What does a donor name add after the relation is supplied?

An auditable component study of creative transfer in Wildcard

christopher robin fiore
20 September 2026
Research manuscript. AI assistance and author roles are disclosed below.

Result and analysis status: In the completed amended panel, adding the correct donor name changed mean QNM@4 by **−0.015625** relative to relation only, with 95% interval **[−0.109375, 0.0625]** and two-sided p = **1.000**. The result is **inconclusive**, not evidence of equivalence. All 64 amended blocks validated. The original preregistered primary remains **halted** because one original judge block failed; the amendment and complete amended results retain that distinction. The same generated sample was evaluated twice, without pooling panels.

Abstract

Wildcard supplies outside cues to a language model attempting a design or engineering task. Earlier bundled-prompt studies leave a component question unresolved: does naming a donor domain add qualified mechanisms after its relation and transfer boundary are supplied? We acquired responses for four conditions on 32 synthetic briefs: strong direct prompting, relation only, correctly named relation, and a mismatched-name control. The outcome counts model-judged qualified mechanism groups absent from a separately acquired finite direct-baseline bank. One invalid block in the original 64-block judge panel halted its primary analysis. A declared amendment acquired one complete panel with constrained reference identities, retaining the generated sample, prompts, rubric, and statistical contrast. All amended blocks validated. The primary correctly named versus relation-only difference was −0.015625 QNM@4 (95% interval [−0.109375, 0.0625], paired mean sign-flip p = 1.000), meeting the prespecified rule for an inconclusive result. Neither secondary contrast met its Holm-corrected test threshold. All arms had sparse bank-relative newness despite high qualified diversity. Separate operational diagnostics passed six of eight pairs under their strict contract. We also report new historical audits: a duplicate and missing Study 2 judgment, exhaustive finite-bootstrap sensitivity for Study 3, and dependence in the legacy sampler and shuffle. The contribution is an inspectable prompt-component study and evidence interface, with an explicitly amended measurement record. It neither validates the full deployed skill nor establishes historical originality, human creativity, or a first-ever prompting method.

1. The problem: an interesting reference is not an implemented idea

A distant reference can change a brainstorm while adding little that can be used. A proposed interface feature might acquire an ecological metaphor without changing its behavior. Conversely, a familiar control principle may produce a useful implementation even when the source name adds nothing. Evaluating whether a response sounds unusual does not separate these cases.

Wildcard's practical premise is to supply a cue from outside the immediate task and ask the model to turn any useful relationship into a concrete target-domain action. This is inference-time context conditioning. It does not modify weights, retrieve identified training examples, inspect activations, or reproduce a human default mode network. The historical skill bundles external selection, source elaboration, mapping, output constraints, and

permission to abstain. A comparison of that bundle with a plain prompt cannot attribute an effect to one component.

The present question is narrower: **when an authored relation and its limitations are held fixed, what does its explicit donor name contribute to qualified candidate mechanisms beyond a finite direct-baseline bank?** The name might activate relevant background associations, add decorative vocabulary, encourage an inappropriate transfer, or have little effect once the relation is explicit. Each possibility is compatible with the general usefulness of analogy. The intervention measures behavior without identifying an internal cognitive process.

The study uses sixteen curated relation cards and thirty-two authored briefs. It does not sample the deployed plugin's full corpus of 378 specialists and 461 concepts, execute its complete skill, or compare sampler versions. Its results cannot validate the revised full skill or the creative benefit of the live sampler. Runtime corpus.

2. Related work and contribution boundary

The emphasis on relations rather than superficial attributes follows a long tradition, including Gentner's structure-mapping account. Analogical prompting already uses model-generated exemplars and relevant knowledge to support reasoning. Component ablations and contrastive distractors also appear in recent work on analogical rule induction. Neither relational abstraction nor counterfactual input testing originates here.

Outside cues and diversity controls are direct prior art. Directed Diversity uses embedding-based prompt selection for human ideation, and the Associative Creativity Sparker recommends words in relation to prior ideas and a target. Creative Thought Embeddings discusses associative generation and filtering, including prompt-based and proposed model-integrated approaches. This repository implements no model-integrated adapter.

The closest recent diversity interventions further limit an originality claim. Agrawal and Goyal inject random concepts in list-generation experiments. Luo and colleagues combine random priming and sentence-level diversions for sustained generation. Zhang, Xin, and Zhong study where diversity is injected and how it is transmitted through generation. Supplying an external relation is therefore not, by itself, a new contribution.

