AUDITING STRUCTURED INFERENCE FOR OPEN-ENDED SEMANTIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured LLM workflows create three separable reliability questions: whether semantic content survives intermediate transformations, whether every stage satisfies its interface contract, and whether the scorer measures evidential support. We audit these questions using TAIWO, a typed six-stage inference graph, on a deterministic 100-document corpus with nine latent concepts assigned to three prevalence strata. Across 39 historical model runs, the frozen v1 task score averages 0.670 (median 0.737; range 0 to 0.883); 36 runs complete all structured stages and three malformed outputs receive task-declared zeroes. Within detailed outputs, recall averages 0.907 for concepts assigned dominant prevalence and 0.481 for those assigned rare prevalence, although concept identity is perfectly confounded with prevalence. The audit finds two measurement defects: an exporter omitted the three zeroes, and the v1 "grounding" term verifies identifier existence rather than evidential support. Exact concepts with deliberately wrong but valid source identifiers can therefore score 1.0. In an exploratory three-model equal-call pilot, the support-aware v2 paired difference between TAIWO and generic six-call decomposition is 0.0068 (95% model-clustered bootstrap CI $[-0.0845, 0.0908]$), providing no evidence of an advantage. A separate human-labelled category-transfer stress test on 100 Reddit mental-health posts also provides no evidence of an advantage: macro source-grounded category recall is 0.289 for TAIWO and 0.413 for generic decomposition (difference -0.1233, 95% CI $[-0.3540, 0.2241]$). The contribution is an auditable failure-bearing evaluation framework. Whether typed state, explicit failure semantics, and provenance contracts constitute useful narrower workflow novelty for TAIWO remains inconclusive because the pilots neither isolate these features nor show stable transfer.

1 INTRODUCTION

Structured prompting decomposes an inference problem into intermediate operations rather than asking a model for a final answer in one call. Chain-of-thought, least-to-most prompting, and decomposed prompting show that this strategy can improve closed tasks with checkable answers (Wei et al., 2022; Zhou et al., 2023; Khot et al., 2023). The situation is less settled for open-ended analysis. A system may produce plausible themes without recovering infrequent concepts, may cite unrelated evidence, or may fail partway through a structured schema. These failures are obscured when evaluation uses only aggregate fluency or an LLM judge.

We study this problem through TAIWO (Thematic Analysis for Interpretive Workflow Optimization), a typed inference graph inspired by thematic analysis (Braun & Clarke, 2006; 2019). The stages successively build a corpus summary, evidence-linked codes, candidate themes, review decisions, refined definitions, and a final synthesis. The workflow is provider-neutral: a provider receives messages and a model specification and returns text plus usage metadata. This design makes the same computational graph executable through a native benchmark proxy or an external API adapter.

Our empirical question is deliberately narrow: *when the same structured workflow is applied to the same controlled corpus, which aspects of semantic discovery remain reliable across a broad set of LLMs?* We answer it with a frozen native benchmark comprising 39 unique model runs. The corpus assigns latent concepts to dominant, medium, and rare prevalence strata, allowing stratified recall to be reported rather than collapsed into a frequency-weighted total.

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.

The audit changes the interpretation of the originally exported summary. The task defines malformed structured output as model behavior with score zero, but the canonicalizer omitted those zeroes because their score occupied a different JSON field. Restoring the task semantics changes the mean from the 36-run complete-case value 0.725 to the 39-run value 0.670. We also show that the term labeled grounding only checks whether a source identifier exists. It does not establish that the source supports a theme. We retain this component solely to reproduce the frozen score and refer to it as *source-ID validity* throughout.

This paper makes three contributions:

1. a candidate narrower workflow design with typed, inspectable stage artifacts and explicit model-output failure semantics, whose distinct value remains unproven;
2. a frozen 39-model diagnostic separating semantic, interface, and measurement reliability, including a 0.426 descriptive recall gap between assigned prevalence strata and a 7.7% structured-output failure rate; and
3. executable scorer attacks, a support-aware v2 metric, and two equal-call pilots that expose effect portability and realized-resource differences while preserving null and negative evidence.

We do not claim that TAIWO outperforms an equal-call baseline, that it is a general-purpose qualitative-analysis validator, or that the observed differences are caused by model architecture. We also do not treat the typed workflow itself as established novelty: the present experiments evaluate one pipeline on two controlled tasks without isolating its proposed design features. That scope is central to the contribution rather than deferred to a footnote.

2 Related Work

Structured test-time inference. Prompt decomposition provides intermediate representations that can constrain later inference. Chain-of-thought elicits intermediate reasoning (Wei et al., 2022), least-to-most prompting orders subproblems (Zhou et al., 2023), and decomposed prompting uses modular subtask handlers (Khot et al., 2023). Self-consistency instead samples multiple paths and aggregates them (Wang et al., 2023). TAIWO's candidate narrower distinction is its target and state: it operates on a set-valued discovery task and passes typed analytical artifacts, including codes, themes, review actions, definitions, and source identifiers, between stages. The current study does not isolate those features from decomposition or additional computation. Its workflow novelty and value therefore remain inconclusive; we report a small exploratory equal-call pilot and retain the larger confirmatory test as future work.

Open-ended evaluation. Lexical overlap metrics are fragile when many outputs can be valid (Novikova et al., 2017; Post, 2018). Learned and LLM-based evaluators provide greater semantic flexibility but introduce judge dependence and possible model-family bias (Liu et al., 2023; Zheng et al., 2023; Fu et al., 2023). Topic-model research likewise shows that apparently standard coherence measures can disagree with human judgments (Hoyle et al., 2021). Our controlled task avoids an LLM judge by generating a corpus from known concepts and scoring recovered themes against deterministic labels. This improves repeatability but creates a different risk: the scoring program itself can encode shortcuts. We therefore treat adversarial scorer checks as empirical tests, not merely implementation unit tests.

LLMs and thematic analysis. Prior work explores LLM support for qualitative and computational social-science analysis (Xiao et al., 2023; Dai et al., 2023; Ziems et al., 2024; Zhang et al., 2025). Recent systems already provide open-source codebook workflows, modular human and AI pipelines, and blinded comparisons with human analysts (Katz et al., 2026; Bai & Finkelstein, 2026; Hill et al., 2026). These results substantially narrow any workflow-level novelty claim for TAIWO. The proposed narrower distinction is measurement-oriented: typed stages expose failures, frozen task semantics preserve malformed outputs, and support-aware tests challenge the scorer. The present evidence does not establish that this combination is itself workflow-novel or beneficial. Reflexive thematic analysis remains an interpretive practice rather than fixed-label prediction (Braun & Clarke, 2019);

our benchmarks measure controlled concept discovery and category transfer, not human-equivalent interpretation. Appendix C summarizes the scope difference.

