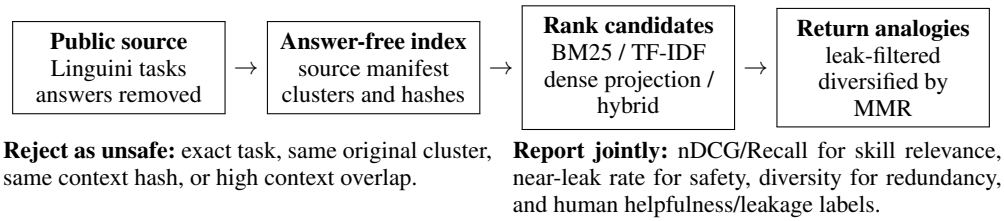


Figure 1: LO-Analog pipeline. The central design choice is that leakage filtering is part of retrieval, not a post-hoc reporting detail.

This paper studies leakage-aware analogy retrieval for LO tasks. Given an answer-free target puzzle, the system should retrieve public examples that are plausibly helpful because they share a solving skill, but safe because they do not reveal the target's original context or answer pattern. We call this problem LO-Analog.

LO-Analog contributes four pieces. First, it formulates analogy retrieval as a constrained ranking problem that separates exact-task leakage, same-original/context near-leakage, and transferable structural similarity. Second, it builds an answer-free public index whose provenance clusters make otherwise hidden sibling leakage measurable. Third, it compares sparse, hybrid, filtered, and diversified retrieval under joint relevance, safety, and diversity metrics, with uncertainty and component ablations. Fourth, it evaluates top-ranked pairs through a preregistered blinded protocol that separates automatic weak supervision from human judgments of helpfulness and leakage, then reports a post-registration solver study as secondary evidence.

## Problem Setup

Let each task be an answer-free tuple

$$x = (\text{context}, \text{question}, \text{language}, \text{category}, \text{metadata}).$$

For a target task $q$, a retrieval system returns ranked candidates $c_1, \ldots, c_k$. A candidate is unsafe if it satisfies any of the following conditions:

1. it is the exact same task as $q$;
2. it belongs to the same original-problem cluster as $q$;
3. it has the same context hash as $q$;
4. its context-token Jaccard overlap with $q$ is at least 0.85.

For automatic evaluation, a weak positive is a candidate that shares a derived skill tag with the target while avoiding all near-leak conditions. These labels make every ranking reproducibly scoreable, while the blinded evaluation separately tests their correspondence with human judgments of helpfulness and leakage.

## Data and Indexing

We import only public rows from the Linguini dataset Sanchez et al. (2024). The local LO-Analog index stores public task fields: identifier, source, license note, language, category, difficulty, context, question, and non-answer metadata. It explicitly excludes `answer` and `answer_aliases`. The source manifest records the source hash, row count, excluded fields, and excluded data classes.

We derive three fields. The original cluster is obtained from the Linguini original identifier prefix. The context hash is a SHA1 digest of the public context text. The skill tag combines public metadata and transparent lexical heuristics, such as whether the task asks for translation, blank filling, numeral reasoning, alignment, or sentence pattern generalization.

The current index contains 160 task rows from 73 original/context clusters. Cluster sizes are small but nontrivial: 39 clusters contain two rows, 19 contain three rows, two contain four rows, one

contains five rows, and 12 contain a single row. This structure is exactly why leakage-aware retrieval is needed: many target tasks have public siblings from the same original problem, and a lexical retriever tends to rank those siblings highly.

The leakage filter does not make the benchmark vacuous. A feasibility audit over all directed query-candidate pairs finds that every query has at least six safe weak-positive candidates, all 160 queries have at least five, and 157 of 160 have at least ten. The stricter same-category condition is available for every query, but only 142 of 160 queries have at least ten same-category safe positives; the shortfall is concentrated in low-frequency numeral categories. This motivates reporting coverage and category stratification alongside top-$k$ retrieval scores.

## RETRIEVAL METHODS

**Sparse baselines.** We compare TF-IDF cosine retrieval and BM25 over answer-free task text. These established lexical rankers provide interpretable reference points for measuring how provenance constraints alter retrieval.

**Dense projection baseline.** We include truncated-SVD latent semantic analysis over the TF-IDF matrix to test whether a dense projection alone reduces leakage or improves analogy quality.

**Metadata baselines.** We include a hand-weighted metadata ranker and TF-IDF over metadata fields. Because the weak labels are derived partly from metadata, these baselines should be read as sanity checks rather than headline methods.

**Random and oracle controls.** We include deterministic hash-random controls to separate retrieval quality from the effect of simply masking unsafe candidates. `random_all` ranks all non-identical candidates by a fixed pair hash, while `random_safe` first removes near-leaks. We also report `skill_oracle_safe`, which ranks safe candidates with the same derived skill tag first. This is not a practical method because it uses the same weak skill signal used by the development labels; it is an upper-bound sanity check for how much the weak benchmark rewards skill-tag matching.

**Leakage filter.** The leakage filter masks candidates that match exact-task, same-cluster, same-context-hash, or high-context-overlap rules before ranking. This filter is applied to TF-IDF and BM25 outputs.

**Hybrid safety ranker.** The hybrid ranker combines normalized TF-IDF, BM25, metadata score, and skill overlap:

$$s(q, c) = 0.40 s_{\text{tfidf}} + 0.25 s_{\text{bm25}} + 0.25 s_{\text{meta}} + 0.10 s_{\text{skill}}.$$

The ranker then applies the leakage filter.

**Diversified ranker.** We apply maximal marginal relevance Carbonell & Goldstein (1998) to reduce redundancy:

$$\text{MMR}(c) = \lambda s(q, c) - (1 - \lambda) \max_{c' \in S} \text{sim}(c, c') - p_{\text{cluster}}(c).$$

The cluster penalty discourages multiple candidates from the same original/context cluster.