Evaluation also has relevant predecessors. NoveltyBench evaluates multiple distinct, high-quality responses; Verbalized Sampling provides another training-free diversity method. We do not compare against either and cannot claim superiority. Lu and colleagues show why creativity metrics and superficial changes require careful interpretation. Our rubric addresses a concrete target-action construct, but model judgments remain fallible.

The scoped contribution combines an explicit-name intervention with held-fixed source-backed relations, an independently acquired finite reference bank, qualification before mechanism counting, counterfactual operational diagnostics, and acquisition provenance. A targeted review through 20 September 2026 found no identical study, but it is not an exhaustive priority search. We make no first-ever claim. The related-work map distinguishes established methods from this study's question.

3. What the historical evidence supports

The new study is motivated by earlier observations, not a reproduction of their treatment. A new audit preserves all 478 historical research and analyzer files byte for byte against commit 04ff0a5. The three original analyzers reproduce their committed results. Each study has ten problem units and three responses per arm per problem; grading rows are repeated measurements rather than independent problems. The audit report, claim ledger, and history manifest document the checks.

3.1 Earlier improvements concerned different outcomes

Study 1 compared self-selected cues, externally selected cues, and a plain prompt. Its selection-entropy difference concerns canonicalized cue labels and externally sampled corpus labels. It does not measure useful output diversity or prove that the model could never reach a cue. The externally selected arm received lower mean structural-genuineness ratings than the self-selected arm, 5.267 versus 5.700, with the original signed-rank p = .0078125. This unfavorable result is part of the record. The small human anchor contains fifteen selected outputs; human/panel correlations for genuineness and novelty are each approximately .01. The pooled cross-dimension correlation does not validate either measure individually. Original Study 1 results.

Study 2 compared revised and earlier skill prompts. The recorded differences favored the revision for genuineness (+0.808889) and usefulness (+0.843333), while novelty decreased (−0.369444). These are changes in model ratings for a bundled revision, not proof that one instruction caused the change or that revised responses were historically original. Original Study 2 results.

3.2 A new Study 2 integrity finding

Study 2 stores 240 grading rows but only 239 unique grader/output pairs. The pair `g3/out-35` occurs twice identically; `g3/out-48` is absent. The manifest/idmap identifies these outputs as `v2-p26-r1` and `v2-p14-r3`. Consequently, the historical analyzer averages five rows for one output and three for another. A row-count check alone concealed this defect. There is no evidence that the duplicate belongs to the missing output, so the original data remain untouched. Historical grades and idmap.

A separate audit sensitivity removes the duplicated pair once and leaves the missing pair absent. The resulting genuineness difference is +0.805556 and the conditional signed-rank p remains .001953125. Considering all seven hypothetical scale values for the missing judgment gives differences from +0.766667 to +0.816667. The original direction is stable in these checks, but matrix completeness is false. Hypothetical values are sensitivity bounds, not recovered observations. New Study 2 audit.

3.3 Study 3: a bounded novelty result

Study 3 compares the recorded full Wildcard configuration with its specified plain-brainstorm comparator. Its primary novelty mean difference is +0.725 on a seven-point model-rating scale. The original paired bootstrap 95% interval is [0.216667, 1.208333], with signed-rank p = .037109375. The preregistered decision required a point estimate of at least +0.30 and an interval excluding zero. It was met. The lower confidence bound does not exceed +0.30, so this does not establish a minimum gain of that size with 95% confidence. Original preregistration and results.

The recorded mean usefulness difference is −0.233333, with a negative percentile interval [−0.447222, −0.025000] and signed-rank p = .140625. These methods concern different statistics and are not inversions of one another. A nonsignificant rank test does not establish equivalent usefulness. Genuineness is also lower in mean, by −0.230556. The primary novelty finding should not become a claim of general output improvement.

The audit evaluates the finite paired bootstrap distribution by integer convolution over all 10^10 ordered problem resamples. The resulting novelty interval is **[0.225, 1.213889]**. This removes Monte Carlo error from that calculation, not uncertainty about sampling, measurement validity, or confounding. Study 3 has 238 of 240 scheduled judgments. All 49 hypothetical combinations for its two missing novelty ratings preserve the original point-plus-positive-interval decision, although signed-rank p ranges from .021484 to .058594. Leaving out one grader at a time preserves positive novelty differences and positive exhaustive-bootstrap lower bounds. These are post-hoc sensitivities of existing observations. Exact-bootstrap and missingness audit.

