FlavourBench: Reliability-Aware Evaluation of Tool-Augmented Culinary Language Models

Josef Chen
Independent Researcher

August 2026

Abstract

Culinary evaluation must separate answer quality from the ability to finish a tool trajectory. FlavourBench uses two blinded tracks: cross-model comparison under fixed tool access and paired within-model evaluation with and without Epicure. Failed generations remain in their assigned denominators. We report nine development collections run on 2–3 August 2026 across 16 exact endpoints, 28 human-authored tasks, and the live Epicure MCP. The panel includes a GPT-5.6 Sol OpenRouter route, the Claude 5 family, Gemini 3, Kimi K3, GLM 5.2, DeepSeek V4, and Cohere Command endpoints. Of 320 scheduled matched pairs, 211 completed. Across complete and partial pairs, the collection retained 524 normalized real answers and 488 tool calls, of which 380 succeeded. No synthetic arm entered the collection. A post-freeze direct QwenCloud smoke of the catalog-observed `qwen3.8-max` mutable alias is reported separately and excluded from these totals and every quality fit.

A source audit quarantined four tasks. The retained review workloads contain 186 uplift pairs and 915 candidate same-task model comparisons constructed from 188 answers on 20 comparison tasks. Answers recur 2–14 times, and five tasks account for 430/915 candidate rows. The specified inference resamples task clusters, keeps reused responses nested within task, and crosses the task resample with rater clusters rather than treating candidate comparison rows as independent. 73 of 480 model-pair-by-family cells remain empty, and zero quality judgments have been admitted. We therefore publish neither an endpoint ranking nor an Epicure answer-quality effect.

A separate retrospective audit covers 2,880 arms from 120 public questions and 12 historical endpoints. Only 393 of 1,486 admitted pairs met the automated-panel consensus rule; tool-available arms had a 12.9-percentage-point higher realized failure proportion. These findings motivate a prospective successor with 240 new tasks, 12,800 primary real response arms, two independent expert ratings per comparison, task-clustered inference, and fail-closed coverage and reproducibility gates.

1 Introduction

A fluent culinary answer can still fail in the kitchen. It may meet the flavor goal while ruining texture, omit an equipment constraint, propose unworkable quantities, or treat suggestive chemical evidence as certainty. Several materially different answers may be good, so exact-match scoring is unsuitable. Pairwise preference fits the task better but inherits rater, position, style, and missing-data effects [7, 43].

Tool access complicates identification. Comparing two tool-using systems ranks them but does not estimate the tool's contribution; comparing a strong tool-available endpoint with a weaker tool-unavailable endpoint confounds treatment with base capability. A successful tool call is not an answer-quality measure: correct retrieval can still lead to poor synthesis [30, 41].

FlavourBench separates two questions. Endpoint comparison holds tool availability constant and varies the endpoint under frozen endpoint-specific decoding contracts. The paired tool-availability track fixes the endpoint, task, base instruction, requested decoding, and answer normalization while changing whether the Epicure registry is available (Figure 1). This pilot asks:

(i) How do complete-case automated preference and realized arm completion differ when Epicure is offered within a fixed endpoint-task cell?

(ii) How do judge completion, orientation agreement, and family-dependent admission alter the retained comparison cohort?

Using culinary troubleshooting as a case study, we apply a matched endpoint-task design to audit target failure and judge admission. The pilot resolves neither an endpoint ordering nor a judge-robust Epicure quality contrast. Generation failures and judge admission are outcomes of the evaluation procedure, not neutral preprocessing. The evidence records preserve prompts, routes, tools, outputs, failures, and costs so that each stage can be audited separately.

1.1 Evidence layers and model currency

Four evidence layers are kept separate. The scored retrospective pilot ran on 16 and 17 July 2026 and retains the exact endpoints that produced its arms. A route registry was frozen on 1 August after consulting provider documentation and the live OpenRouter catalog [4, 12, 15, 18, 28, 42]. It covers 16 exact routes; 15 have a complete contract receipt comprising two provider generations, one live Epicure call, a schema-valid final response, and provider identity evidence. These checks measure interface compatibility, not culinary quality. The third layer comprises

nine development collections made on 2–3 August over a 16-endpoint panel, all with real provider outputs and live Epicure access. Two core-panel runs share a strict execution policy and 12 tasks. Two full-panel high-resource runs use eight disjoint tasks; two targeted high-resource replenishments use eight further tasks to raise the least observed endpoints to the frozen completion floor. Three direct Cohere collections extend the high-resource stratum with two exact endpoints on 12 of its tasks. These protocol strata are reported separately. Together they produced 211 operationally complete pairs before the post-collection task audit, but no admitted quality judgment. The prospective season is a fourth layer and has not been collected. Records are not pooled across evidence layers or execution-policy strata.

2 Related work

Subjective model evaluation. Chatbot Arena established anonymous pairwise voting and Bradley-Terry (BT) aggregation at scale [6,7]. MT-Bench documented position and judge biases in language-model-based evaluation [43]. WildBench uses real-user prompts with task-specific checklists, while HELM treats evaluation as a matrix of scenarios and metrics rather than one accuracy number [19, 20]. These systems motivate blinded comparison, but prior work also finds that agreement among model judges need not imply agreement with humans [23]. We study a related selection problem: successful dual-orientation judgments and family-dependent consensus rules determine which model comparisons remain observable.

Benchmark item validity. GPQA combines expert authorship with expert and non-expert validation; MMLU-Pro subjects questions to repeated review; IFEval reduces grader discretion with explicit constraints [33,38,44]. SWE-bench scores hidden repository tests, while its Verified subset adds repeated professional review of task specification and test validity [17,24]. LiveCodeBench versions newly released contest tasks by time, and LiveBench periodically refreshes objective items [16, 39]. Culinary advice cannot be wholly executable, but the same discipline still calls for a construct map, independent item review, hidden reserves, and objective checks where the task permits. Human review is not a permanent guarantee: later audits reported residual specification and test defects in both SWE-bench Verified and SWE-bench Pro [25,26]. We therefore treat task validity as a versioned evidence record, not a one-time label. BetterBench places the same concern in a full benchmark lifecycle that includes construct design, implementation, documentation, uncertainty, replicability, and maintenance [34]; Season 1 makes those properties release conditions rather than post-publication aspirations.

Tool interventions. ToolBench evaluates tool selection and execution [30]. BFCL combines abstract syntax tree and execution checks with abstention and stateful cases; τ-bench and τ^2-bench score multi-turn trajectories in controlled environments [5, 29, 41]. MCP supplies a common schema for tool discovery and calls [22]. Our paired estimand is narrower: it measures the contrast associated with offering one fixed culinary tool bundle on the final answer. It does not identify whether any effect comes from retrieval, representation quality, additional context, tool-use policy, or answer length.

Culinary evaluation. CookingSense and FoodBench test culinary decision support with retrieved knowledge [8]; PizzaCommonSense makes intermediate recipe state explicit [13]; and Recipe2Plan tests parallel scheduling under temporal constraints [40]. FlavourBench instead samples open-ended troubleshooting and separates endpoint comparison from a paired tool-availability contrast. This pilot does not establish general culinary competence.

Culinary representations. Ingredient networks, FlavorDB, and Recipe1M show how culinary relations can be represented computationally [1, 14, 35]. Related Epicure work analyzes structure in ingredient representations [31, 32]. In this study, similarity is evidence available to a model, not ground truth for judging an Epicure-assisted answer.

3 Real-call exact-frontier study

3.1 Panel, tasks, and routing

The development study fixes 16 named endpoint contracts across nine real-call collections. The first two collections run the 14-endpoint core panel under one strict execution policy and contain 12 distinct tasks, three per family. Two core-panel collections use a high-resource policy and eight different tasks, two per family. Two completion-floor replenishments retain that policy but schedule only Grok 4.5, MiniMax M3, and Mistral Medium 3.5 on four further tasks, followed by Mistral Medium 3.5 on four new tasks. A direct-Cohere extension runs Command A Plus and Command A Reasoning on 12 tasks from the same high-resource task set in three further collections. All 28 prompts came from human-authored public troubleshooting posts selected before the corresponding generation run. They are development candidates, not confirmatory items. The post-collection audit holds one task for a pending source-rights decision and three for context or construct defects. No synthetic prompt, response, or tool trace enters the study.