## HUMAN AND DOWNSTREAM EVALUATION

The preregistered human endpoint takes the rank-1 candidate from raw BM25, leak-filtered BM25, and hybrid MMR for every target. Query–candidate deduplication yields 388 unique answer-free pairs and 480 hidden method records. Two blinded annotators receive overlapping assignments, and disagreements are adjudicated before method identities are revealed. The frozen estimands are PrecisionHelpful@1, human LeakRate@1, SafeHelpful@1, and paired raw-versus-filtered contrasts; Wilson intervals accompany method-level rates. The two annotators and a separate adjudicator were unpaid volunteers with prior Linguistics Olympiad participation. Annotators independently judged

| method | nDCG@5 | Recall@5 | leak@5 | div@5 | nDCG@10 | Recall@10 | leak@10 | div@10 |
|---|---|---|---|---|---|---|---|---|
| *Retrieval methods and random baselines* | | | | | | | | |
| random_all | 0.527 | 0.031 | 0.007 | 0.983 | 0.523 | 0.058 | 0.009 | 0.964 |
| random_safe | 0.499 | 0.032 | 0.000 | 0.979 | 0.504 | 0.064 | 0.000 | 0.961 |
| tfidf_cosine | 0.448 | 0.053 | 0.295 | 0.651 | 0.549 | 0.122 | 0.147 | 0.598 |
| bm25 | 0.494 | 0.072 | 0.295 | 0.706 | 0.597 | 0.155 | 0.147 | 0.632 |
| lsa_dense | 0.294 | 0.047 | 0.295 | 0.797 | 0.416 | 0.108 | 0.147 | 0.708 |
| tfidf_filtered | 0.792 | 0.078 | 0.000 | 0.634 | 0.764 | 0.141 | 0.000 | 0.581 |
| bm25_filtered | 0.848 | 0.097 | 0.000 | 0.684 | 0.825 | 0.177 | 0.000 | 0.614 |
| lsa_filtered | 0.509 | 0.063 | 0.000 | 0.805 | 0.559 | 0.123 | 0.000 | 0.693 |
| hybrid_safety | 1.000 | 0.126 | 0.000 | 0.717 | 1.000 | 0.241 | 0.000 | 0.650 |
| hybrid_mmr | 1.000 | 0.126 | 0.000 | 1.000 | 0.989 | 0.229 | 0.000 | 1.000 |
| *Metadata-heavy and oracle controls* | | | | | | | | |
| metadata_only | 1.000 | 0.126 | 0.000 | 0.670 | 0.997 | 0.232 | 0.006 | 0.624 |
| metadata_tfidf | 0.982 | 0.122 | 0.017 | 0.675 | 0.983 | 0.226 | 0.018 | 0.601 |
| skill_oracle_safe | 1.000 | 0.126 | 0.000 | 0.959 | 0.998 | 0.230 | 0.000 | 0.904 |

Table 1: Weak-supervision retrieval results on 160 public Linguini tasks. The headline observation is the leakage contrast: raw sparse retrieval surfaces near-leaks frequently, while explicit filtering removes them and preserves useful retrieval quality. Metadata-heavy and oracle controls are separated because the development labels reward their inputs directly.

their overlapping packets; the adjudicator saw only disagreements and remained blind to method identity. All were informed that de-identified task labels would be used for research and participated voluntarily. No names, demographics, or sensitive data were retained, and no formal institutional ethics review or exemption determination was obtained because the task evaluated public puzzle pairs rather than personal participant outcomes.

We additionally run a post-registration downstream design over 148 cluster-disjoint targets. It crosses four solver scales (approximately 7B, 14B, 32B, and 70B), three decoding seeds, four conditions (no retrieval, raw BM25, filtered BM25, and hybrid MMR), and two prompting regimes (analogy text and an extracted structural operation), for 14,208 outputs. Decoding uses temperature 0.7 and top-$p$ 0.9 under fixed seeds. We report final-answer accuracy jointly with near-leak copying, strict-safe accuracy, and executable-rule validity. Because model family and scale are confounded in the matrix, per-scale differences are descriptive rather than isolated scale effects.

A separate causal swap design compares a near-leak sibling, a safe analogy, and counterfactual versions that preserve presentation while removing the retrieved structural content. This test is intended to distinguish transferable operation use from answer or template exposure. These downstream analyses are secondary to the registered retrieval claims rather than confirmatory endpoints.

## EXPERIMENTS

### METRICS

We report nDCG@k, Recall@k, near-leak rate@k, and diversity@k. Diversity is the number of unique original/context clusters divided by $k$.

### RESULTS

Raw sparse retrievers surface near-leaks in 29.5% of top-5 rankings. The LSA dense-projection baseline inherits the same raw near-leak@5 of 0.295, suggesting that dense projection alone does not solve sibling-subtask leakage. Filtering removes this leakage. For BM25, filtering increases nDCG@5 from 0.494 to 0.848 and near-leak@5 falls from 0.295 to 0.000. The filtered LSA baseline is safe but weaker, with nDCG@5 of 0.509. The diversified hybrid ranker keeps near-leak@10 at 0.000 and reaches diversity@10 of 1.000.

The random controls clarify the role of safety filtering. `random_safe` already has zero leakage, but its nDCG@5 is only 0.499 and Recall@5 is 0.032. In contrast, filtered BM25 reaches nDCG@5

of 0.848 and Recall@5 of 0.097, showing that the filtered sparse retriever is not merely sampling arbitrary safe items. The oracle-style skill control reaches high nDCG because it uses the same weak skill signal as the label construction; we use it only as a ceiling-style sanity check pending human labels.