Fabrication flags require precise units. Wildcard has 0 positive flags among 120 judgments, covering 30 outputs. Plain has 6 positive flags among 118 judgments, concentrated in 2 of 30 outputs. Those are model flags, not six independently verified factual errors. A spot check found that normalization changed a distinction between PostgreSQL materialized views and TimescaleDB continuous aggregates, preventing automatic attribution of the resulting flag to the generator alone. Zero observed flags is not a no-fabrication guarantee. Flag and normalization audit.

3.4 Dependence in selection and shuffle

Every recorded Study 3 treatment seed is 22 bytes and satisfies `modeIndex === lensIndex % 2`. A probe of 100,000 equal-length seeds reaches 189 of 378 specialist leaves conditional on automatic specialist mode, and four of eight lenses within each mode. A separate historical shuffle probe over 100,000 six-byte seeds with tag `audit` reaches 12 of 24 permutations for length four and 2,520 of 40,320 for length eight. Equal-length tagged CRC streams couple low bits used for selection. Marginal balance and browser/shell parity are insufficient checks for independence. These probes do not recover unrecorded historical grader presentation orders. Sampler and shuffle audit.

The historical estimate applies to the whole recorded configuration, including its legacy sampler. A corrected sampler creates a changed treatment whose effectiveness must be measured separately. The sign and size of any historical bias relative to that replacement are unknown. Requested model names and repository commit order are retained, but missing request envelopes, returned model identities, decoding settings, grading-order receipts, and usage records limit stronger provenance claims. Provenance boundary.

4. New study design

The protocol and literal prompts govern acquisition and analysis. This manuscript explains them without replacing their frozen definitions.

4.1 Materials and allocation

The task set has 32 synthetic main briefs, eight each in interaction/accessibility, software systems, operational workflows, and information/creative tooling. Every brief specifies constraints, success criteria, and non-goals. Four disjoint development tasks exercise transport and measurement. An AI assistant authored the tasks knowing the research question; task construction is not blinded or a probability sample of customer work.

Sixteen donor cards contain a domain name, an authored relation, a transfer boundary, primary-source links, and a source-fact note. The source supports the donor fact, not a target intervention. Cards include familiar systems patterns as well as less familiar sources; equal semantic distance is not claimed. For example, a directional lighthouse signal motivates a possible relation between position and relevant warning. Its source does not establish that an analogous interface is effective or safe.

A deterministic SHA-256 ordering with public seed `wildcard-transfer-2026-09-20-v1` assigns each card to two tasks. A fixed offset supplies another card's label for the mismatch condition. This schedule is recoverable from the manifest. It is not an inference RNG seed, a uniform sample of semantic space, or a guarantee of provider-level statistical independence.

Arm	Input beyond the common strong direct prompt	Interpretation
S	None	Direct generation comparator
R	Relation and transfer boundary	Added relational cue bundle
LR	The same relation and boundary, plus its correct donor name	Primary name intervention
XR	The same relation and boundary, plus a different donor name	Conflicting-name diagnostic contrast

The common prompt requests executable, constraint-respecting actions, materially different mechanisms, implementation steps, and falsifiable checks. The cue is optional and may be ignored. R, LR, and XR share byte-identical relation and boundary text. The source name is omitted from the required output in every arm. Removing the explicit label does not erase all domain information; a relation may make its source guessable. XR changes label congruence as well as labeling and does not isolate a pure effect of familiarity.

4.2 Reference bank and candidate acquisition

Before main generation, each task receives two separate strong-direct bank calls with up to eight actions each. Fresh contexts are used; the bank is never shown to main generators. The bank's realized size is recorded and the same bank evaluates all four arms, including S. It is a finite sample from direct prompting, not a catalogue of all obvious, published, or previously learned ideas.

Each task then receives one generation in each arm, with up to four actions and a maximum of 110 whitespace-delimited words per action. An action contains `action`, `mechanism`, `implementation`, `check`, and `risk`. Fewer actions and explicit abstention are allowed. No response is selected or rewritten after generation.

The requested generator is `gpt-6-astra`, with requested judges `gpt-6-astra` and `gpt-5.5`, all using the low reasoning configuration through the locally authenticated Codex CLI. These are requested aliases. The transport does not expose a verified provider-returned model snapshot. Temperature, top-p, inference seed, and a hard output-token cap are unknown or unset and recorded as null. Budgets match action counts and word limits, not tokens. Usage and latency are logged where emitted; marginal dollar cost is unavailable.