Every scheduled endpoint-task cell ran once without Epicure and once with required Epicure access. The paired arms used the same task, base instruction, final-answer contract, decoding request, and endpoint route. Only the treatment arm received the tool catalog. The two arms were dispatched concurrently. A treatment response was normalized only when its immutable trace

Endpoint	Strict	High	Combined	Epicure ok/all	Exposure ($)
GPT-5.6 Sol (OR pro)	7/12	6/8	13/20	19/20	5.409
Claude Fable 5	10/12	7/8	17/20	43/59	6.684
Claude Opus 5	11/12	8/8	19/20	33/39	5.763
Claude Sonnet 5	11/12	7/8	18/20	37/51	1.709
Gemini 3.1 Pro	8/12	6/8	14/20	15/16	1.297
Gemini 3.6 Flash	10/12	7/8	17/20	18/20	0.354
Grok 4.5	0/12	8/12	8/24	20/22	0.878
Kimi K3	9/12	7/8	16/20	27/34	25.452
GLM 5.2	5/12	5/8	10/20	28/35	0.367
DeepSeek V4 Pro	5/12	5/8	10/20	24/36	1.730
DeepSeek V4 Flash	11/12	7/8	18/20	42/51	0.048
MiniMax M3	2/12	7/12	9/24	14/36	0.286
Nemotron 3 Ultra	8/12	6/8	14/20	21/26	0.911
Mistral Medium 3.5	4/12	4/16	8/28	9/9	2.820
Command A Plus	—	12/12	12/12	12/15	not returned
Command A Reasoning	—	8/12	8/12	18/19	not returned
Total	101/168	110/152	211/320	380/488	53.708 + 2 unpriced

Table 1: Real exact-frontier development collections made on 2–3 August 2026. Rows retain frozen manifest order. Strict and High give complete matched Epicure-off/on pairs out of the tasks scheduled for each endpoint; the Combined column reports their sum. Epicure ok/all counts live successful and total calls across both strata. Exposure is the known US-dollar subtotal over admitted attempts, including unresolved full allowances and a separately identified direct-Kimi rate-card estimate. The two direct-Cohere routes returned token usage but no provider charge, so their exposure cells read `not returned` and the total names both unpriced endpoints. A dash records that a Cohere endpoint was not run in the earlier strict stratum. Every endpoint has at least eight complete pairs in the combined corpus. These are operational measurements; every row has zero human quality judgments and is unranked.

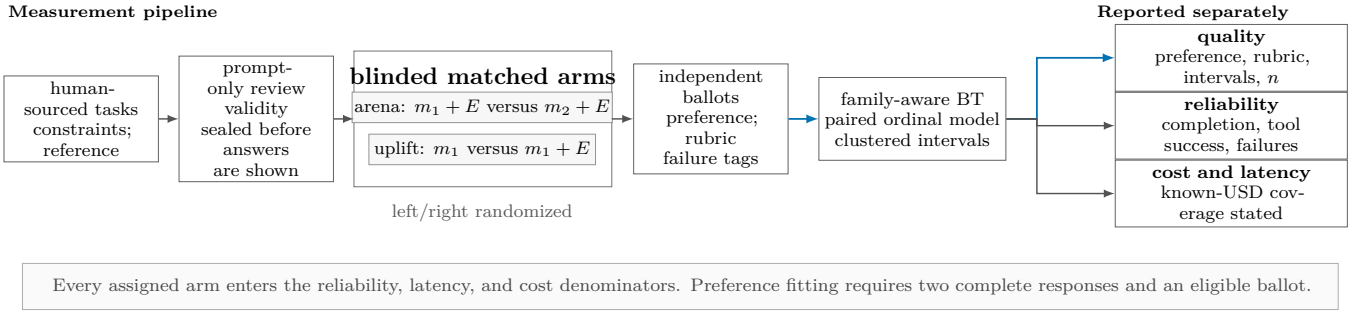

Figure 1: Measurement architecture. Human-sourced tasks undergo answer-hidden validity review before matched arms are generated and randomized. Independent ballots feed family-aware Bradley-Terry and paired ordinal estimators with task, response, and rater clustering. Quality, reliability, and cost or latency are reported separately. At the present cutoff, the human-ballot branch is empty, so this paper reports operational evidence without a quality ranking.

contained at least one successful call to the live Epicure MCP service. Failed and partial pairs remain in the scheduled denominator.

Kimi K3 used the direct Kimi coding endpoint with requested model ID `k3`. Command A Plus and Command A Reasoning used Cohere's V2 Chat endpoint with their exact returned model IDs. The Cohere catalog advertised chat, tools, and reasoning support for both routes. Because these models rejected the API's native required-tool selector, the frozen adapter omitted that field, inserted a content-hashed tool-use instruction on required tool turns, and rejected any treatment answer without a successful Epicure trace. This route-specific translation is disclosed because it can affect behavior [9–11]. The routing order was direct Kimi for Kimi K3, direct Cohere for the two Command endpoints, then an exact Bedrock route, then one exact OpenRouter route. A fresh `us-west-2` Bedrock audit found global inference profiles for Claude Fable 5, Claude Opus 5, and Claude Sonnet 5, but all three runtime smokes failed before generation with `AccessDeniedException`. The other ten non-Kimi releases were absent from the Bedrock catalog. The 13 OpenRouter routes were therefore frozen before collection as explicit fallbacks; there was no generation-time provider substitution. OpenRouter fallbacks were disabled, requested parameters were required, and provider data collection was denied. Each normalized response records its requested and returned model identity, actual provider, generation identifiers, prompt and schema hashes, decoding settings, full MCP trace, token counts, latency, and cost evidence [27, 28]. The 14 core endpoint contracts are unchanged across the two protocol strata; the Cohere endpoints occur only in the high-resource extension.

Post-freeze QwenCloud route check. On 8 August, a separate engineering smoke exercised the catalog-observed `qwen3.8-max` alias through direct QwenCloud

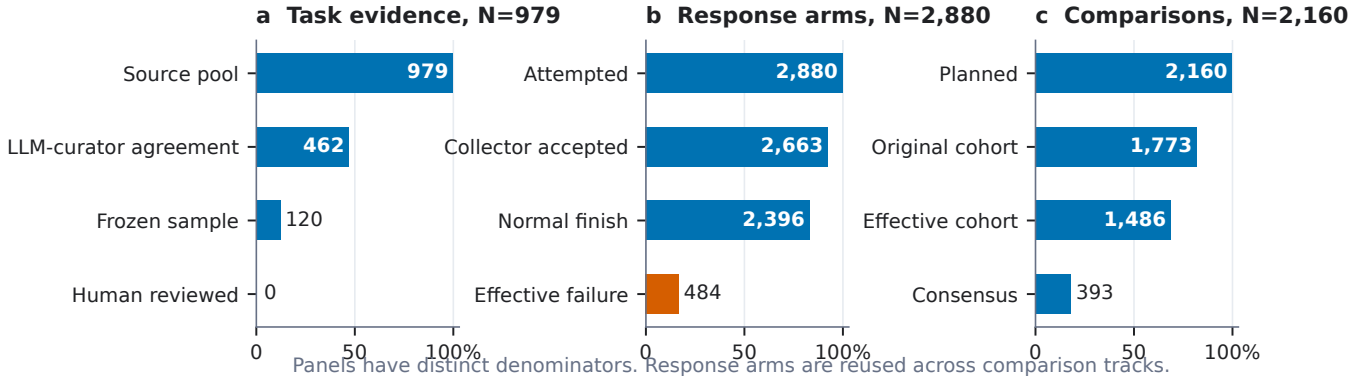

Figure 2: Evidence flow for the observed pilot. Each panel uses its own denominator. Public task provenance and LLM-curator agreement are distinct from qualified human review, which was incomplete for all 120 retained items. A completed arm must have a normal provider finish reason in addition to a reconciled, non-empty final answer. Response arms were reused across comparison tracks, so response and comparison counts are not nested. The completion audit removed 287 previously admitted comparisons; 393 of 2,160 planned comparisons reached primary consensus.