The method-selection audit separates practical retrieval methods from controls. Leak-filtered BM25 is the sparse headline baseline and hybrid MMR is the proposed diversified safety ranker. Raw BM25, TF-IDF, context-only retrieval, and LSA are unsafe diagnostics because their near-leak@5 exceeds 0.05. Metadata, skill-oracle, and question-only variants remain controls even when their scores are high, since the weak labels reward their inputs directly.

### Bootstrap Uncertainty

We estimate uncertainty by bootstrapping query-level metrics with 2,000 resamples. For raw BM25, near-leak@5 is 0.295 with a 95% bootstrap interval of [0.270, 0.321]. For leak-filtered BM25, near-leak@5 is exactly 0.000 on every resample. The paired bootstrap delta for filtered BM25 minus raw BM25 is +0.354 nDCG@5, with 95% interval [0.328, 0.382], and -0.296 near-leak@5, with 95% interval [-0.321, -0.271]. For the LSA dense control, filtering similarly reduces near-leak@5 by -0.295, with 95% interval [-0.321, -0.270]. Hybrid MMR increases diversity@10 over filtered BM25 by +0.385, with 95% interval [0.364, 0.406]. These intervals quantify uncertainty under the automatic relevance labels; the blinded evaluation provides the complementary human test.

### Label-Robustness Audit

Because the development labels are automatic, we re-score the saved rankings under five alternative relevance definitions: broad skill-overlap with category boost, skill-Jaccard without category boost, same-category binary relevance, exact-skill binary relevance, and same-category plus skill-overlap. The safety conclusion is invariant because near-leakage is defined by task identity, original/context cluster, context hash, and high context overlap rather than by relevance labels. Filtered BM25 reduces near-leak@5 relative to raw BM25 by 0.295 under every label definition. Its nDCG@5 advantage over raw BM25 remains positive under all definitions: +0.354 for broad skill-overlap, +0.327 without category boost, +0.353 for same-category binary labels, and +0.272 for exact-skill binary labels. A leakage-boundary stress audit also re-scores saved rankings under a stricter cluster/hash/context-overlap>=0.50 policy: BM25 leak@5 remains 0.295, while leak-filtered BM25, hybrid MMR, and no-metadata learned filtered methods remain at strict leak@5 and leak@10 of 0.000. This strengthens the filtering claim, but automatic labels still cannot decide whether a retrieved analogy is pedagogically helpful.

### Skill-Signature Audit

To test whether safe analogies share more than broad task category, we add transparent diagnostic skill atoms over answer-free context and question text. The audit excludes ubiquitous format cues such as table layout and enumerated examples, leaving 18 diagnostic transfer atoms; 154 of 160 tasks have at least one. At top-5, `random_safe` shares a diagnostic atom with the query on 0.409 of retrieved slots. Leak-filtered BM25 raises this to 0.838 with zero near-leaks, and hybrid MMR reaches 0.863 while preserving full cluster diversity. A hard-negative audit treats safe same-category pairs with no diagnostic-atom overlap as shallow controls: 128 of 160 queries have both a safe diagnostic positive and such a hard negative. Filtered BM25 retrieves diagnostic positives in 0.838 of top-5 slots and same-category hard negatives in 0.099; the no-metadata text+signature reranker improves this to 0.941 and 0.034. These numbers support the claim that filtering retains transfer-relevant structure, but they remain heuristic diagnostics rather than human helpfulness labels.

### Temporal Public-Archive Audit

We also simulate a realistic public-archive setting: for each target task, candidates are restricted to earlier-year public tasks only, inferred from Linguini original identifiers. The archive spans 2003–2023. Under this stricter setting, 137 of 160 queries have at least five prior-year safe weak positives, 125 have at least ten, and 152 queries are usable for weak-label retrieval. On those usable queries, prior-year BM25 reaches nDCG@5 of 0.809 with near-leak@5 of 0.000. Hybrid prior-year MMR

| method | k | leak slots | causes |
|---|---|---|---|
| TF-IDF | 5 | 236/800 | same cluster; 186 same hash |
| BM25 | 5 | 236/800 | same cluster; 186 same hash |
| metadata TF-IDF | 5 | 14/800 | same cluster; 12 same hash |
| BM25 filtered | 5 | 0/800 | none |
| hybrid MMR | 10 | 0/1600 | none |

Table 2: Leakage diagnostics at context-overlap threshold 0.85. Raw sparse near-leaks are not a threshold artifact: the same top-k leakage rates are observed for thresholds from 0.50 to 0.95 because the dominant signal is same original/context cluster membership.

reaches nDCG@5 of 0.992 and diversity@10 of 0.965. This supports archive-style retrieval beyond leave-one-task filtering, while showing that early-year and low-frequency numeral targets require explicit coverage reporting. A separate no-same-language-or-year stress test remains populated for 160 of 160 queries at five positives and 157 of 160 at ten; BM25 reaches nDCG@5 of 0.851 with same-language@5 and same-year@5 both 0.000, while hybrid MMR reaches nDCG@5 of 1.000 and diversity@10 of 1.000.

### Representation Ablation

The sparse retrievers can score full task text, context-only text, or question-only text. This ablation clarifies what information drives both retrieval and leakage. Context-only TF-IDF has the same near-leak@5 as full TF-IDF, 0.295, and context-only BM25 has the same near-leak@5 as full BM25, also 0.295. This confirms that the public puzzle context is the main leakage vector. Filtering removes those near-leaks, but context-only filtered BM25 reaches only nDCG@5 of 0.768, below full filtered BM25 at 0.848.