Calls use ephemeral contexts, frozen replacement system instructions, ignored user configuration, an empty working directory, and disabled tools, web search, memory, plugins, apps, and skill discovery. The CLI still introduces common infrastructure. Requests, responses, events, status, attempts, timing, usage, and hashes are retained. A call has a 240-second timeout and at most two transport attempts. A delivered invalid response is not regenerated; only unsuccessful or incomplete transport permits a retry. A tool-policy violation is an error.

The main design requires 64 bank calls, 128 candidate calls, and 64 judge blocks. Sixteen diagnostic calls and separately recorded development/calibration calls are additional. The ledger determines actual counts. An invalid bank prevents main acquisition. Main acquisition failures and valid abstentions score zero; a missing or invalid required judge block prevents primary analysis. This defines an acquisition-inclusive estimand rather than selecting successful responses.

Development and pre-main amendment. The first completed development pass used the four disjoint development briefs. All eight bank responses and eight judge blocks passed validation. Eleven of sixteen candidate responses passed; five whole responses failed because an action exceeded the 110-word cap. The offending action lengths were 112, 112, 112, 113, and 117 words. This 31.25% failure rate concerns development-format compliance, not the main treatment effect. Raw failures and the development analysis remain archived.

Before any main-cohort acquisition, the common generation instruction was clarified to target 60-80 words total across the five field values, with suggested budgets of 8 words for action, 16 for mechanism, 22 for implementation, 20 for check, and 10 for risk. These budgets guide generation; the unchanged hard validation rule remains a maximum of 110 words across all five values. The same amended instruction applies to bank calls and all arms. A single declared rerun on the same four development briefs produced **32 valid calls out of 32: eight banks, sixteen candidate responses, and eight judge blocks**. The complete dataset passed the strict analyzer. No further prompt tuning followed. Final development record and frozen prompt source.

This adaptation uses observed development outputs and is disclosed as such. It neither repairs those outputs nor licenses shortening later responses after delivery. The reused development set tests instrumentation and supplies no main observations or fresh efficacy evidence. An earlier bank-only prefreeze check retained eight successful calls but stopped after source drift, before any candidate or judge requests; exact hash-verified source snapshots were recovered and preserved. The main materials and their separate manifest were committed after these development records. The main zero-for-failure rule remained unchanged. All 128 main candidate responses passed, so format failures did not contribute zero scores to this realized main comparison.

4.3 Masked measurement

For each task, each judge receives the task, the realized bank, and all candidate actions under masked IDs. Arm, cue, and source metadata are omitted. Candidate fields are carried verbatim, avoiding the historical normalization stage. The first order is shuffled; the second judge sees its reverse. Content can still reveal a condition, and joint presentation can produce context effects. Masking is a documented procedure, not proof that treatment is unknowable.

Each candidate receives three qualification flags: constraint validity, feasibility using the stated resources, and actionability with an observable check. All must be true. The judge also assigns a mechanism group and identifies a matching bank action, if any. Cosmetic names, metaphor vocabulary, and parameter changes alone do not establish a new mechanism. Judgments and concise reasons remain inspectable.

QDM@4 counts distinct groups with at least one qualified action in a set. **QNM@4** counts those groups only when they are absent from the bank according to the judge. Matching is propagated conservatively across the whole task/judge block: if any candidate in a group matches the bank, that group cannot count as new in any arm. This avoids awarding the same mechanism novelty in one arm after matching it to the bank in another. It can also spread a mistaken grouping or match. Raw contradictions are preserved and reported. Both scores range from zero to four.

QNM measures **model-judged qualified mechanisms absent from this finite bank**. It is not an estimate of training-data novelty, patentability, human creativity, implemented performance, or truth of every factual statement. QDM distinguishes absence from the bank from general qualified diversity.

4.4 Estimand and inference

For task i, arm scores are averaged across the two judges. The primary paired difference is QNM(LR) minus QNM(R); its mean across all 32 tasks is the primary estimate. Actions, groups, and judge calls are not extra independent problem units. One generation per task/arm does not separately estimate within-task generation variability.

A 95% percentile bootstrap interval uses 50,000 paired task resamples within the four fixed eight-task family strata. The primary two-sided mean sign-flip test samples 100,000 sign vectors and uses the plus-one correction, `(1 + extreme_count) / 100001`, including ties. Its null relies on paired exchangeability or sign symmetry. Independent named SHA-256-counter streams specify analysis randomness in the analysis implementation. Resampling conditions on the fixed card library and does not establish population generalization over all briefs or cues.