Attempt	Arms	Resp./reject	Epicure ok/all	Held ceiling
Predecessor	1/2	4/1	0/0	USD 2
Successor	2/2	6/0	2/2	USD 2

Table 2: Post-freeze direct-QwenCloud operational evidence. "Arms" gives delivered/requested response arms; provider requests were 5 and 6, respectively. The held ceiling is conservative exposure, not a provider charge. The alias is mutable and unranked. These rows are excluded from Table 1, both scoring pools, and every quality analysis.

on one composition task (Table 2). The first attempt delivered only the Epicure-off arm; the Epicure-on arm stopped at its first tool round with HTTP 400. The recorded attribution to the required tool selector is an inference, not a provider attestation. A successor delivered both real arms with six provider responses, no retries, and two successful live Epicure calls (`list_targets` and `flavour_correlations`); both final responses ended with `stop`. QwenCloud returned token usage but no rate or charged amount. The full USD 2 ceiling was retained for each attempt, USD 4 in total, so the recorded zero cost means unknown rather than free. This mutable alias was pinned only at one catalog observation, not to an immutable release. The addendum has no quality judgment and enters neither the 186-pair uplift pool, the 915-comparison arena pool, nor any ranking or quality fit.

3.2 Execution policies

The strict policy allowed two tool rounds, four calls per round, eight calls in total, 4,096 intermediate tokens, 8,192 final-answer tokens, and 32 KiB of cumulative tool results. The high-resource policy raised these limits to three rounds, six calls per round, 12 calls in total, 8,192 intermediate tokens, and 64 KiB of cumulative tool results. Both policies allowed one tool argument repair and at most two provider attempts for a request rejected before delivery. An ambiguous delivery was not replayed. Both used the same matched-evidence protocol and required the first treatment tool turn. The policies have different content hashes, and their task sets are disjoint. Both policies explicitly requested low reasoning effort for the intermediate tool-selection phase and the normalized final answer. The exact-frontier comparisons therefore describe low-effort endpoint configurations, not provider-default or high-effort configurations. Earlier reasoning-effort route checks recovered no usable contrast. Their accounting and identifier closures appear in Appendix B.

On 8 August 2026, a later governed block completed one real Claude Sonnet 5 Epicure-off/on pair before a post-generation audit fault stopped the block. The pair contains seven reconciled provider generations at USD 0.100547 and ten live Epicure calls, seven of them successful. An append-only V9 recovery admitted that source record and released 27 reservations whose requests had never started. The recovery made no provider, catalog, or Epicure calls and created no reservation. This single low-effort pair has no human quality judgment and no default or high-effort counterpart. It supplies neither an answer-quality estimate nor a reasoning-effort estimate, and it remains unranked.

3.3 Operational results

The strict stratum produced 262 normalized arms from 168 scheduled pairs. It preserved 964 provider generation identifiers and 215 live tool calls, of which 173 succeeded. Its known USD exposure subtotal was $26.546. The high-resource stratum produced 262 normalized arms, retained 919 generation identifiers, and recorded 207 successful calls among 273 total calls. Its known USD exposure subtotal was $27.162; provider charges were unavailable for 2 direct-Cohere endpoints.

Pair completion varied substantially across endpoints

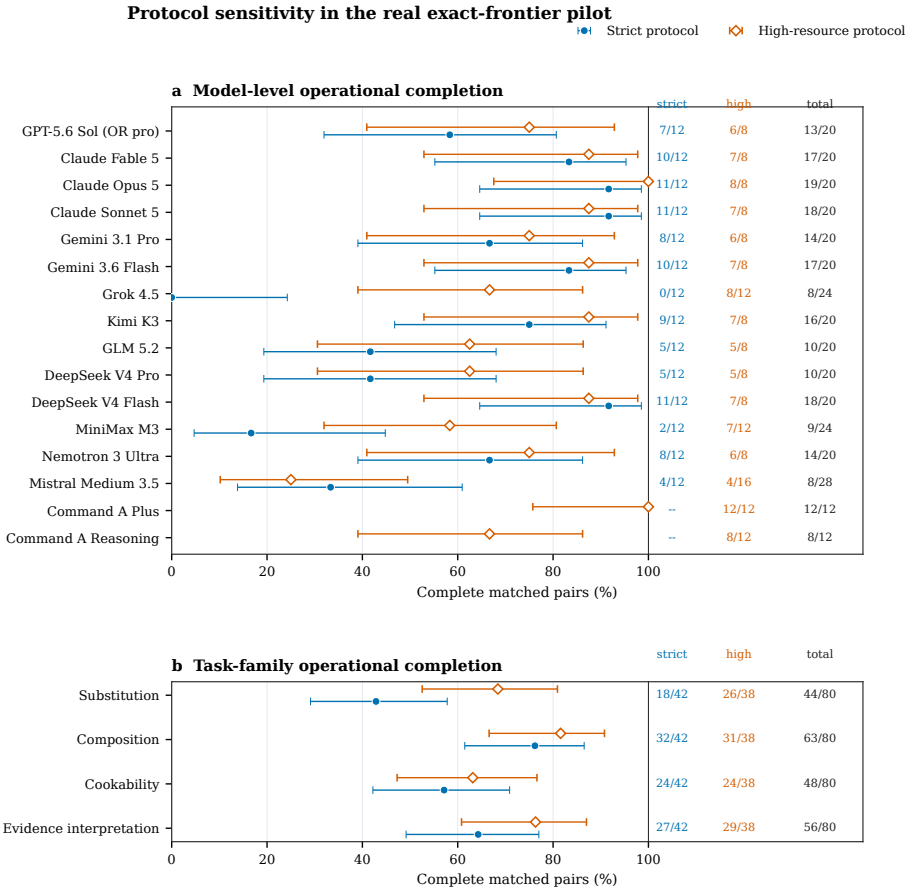

Figure 3: Operational completion under two frozen execution protocols. Points report the fraction of scheduled matched Epicure-off/on pairs for which both normalized responses were recovered and the treatment arm contained at least one successful Epicure call; whiskers are 95% Wilson intervals. The strict stratum completed 101/168 pairs and the high-resource stratum completed 110/152. The task sets are disjoint, so differences are descriptive and should not be read as causal resource effects or model-quality scores. The right-hand count columns show strict, high-resource, and combined completion. The two Cohere endpoints were not run in the earlier strict stratum and are marked accordingly; every endpoint reaches at least eight complete pairs in total.

(Figure 3). Claude Opus 5 completed 19/20 pairs, Claude Sonnet 5 and DeepSeek V4 Flash each completed 18/20, and Kimi K3 completed 16/20. The targeted runs raised Grok 4.5 to 8/24, MiniMax M3 to 9/24, and Mistral Medium 3.5 to 8/28. In the Cohere extension, Command A Plus completed 12/12 pairs and Command A Reasoning completed 8/12; the four failures remain in the reliability denominator. Every endpoint therefore has at least eight complete real pairs. These values describe the realized execution workload. Because protocol limits and tasks changed together, the study does not attribute the differences to additional resources.

3.4 Post-collection task audit and frozen scoring workloads

The initial response audit yielded 211 complete same-model pairs and 1024 possible cross-model comparisons from 218 Epicure-on answers across 24 tasks. A later source-completeness review quarantined four tasks before admitting any research ballot. `fb-s0-composition-006` depends on a linked recipe and product that are absent from the stored prompt. `fb-s0-composition-008` has an unresolved composition construct and source rights question. `fb-s0-composition-009` asks a terminology and sensory question without a clear composition decision. `fb-s0-cookability-003` omits four referenced images and includes an edit that advances the author's own causal resolution. These are conservative holds pending qualified adjudication, not retrospective invalid labels. The response records were not edited.