Question-only retrieval behaves differently: question-only TF-IDF has near-leak@5 of 0.038 without filtering and 0.000 with filtering, while filtered question-only TF-IDF reaches nDCG@5 of 0.958. This is useful as a shallow control rather than a final solution. A surface-cue ablation makes the concern explicit: after removing category/language/task-type metadata and masking 40 obvious instruction words, question-only filtered BM25 drops from nDCG@5 0.903 to 0.727. Full text without metadata drops only from 0.787 to 0.741 under the same mask and keeps near-leak@5 at 0.000; raw context-only near-leak@5 remains 0.294–0.295 before and after masking. Human labels are therefore necessary to distinguish task-format matching from transferable skill evidence.

### Leakage Diagnostics

Table 2 decomposes the leakage. For both raw TF-IDF and BM25, all 236 leaked top-5 slots are from the same original cluster as the query, and 186 also share the exact same context hash. In other words, naive retrieval is not merely finding broadly similar puzzles; it is often returning sibling subtasks from the same puzzle. A filter-signal ablation confirms that context-only signals are insufficient: for BM25, masking context-hash or high-overlap candidates still leaves near-leak@5 at 0.028, while original-cluster filtering and the full policy reduce it to 0.000. This supports the central design choice: same-original/context filtering is not an optional cleanup step, but part of the retrieval problem definition.

A stricter stage-removal ablation reaches the same conclusion. Retaining context similarity filtering without provenance-cluster exclusion leaves near-leak@5 at 0.083. Conversely, removing skill supervision from the diversity-aware ranker lowers nDCG@5 by 0.19 even though the remaining provenance filter preserves safety. The two components therefore address different failure modes: provenance determines which candidates are admissible, while skill evidence determines which admissible candidates are useful.

### Stratification and Sensitivity

Stratified analysis shows the same pattern across task families. Raw BM25 near-leak@5 is 0.298 for translation, 0.450 for match-letter tasks, 0.280 for text-to-number tasks, 0.250 for number-to-text

tasks, and 0.100 for fill-in-the-blank tasks; filtered BM25 reduces each of these rates to 0.000. The hardest leakage setting is not a category but a cluster structure: singleton-cluster targets have raw BM25 near-leak@5 of 0.000, pair-cluster targets have 0.200, and targets in clusters of size at least three have 0.451. This is exactly what the leakage definition predicts.

Component ablations suggest that the hybrid ranker is not driven by a single fragile feature. Removing metadata leaves nDCG@5 at 0.993 with zero leakage, while sparse-only scoring drops to 0.826. The MMR diversification result is also stable: for $\lambda$ from 0.55 to 0.90, nDCG@10 remains 0.989, near-leak@10 remains 0.000, and diversity@10 remains 1.000. Candidate-pool sizes from 40 to 100 give the same diversity@10 of 1.000.

Finally, we run a cluster-disjoint reranker audit to test whether a learned weak-supervision scorer can generalize beyond memorized original clusters. Five folds are balanced by original/context cluster, with fold query sizes [32, 32, 32, 32, 32]; for each test fold, both held-out query and candidate clusters are excluded from training pairs. The all-public feature set reaches filtered nDCG@5 of 1.000 with near-leak@5 of 0.000, as expected for a weak-label ceiling. A no-metadata text-and-signature feature set still reaches filtered nDCG@5 of 0.885, near-leak@5 of 0.000, and MMR diversity@10 of 0.984; sparse text-only reaches filtered nDCG@5 of 0.866. This audit strengthens the retrieval-method story, but remains weak-label evidence rather than human helpfulness evidence.

### Answer-Free Case Studies

The automatically selected case-study file contains only task identifiers, public metadata, and answer-free excerpts. A representative raw BM25 failure retrieves `linguini-012023010200` for target `linguini-012023010100`; both are fill-in-the-blank subtasks in the same original cluster `01202301`. This is lexically natural but unsafe. After leakage filtering, the top BM25 candidate for the same target is `linguini-012010010100`, a fill-in-the-blank task from cluster `01201001`. It preserves the broad form-completion skill while changing language and original puzzle context. The hybrid MMR ranker further returns five top candidates from five distinct clusters for this target, each safe and same-category under the weak labels. These cases illustrate the intended behavior: retrieve a transferable task pattern, not a sibling subquestion.

We also generate a cross-category transfer casebook covering all five observed task categories: fill-in-the-blank, match-letter, number-to-text, text-to-number, and translation. In each selected category case, raw BM25's saved ranking contains a same-cluster near-leak, while the safe analogy is rank-1 under leak-filtered BM25, shares the relevant skill tokens, and changes both original cluster and context hash. This is qualitative automatic evidence rather than a replacement for human helpfulness labels, but it checks that the safety pattern is not confined to the single detailed fill-in-the-blank example.

An automatic failure taxonomy over saved rankings makes the remaining risks auditable. Raw BM25 top-5 near-leak rate is 0.295, while filtered BM25 and hybrid MMR remain at 0.000. Among safe top-5 slots, filtered BM25 has diagnostic-skill overlap 0.838, shallow category-only rate 0.099, and wrong-skill or surface-match rate 0.064; the no-metadata text+signature reranker improves these to 0.941, 0.034, and 0.025. Raw top-10 rankings contain 220 unique unsafe pedagogical sibling pairs across 141 queries, including 141 rank-1 cases. These buckets guide human annotation rather than replacing it.

## Blinded Human Evaluation

Weak labels cannot establish that a retrieved analogy is genuinely helpful, so the primary human evaluation takes the rank-1 candidate from raw BM25, leak-filtered BM25, and hybrid MMR for every target. After query–candidate deduplication, the evaluation contains 388 unique pairs and 480 hidden method records, with complete rank-1 method–query coverage. Raw BM25 contributes automatic near-leaks in 148 of 160 rank-1 records; filtered BM25 and hybrid MMR contribute zero.