Fine-grained provenance. Recent provenance systems bind generated claims to evidence units and mechanically audit citations (Wei et al., 2026; Xu et al., 2026). Their findings reinforce the distinction between syntactically valid citations and supported claims. We apply that principle to set-valued semantic discovery: source-ID validity is retained only as a historical diagnostic, while v2 requires concept-specific source support.

Reproducible benchmarking. Machine-learning results depend on implementation, environment, randomization, and reporting choices (Pineau et al., 2021; Henderson et al., 2018; Dodge et al., 2019). Our artifact freezes run identifiers, the corpus hash, task-source hash, model panel, and canonical tables. The audit demonstrates that hashes are necessary but insufficient: a deterministic transformation can still omit scientifically meaningful failures.

3 Auditable Structured Inference

Let $D = \{(i_j, x_j)\}_{j=1}^{n}$ denote documents with immutable source identifiers. TAIWO is the experimental instrument and a candidate narrower design: a failure-bearing graph of six typed transformations. The familiar operations are not novel, and the distinct value of their typed composition is not established here,

$$z_s = f_s(D, z_{<s}; m, \tau, b, r), \qquad s \in \{1, \ldots, 6\}, \tag{1}$$

where m is the model route, τ is temperature, b is the per-call completion budget, and r is the seed. Each call requests a JSON object. A parser validates the returned fields before the artifact becomes input to the next stage.

Stage 1: familiarization. The model reads the tagged corpus and produces a concise summary. Document text is delimited and explicitly designated as data rather than instruction, reducing direct prompt-injection risk.

Stage 2: initial coding. The model emits fine-grained code objects with labels, descriptions, verbatim evidence, and source identifiers. This stage shifts the representation from raw text to evidence-linked observations.

Stage 3: theme identification. Codes are grouped into candidate themes. Each theme contains a name, definition, code labels, evidence, and source identifiers. The transformation is many-to-many: related codes may support one theme and a code may inform more than one candidate.

Stage 4: theme review. The model reviews candidates against the corpus for coherence, distinctiveness, and evidence alignment. It records an explicit REMAINED, MERGED, or SPLIT decision and a rationale for each reviewed theme.

Stage 5: definitions and naming. Reviewed themes receive refined names, definitions, relevance statements, and preserved evidence. This separates discovery from presentation and makes the final semantic objects directly inspectable.

Stage 6: synthesis. The final call produces a concise synthesis grounded in the theme set. The benchmark score uses the themes, not prose quality in this synthesis. Across successful frozen runs, the ledger records exactly six calls, confirming that every scored completion traversed the full pipeline.

The implementation separates the provider from the pipeline. A model response contains text, model and provider names, token usage, cost, latency, response identifier, finish reason, and provider metadata. The benchmark adapter invokes the native model proxy; the external adapter uses LiteLLM-compatible routes. This abstraction supports future OpenRouter replication, but provider rerouting is a new access condition and cannot be treated as identical to the frozen native-proxy runs.

4 CONTROLLED BENCHMARK

4.1 CORPUS AND PROTOCOL

The benchmark deterministically generates 100 documents with seed 2027. Nine latent concepts are sampled independently from fixed prevalences. Three dominant concepts have prevalences 0.30, 0.26, and 0.18; three medium concepts have prevalences 0.11, 0.08, and 0.055; and three rare concepts have prevalences 0.028, 0.018, and 0.012. Each active concept contributes one of three natural-language templates, and neutral distractors may be inserted. The model receives document IDs and text, but not the latent labels stored by the generator. Appendix B lists all concepts and anchors.

Every run uses the same corpus, temperature 0.2, maximum 3,000 completion tokens per call, and the full six-stage workflow. The frozen cohort contains 39 model routes from multiple providers and model families. It is a breadth snapshot rather than a random sample from a defined model population; inferential claims therefore concern these runs, not all LLMs. Thirty-one rows report package v0.6.3 and eight report v0.7.1. The recovered v0.7.1 and v0.7.2 scientific modules are byte-identical, but exact v0.6.3 source is unavailable. Model composition and package version are thus confounded.

4.2 HUMAN-LABELLED TRANSFER PROTOCOL

We separately evaluate transfer on Part B of the public Reddit Mental Health Dataset (RMHD) (Rani et al., 2024). Its authors assign each post one of four root-cause categories (personality factors, drug and alcohol, early life, or trauma and stress) and report Cohen's $\kappa = 0.869$. The published files contain 823 rows but 23 blank texts. After removing one conflicting duplicate text and two same-label exact duplicates, 796 unique label-consistent posts remain. A deterministic seed-2027 sample selects 25 posts per category. Labels are kept outside model-facing documents. We do not use TripAdvisor star ratings (Alam et al., 2016) as thematic ground truth because they do not annotate review themes.

The transfer pilot uses the same three routes and three seeds as the synthetic paired pilot, with six calls per condition and a 6,000-token per-call ceiling. Its primary metric is macro source-grounded category recall: discovered themes are assigned one-to-one to the four reference categories, and a matched category receives the fraction of its cited source IDs carrying that human label. A semantic threshold of 0.45 was frozen from 12 paper-derived positive paraphrases and eight out-of-taxonomy negative controls before the final v3 run; all 20 calibration cases were classified correctly. Holding the lexical threshold at 0.50, a fixed sensitivity sweep retains balanced accuracy 1.0 at embedding thresholds 0.35, 0.40, and 0.45; it falls to 0.875, 0.833, and 0.708 at 0.50, 0.55, and 0.60 as positive sensitivity declines while specificity remains 1.0. The controls were fixed independently of run outcomes but are small. This metric is specific to the curated four-category root-cause taxonomy, not an exhaustive thematic or clinical-validity score.