The quarantine removes 32 uplift pairs, 33 unique Epicure-on answers, and 148 arena comparisons. The retained uplift pool contains 179 same-model pairs with condition identity concealed and Epicure-on placement deterministically balanced. The retained model-arena pool

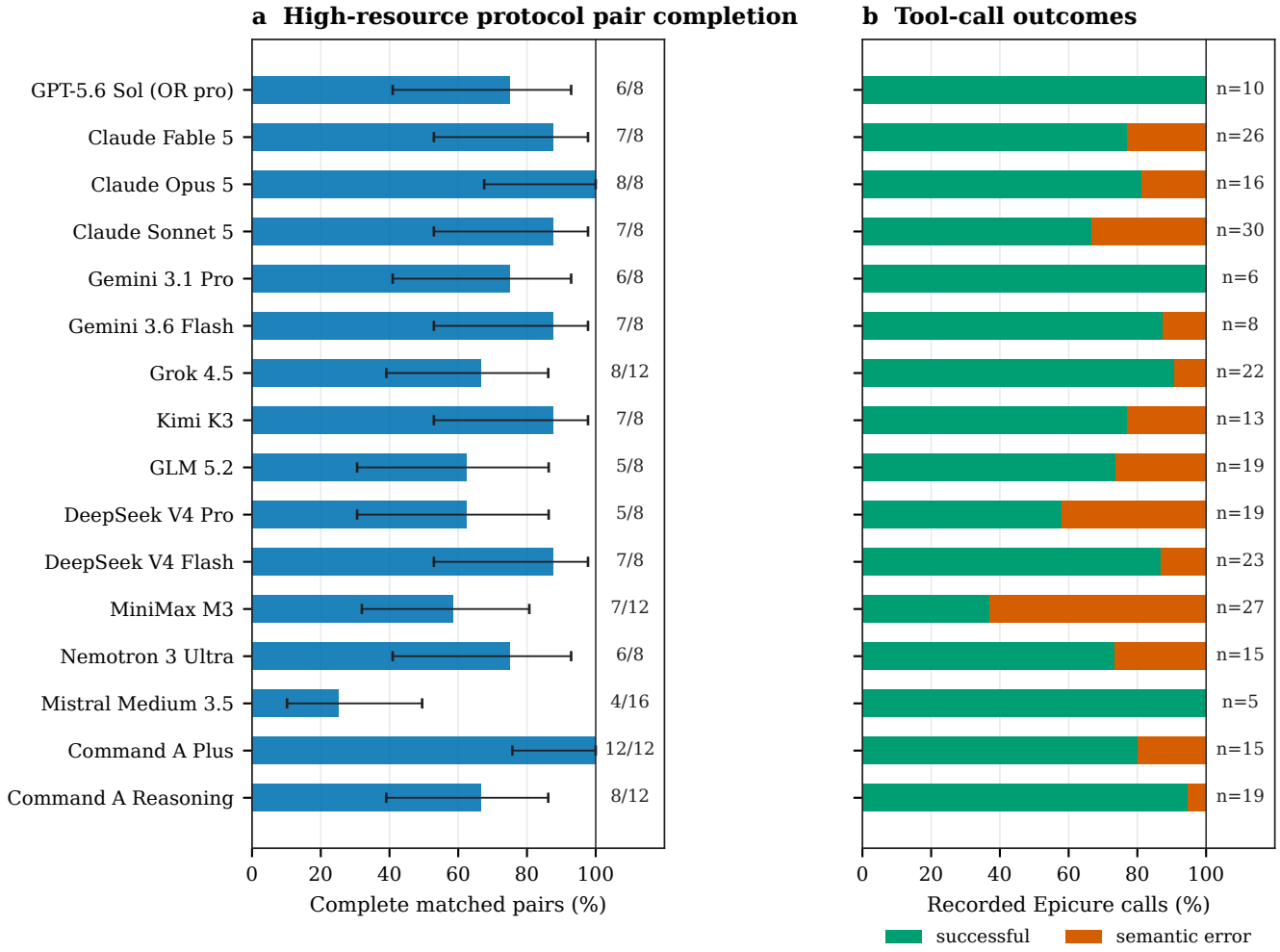

Figure 4: Execution reliability in the high-resource real-call stratum. Panel (a) gives complete matched pairs in frozen manifest order; whiskers are 95% Wilson intervals and the right column gives exact counts. Panel (b) separates successful from error-marked Epicure results across all recorded treatment attempts. Calls made before a later failed response remain in the trace. Pair completion and tool-call success are operational outcomes, not answer-quality scores.

uses 185 valid Epicure-on answers from 20 tasks that have at least two available endpoints. Within each task and execution stratum it includes every available unordered pair of different endpoints, producing 876 anonymous comparisons. The comparison graph contains all 16 models in one connected component. Model exposure ranges from 43 to 157 comparisons. All endpoints meet the paper's 40-comparison provisional workload floor, although no endpoint meets the separate ballot requirement yet.

An append-only source reconstruction subsequently admitted only recovery records whose route identity, finish state, generation accounting, response envelope, and complete EPICURE trace replayed from the raw artifacts. It added 7 real Epicure-on answers, 39 arena comparisons, and 7 matched uplift pairs. The current development pools therefore contain 192 Epicure-on answers, of which 188 enter 915 anonymous comparisons, and 186 matched uplift pairs. The remaining 4 answers have no same-task peer and are excluded from the arena workload. Failed and incomplete attempts remain in reliability accounting. This correction supersedes the coverage snapshot, not the immutable operational record, and it creates no quality judgment.

The uplift count is an append-only successor to the initial 187-pair reconstruction. Release review placed one cross-source pair on hold. Both arms record the same execution-policy digest, model, provider, task, and prompt, but the normalized pair does not attest that every condition-specific protocol difference is an allowed treatment delta. The source records remain unchanged