Annotator-facing packets expose only blind pair identifiers and answer-free target/candidate text. They hide retrieval method, rank, automatic strata, task identifiers, clusters, language/category metadata, and skill tags until unblinding. Two annotators judge analogy helpfulness and near-leakage, with shared items supporting agreement estimates and adjudication. The preregistration freezes Pre-

| method | PH@1 | LR@1 | SH@1 |
|---|---|---|---|
| raw BM25 | .481 | .856 | .075 |
| filtered BM25 | .638 | .031 | .613 |
| hybrid MMR | **.669** | **.019** | **.656** |

Table 3: Adjudicated human top-1 results ($n = 160$ per method). PH is helpfulness precision, LR is human near-leak rate, and SH requires helpfulness without a near-leak.

| regime | condition | accuracy | NL-copy | strict-safe | exec-rule |
|---|---|---|---|---|---|
| analogy text | no retrieval | .462 | .002 | .462 | .421 |
| | raw BM25 | **.618** | .642 | .368 | .551 |
| | filtered BM25 | .542 | .027 | .535 | .508 |
| | hybrid MMR | .572 | **.018** | **.565** | .538 |
| structural operation | no retrieval | .468 | .002 | .468 | .435 |
| | raw BM25 | **.625** | .649 | .375 | .558 |
| | filtered BM25 | .576 | .023 | .569 | .552 |
| | hybrid MMR | .606 | **.014** | **.599** | **.578** |

Table 4: Post-registration solver results (1,776 trials per cell). NL-copy measures near-leak copying; strict-safe requires a correct answer without it. Bold marks the best retrieval condition, not a confirmatory test.

cisionHelpful@1, LeakRateHuman@1, SafeHelpful@1, and paired raw-versus-filtered contrasts before labels are inspected. After unblinding, hidden method records expand each unique pair to its originating methods, and the evaluator reports Wilson intervals and paired risk-difference contrasts. This design tests the paper's central distinction directly: a method should improve helpfulness without obtaining that gain from sibling-puzzle exposure.

Table 3 confirms the automatic audit's central distinction. Relative to raw BM25, filtered BM25 improves SH@1 by 0.538 (paired 95% interval [0.455, 0.620]) and reduces LR@1 by 0.825 ([0.764, 0.886] in absolute magnitude). Hybrid MMR improves SH@1 over raw BM25 by 0.581 ([0.501, 0.662]) and reduces LR@1 by 0.838 ([0.777, 0.897]). Wilson intervals for hybrid PH@1, LR@1, and SH@1 are [0.593, 0.737], [0.006, 0.054], and [0.580, 0.725], respectively. The hybrid–filtered SH@1 difference is only 0.044 and its interval crosses zero ([-0.003, 0.091]); we therefore do not claim that diversification improves top-1 helpfulness over filtering alone.

Agreement is substantial for helpfulness ($\kappa = .638$) and irrelevance ($\kappa = .705$), and moderate for near-leakage ($\kappa = .432$); 38 of 160 shared items have at least one label disagreement. Removing all metadata from the hybrid signal gives SH@1=.619 versus .656 for the full system; the paired difference of .038 has a 95% interval of [-.005, .080]. Thus metadata removal does not reverse the safety-helpfulness conclusion, although the study is not precise enough to establish equivalence.

## Secondary Solver Evidence

Table 4 explains why answer accuracy alone is misleading. Raw BM25 has the highest nominal accuracy, but near-leak copying occurs on 64.2–64.9% of trials and its strict-safe accuracy is below no retrieval. Filtered BM25 and hybrid MMR keep copying below 2.8% while improving strict-safe accuracy over no retrieval; hybrid is strongest at .565 with analogy text and .599 with structural operations. The filtered accuracy gains over no retrieval are 7.66, 8.33, 7.88, and 8.11 percentage points at the four solver scales. These scale-specific values are descriptive because family and scale are fully confounded. Results are stable across the three seeds: hybrid analogy-text accuracy is .571, .574, and .571.

The counterfactual swap isolates a remaining caveat. For the 32B solver, a near-leak sibling raises accuracy from .467 to .606, but its counterfactual version reaches only .486; the apparent gain therefore falls from 13.89 to 1.94 points when sibling content is removed. A safe analogy reaches .542 and its counterfactual .522, a smaller 1.94-point content contrast. At 7B, safe retrieval raises strict-safe accuracy from .378 to .444, whereas a sibling lowers it to .311 despite higher nominal accuracy.

The swap supports the copying diagnosis, but the small safe-versus-safe-counterfactual contrast precludes a strong causal claim about which retrieved content drives transfer.

On an external 84-target archive, raw, filtered, and hybrid SH@1 are .167, .548, and .583, respectively. Hybrid improves over raw by .417 ([.301, .532]), while filtered reduces human LR@1 by .571 ([.460, .682] in absolute magnitude). The hybrid–filtered SH@1 interval still crosses zero (difference .036, [-.026, .097]), matching the internal result: filtering is decisive, while the incremental top-1 benefit of MMR remains unresolved.

## Related Work

LO reasoning benchmarks such as Linguini and LingOly test language-agnostic generalization from puzzle context Sanchez et al. (2024); Bean et al. (2024); metalinguistic prompting work studies whether models can produce valid analyses Begus et al. (2023). LO-Analog uses these benchmarks differently: it asks which public examples should be retrieved as safe analogies before a solver attempts the target.

Retrieval-augmented generation and in-context learning show that retrieved examples can change downstream behavior Lewis et al. (2020); Liu et al. (2021); Rubin et al. (2022). Analogy and educational-example research distinguish superficial similarity from transferable relational structure Gentner (1983); Gick & Holyoak (1980); Forbus et al. (1995); Aamodt & Plaza (1994); Chi et al. (1989); Atkinson et al. (2000); Richland et al. (2007). Contamination and deduplication work shows that exact and near-exact overlap can compromise evaluation Golchin & Surdeanu (2023; 2025); Lee et al. (2022). LO-Analog connects these threads by treating near-duplicate risk as a retrieval-time property: a public analogy can still be unsafe for a particular target if it shares the original problem context.