4.3 FROZEN OPERATIONAL SCORE

For each output theme t and latent concept c, the scorer combines lexical overlap with claimed-source support. The best concept for each theme is accepted above a fixed threshold. Let R_g be concept recall within rarity group g, P_T the number of distinct matched concepts divided by the number of output themes, and V_I the mean fraction of each theme's source IDs that exist in the corpus. The frozen score is

$$R_{\text{bal}} = \frac{1}{3} \sum_{g \in \{\text{dom,med,rare}\}} R_g, \tag{2}$$

$$S_{\text{v1}} = 0.60 R_{\text{bal}} + 0.25 P_T + 0.15 V_I. \tag{3}$$

Duplicate themes reduce P_T, and equal weighting of rarity strata prevents dominant concepts from numerically overwhelming rare concepts. However, V_I checks identifier validity, not semantic support. We preserve it only to reproduce the frozen task.

The task distinguishes infrastructure errors from model-output errors. Proxy failures are raised for platform handling. Other exceptions, including malformed structured outputs, record the exception and return $S_{\text{v1}} = 0$. This distinction is important because the exporter originally treated three model-output zeroes as missing.

Table 1: Frozen breadth results. Score summaries in the first row include all three task-declared model-output failures as zero. Component summaries require detailed theme outputs and therefore use the 36 successful structured completions.

Measure	n	Mean	SD	Median	Min	Max
Task score (all runs)	39	0.670	0.246	0.737	0.000	0.883
Task score (detailed)	36	0.725	0.155	0.757	0.190	0.883
Balanced recall	36	0.691	0.178	0.722	0.222	0.889
Theme precision	36	0.662	0.144	0.700	0.200	0.875
Valid source-ID rate	36	0.973	0.165	1.000	0.011	1.000

4.4 Provenance and audit procedure

Scientific freeze `KBF-001` records 39 run IDs, 39 model routes, one corpus hash, one server task version, the task-source hash, and a post-freeze policy that forbids changing task source, score, corpus, or prompts. Canonical CSV and Parquet tables agree on run IDs and scientific values to floating-point precision. We independently recomputed all reported summaries from the CSV.

We then performed executable adversarial checks. First, we passed an exact task-failure payload through the canonicalizer and observed a missing score, identifying the fallback-field defect. Second, we constructed perfect concept labels with (i) no source IDs and (ii) valid IDs selected from documents known not to contain the matched concept. Regression tests additionally cover fabricated IDs, mixed correct and irrelevant sources, label-only leakage, duplicated themes and evidence, and empty output. These probes test whether the score actually requires evidential support.

5 Results

5.1 Historical breadth performance and structured-output failures

Table 1 separates all-run task scores from metrics available only for the 36 detailed completions. With the three task-declared zeroes restored, the all-run mean is 0.670, median 0.737, and sample standard deviation 0.246. The full range is 0 to 0.883. The structured-output completion rate is $36/39 = 92.3\%$.

The distinction is not cosmetic. The complete-case mean exceeds the all-run mean by 0.0558, or 8.33% relative to the all-run value. The three failures are Claude Opus 5, DeepSeek R1 0528, and Qwen3 Next 80B Thinking, all among the eight rows labeled v0.7.1. The canonical table preserves the raw-file hashes but not the failure text because of the exporter defect. We therefore interpret them only as structured model-output failures, not as evidence that a particular internal stage or provider caused them.

5.2 Recovery differs across assigned prevalence strata

Among the 36 detailed completions, mean recall is 0.907, 0.685, and 0.481 for concepts assigned dominant, medium, and rare prevalences, respectively (Table 2). The dominant-to-rare descriptive gap is 0.426. No detailed run recovers all rare-assigned concepts: observed recall ranges from 0 to 0.667, while dominant-assigned recall reaches 1.0 for most runs. Realized support is not identical to nominal prevalence: the nine concepts occur in 33, 22, 19, 15, 4, 5, 2, 3, and 1 documents. Concept identity is fixed to stratum, so these data cannot distinguish a prevalence effect from differences in concept wording or semantics. Appendix Figure 1 shows the full distributions.

The result motivates, but does not identify, a frequency-dependent discovery bottleneck. A deconfounded experiment must rotate concept-to-prevalence assignments; the released generator now defines nine deterministic rotations for this future test.

Table 2: Recall by assigned prevalence stratum among 36 detailed runs. "Support" lists realized document counts for the three concepts in each stratum.

Stratum	Support	Mean	SD	Median	Min	Max
Dominant	33/22/19	0.907	0.171	1.000	0.333	1.000
Medium	15/4/5	0.685	0.264	0.667	0.000	1.000
Rare	2/3/1	0.481	0.232	0.667	0.000	0.667

5.3 Scorer audit

The first adversarial probe demonstrates aggregation bias. The task's model-behavior failure payload stores `score: 0.0` at the top level. The exporter searched for `leaderboard_score` and then nested `metrics.score`; both are absent in that payload, producing missing values. A regression-tested repair now falls back to the top-level score and retains failure type and message. We do not overwrite the frozen table; Table 1 applies the task's declared semantics as a transparent sensitivity correction.

The second probe demonstrates construct undercoverage. Exact labels and definitions for all nine concepts score 0.85 with no source IDs: balanced recall and precision can saturate without evidence. Adding arbitrary valid IDs raises the score to 1.0 even when each ID is selected from a document whose latent labels exclude the corresponding concept. Consequently, V_I is not an evidence-grounding metric. The high observed mean of 0.973 says that models usually emitted valid identifiers, not that citations support their themes.

5.4 Human-labelled RMHD transfer

All 18 final RMHD v3 runs completed and passed independent artifact validation. Mean source-grounded category recall was 0.289 for TAIWO and 0.413 for generic decomposition. The mean paired difference was -0.1233 (median -0.2083), with two-sided exact paired permutation $p = 0.3008$ and model-clustered bootstrap 95% CI $[-0.3540, 0.2241]$. Thus the point estimate favors generic decomposition, but the interval is wide and crosses zero. The three model-level mean differences were -0.2913 for Gemini 3.6 Flash, $+0.2356$ for Qwen3 Coder, and -0.3144 for GLM-5, showing that the aggregate is not a uniform model-family effect. Table 7 reports every pair. These results do not support a TAIWO advantage on the human-labelled transfer task; Appendix Figure 2 compares this heterogeneity with the synthetic paired pilot. They also do not establish that TAIWO's typed composition constitutes useful narrower workflow novelty.