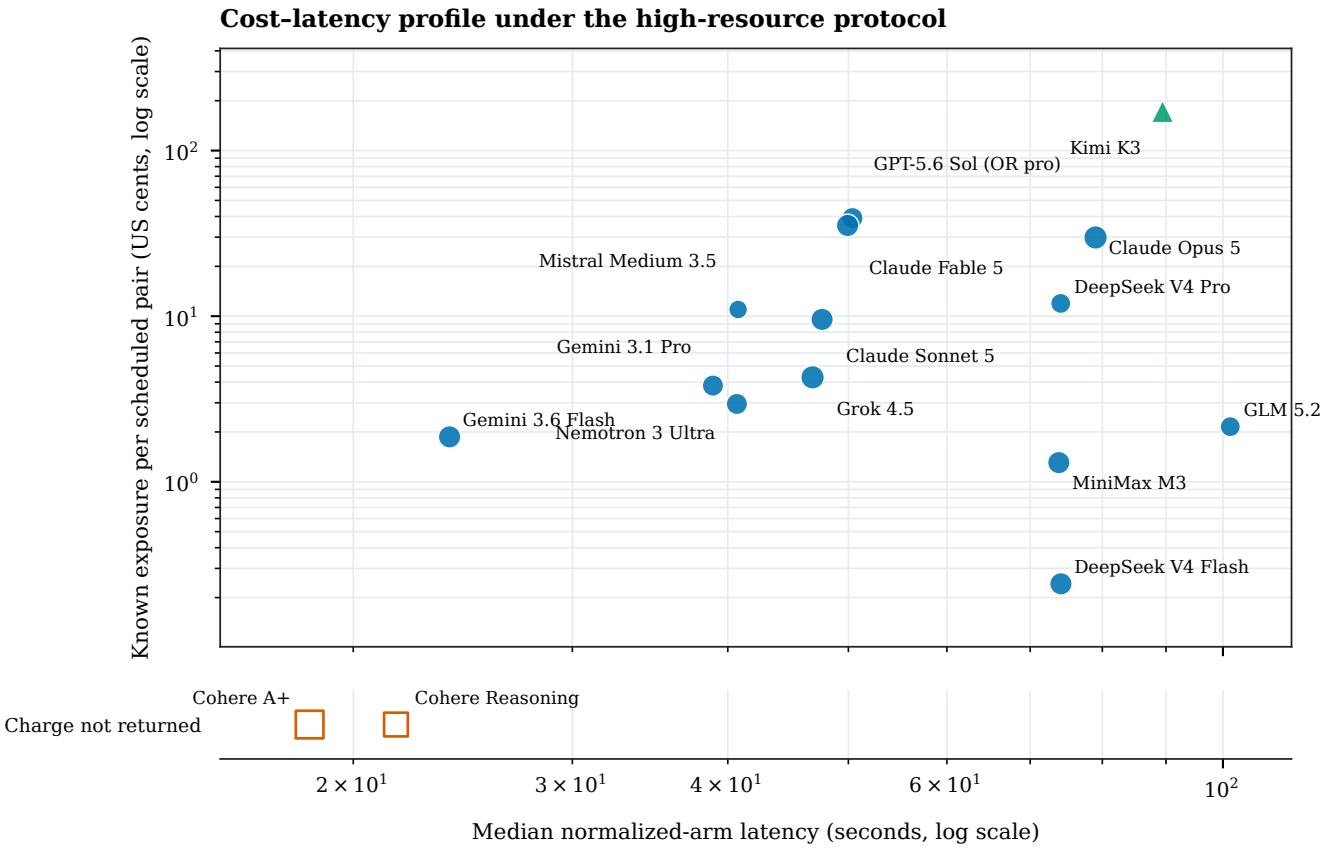

Figure 5: Cost and latency in the high-resource stratum. Each filled point is an exact endpoint with a known USD exposure. The horizontal axis is median latency over normalized response arms; the vertical axis is known exposure per scheduled matched pair. Both axes are logarithmic, and marker area scales with the number of complete pairs. The triangle denotes the direct Kimi route. Open squares in a separate categorical strip retain the measured latency of both direct-Cohere routes. They have no vertical cost coordinate because the provider charge was not returned. The figure combines model, route, output length, provider load, and tool trajectory, so it is not a causal provider or efficiency ranking.

Evidence unit	Gross	Held	Retained
Tasks in arena	24	4	20
Epicure-on answers	218	33	185
Arena comparisons	1,024	148	876
Matched uplift pairs	211	32	179

Table 3: Gross and retained units after the post-collection task quarantine. Excluded records remain in append-only operational storage and reliability accounting, but cannot enter a preference fit unless qualified adjudication supersedes the hold.

and continue to count toward operational reliability; the pair cannot enter an uplift fit without a formal pair-policy attestation.

The 915 candidate comparison rows reuse each compared answer between 2 and 14 times, with median 9. Consequently, a comparison row is not an independent sampling unit. The specified estimator standardizes across the four task families and uses a task-cluster by rater-cluster bootstrap. All rows containing a reused answer remain inside the same task cluster. Each answer belongs to exactly one task, so the task resample subsumes the nested answer clusters. This preserves the dependence induced by answer reuse while allowing raters to contribute across tasks.

Family support is still incomplete. 73 of 480 model-pair-by-family cells contain no comparison: 3 in composition, 20 in cookability, 27 in evidence, and 23 in substitution. The graph is connected globally, but the least-supported endpoint-family cells rest on one task. These data cannot support a family-specific ranking. Another 146 cells contain only one candidate row, so 219/480 cells have at most one row. Additional coverage is admitted only prospectively, in endpoint-bounded batches, and only observed usable arms may change these counts.

For each assigned comparison, task validity is sealed before either answer appears. The reviewer then scores task completion, constraint compliance, coherence, sen-

sory promise, cookability, clarity, originality, evidence use, and calibration, followed by left, right, tie, or both-bad preference. Model and condition identities remain hidden until submission. The uplift and model-arena ballots are stored as different tracks and the strict and high-resource execution strata remain explicit. At the evidence cutoff, both pools contain zero judgments admitted to the research analysis. Their hashes commit the candidate sets and side layouts, but they do not constitute a leaderboard.

4 Prospective Season 1 contract

The retrospective audit exposed three avoidable sources of ambiguity: incomplete item review, post hoc admission choices, and a tool runtime that can now be reconstructed technically from public inputs but still lacks recovered training lineage, attested payload rights, and an independent reproduction. Season 1 therefore has a content-addressed study design that predates scored collection. It prohibits pilot records and generated test fixtures from entering any leaderboard. Passing a route contract also remains separate from measuring answer quality.

Tasks and splits. The confirmatory bank requires 240 newly written tasks, 60 in each family. Each family is divided into five construct cells with 12 tasks per cell: eight scored, two development, and two private reserve. At least 20 verified people must contribute tasks. For each item, the author, two source validators, adjudicator, validator reviewer, and contamination reviewer are 6 distinct people. A source validator first solves and classifies the prompt without seeing the author pack, then reconciles the sealed solution against the author's construct labels, constraints, and validator plan. The adjudicator sees both source records but no model output and freezes the criterion pack later shown to response raters. Separate qualified reviewers reproduce the executable-validator and contamination receipts. The bank therefore requires 1,680 human admission records: 480 prompt-only solutions, 720 reconciliations and adjudications, and 480 evidence inspections. At least 96 tasks must have item-specific executable constraint diagnostics. These diagnostics test prespecified surface constraints and are reported separately; they do not establish culinary correctness. Their blind mutation suite must reach 0.95 precision and 0.90 recall before use. Exact, fuzzy, word-ngram, deterministic semantic, and captured-web overlap checks are replayed from one content-addressed corpus; a high-similarity match rejects the item. The detector must first pass 150 independently labeled exact-copy, paraphrase, and unrelated cases, reaching at least 0.95 precision, 0.90 recall, and 0.85 paraphrase recall. The bank is rejected if cell or difficulty quotas, author concentration, six-person role separation, rights records, lifecycle ordering, contamination replay, or specialist exclusions fail. After release, a public content-addressed challenge requires two qualified adjudicators who held no role on the original task; a confirmed error retires the item and forces snapshot recomputation.

Post-collection validity and robustness. Model behavior can expose ambiguity that precollection review misses. Before release, two new reviewers therefore inspect every anomaly-flagged task and a seed-committed random sample of at least 60 tasks. They receive the prompt, criterion pack, validator, response and tool traces, and observed disagreement patterns. A material defect blocks release until the task is retired and all affected snapshots are recomputed. Separately, 24 cookability tasks contribute one preregistered blinded output each to 48 kitchen executions, two independent cooks per output. Completion, elapsed time, deviations, yield, acceptability, and their association with the review rubric test construct validity; these measurements do not enter the preference ranking.

Runs and identities. The controlled collection fixes 3,200 model-arena battles and 3,200 paired Epicure contrasts, for 12,800 real response arms. The arena gives every endpoint at least 400 appearances; the uplift track gives every endpoint at least 200 pairs and 50 pairs in each task family. Each ranked row binds a dated model identifier, one provider endpoint, a decoding contract, and the released Epicure package. Fallbacks and provider substitution are not rank eligible. Failures remain in the assigned denominator and never become preference votes.

A stratified 20-task subset is run 3 times for every endpoint and condition, yielding 1,920 panel arms, of which 1,280 are additional to the primary collection. Provider retries are not repetitions. Pass-at-1, all-three-valid reliability, within-endpoint variance, pairwise outcome consistency, and tool-trajectory consistency are reported separately. A development-only audit adds 480 real arms across three frozen prompt variants and a provider- and family-stratified eight-endpoint subset. Variant selection after results is prohibited, and this audit cannot alter the primary ranking. Across primary and supplemental studies the design schedules 14,560 real model arms.

Human measurement and analysis. The primary expert cohort requires at least 800 unique comparisons in each track and 2 independent, qualification-matched raters per comparison. Public arena votes form a separate observational cohort. Model-arena estimation uses the frozen `arena-rank` 0.1.1 weighted BT fit and 5,000 family-stratified task-cluster bootstrap replicates [37]. The paired Epicure analysis reports a family-standardized half-win share, assigned-arm completion, conditional preference, both-bad frequency, and best and worst missing-preference bounds. Preference, constraint pass rate, reliability, latency, and cost remain separate outcomes.