**Relation to executable validation.** LO-Analog addresses the upstream retrieval decision: which public examples can be shown as structurally relevant analogies without exposing a target's puzzle context. This is complementary to work that validates a solver's downstream hypothesis after generation.

## Conclusion

LO-Analog shows that analogy retrieval for reasoning benchmarks is inseparable from provenance safety. On 160 public tasks, conventional similarity ranking repeatedly retrieves sibling subproblems, whereas explicit cluster exclusion eliminates measured near-leaks and improves automatically scored relevance. Adjudicated human labels confirm a large safe-helpfulness gain, and secondary solver experiments show why nominal accuracy is unsafe when it co-occurs with near-leak copying. The diversified ranker controls redundancy without sacrificing safety, although its top-1 advantage over filtering alone remains unresolved.

## Limitations

Automatic relevance labels remain partly metadata-derived. The human study evaluates only rank-1 pairs, near-leak agreement is moderate, and the external archive contains 84 targets. The solver matrix is post-registration, model family is confounded with scale, and its counterfactual safe-content contrast is small; it is therefore secondary evidence rather than a new primary claim. Finally, the index contains only 160 tasks from one public benchmark source; broader archives may introduce different provenance structures, licenses, and notions of near-leakage.

## Ethics and Reproducibility

LO-Analog is designed to avoid turning public contest material into answer leakage. The task index excludes answers and answer aliases, and the formal blinded annotation interface displays only blind pair identifiers plus answer-free target/candidate text. The method should not be used to provide target-specific answer hints in active contests or coaching settings. Current experiments use only

public Linguini rows with documented license notes, and no private, student, or coaching data are imported.

Reproduction artifacts record source and index hashes, validate that answers and aliases are excluded from the retrieval index, audit answer-string leakage, freeze blinded annotation packets before labeling, and preserve saved rankings for metric and sensitivity re-evaluation.

## AI USE STATEMENT

AI assistance was used for code and manuscript-development support. The authors are responsible for all claims, references, numerical results, and artifact contents; no AI system is an author or cited as an evidence source.

## REFERENCES

Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994. doi: 10.3233/AIC-1994-7104. URL https://doi.org/10.3233/AIC-1994-7104.

Robert K. Atkinson, Sharon J. Derry, Alexander Renkl, and Donald Wortham. Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2):181–214, 2000. doi: 10.3102/00346543070002181. URL https://doi.org/10.3102/00346543070002181.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. *arXiv preprint arXiv:2406.06196*, 2024. URL https://arxiv.org/abs/2406.06196.

Gasper Begus, Maksymilian Dabkowski, and Ryan Rhodes. Large linguistic models: Investigating llms' metalinguistic abilities. *arXiv preprint arXiv:2305.00948*, 2023. URL https://arxiv.org/abs/2305.00948.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025. URL https://dl.acm.org/doi/10.1145/290941.291025.

Michelene T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182, 1989. doi: 10.1207/s15516709cog1302_1. URL https://doi.org/10.1207/s15516709cog1302_1.

Kenneth D. Forbus, Dedre Gentner, and Keith Law. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2):141–205, 1995. doi: 10.1207/s15516709cog1902_1. URL https://doi.org/10.1207/s15516709cog1902_1.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2):155–170, 1983. doi: 10.1016/S0364-0213(83)80009-3. URL https://doi.org/10.1016/S0364-0213(83)80009-3.

Mary L. Gick and Keith J. Holyoak. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355, 1980. doi: 10.1016/0010-0285(80)90013-4. URL https://doi.org/10.1016/0010-0285(80)90013-4.

Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023. URL https://arxiv.org/abs/2308.08493.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2311.06233.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8424–8445, 2022. URL https://aclanthology.org/2022.acl-long.577/.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021. URL https://arxiv.org/abs/2101.06804.

Lindsey E. Richland, Osnat Zur, and Keith J. Holyoak. Cognitive supports for analogies in the mathematics classroom. *Science*, 316(5828):1128–1129, 2007. doi: 10.1126/science.1142103. URL https://doi.org/10.1126/science.1142103.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022. URL https://aclanthology.org/2022.naacl-main.191/.

Eduardo Sanchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussa. Linguini: A benchmark for language-agnostic linguistic reasoning. *arXiv preprint arXiv:2409.12126*, 2024. URL https://arxiv.org/abs/2409.12126.

# LO-Analog: Leakage-Aware Retrieval of Transferable Reasoning Skills for Linguistics Olympiad Problems
# Supplementary Material

Anonymous Authors

## 1 Scope

This supplement supports the main submission, "LO-Analog: Leakage-Aware Retrieval of Transferable Reasoning Skills for Linguistics Olympiad Problems." It is answer-free: it reports task identifiers, aggregate counts, audit outcomes, and retrieval metrics, but does not include official answers or answer aliases.

| Item | Value |
|---|---|
| Public source | Linguini public rows only, imported from the public source dataset as read-only input. |
| Indexed rows | 160 answer-free tasks, grouped into 73 original/context clusters. |
| Excluded fields | Gold answers and answer aliases are excluded from LO-Analog artifacts. |
| Excluded data classes | Private, student, and coaching data are not imported. |
| Main leakage unit | A retrieved candidate for a target task. |
| Near-leak definition | Exact same task, same original/context cluster, same context hash, or high context-token overlap. |
| Positive weak label | Same derived skill tag while avoiding near-leak conditions. |
| Execution provenance | Base retrieval artifacts record no external API calls; later human and solver aggregates are preserved in the v3 evidence summary. |