5.5 Post hoc portability and realized-resource audit

The common call counts and completion ceilings did not equalize realized computation. Relative to generic decomposition, TAIWO used 24.4% fewer total tokens and cost 16.8% less on the synthetic task, then used 43.7% fewer tokens and cost 41.6% less on RMHD; latency was nevertheless 12.2% and 24.0% higher. Quality effects were sign-concordant in only three of nine matched model and seed pairs across tasks (Spearman $\rho = -0.45$). Component differences also changed interpretation: synthetic rare-stratum recall rose 0.148 while dominant and medium recall fell 0.037 and 0.074; on RMHD, category recall fell 0.167 while TAIWO emitted 1.67 more themes and 2.33 more out-of-taxonomy themes. These post hoc, nine-pair diagnostics show context compression and exploratory output expansion without stable task transfer, not a general efficiency or quality advantage. They leave the narrower workflow-novelty question unresolved.

6 Discussion

The evidence separates three reliability layers. *Semantic reliability* asks whether target content is recovered and retained. Recall differs sharply across the frozen assigned-prevalence strata, but concept identity and prevalence are confounded; the result is a diagnostic, not a causal frequency effect. The released prospective utility can measure concept survival and conditional attrition at

each stored stage, but the historical breadth artifact lacks those intermediates and cannot support a retrospective stage-loss claim.

Interface reliability is part of model performance in a staged system. A model that produces an unusable intermediate object prevents downstream inference. Treating such runs as missing estimates quality conditional on success and removes a practical failure mode. The 8.33% gap between complete-case and all-run means in this cohort shows that failure semantics should be specified before aggregation.

Measurement reliability is distinct from both. Determinism and provenance do not imply construct validity. V_I is exactly reproducible, but it rewards syntactically valid citations whether or not they support a theme. A corrected score should match each claimed source against the latent concept associated with the theme, penalize unsupported evidence, and receive a new metric version. Historical v1 scores remain immutable and are interpreted only as diagnostic evidence.

We additionally ran an exploratory route-transfer pilot on one frozen representative from each breadth stratum and seeds 2027 to 2029. Full TAIWO and generic decomposition each received six calls with the same 6,000-token per-call ceiling, corpus, route, and temperature. After excluding earlier transport-development attempts, the frozen v3 run completed all 18 conditions and nine pairs without failure. Mean v2 score was 0.610 for TAIWO and 0.603 for generic decomposition; the mean paired difference was 0.0068 (median 0.0548), with two-sided exact paired permutation $p = 0.859$ and a model-clustered bootstrap 95% CI of $[-0.0845, 0.0908]$. Thus this small pilot provides no evidence that TAIWO's named stages outperform generic iterative computation. The next decisive experiment remains the prespecified 36-run comparison on all six frozen models below; the pilot is not confirmatory because it has only three model clusters and its output-transport protocol was revised after failed infrastructure trials. Accordingly, the typed artifact flow is a candidate design contribution whose novelty and utility remain inconclusive, not a validated workflow advance.

7 Minimum Confirmatory Test

The breadth study cannot determine whether TAIWO's named stages add value beyond six generic calls, so narrower workflow novelty remains inconclusive. The minimum confirmatory design therefore pairs full TAIWO with generic six-call decomposition for six frozen models and three seeds, yielding 18 model by seed pairs and 36 runs. Within each pair, condition order is randomized while corpus, route, temperature, call count, and aggregate completion-token budget remain fixed; malformed analytical outputs score zero and infrastructure failures remain explicit missing data.

The primary outcome is the support-aware v2 score frozen before execution, and the estimand is the mean within-pair difference. Evidence for an advantage requires a positive mean whose model-clustered 95% interval excludes zero, without reversal from excluding failures or a material budget violation. Every pair, a paired permutation test, and secondary component outcomes must also be reported. Appendix J gives the full operational specification.

That paired test does not deconfound prevalence. A separate prospective design rotates all nine concepts through all nine prevalence profiles and records every stage. It can estimate stage survival and conditional attrition by assigned stratum. Stage necessity should be tested by resource-audited representation substitution rather than deletion, so removing a named operation does not also remove a model call. Equal call counts and token ceilings must be reported separately from realized token use, cost, and latency.

8 Implications for Benchmark Design

The audit yields four requirements. First, evaluation contracts must distinguish model-output failures from infrastructure failures before outcomes are observed and must not silently discard either. Second, discovery benchmarks should stratify targets by prespecified prevalence and counterbalance concept identity so aggregate recall cannot hide minority patterns or confound frequency with semantics. Third, provenance must follow raw responses through parsing, scoring, and canonical tables; hashes at only the endpoints did not prevent the observed cohort error. Fourth, scorer validation needs adversarial controls for absent, incorrect, duplicated, and fabricated evidence. These requirements turn reproducibility from file preservation into an executable test of the measurement claim.

9 LIMITATIONS

The corpus is synthetic, small, and generated from nine known concepts. It enables deterministic measurement but cannot represent the ambiguity, positionality, or context dependence of human qualitative inquiry. Template phrases may also make lexical recovery easier than naturalistic discovery. Results should not be generalized to human-equivalent thematic analysis.

The cohort is a convenience breadth snapshot rather than a probability sample of models. It mixes package versions and changing, potentially correlated provider routes. Exact v0.7.1 and v0.7.2 scientific code is available and identical, but exact v0.6.3 source is not. Model family, route, date, and package version cannot be disentangled, and one run per model is inadequate for model-specific uncertainty.

Concept identity is perfectly confounded with assigned prevalence, and nominal probabilities differ materially from realized supports in this 100-document sample. The breadth artifact also lacks raw stage outputs and all path-referenced metric files. Consequently, we cannot rescore the 39 historical runs under v2, reconstruct where concepts disappeared, or estimate stage-wise attrition. Counterbalancing and stage survival are prospective protocols, not completed results.

The attached provenance bundle omits the 39 raw metric files referenced by its canonical rows and omits the partial rerun tables. Source-file hashes therefore cannot be checked against the original bytes in this artifact. The canonical CSV and Parquet agree, and the task source matches its frozen hash, but the chain of custody remains incomplete.

Finally, the v1 score uses lexical anchors and source support defined from generator truth. Participants with task-source access could construct target labels directly. The benchmark is useful as a controlled execution stress test, not as a secure hidden test or a general quality metric.