The official global arena fit is withheld unless all 160 scored tasks are admitted, with 40 in each family; every endpoint has at least 20 distinct task clusters in

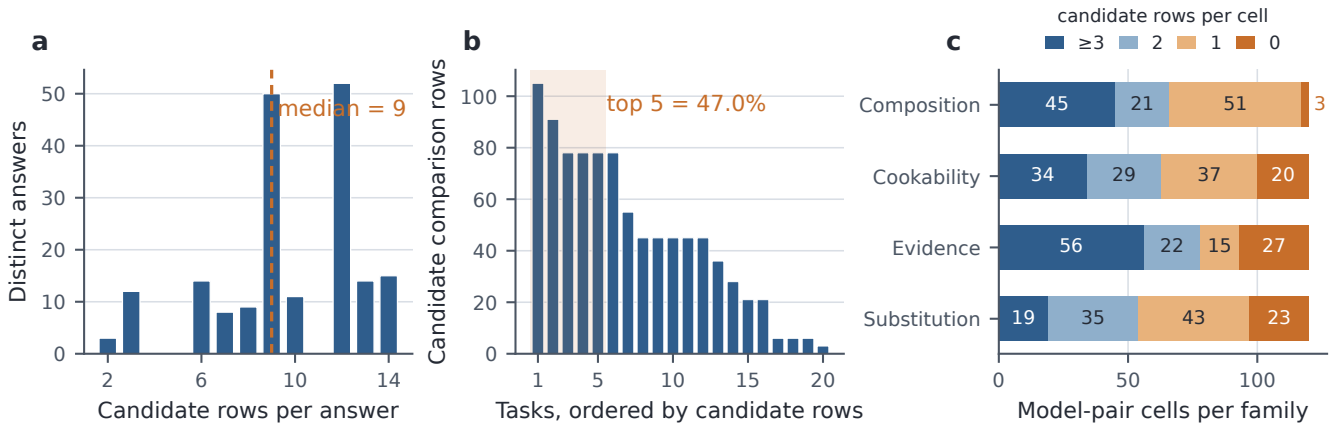

Figure 6: Effective support of the corrected real-output arena candidate pool. Panel (a) shows how often each of the 188 compared answers recurs across the 915 candidate rows. Panel (b) orders the 20 comparison tasks by candidate-row count; the five largest tasks account for 430/915 rows. Panel (c) separates the 480 model-pair-by-family cells by support depth: 73 have no candidate row, 146 have one, 107 have two, and 154 have at least three. The counts describe a review workload with zero admitted quality judgments, not independent observations or model scores.

Scenario	Truth	Mean	Coverage	Check	Rate
Null with 20% ties	0.500	0.500	95.2%	type I	4.8%
+0.10 half-win effect with 20% ties	0.600	0.600	94.7%	power	88.9%

Table 4: Finite-sample checks for the family-standardized paired estimator and its identifiable two-endpoint BT subproblem. Each scenario uses four families with 50 task clusters per family. These Monte Carlo records test the analysis code only. They contain no prompts, model answers, Epicure effect observations, or leaderboard evidence.

every family and at least 100 unique comparisons overall; every comparison has two independent raters; the global and four family graphs are connected; and at least 99% of global and family bootstrap resamples remain connected. Any unresolved material task defect also withholds the fit. A family-specific fit uses the same 40-task, 20-cluster, 100-comparison, two-rater, connectivity, and 99% resample floors. Pair-specific intervals are suppressed below ten shared task clusters. Insufficient layouts return `withheld_insufficient_task_clusters` with no ratings or intervals.

Method checks. Before model collection, the estimator primitive was tested on 2,000 Monte Carlo datasets per scenario with 999 task-cluster bootstrap replicates per dataset. Under a symmetric null with 20% ties, interval coverage was 95.2% and the two-sided type I error was 4.8%. For a prespecified 0.10 increase in half-win probability, coverage was 94.7% and one-sided power was 88.9%.

This balanced two-endpoint check is not the production-layout gate. The frozen 16-endpoint gate requires 16,000 simulated method-validation datasets across eight adversarial scenarios, 5,000 clustered resamples per dataset, exact duplication and sparse-graph checks, 93–97% interval coverage, 4–6% type I error, at least 80% power for a 50-Elo shift, and probability-scale absolute bias below 0.02. Its current status is `required_not_yet_executed`. No official Season 1 ranking may be published until the complete content-addressed result passes and reproduces.

Distributed-contract revision 2 binds the same estimator source bundle and explicitly forbids a surrogate implementation. It replaces an over-broad build hash with a narrow projection of the exact worker sources, dependency lock, image definition, interpreter, installed versions, and thread settings; revision 1 remains preserved as retired evidence. The 16,000-dataset manifest was materialized deterministically. A bounded Modal measurement stopped locally for missing authentication before cloud submission, so no cloud shard, simulated dataset, or clustered resample has run. The aggregate therefore remains empty and cannot be cited as inferential validation or model-quality evidence.

An official snapshot must reproduce this check, pass the disconnected-graph rule, bind all raw records and analysis sources into a signed deterministic archive, and state the exact task, endpoint, provider, prompt, schema, and Epicure hashes. The same archive must contain and name the hashes of the post-collection item audit, reliability panel, prompt-sensitivity audit, and kitchen execution study. Until the task bank, public Epicure release, independent ratings, and real collection exist, Season 1 has a protocol but no quality result.

5 Retrospective pilot design

5.1 Task sample

The source is Seasoned Advice, the cooking site in the Stack Exchange network. Candidate questions were re-

quired to have an accepted forum answer and metadata sufficient for attribution under the applicable CC BY-SA version [36]. Two disjoint pools contained 480 historical and 499 newer questions. Claude Sonnet 4.6, requested through Amazon Bedrock as `global.anthropic.claude-sonnet-4-6`, and Gemini 3.1 Flash Lite, requested through OpenRouter as `google/gemini-3.1-flash-lite` on `google-vertex/global/flex`, independently scored all 979 candidates without seeing accepted-answer text. They assessed specificity, difficulty, family fit, multi-step reasoning, self-containment, Epicure relevance, and specialist risk.

Admission required exact family agreement, two judgments, difficulty at least 2 from each curator, family fit at least 3 from each, mean Epicure relevance at least 2, specificity at least 2 from each, self-containment, and no specialist-risk flag. The two pools yielded 246 and 216 strict-consensus candidates. Exact-family agreement was 0.769 and 0.816; Cohen's κ was 0.708 and 0.744. Within each family, candidates were ordered by mean curation score, number of multi-step votes, accepted-answer site score, question site score, recency, and source ID. The deterministic top 30 formed each intended family: substitution, composition, cookability, and evidence interpretation.

The accepted forum answer was stored and later shown to automated judges as an orientation aid. It was not an automatic ground truth. Showing accepted forum answers may have introduced reference anchoring. None of the 120 principal-investigator item-review rows was complete when generation began. The current scope is therefore public English culinary troubleshooting, not general culinary expertise.

5.2 Historical endpoints and routes

The historical manifest contains 12 canonical endpoints, listed in Appendix D. The roster assigned four slots to proprietary model families, four to open-weight families, two to efficiency-oriented endpoints, and two to reasoning-oriented endpoints; it is neither a probability sample nor a universal frontier roster. Seven routes used Amazon Bedrock and five used OpenRouter. OpenRouter routes fixed one provider endpoint and disabled fallback substitution. All 1,140 collector-accepted OpenRouter arms returned the expected dated canonical model identity and provider name under the fixed requested route. Those canonical identities differ from the human-facing request aliases and are reported in Appendix D. Bedrock records retain the requested endpoint and hashed AWS request identifiers; the Converse responses did not return a model identity, so they establish contracted route identity rather than the physical model that executed the request. The historical transport and retry limitations are reported in Appendix A.