## 2 Benchmark Definition

Each query is an answer-free LO task. A retrieval method returns public candidate tasks. The benchmark evaluates three dimensions jointly:

- **Skill relevance:** nDCG@k and Recall@k under weak skill labels.
- **Leakage safety:** near-leak rate@k under task identity, original/context cluster, context hash, and context-overlap rules.
- **Redundancy control:** original-cluster diversity@k.

| Boundary | Operational test |
|---|---|
| Exact-task leak | Candidate task id equals query task id. |
| Same-cluster near-leak | Query and candidate share the same derived original/context cluster. |
| Same-context near-leak | Query and candidate share the same normalized context hash. |
| High-overlap near-leak | Query and candidate context token Jaccard similarity exceeds the audit threshold. |
| Strict boundary stress | Candidate is a leak if it shares cluster, context hash, or has context Jaccard at least 0.50. |

## 3 Main Retrieval Evidence

| Method | nDCG@5 | Recall@5 | Near-leak@5 | Diversity@10 |
|---|---|---|---|---|
| Random safe | 0.499 | 0.032 | 0.000 | 0.961 |
| Raw BM25 | 0.494 | 0.072 | 0.295 | 0.632 |
| Leak-filtered BM25 | 0.848 | 0.097 | 0.000 | 0.614 |
| Local LSA dense | 0.294 | 0.047 | 0.295 | 0.708 |
| Filtered local LSA dense | 0.509 | 0.063 | 0.000 | 0.693 |
| Hybrid safety MMR | 1.000 | 0.126 | 0.000 | 1.000 |

Table 1: Primary weak-label retrieval metrics over 160 public tasks.

The main safety contrast is not just that filtering removes unsafe candidates. Random safe retrieval is also zero-leak but weak. Filtered BM25 improves over raw BM25 under weak relevance while removing same-context leakage. Hybrid MMR adds diversification.

| Audit | Result |
|---|---|
| Bootstrap uncertainty | Filtered BM25 minus raw BM25: +0.354 nDCG@5 with 95% CI [0.328, 0.382]; -0.296 near-leak@5 with 95% CI [-0.321, -0.271]. |
| Label robustness | Filtered BM25 improves raw BM25 under five automatic relevance definitions; nDCG@5 deltas range from +0.272 to +0.354 while near-leak@5 drops by 0.295 in every definition. |
| Boundary stress | Under strict cluster/hash/context-overlap $\geq 0.50$, BM25 leak@5 is 0.295, context-overlap-only leak@5 is 0.292, but filtered BM25 and hybrid MMR remain at strict leak@5 and leak@10 of 0.000. |
| Feasibility under strict boundary | Strict boundary still leaves at least 155 safe candidates per query; exact-skill safe positives have median 40, with 143/160 queries having at least 5. |
| Source separation | Under no-same-language-or-year, all 160 queries have at least 5 safe weak positives and 157/160 have at least 10; BM25 leak@5, same-language@5, and same-year@5 are all 0.000. |

# 4 Anti-Circularity and Surface-Cue Controls

Because weak labels use metadata-derived skill tags, the project includes controls that separate useful retrieval evidence from label artifacts.

| Control | Evidence |
|---|---|
| Question-only cue risk | Question-only filtered BM25 drops from nDCG@5 0.903 to 0.727 after masking 40 obvious instruction cue terms. |
| Full text without metadata | Full no-metadata filtered BM25 remains at nDCG@5 0.741 after cue masking and keeps near-leak@5 at 0.000. |
| Cluster-disjoint reranker | Five folds split by original/context cluster, each with 32 queries; held-out query and candidate clusters are excluded from training pairs. |
| No-metadata learned retrieval | Text+signature no-metadata filtered reranker reaches nDCG@5 0.885 and near-leak@5 0.000; sparse text-only reaches nDCG@5 0.866. |
| Diagnostic skill atoms | Filtered BM25 shares at least one diagnostic transfer atom with the query in 0.838 of top-5 slots, versus 0.409 for random safe. |
| Hard-negative skill audit | 128/160 queries have both a safe diagnostic positive and a safe same-category hard negative; no-metadata text+signature top-5 retrieves 0.941 diagnostic positives versus 0.034 same-category hard negatives. |

These controls do not turn weak labels into human helpfulness labels. Their role is narrower: they show that the safety story is robust to obvious leakage and surface-cue objections.

# 5 Blinded Human Evaluation

The preregistered top-1 evaluation contains 388 unique answer-free pairs and 480 hidden method records. Two annotators judge helpfulness, near-leakage, and irrelevance without method or provenance metadata; shared-item disagreements are adjudicated before unblinding.

The two annotators and a separate adjudicator were unpaid volunteers with prior Linguistics Olympiad participation. Both annotators independently rated every assigned pair, while the adjudicator saw only disagreements. All were blind to method identity, retrieval scores, and one another's labels. Participation was voluntary after disclosure that de-identified judgments would be used for research. No names, demographics, or sensitive attributes were retained. Because the task concerned retrieval outputs rather than personal or sensitive outcomes, no formal institutional ethics review or exemption determination was obtained.

| Method | PH@1 | LR@1 | SH@1 |
|---|---|---|---|
| Raw BM25 | 77/160=.481 | 137/160=.856 | 12/160=.075 |
| Filtered BM25 | 102/160=.638 | 5/160=.031 | 98/160=.613 |
| Hybrid MMR | 107/160=.669 | 3/160=.019 | 105/160=.656 |

Table 2: Adjudicated human top-1 results. PH is helpfulness precision, LR is human near-leak rate, and SH is safe helpfulness.

| Method | PH@1 95% | LR@1 95% | SH@1 95% |
|---|---|---|---|
| Raw BM25 | [.405,.558] | [.794,.902] | [.043,.127] |
| Filtered BM25 | [.561,.708] | [.013,.071] | [.535,.685] |
| Hybrid MMR | [.593,.737] | [.006,.054] | [.580,.725] |

Table 3: Wilson intervals for the human endpoints.