The RMHD extension is also narrow. Its four mutually exclusive labels represent a curated root-cause taxonomy rather than exhaustive themes or diagnoses. The balanced 100-post sample omits much of the corpus, and threshold calibration uses a small fixed set of paper-derived paraphrases and negative controls. With only three model clusters, the permutation test and clustered interval are exploratory. Qwen also showed unstable structured-output compliance during excluded feasibility attempts, requiring a frozen forced-tool transport and auditable condition-level retries before the final run.

Neither paired pilot isolates typed state, explicit failure handling, provenance contracts, or any individual stage. Generic decomposition is therefore an equal-call comparator, not an ablation of the proposed narrower design. These data cannot decide whether that design is novel in effect, necessary, or beneficial.

10 CONCLUSION

TAIWO makes a structured inference graph inspectable, but its narrower workflow novelty remains inconclusive: the experiments neither isolate its proposed design features nor show that its named stages outperform generic iteration. Across 39 frozen historical runs, the graph exhibits a 7.7% interface-failure rate and weaker recovery in the concepts assigned low prevalence. Auditing changes the benchmark's own headline: model-output zeroes must remain zeroes, and valid source identifiers do not establish evidential support. Evaluation of structured open-ended inference must therefore test semantic retention, interface contracts, and measurement validity separately. Both paired pilots remain inconclusive; the strongest demonstrated contribution is the auditable evaluation framework and its explicit record of what the available artifacts cannot establish, not a validated workflow innovation.

ETHICS STATEMENT

The breadth benchmark uses procedurally generated synthetic text. The transfer pilot uses publicly released, human-authored Reddit mental-health posts under the dataset's CC0-1.0 license; public availability does not remove their sensitivity. The analysis transmitted the deterministic 100-post sample to the named model provider, retained raw response journals privately for audit, and redistributes neither posts nor response journals. Model-facing records exclude human labels and direct account identifiers. We make no diagnosis, clinical-validity claim, or claim that automated themes replace expert interpretation.

REPRODUCIBILITY STATEMENT

Appendix A specifies the task, Appendix E lists all 39 scores, and Appendix F records hashes and omissions. The anonymized supplement provides the scientific code, configuration, canonical tables, an account-redacted task copy, and 51 passing scientific, provider, audit, and provenance tests. The RMHD extension additionally records four source-file hashes, deterministic exclusion counts, corpus and truth hashes, numeric summaries for all 18 runs, nine paired outcomes, route metadata, a threshold-sensitivity artifact, and an independently recomputed audit. A deterministic aggregate records the cross-task and realized-resource synthesis. Sensitive posts, raw run objects containing quotations, and private response journals are excluded.

AI USE STATEMENT

Generative AI tools were used to support codebase review, language editing, structural review, consistency checking, and reproducibility documentation. They were not used to create the frozen synthetic corpus or generate the historical benchmark outputs reported here. All numerical claims were recomputed with deterministic scripts or traced to persisted notebook output as explicitly labeled. The authors independently executed the reported tests, checked cited sources, verified the mathematical arguments, numerical results, references, code, and final wording, and take full responsibility for the content, claims, and artifacts of the manuscript.

REFERENCES

Md. Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Information Sciences*, 339:206 to 223, 2016. doi: 10.1016/j.ins.2016.01.013. URL https://doi.org/10.1016/j.ins.2016.01.013.

Yunbing Bai and Joseph Finkelstein. Multimodule human and artificial intelligence collaboration pipeline for large language model assisted thematic analysis across digital health interview studies: Comparative evaluation study. *JMIR Medical Informatics*, 14:e96129, 2026. doi: 10.2196/96129. URL https://doi.org/10.2196/96129.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77 to 101, 2006. URL https://doi.org/10.1191/1478088706qp063oa.

Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589 to 597, 2019. URL https://doi.org/10.1080/2159676X.2019.1628806.

Zhe Dai, Lingzhen Chen, Peter Reif, and Matthias Grabmair. Can large language models perform thematic analysis? *arXiv preprint arXiv:2309.04558*, 2023. URL https://arxiv.org/abs/2309.04558.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2185 to 2194, 2019. URL https://aclanthology.org/D19-1224.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023. URL https://arxiv.org/abs/2302.04166.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1709.06560.

Callum Hill, Arun Dahil, Glenn Simpson, David Hardisty, Jacob Keast, Cameron Kumar Pinn, and Hajira Dambha-Miller. Large language models for thematic analysis in healthcare research: A blinded mixed-methods comparison with human analysts. *PLOS Digital Health*, 5(4):e0001189, 2026. doi: 10.1371/journal.pdig.0001189. URL https://doi.org/10.1371/journal.pdig.0001189.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2018 to 2033, 2021. URL https://arxiv.org/abs/2107.02173.

Andrew Katz, Gabriella Coloyan Fleming, and Joyce B. Main. Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. *Humanities and Social Sciences Communications*, 13(1):209, 2026. doi: 10.1057/s41599-026-06508-5. URL https://doi.org/10.1057/s41599-026-06508-5.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02406.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511 to 2522, 2023. URL https://arxiv.org/abs/2303.16634.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241 to 2252, 2017. URL https://aclanthology.org/D17-1238.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164):1 to 20, 2021. URL https://www.jmlr.org/papers/v22/20-303.html.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186 to 191, 2018. URL https://aclanthology.org/W18-6319.

Saima Rani, Khandakar Ahmed, and Sudha Subramani. From posts to knowledge: Exploring mental health discourse on Reddit. *Applied Sciences*, 14(4):1547, 2024. doi: 10.3390/app14041547. URL https://doi.org/10.3390/app14041547.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824 to 24837, 2022. URL https://arxiv.org/abs/2201.11903.

Jingxuan Wei, Xingyue Wang, Yanghaoyu Liao, Jie Dong, Yuchen Liu, Caijun Jia, Bihui Yu, and Junnan Zhu. GenProve: Learning to generate text with fine-grained provenance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5027 to 5048, 2026. doi: 10.18653/v1/2026.acl-long.228. URL https://aclanthology.org/2026.acl-long.228/.

Chacha Xiao, Shi Feng, John Bertenthal, Karla Badillo-Urquiola, Michael Wornow, Karthik Prasad, Zachary Talbot, Niloufar Salehi, Gina Neff, Max Heron, Aditya Kandala, and Katherine Tsui. Supporting qualitative analysis with large language models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 75 to 78, 2023. URL https://dl.acm.org/doi/10.1145/3581754.3584136.