The evidence rule reports a direction only when the mean has that sign, the primary p-value is at most .05, and the 95% interval excludes zero in the same direction. Otherwise the result is inconclusive, not equivalent. There is no prespecified minimum practically useful effect or power justification. The secondary QNM contrasts R−S and XR−LR form a two-test Holm family. They cannot replace an unsuccessful primary contrast. QDM, individual judges, per-family patterns, qualification, failures, costs, ceilings, and disagreement are descriptive.

A complete-success sensitivity includes only tasks where both contrasted calls produced valid nonempty actions and explicitly reports that selected denominator. The all-task analysis remains primary. A bank-size sensitivity requires separate judging or complete bank-match annotations. One chosen match ID cannot recover all possible matches against a reduced bank. That sensitivity was not acquired and is reported as unexecuted.

4.5 Counterfactual operational diagnostics

Eight paired fixtures hold a toy target and synthetic donor label fixed while changing the supplied relation. Expected decisions and observable traces are specified before acquisition. Some changes reverse a target rule; others make literal transfer violate an explicit constraint. Each variant must match its expected decision and trace, and one fixture also requires an exact command token. A pair passes only if both variants pass. Failures remain visible and do not trigger redesign of the test.

These cases distinguish some forms of operational sensitivity from simply copying a donor name. They can also be solved through literal instruction-following. They do not establish broad analogical ability, explain the main contrast by causal mediation, or constitute an independent creativity benchmark. Qualitative explanation review cannot override the exact diagnostic score.

5. Acquisition outcome and measurement amendment

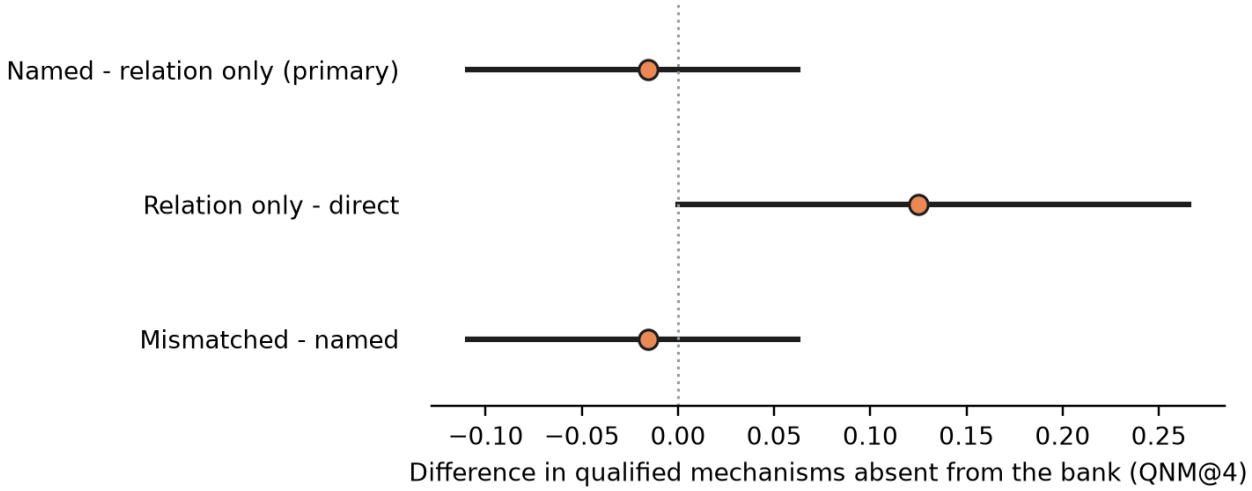

Figure 1. Prespecified contrasts evaluated in the amended measurement panel on the 32 authored briefs. The original primary remains halted. Points show means; lines show unadjusted 95% family-stratified bootstrap intervals. The primary test and Holm-adjusted secondary tests are reported in the text. This is bank-relative, model-judged yield.

5.1 The original primary analysis halted

The original cohort completed all 272 scheduled calls in 272 transport attempts. All 64 bank calls, 128 candidate calls, and 16 diagnostic calls passed validation. The bank contains 448 realized actions. Of the 64 judge blocks, 63 passed and one failed. These are acquisition outcomes, not a completed primary efficacy analysis. Original call records and judge manifest.