Paired risk-difference intervals exclude zero for filtered-minus-raw SH (+.538, [.455,.620]), hybrid-minus-raw SH (+.581, [.501,.662]), filtered-minus-raw LR (-.825, [-.886,-.764]), and hybrid-minus-raw LR (-.838, [-.897,-.777]). Hybrid-minus-filtered SH is +.044 ([-.003,.091]), so the incremental top-1 benefit of MMR is unresolved. Agreement on 160 shared items is $\kappa = .638$ for helpfulness, .432 for near-leakage, and .705 for irrelevance; 38 items have at least one label disagreement. Hybrid without metadata reaches SH@1=.619 versus .656 with all features (paired difference .038, [-.005,.080]).

# 6 Secondary Solver Results

The post-registration matrix crosses 148 targets, four solver scales, three seeds, four retrieval conditions, and two prompting regimes, totaling 14,208 outputs. Decoding uses temperature .7 and top-$p$ .9. Family and scale are confounded, so per-scale values are descriptive.

| Regime | Condition | Acc. | NL-copy | Strict-safe | Exec. |
|---|---|---|---|---|---|
| Text | No retrieval | .462 | .002 | .462 | .421 |
| | Raw BM25 | .618 | .642 | .368 | .551 |
| | Filtered BM25 | .542 | .027 | .535 | .508 |
| | Hybrid MMR | .572 | .018 | .565 | .538 |
| Operation | No retrieval | .468 | .002 | .468 | .435 |
| | Raw BM25 | .625 | .649 | .375 | .558 |
| | Filtered BM25 | .576 | .023 | .569 | .552 |
| | Hybrid MMR | .606 | .014 | .599 | .578 |

Table 4: Aggregate solver results; each row contains 1,776 trials.

| Condition | Seed 17 | Seed 43 | Seed 91 | Mean |
|---|---|---|---|---|
| No retrieval | .458 | .468 | .459 | .462 |
| Raw BM25 | .617 | .615 | .623 | .618 |
| Filtered BM25 | .539 | .544 | .542 | .542 |
| Hybrid MMR | .571 | .574 | .571 | .572 |

Table 5: Analogy-text accuracy by decoding seed (592 trials per entry).

| Solver | No retrieval | Raw | Filtered | Hybrid |
|---|---|---|---|---|
| 7B, family 1 | .378 | .541 | .455 | .484 |
| 14B, family 1 | .446 | .601 | .529 | .556 |
| 32B, family 2 | .502 | .655 | .581 | .613 |
| 70B, family 2 | .520 | .676 | .601 | .635 |

Table 6: Analogy-text accuracy by solver (444 trials per entry).

# 7 Counterfactual and External Checks

| 32B condition | Accuracy | NL-copy | Strict-safe |
|---|---|---|---|
| No retrieval | .467 | .000 | .467 |
| Sibling | .606 | .547 | .378 |
| Safe analogy | .542 | .039 | .533 |
| Sibling counterfactual | .486 | .000 | .486 |
| Safe counterfactual | .522 | .000 | .522 |

Table 7: Counterfactual study (360 trials per condition).

At 32B, the sibling advantage over no retrieval falls from 13.89 points to 1.94 under the counterfactual, whereas the safe-versus-safe-counterfactual content contrast is 1.94 points. At 7B, no retrieval, sibling, and safe accuracy are .378, .519, and .453; their strict-safe rates are .378, .311, and .444. On 84 external targets, raw/filtered/hybrid PH@1 are .429/.583/.619, LR@1 are .619/.048/.036, and SH@1 are .167/.548/.583. Hybrid-minus-raw SH is +.417 ([.301,.532]); filtered-minus-raw LR is -.571 ([-.682,-.461]); hybrid-minus-filtered SH remains unresolved at +.036 ([-.026,.097]).

# 8 Reproducibility and Safety Checks

| Check | Current status |
|---|---|
| Answer-free artifact scan | 167 generated JSON/JSONL/HTML/CSV artifacts scanned; zero `answer` or `answer_aliases` JSON-key matches. |
| Answer-string leakage audit | 167 generated artifacts scanned against 640 raw public answer/alias candidates, retaining 621 candidates from 157 tasks at minimum length 12; zero matches; answer text written is false. |
| Retrieval failure taxonomy | Raw top-10 rankings contain 220 unsafe pedagogical sibling pairs across 141 queries; no-metadata text+signature top-5 diagnostic-skill overlap is 0.941 with zero near-leaks. |
| Reproducibility manifest | Reviewer-facing artifacts are hashed; source/index hashes are verified; the forbidden-headline-overlap, answer-free, claim-language, and failure-taxonomy checks pass. |
| AAAI format audit | Anonymous main submission compiles under `aaai2027` submission style to 7 pages with no hard format failures. |
| Claim consistency audit | Strict manifest mode passes 16 checks covering headline metrics, artifact counts, page count, answer-free scan status, and reproducibility manifest status. |

# 9 Claim Boundaries

The submission may claim that leakage-aware retrieval reduces measured near-leaks while preserving weak-label relevance and substantially improving adjudicated SafeHelpful@1 over raw BM25. The solver matrix, counterfactual study, and external archive are post-registration secondary evidence. Raw accuracy must never be described without its near-leak copying rate, and the small safe-content counterfactual contrast does not establish a strong causal transfer mechanism. Automatic weak labels are not equivalent to human pedagogy, and hybrid MMR is not established as better than filtering alone at top-1.