Ruoyu Xu, Gaoxiang Li, and Victor S. Sheng. GAVEL: Evidence-contract debate with mechanized scrutiny for provenance-grounded fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35907 to 35920, 2026. doi: 10.18653/v1/2026.findings-acl.1789. URL https://aclanthology.org/2026.findings-acl.1789/.

He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M. Carroll. Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans*, 4:100144, 2025. doi: 10.1016/j.chbah.2025.100144. URL https://doi.org/10.1016/j.chbah.2025.100144.

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. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.10625.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237 to 291, 2024. URL https://arxiv.org/abs/2305.03514.

A Detailed Protocol

Table 3: Frozen task configuration.

Field	Value
Task	`taiwo-structured-discovery`
Scientific freeze	`KBF-001`
Server task version	1
Corpus size	100 documents
Corpus seed	2027
Temperature	0.2
Calls per completed run	6
Maximum completion tokens per call	3,000
Model runs	39
Detailed structured completions	36
Task-declared model-output failures	3

All prompts include a system boundary declaring tagged documents to be data rather than instructions and requiring valid JSON. Stage schemas are implemented as typed objects. Code objects contain label, description, evidence, and source IDs. Theme objects contain name, definition, code labels, evidence, source IDs, relevance, and provenance. Review actions are restricted to REMAINED, MERGED, or SPLIT.

The corpus digest is computed over newline-joined `doc_id`, tab, and document text records. Metadata containing latent labels is not included in prompt rendering. The same digest appears in every canonical breadth row.

B Synthetic Corpus

Table 4: Latent concepts, nominal generation prevalence, and realized document support in the frozen 100-document corpus.

Stratum	Concept	Nominal p	Realized n
Dominant	access friction	0.300	33
Dominant	support quality	0.260	22
Dominant	time pressure	0.180	19
Medium	trust uncertainty	0.110	15
Medium	adaptation benefit	0.080	4
Medium	coordination breakdown	0.055	5
Rare	privacy concern	0.028	2
Rare	rare recovery path	0.018	3
Rare	contradictory success signal	0.012	1

For each document, each concept is sampled independently at its prevalence. An active concept contributes one uniformly selected template from a set of three. Documents with no active concept receive a neutral distractor; documents with active concepts receive a distractor with probability 0.22. Pieces are shuffled before rendering.

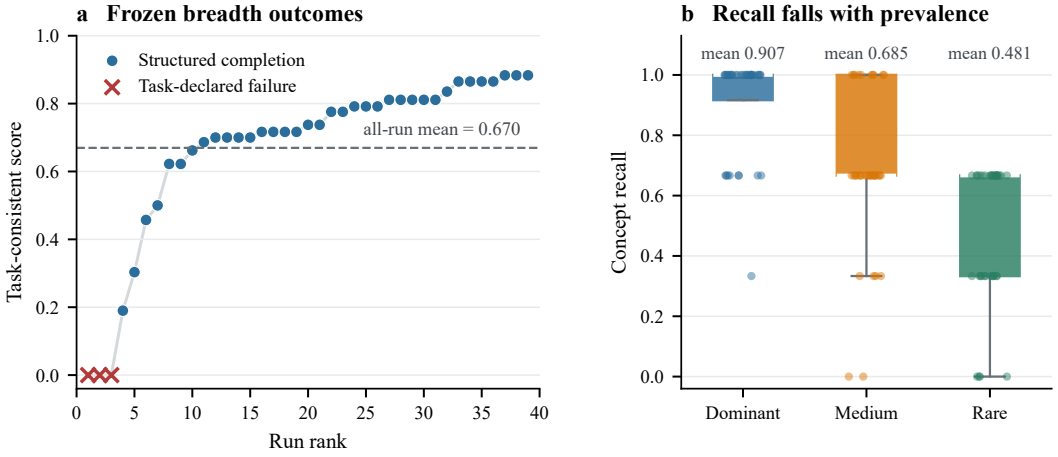

Figure 1: Breadth-study diagnostics. **a**, task-consistent scores for all 39 runs, including the three task-declared structured-output failures at zero; the dashed line is the all-run mean. **b**, dominant-, medium-, and rare-concept recall for the 36 detailed completions. Boxes show quartiles and medians; points are runs. The visual separation describes the assigned strata but does not isolate prevalence from concept identity or imply a population-level model comparison.

C Scope Relative to Recent Work

Work	Setting	Reference or evidence check	Primary distinction
GATOS (Katz et al., 2026)	Open-source inductive codebooks	Synthetic case studies	Workflow for scalable qualitative analysis
Bai and Finkelstein (Bai & Finkelstein, 2026)	Three digital-health interview studies	Human-theme embedding alignment	Modular human and AI workflow comparison
Hill et al. (Hill et al., 2026)	Health focus-group transcript	Expert consensus and manual quote verification	Blinded human/model comparison
GenProve; GAVEL (Wei et al., 2026; Xu et al., 2026)	Generation and fact-checking	Fine-grained or mechanized provenance	Evidence binding beyond citation validity
This work	Synthetic discovery plus RMHD category transfer	Latent labels, source support, and attack tests	Proposed joint audit; narrower workflow novelty inconclusive

D Scoring and Adversarial Audit

Theme text concatenates name, definition, and relevance. After stop-word removal, the lexical term for concept c is the larger of anchor coverage and Jaccard similarity between theme and concept tokens. Source support is the fraction of valid claimed documents whose latent labels contain c. Candidate match strength is

$$M(t, c) = 0.65L(t, c) + 0.35E(t, c). \tag{4}$$

A match requires $M(t, c) \geq 0.45$ and either $L(t, c) \geq 0.20$ or $E(t, c) \geq 0.50$. Only the best concept per theme is retained. Distinct matched concepts determine rarity recall and theme precision.

The final $0.15V_I$ term is separate from $E(t, c)$. It checks only whether IDs are members of the corpus ID set. The support-aware v2 metric replaces that term with concept-specific source support but cannot be retroactively applied to missing breadth outputs. Representative executable probes produced:

Probe	v1 score	v2 score	v2 support
Exact concepts, correct IDs	1.000	1.000	1.000
Exact concepts, no IDs	0.850	0.850	0.000
Exact concepts, fabricated IDs	0.850	0.850	0.000
Exact concepts, wrong valid IDs	1.000	0.850	0.000

This is why the main text treats v1 as historical and does not interpret V_I as support. Additional tests reject bare label leakage and cover mixed support, duplicate spam, duplicated fabricated evidence, and empty output.