All arms received the same base instruction, task text, and Markdown answer contract. Tool-available arms also received the Epicure treatment instruction and at most eight tool rounds. The intended repair rule allowed one model turn after an invalid tool call. The executed collector instead used one arm-wide error counter and stopped after the second error-marked MCP result. With parallel calls, that stop could occur before the model received the first error. This protocol deviation is an observed property of the intervention, not a model-only failure. A target arm was never replayed after possible delivery under the same generation identity. All target failures remained in the reliability denominators. We call a target arm *completed* when its reconciled, rank-eligible record contains a non-empty final answer and the provider finish reason is `stop`, `stop_sequence`, `end_turn`, or `completed`. The historical collector instead accepted any non-empty answer. A post-run audit therefore reclassified 267 arms: 202 ended at `length`, 63 at `max_tokens`, and two at `content_filter`. The raw records are unchanged. The derived correction excludes those arms from preference fitting while retaining their outputs, traces, latency, and cost as operational evidence.

5.3 Epicure intervention

The treatment was the *offer* of access to a fixed 13-tool Epicure MCP contract. Tool-available arms could inspect ingredients, similarity evidence, substitutions, and composition relations, but were not forced to call a tool. Tool-unavailable arms received the same task and final-answer contract without those tools. Within each endpoint-task pair, both conditions requested the same route and decoding parameters.

This is a package estimand. It bundles the tool schema, the Epicure representation, tool-selection behavior, returned evidence, and synthesis. The executed runtime is `exploratory-unmatched-1790-runtime`. A recovery inventory now enumerates all 11 bundled data files, the 1,790 by 300 embedding shape, 28 installed Python source files, the observed Python environment, and the exact tool contract. A loopback runtime attestation matches bundle SHA-256 `98d0403115bf8eb4fb71dbb89a53362e9b9acafda7494c464763df7d764174d1` and application SHA-256 `be4216ae799f330b76f1f8c009ad2e79ce891d5dec452854dbc449e66e2eb313`. The observed 41-package runtime environment is now pinned to selected wheel hashes and described by a CycloneDX 1.5 software bill of materials. A clean private reinstall with no package index reproduced the same dependency payload, data-bundle identity, and application identity. This same-operator reconstruction does not recover the original training run, establish payload rights, constitute independent reproduction, or supply an immutable image. All 28 exact application files are now preserved in a deterministic source-only archive under the existing MIT licence. Its SHA-256 is `d08fb475e9c325a8c41daf5b789e6b4bca547228139eece4578f9b06c324703c`; it contains no runtime data, model payload, dependency wheelhouse,

credentials, training corpus, or new licence. All 11 exact data files are publicly readable at immutable Git commit `14ddf04aba81a76b75efa6554041f6bff48992c6`; a per-file byte and SHA-256 audit matches the frozen runtime manifest. Public availability is not evidence of a payload licence or upstream source rights. A later 63,814-byte public-input kit binds that source archive, the immutable data map, the 41-wheel lock, SBOM, tool catalog, rights boundary, and fixed parity probes. In a clean CPython 3.12/Linux x86-64 workspace, the study operator fetched all 11 data files from the frozen Git commit and all 41 exact wheels from the public PyPI index, installed the wheels offline, verified their declared RECORD hashes and the frozen importable-payload aggregate, and reproduced the application and bundle identities plus the fixed probes. The kit, execution receipt, and release-candidate records are content-addressed as `bc5b109840c3c7d468fdc050eff7b907b66a0bc3a31899bd4c39ea8bcb1ae4b6`, `e1140807dfd2feb0286318f6ab6a8a60246273b06b6f06b14ca7baccf2750cef`, and `1757620645fabc9f758152315b7c09a802143f5790db58b5844b647f8ca22e59`. This establishes technical reconstruction from public inputs on the frozen target. Operator independence was not adjudicated; the receipt explicitly records `independent_reproduction: false`, `payload_redistribution_cleared: false`, and `rank_eligible: false`. A subsequent private-repository audit recovered the May 2026 payload-import commit and a plausible Cooc run identifier, job specification, default seed, training implementation, and graph-source record. It did not establish identity: the candidate export was recorded from a dirty worktree, its source embedding is absent, and its recorded cosine extrema and mean differ from the deployed matrix. We therefore reject that candidate as the exact lineage rather than relabelling the runtime. The public archive now includes a portable, content-addressed reconstruction packet and a standalone fail-closed verifier. An independent reader can check the archive metadata, runtime identity links, dependency lock and SBOM, private rebuild receipt, and rejection of the candidate training lineage without network access. Packet v3 verifies every member of the exact application source archive and binds immutable raw URLs, byte counts, and hashes for the exact data files; optional offline-directory and network-stream modes verify those data bytes. The research archive does not mirror the data, wheels, or later public-input kit. The payload and source-rights matrix, original training input and full corpus lineage, signed clean release, matching immutable image, public runtime provenance, and independent reproduction receipt are also absent. The evidence therefore supports exact attribution, technical reconstruction from public inputs on the frozen target, and independently verifiable archive integrity. It does not support independent reproduction, payload redistribution, or official ranking. Epicure similarity values are therefore explanatory inputs, never answer-quality labels.

5.4 Execution record

Target collection ran from 2026-07-16 12:18:25 UTC to 14:45:27 UTC. Automated judging and its single permitted recovery pass ran from 2026-07-16 14:48:24 UTC to 2026-07-17 03:05:09 UTC. Target work items were ordered by a deterministic SHA-256 scheduling key rather than randomized at dispatch. Within the 1,440 paired cells, the tool-unavailable arm started first in 749 and the tool-available arm started first in 691. This count balance does not remove possible time or provider-load effects. Target arms used a 2,048-token ceiling and temperature 0.2 where the endpoint supported it; otherwise the provider default in force at execution was used and its semantic value was not recorded. Judge calls used temperature 0, an 8,192-token ceiling, and provider-enforced structured output. Exact instructions, prompt construction, retry behavior, recovery limits, normalization, and SHA-256 identifiers are given in Appendix A.

5.5 Comparison tracks

For each of 120 tasks and 12 endpoints, the collector attempted one tool-available and one tool-unavailable arm, giving $120 \times 12 \times 2 = 2{,}880$ target arms. The paired tool-availability track planned 1,440 within-endpoint comparisons. The endpoint-comparison track planned 720 comparisons among tool-available arms using a fixed balanced pair schedule. Left/right placement was the parity of SHA-256 over the frozen track, task, and model or condition identity; the resulting comparison manifest is reported in Appendix A. Model and condition identity were absent from judge prompts. One otherwise successful answer named Epicure. Its occurrence in one endpoint comparison and one paired tool-availability comparison caused two exclusions before judging.

6 Judging and estimation

6.1 Automated panel

Four Bedrock-hosted judge endpoints were assigned each available pair: Claude Haiku 4.5, Claude Sonnet 4.6, Devstral 2 123B, and Qwen3 Next 80B. Each assignment comprised two calls, one in each orientation. Successful calls returned a choice (left, right, tie, or both bad), failure tags, and nine rubric scores on a 1 to 5 scale, defined in Appendix G.

A judge vote entered the primary cohort only if both orientations completed and mapped to the same normalized choice. A judgment was excluded only when the judge was the exact target endpoint; other members of the same canonical model family remained eligible. Consensus required a strict majority among at least two eligible endpoint-excluded votes; a two-vote agreement could

therefore constitute consensus. The sensitivity analysis first collapsed unanimous votes within each canonical judge family and then required a strict majority across families. The analysis contains no human ratings. Claude Sonnet 4.6 served as curator, target endpoint, and judge; its exact-endpoint self-judgments were excluded, but those overlapping roles and same-family dependence remain design limitations.

Hereafter, *primary* names the main retrospective specification. It does not imply preregistration or prospective approval.