Original phase	Completed calls	Valid responses	Validation failures	Transport attempts
Reference bank	64	64	0	64
Candidate generation	128	128	0	128
Operational diagnostics	16	16	0	16
Judge panel	64	63	1	64

The invalid block, `i07-judge-2`, returned all sixteen candidate ratings but placed candidate ID `73e7cd8ccf86c466` in two `baseline_match` fields. It was not any of the fifteen permitted bank IDs; one occurrence was a self-reference. The original structured-output schema allowed arbitrary strings in those fields, while the subsequent validator correctly rejected the references. Transport had completed, so the frozen no-regeneration rule applied. The original status remains `validation_error`, and its primary analysis remains halted. We neither converted the links to null nor inferred their intended targets. Failed record, raw response, and request.

No aggregate main effect estimate, interval, or hypothesis test was calculated before choosing the amendment. The failed block, individual examples, and raw diagnostic outputs had been inspected. This is therefore a post-failure instrument amendment, not an outcome-blind change or an untouched preregistration.

5.2 One new panel under constrained identities

The amendment authorized exactly one new panel of all 64 judge blocks. Its manifest was frozen at 20:42:49 UTC and committed and published before amended acquisition. Per-block output schemas restrict candidate IDs to the supplied candidate set and bank references to the actual bank IDs or null. Prompt bytes, bank and candidate texts, presentation order, requested models, reasoning settings, rubric, and scoring remain unchanged. No generated candidate, bank action, or diagnostic response is reacquired.

All 64 blocks were measured again under the same amended interface and passed validation on their first transport attempt. The 63 valid original blocks are not pooled with the new panel, and panels cannot be selected according to their results. The amended panel alone determines its primary estimate using the already prespecified LR−R contrast and statistical procedures. Delivered invalid responses are still not regenerated, and any missing or invalid required block halts the amended analysis. The amendment authorizes no additional panel.

The new panel is a changed measurement procedure applied to the same generated sample. It is not an independent generator replication or a second set of 32 problems. Constraining IDs prevents this reference-domain error but does not establish semantic reliability. The complete amended results contain every task score, both acquisition ledgers, and their source hashes. The amended panel contains no omitted blocks and required no repairs or retries.

5.3 A separate post-hoc metric-invariance sensitivity

The two invalid links belong to the same mechanism group, which also contains valid bank matches in the original response. An independent diagnostic enumerated all 256 hypothetical assignments of fifteen valid bank IDs or null to those two fields, holding all other ratings and group assignments fixed. Every assignment yields the same per-arm QNM and QDM vector under the frozen global propagation rule. A second implementation check reproduced that invariance without changing any source record.

This establishes score invariance to that narrowly defined missing-identity choice. It does not establish that the remaining matching or grouping judgments are semantically correct. No intended link was recovered, and the original response remains invalid. Any aggregate estimate using this invariance must be labeled **post-hoc original-panel sensitivity**, reported separately, and cannot reinstate the halted primary or replace the amended panel. The exhaustive proof, sensitivity implementation, and amendment expose the assumptions and boundary.

After the amendment manifest was frozen, the separately labeled original-panel sensitivity was computed for all 32 tasks. Original-panel mean QNM was 0.03125 for S, 0.15625 for R, and 0.140625 for both LR and XR. The following estimates are conditional on the unchanged original qualification and grouping judgments. They are not the amended panel's results. Post-hoc sensitivity artifact.

Original-panel sensitivity contrast	Mean QNM difference	95% percentile interval	Two-sided mean sign-flip p	Holm-adjusted p
LR−R	−0.015625	[−0.09375, 0.03125]	1.0000	Not applicable
R−S	+0.125000	[0.015625, 0.25000]	0.093939	0.187878
XR−LR	0.000000	[−0.09375, 0.09375]	1.0000	1.0000

The R−S bootstrap interval excludes zero while its paired mean sign-flip test and Holm-adjusted p do not meet .05. These procedures are not inversions of each other, especially with sparse, discrete differences and stratified resampling. This sensitivity supplies no restored primary decision or significant secondary claim. The name contrast's interval includes zero; it does not establish equivalence or demonstrate that names never matter. The new panel remains the sole source of amended estimates.