E ALL FROZEN MODEL RUNS

Table 5: All frozen runs, ordered by task-consistent score. "Detail" indicates whether component metrics are present.

Model route	Score	Dominant	Rare	Detail
google/gemini-3.6-flash	0.883	1.000	0.667	yes
google/gemini-3.7-flash	0.883	1.000	0.667	yes
openai/gpt-5.4-mini-2026-03-17	0.883	1.000	0.667	yes
anthropic/claude-haiku-4-5@20251001	0.865	1.000	0.667	yes
google/gemma-4-31b	0.865	1.000	0.667	yes
openai/gpt-5.5-2026-04-23	0.865	1.000	0.667	yes
openai/gpt-5.6-luna	0.865	1.000	0.667	yes
anthropic/claude-sonnet-4-6@default	0.835	1.000	0.667	yes
anthropic/claude-opus-4-6@default	0.811	1.000	0.333	yes
anthropic/claude-opus-4-8@default	0.811	1.000	0.667	yes
anthropic/claude-sonnet-4@20250514	0.811	1.000	0.667	yes
openai/gpt-5.4-2026-03-05	0.811	1.000	0.667	yes
zai/glm-5	0.811	1.000	0.667	yes
anthropic/claude-opus-4-5@20251101	0.792	1.000	0.667	yes
google/gemini-2.5-flash	0.792	1.000	0.333	yes
google/gemini-3.8-flash	0.792	1.000	0.667	yes
anthropic/claude-opus-4-7@default	0.776	1.000	0.667	yes
openai/gpt-5.6-terra	0.776	1.000	0.667	yes
anthropic/claude-sonnet-5@default	0.737	0.667	0.667	yes
xai/grok-4.20-0309-reasoning	0.737	1.000	0.333	yes
anthropic/claude-sonnet-4-5@20250929	0.717	1.000	0.333	yes
google/gemini-2.5-pro	0.717	0.667	0.333	yes
google/gemini-3-flash-preview	0.717	0.667	0.667	yes
xai/grok-4.20-0309-non-reasoning	0.717	1.000	0.333	yes
google/gemini-3.5-flash	0.700	1.000	0.333	yes
google/gemini-3.5-flash-lite	0.700	1.000	0.333	yes
google/gemma-4-26b-a4b	0.700	1.000	0.333	yes
openai/gpt-5.4-nano-2026-03-17	0.700	1.000	0.333	yes
google/gemini-3.1-pro-preview	0.686	1.000	0.667	yes
anthropic/claude-opus-4-1@20250805	0.662	0.667	0.333	yes
openai/gpt-oss-120b	0.622	1.000	0.000	yes
qwen/qwen3-coder-480b-a35b-instruct	0.622	0.667	0.667	yes
google/gemini-3.1-flash-lite-preview	0.500	0.667	0.000	yes
qwen/qwen3-235b-a22b-instruct-2507	0.457	0.667	0.000	yes
qwen/qwen3-next-80b-a3b-instruct	0.303	0.667	0.000	yes
openai/gpt-oss-20b	0.190	0.333	0.333	yes
anthropic/claude-opus-5@default	0.000	NA	NA	no
deepseek-ai/deepseek-r1-0528	0.000	NA	NA	no
qwen/qwen3-next-80b-a3b-thinking	0.000	NA	NA	no

F PROVENANCE RECORD

Scientific freeze. `KBF-001`, frozen 2026-09-19T04:33:43Z.

Corpus SHA-256. `c11e68ea03e0ef2f7bbc3595708ee2e9c8d49235a695d2c5e9beb8bc79e96ad3`.

Task SHA-256. `06dde698c77d8dd14208ab1585317458 5d3ae1aae026d58a7d2d66cae7ddbb84`.

Canonical CSV SHA-256. `d00fc949d16dfb3b8f61ca3536e0ae56 91943e3a9f4e27cbadd291c2561efde2`.

Canonical Parquet SHA-256. `f4047e90a8ebbc8bd8b416c71785cac0 088ef82ae048893b37a525aa6d94dc62`.

Manifest file SHA-256. `9b456f949c32d9f41295779c47a285ea 6b8fd4bee7f3e183f2d752248ec3d3fa`.

Manifest self-hash. `373666982a28c4d44649d64ce7663213 57feeb1203620a613ff21f627b906a36`.

The attached bundle lacks all 39 path-referenced raw metric files and the partial rerun run/pair/report files. These omissions prevent source-byte verification and independent recomputation of the paired rerun summary. They are disclosed rather than reconstructed from unverified values.

G Exploratory Paired Pilot

The validated v3 route-transfer pilot uses `google/gemini-3.6-flash`, `zai/glm-5`, and `qwen/qwen3-coder-480b-a35b-instruct`. Provider-specific structured-output transport was frozen before v3: forced schema-bound tools for Gemini and GLM, and native JSON schema for Qwen. Both conditions retained the same six-call and 6,000-token per-call ceilings. Table 6 reports every pair; positive values favor TAIWO.

Table 6: All exploratory v3 paired outcomes under support-aware score v2.

Model	Seed	TAIWO	Generic	Difference
Gemini 3.6 Flash	2027	0.6611	0.5283	0.1328
Gemini 3.6 Flash	2028	0.7000	0.6473	0.0527
Gemini 3.6 Flash	2029	0.7367	0.6474	0.0893
Qwen3 Coder	2027	0.5433	0.6250	-0.0817
Qwen3 Coder	2028	0.5533	0.4960	0.0574
Qwen3 Coder	2029	0.5615	0.6441	-0.0826
GLM-5	2027	0.6375	0.5689	0.0686
GLM-5	2028	0.5500	0.4952	0.0548
GLM-5	2029	0.5424	0.7726	-0.2302

The 108 experimental calls used 400,094 prompt tokens and 136,899 completion tokens, cost $0.509719, and accumulated 1,323.8 seconds of provider latency. The artifact contains 18 raw result files, 18 score rows, nine pair rows, and 111 hashed response journals (108 experimental plus three excluded preflight calls). Both validators pass; the manifest self-hash is `0a62c1c5521cf5164b8277c2aecd710135210d15f722eb4bf3bbd20307d90dd0`. All earlier incomplete attempts remain excluded.