5.4 Strict operational diagnostics: six of eight pairs

All sixteen diagnostic calls returned valid structured responses. Twelve of sixteen variants, forming **six of eight complete pairs**, matched the frozen exact decision-and-trace contract. Both variants of x03 and x04 failed because their traces added actor or time labels. For example, x03-a returned `A: GRANT` and `B: WAIT` instead of `GRANT` and `WAIT`; x04-a returned time-prefixed `UNKNOWN` values. The decisions matched their expected choices, and the prefixed state values are consistent with the intended operations. This descriptive observation does not change the strict score. No labels were stripped to turn failures into passes. x03-a response, x04-a response, and frozen expectations.

These designed cases can be solved by literal instruction-following. Six passing pairs therefore do not establish general analogical ability or explain the mechanism behind the main contrast. Their failures also illustrate sensitivity to an exact trace-format contract, which must remain visible alongside operational interpretation.

5.5 Amended estimates: no established name benefit

The amended analysis includes all 32 tasks and only the new panel's judgments. Correctly named relations averaged 0.15625 QNM@4, versus 0.171875 for relation only. The primary difference is **−0.015625**, with 95% interval **[−0.109375, 0.0625]** and paired mean sign-flip p = **1.000**. Twenty-nine of thirty-two paired differences are zero; three are nonzero. The frozen decision rule returns **inconclusive**. This does not establish equivalence, demonstrate that donor names never matter, or support removing provenance information from a human interface. Complete amended analysis and paired task scores.

Amended contrast	Mean QNM difference	95% percentile interval	Two-sided mean sign-flip p	Holm-adjusted p
LR−R, primary	−0.015625	[−0.109375, 0.0625]	1.000000	Not applicable
R−S, secondary	+0.125000	[0, 0.265625]	0.139459	0.278917
XR−LR, secondary	−0.015625	[−0.109375, 0.0625]	1.000000	1.000000

Neither secondary contrast meets its Holm-corrected .05 threshold. R−S has a positive point estimate, but its interval includes zero exactly at the lower endpoint. It does not establish a relational-prompt benefit. The mismatched-name comparison also remains inconclusive. These contrasts cannot replace the primary question.

The generated sample contains 478 candidate actions. All 128 candidate responses were valid and nonempty, so the complete-success sensitivity retains all 32 tasks and gives the same differences. The reference bank contains 448 actions, with 10-16 per task and a mean of 14.

Arm	Candidate actions	Mean QNM@4	Mean qualified distinct mechanisms, QDM@4
S	118	0.046875	3.671875
R	121	0.171875	3.765625
LR	120	0.156250	3.750000
XR	119	0.140625	3.718750

Bank-relative yield is sparse in every arm, while qualified diversity lies near the maximum of four. Most proposed mechanisms were judged qualified but already represented in the direct-prompt bank. This pattern is conditional on the authored briefs, finite bank, and conservative grouping rules. It may reflect strong direct-prompt coverage, broad matching, or limited intervention effects; this study does not separate those explanations. No bank-size sensitivity was executed, and sparse differences constrain interpretation of the small name contrast.

The requested `gpt-6-astra` judge gives a primary mean difference of 0; the requested `gpt-5.5` judge gives −0.03125. Descriptive family differences are 0 for interaction/accessibility and information/creative tooling, −0.125 for operational workflows, and +0.0625 for software systems. These eight-task subgroup patterns are not separate confirmatory findings.

Across 478 candidate actions, qualification agrees for 476 (99.58%); feasibility and actionability individually agree for all 478. Bank-newness classifications agree for 471 (98.54%). Pairwise mechanism grouping agrees for 3,334 of 3,373 within-task candidate pairs (98.84%), with positive-link agreement of 96.64%. Pair observations are correlated, and the all-pair percentage benefits from many nonlinks. These are agreement statistics, not accuracy or human validation. The high qualification rates and QDM ceilings can also limit the rubric's discrimination.

One task/judge mechanism group, in `i07` under judge 2, contains both a valid bank match and a null match. The declared global propagation rule treats the entire group as already represented across arms; the raw disagreement remains recorded. The mean absolute QNM difference between judges is 0.0546875 across task/arm sets. Agreement and contradiction records.

The separately calculated original-panel sensitivity has the same LR−R point difference but a different interval. This is consistency on the same generated sample under related measurement procedures, not independent replication. The amended panel's estimates above remain the sole amended result; no panel was selected for its outcome.