H Human-Labelled RMHD Transfer Pilot

The final RMHD v3 artifact contains 18 validated runs and nine complete pairs under metric v2. Positive differences favor TAIWO. The 108 included experimental calls used 2,964,255 prompt tokens and 183,781 completion tokens and cost $1.900087. The private audit verifies 129 response journals: 111 included responses (108 experimental and three route preflights) and 18 excluded feasibility or superseded responses. Its manifest self-hash is `ade97f63cc4c787e5527c72cac2b07dc54da0646c1f804245be3ba60e1bf6e2f`.

Table 7: All exploratory RMHD v3 paired outcomes under the human-labelled category-transfer metric.

Model	Seed	TAIWO	Generic	Difference
Gemini 3.6 Flash	2027	0.4167	0.6250	-0.2083
Gemini 3.6 Flash	2028	0.3333	0.6000	-0.2667
Gemini 3.6 Flash	2029	0.1667	0.5655	-0.3988
Qwen3 Coder	2027	0.5208	0.1250	0.3958
Qwen3 Coder	2028	0.5833	0.0833	0.5000
Qwen3 Coder	2029	0.0000	0.1889	-0.1889

Model	Seed	TAIWO	Generic	Difference
GLM-5	2027	0.2500	0.7393	-0.4893
GLM-5	2028	0.1250	0.3785	-0.2535
GLM-5	2029	0.2083	0.4087	-0.2004

The observed human labels are never included in prompts. Exact duplicates and every text hash associated with conflicting labels are excluded before sampling. The final point estimate is negative and its interval includes zero; this pilot therefore cannot support a superiority claim for TAIWO.

I Post Hoc Artifact Synthesis

Table 8 audits realized resources rather than inferring compute equivalence from equal call counts and completion ceilings. Table 9 decomposes the paired outcomes. All values are TAIWO minus generic decomposition or condition means over the same nine pairs. The released JSON records full-precision results and the descriptive cross-task correlation.

Table 8: Outcome and realized-resource synthesis. T/G gives TAIWO and generic means.

Task	Δ outcome	Tokens T/G	Δ tokens	Δ cost	Δ latency
Synthetic	0.0068	25,687/33,979	-24.4%	-16.8%	+12.2%
RMHD	-0.1233	125,966/223,816	-43.7%	-41.6%	+24.0%

Table 9: Selected mean paired component differences.

Task	Component	Difference
Synthetic	Dominant / medium / rare recall	-0.037 / -0.074 / +0.148
Synthetic	Theme count / precision / source support	-0.44 / -0.002 / -0.0004
RMHD	Category / quote-grounded category recall	-0.167 / +0.042
RMHD	Theme count / out-of-taxonomy count	+1.67 / +2.33

Only three of nine pairwise effects have the same sign across tasks; only one of three model-level mean effects does. Pearson and Spearman correlations are -0.190 and -0.45, respectively. These post hoc statistics describe two fixed tasks, nine pairs, and three model clusters. They do not estimate cross-task generalization in a model population or establish the proposed typed workflow as a validated novelty.

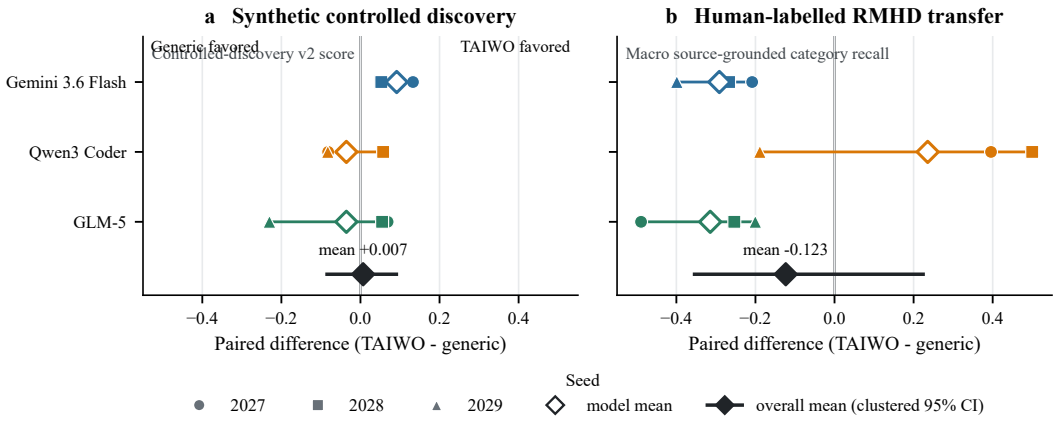

Figure 2: Paired TAIWO-minus-generic effects in the two exploratory pilots. Points show seeds, colored lines show each model's observed seed range, and open diamonds show model means. Black diamonds and lines show overall means and model-clustered bootstrap 95% intervals. The metrics differ by panel and are named inside each panel. Both intervals cross zero, with only three model clusters.

J MINIMUM CONFIRMATORY EXPERIMENT

The preregistered unit should be a model by seed pair. For each of six frozen panel models and three seeds, execute full TAIWO and a generic six-call decomposition on the identical corpus. Randomize within-pair order and hold the provider route, temperature, maximum calls, and aggregate completion-token ceiling fixed. Score all outputs, including malformed outputs as zero, with a support-aware v2 metric fixed before any run. Report paired differences, a two-sided paired permutation test, a model-clustered bootstrap 95% interval, all individual outcomes, and all exclusions. No early stopping or post-hoc model replacement is permitted.

The separate prevalence-identification design uses the released nine rotations, each of which assigns every concept to a different nominal prevalence profile. For stored stage artifact s, let $A_{r,c,s} = 1$ when any stage item semantically matches reference concept c in run r. For stratum g, prospective survival and conditional attrition are

$$S_g(s) = \frac{1}{|R||C_g|} \sum_{r \in R} \sum_{c \in C_g} A_{r,c,s}, \tag{5}$$

$$H_g(s) = \frac{\sum_{r,c \in C_g} A_{r,c,s-1}(1 - A_{r,c,s})}{\sum_{r,c \in C_g} A_{r,c,s-1}}, \tag{6}$$

with $H_g(s)$ undefined when no concept in g is present at the preceding stage. The released utility evaluates this definition on future stored artifacts; it does not reconstruct absent historical stages. Tests of stage contribution replace a stage's representation with a resource-audited alternative while preserving call and token ceilings and reporting realized use.