5.6 Acquisition and observed resource use

The original main cohort and amended panel comprise **336 logical calls in 336 transport attempts**: 64 bank calls, 128 candidate calls, 16 diagnostics, 64 original judge blocks, and 64 amended judge blocks. Of these, 335 responses passed validation and one original judge block failed. The amendment adds evaluation effort without adding generated problems or candidate responses. The following usage counters are summed from retained attempts, including the invalid original block. Complete acquisition and usage ledgers.

Phase	Calls / attempts	Reported input tokens	Reported output tokens	Sum of call elapsed seconds
Reference bank	64 / 64	456,274	60,729	2,169.04
Candidate generation	128 / 128	925,620	71,686	2,865.67
Operational diagnostics	16 / 16	111,965	1,142	107.16
Original judge panel	64 / 64	611,294	77,296	2,094.70
Amended judge panel	64 / 64	651,120	81,294	2,136.86
Total	336 / 336	2,756,273	292,147	9,373.44

The transport additionally reports 1,102,720 cached input tokens and 49,422 reasoning-output tokens across these calls. These counters are not added again to the input/output totals. Summed call duration is not wall-clock project duration because acquisition used concurrent workers. The amended judge panel reports zero cached input tokens, while the original panel reports 178,816; token or timing differences therefore should not be read as controlled efficiency effects. Marginal dollar cost is unavailable and is not estimated.

A further **74 preserved development/calibration calls** comprise eight prefreeze bank checks, thirty-two calls in the first completed development pass, thirty-two in the final development pass, and two known-answer judge calibration calls. Together these ledgers retain 410 calls in 410 attempts. Development and calibration contribute no main task observations, and their usage is separate from the 336-call table. This accounting does not include the AI-assisted research, writing, and software work surrounding the experiment.

6. Limitations and product implications

The synthetic tasks favor explicit constraints, state, verification, and plausible implementation. Their author knew the intended construct. A result on this set does not generalize to visual taste, unconstrained art, expert professional judgment, natural user projects, or long-term team creativity. Curated relations can favor mechanisms already natural to such tasks, and two uses per card do not support broad card-specific conclusions.

The bank-relative score depends on reference coverage and matching granularity. A larger bank may reduce apparent newness; a broad cluster may merge meaningful alternatives, while a narrow cluster may inflate diversity. Same-provider judges can share preferences and blind spots, particularly when one requested alias is also the generator. Conservative rules reduce some obvious inconsistencies but do not make model judgments objective. No new human ratings or implementation trials are included.

The measurement amendment adds a further limitation: the output interface changed after an observed validation failure. Exact prompt content and criteria are preserved, but constraining the available reference strings can still affect model behavior. Agreement with the original-panel sensitivity is measurement consistency on the same outputs, not independent replication. Differences between panels remain reported rather than resolved by selecting a favorable panel.

The primary intervention adds text to a model prompt. It does not test whether showing a donor name to a person improves their decisions or interface comprehension. A product can expose source provenance for

accountability regardless of the model-level effect. Likewise, a draw receipt demonstrates selection and replay; it cannot certify that a resulting suggestion is useful.

The historical audit supports a design principle with a narrower basis than a creativity guarantee: make the distinction between **source fact, authored relation, proposed target action, and observed evaluation** visible. Readers should be able to inspect the reference, reproduce the selection, and follow a claim back to a measured artifact. The corrected live sampler and historical replay must identify their versions. Neither should inherit effectiveness evidence from a different configuration.

7. Authorship, assistance, and reproducibility

christopher robin fiore directs this project and its research-to-product presentation. AI assistants helped review prior work, audit historical artifacts, author the synthetic materials, implement acquisition and analysis, and build the interface. Task authorship was not blinded. The new benchmark used two recorded model-judge configurations; it contains no new human rating or user study. Human authorship of the project should not be confused with human adjudication of generated candidates.

The repository preserves historical sources separately from the new audit and acquisition. The protocol, prompts, tasks, cards and primary sources, diagnostics, collector, original validator, statistical analysis, and measurement amendment make the procedure inspectable. Each run manifest and retained source snapshot identifies the version actually executed; the amendment does not overwrite the original record. Public attribution is christopher robin fiore; the repository is hosted at globalanomalyindex/wildcard.

The completed amended analysis did not establish a donor-name benefit under the prespecified rule. It also provides a reproducible record of sparse bank-relative yield and a measurement failure that remained visible through correction. The product implication is to support inspection of source facts, relations, boundaries, and candidate actions without promising improved creativity. Reproducible arithmetic, complete measurement, valid constructs, and useful deployed behavior remain separate achievements.