6.2 Estimands and uncertainty

We fit the retrospectively specified global `arena-rank` 0.1.1 BT model with sandwich intervals and counted ties as half-wins. We retain that output as a software-compatibility diagnostic; its sandwich intervals do not cluster by task and are not the substantive uncertainty statement [37]. For a task-resampling stability analysis, we resampled the 89 represented tasks 1,000 times with NumPy `default_rng` seed 20260721 and refit a ridge-regularized BT model with $\lambda = 10^{-6}$ in each replicate. `both_bad` was excluded, and all replicate graphs were not assumed connected: 985 replicates were fitted and 15 were withheld. Figure 8 reports medians and percentile ranges in frozen manifest order. Rows with fewer than 100 observed comparisons are labeled provisional; this is a reporting convention, not a preregistered error guarantee. Under complete separation, the unpenalized BT maximum-likelihood estimate is unbounded.

The primary paired summary is the complete-case, cell-weighted, tie-adjusted automated-panel preference share

$$\widehat{p} = \frac{W + T/2}{W + T + L}, \tag{1}$$

where W is a tool-available win and L a tool-unavailable win. Cell weighting means that a task or endpoint with more surviving contrasts contributes more. Its interval resamples the 102 observed tasks 5,000 times with `default_rng` seed 20260716 while retaining all surviving endpoint contrasts within each sampled task. Two post-run weighting sensitivities use seed 20260722: one averages the 102 task means equally, and one averages the 12 endpoint means equally within each task resample. The latter retained 4,967/5,000 resamples in which every endpoint was represented.

The 120 tasks and their retained subsets were selected deterministically rather than sampled from a defined population. We therefore call all bootstrap ranges in this paper *observed-task resampling intervals*; they are stability summaries, not design-based population confidence intervals.

Both estimands condition on two completed, reconciled, rank-eligible, identity-leak-free target arms, followed by a surviving judge consensus. They are not end-to-end effects under treatment-dependent failure. We omit rubric-dimension and task-family inference because the items and scoring criteria have not received qualified validation and no multiplicity-adjusted analysis was planned.

We therefore report a separate paired reliability estimand over all 1,440 endpoint-task cells: the tool-available minus tool-unavailable difference in realized target-arm success proportions. Its stability range resamples the 120 tasks 5,000 times with seed 20260723, retaining all 12 fixed endpoints within each sampled task. As above, this is an observed-task resampling interval.

We also report a bounded-score completion sensitivity for the planned paired population, following the logic of worst-case response bounds [21]. Let $Y = 1$ for a tool-available win, $Y = 0.5$ for a tie, and $Y = 0$ for a tool-unavailable win. Cells that fail before two answers exist have no defined answer-preference outcome; exclusions and judge non-consensus also leave no observed score. The sensitivity assigns every such unresolved or undefined cell a hypothetical score in $[0, 1]$. No `both_bad` result entered the observed paired cohort. With $S = W + T/2 = 103.0$, $n = 241$ observed scores, and $N = 1{,}440$ planned cells, this completion rule gives

$$\left[\frac{S}{N}, \frac{S + N - n}{N}\right] = [0.072, 0.904]. \tag{2}$$

This is a sensitivity range, not a confidence interval or an identified answer-preference estimand. It shows how little the observed complete cases constrain a hypothetical all-cell score.

7 Historical task validity status

The bank contains 120 distinct public source IDs and 120 distinct retained prompt hashes. No task has a completed qualified review, independent approval, task-specific rubric, or executable validator. Item `fb-s0-cookability-003` depends on images absent from the retained question; the source page displays six images. After the completion correction, the item contributed 15 admitted comparisons and seven primary consensuses, including five paired-tool ties; Figure 9 reports the leave-one-task-out result. These records support an audit of the executed pilot, but they do not validate its task set.

Contamination and representativeness are also unresolved. All 120 questions are public Stack Exchange posts, 111 predate 2025, and all come from one English website. Accepted-reference authorship is concentrated: the five most frequent authors wrote 48/120 references. The study population is a convenience sample of Seasoned Advice troubleshooting, not global culinary reasoning.

8 Retrospective pilot results

8.1 Target and judgment attrition

Of 2,880 attempted target arms, the collector accepted 2,663 and rejected 217. After the finish-reason audit,

2,396 met the completed-arm definition and 484 failed. The task bank comprises retained public forum prompts. All reported response-arm and judgment counts derive from recorded provider executions. The traces contain 1,387 Epicure calls, including successful calls made before a later arm failure. Of 2,160 planned comparisons, the source manifest sent 1,773 pairs to judging after 385 failed-arm and two identity-leak exclusions. The completion audit then removed 287 of those pairs, leaving 1,486 in the effective cohort. The judging workload created 14,184 terminal judgment identities and 15,839 provider-attempt records, but those operational counts are not the inferential sample size. Only 393 comparisons entered the primary analysis (26.4% of the effective cohort), split into 152 endpoint comparisons and 241 paired tool-availability comparisons (Figure 2).

8.2 Judge completion and orientation agreement

Completion of both orientations ranged from 0.5% for Qwen to 96.4% for Haiku (Figure 7). Conditional agreement ranged from 61.0% to 100%; Qwen's 100% corresponds to only 7/7 completed pairs. Overall exact-endpoint-excluded eligible-vote yield ranged from 0.4% to 58.8%. Qwen supplied only six eligible votes, chiefly because few pairs completed twice. Of the 393 primary consensuses, 301 had exactly two eligible votes. Haiku and Sonnet belong to the same model family, so ordinary majority can overrepresent shared family behavior. Agreement among surviving votes does not remove this selection effect.

The longer answer was selected in 74.9% of 239 directional consensuses with unequal lengths. The median absolute difference was 122 words. This association may reflect answer quality, verbosity, style, or their interaction. The current data do not separate these explanations.

8.3 Uncertainty in endpoint comparisons

Observed primary-consensus comparison count n ranges from 4 to 41; no endpoint reaches the 100-comparison reporting threshold. Fifteen of the 1,000 task resamples produced a disconnected graph and were withheld. Devstral's 0/0/41 record is complete separation, so an unpenalized finite maximum-likelihood distance does not exist. Global connectivity does not resolve sparse uncertainty or separation. The sample does not resolve an endpoint ordering. The numerical output of the specified compatibility estimator is reported in Appendix E; complete separation prevents interpreting it as a resolved ordering.

8.4 Weighting and admission sensitivity

The retrospective primary cell-weighted rule selected 241 paired comparisons, with W/T/L 31/144/66 and automated-panel preference share 0.427 [0.388, 0.466]. Surviving cells ranged from one to seven per represented task and from five to 42 per endpoint. Equal weighting over the 102 represented task means gave 0.418 [0.370, 0.466]; equal weighting over the 12 represented endpoint means gave 0.487 [0.450, 0.521]. Thus the magnitude and apparent separation from 0.5 depend on the target weighting.

The post hoc cross-family rule retained 133 of the primary cohort's 241 comparisons and admitted no new comparison. None of the retained choices changed. Their W/T/L was 16/102/15, giving 0.504 [0.465, 0.545]; the 108 excluded comparisons had W/T/L 15/42/51 and a tie-adjusted share of 0.333. The estimate changed because the rule selected a strict subset, not because any retained decision changed. Within the primary cohort, 73 paired comparisons used only Claude Haiku and Claude Sonnet votes. Their W/T/L was 15/7/51 (share 0.253); the remaining 168 comparisons had W/T/L 16/137/15 (share 0.503). This descriptive contrast shows that judge composition is consequential. Removing the image-dependent item left 236 paired comparisons, all five removed outcomes having been ties, and changed the cell-weighted estimate from 0.427 [0.388, 0.466] to 0.426 [0.386, 0.464]. Task-family slices are omitted because the labels had not undergone qualified content validation.

The 241 observed preference scores sum to 103.0. If each of the other 1,199 planned paired cells is allowed any score in $[0, 1]$, the planned-cell mean lies in [0.072, 0.904]. A mean score of 0.515 among those missing cells would move the completed score to 0.5. The observed complete-case preference therefore does not identify the direction of a planned-cell contrast.

8.5 Operational evidence

Tool-available and tool-unavailable arms had different realized completion rates. Among 1,440 matched endpoint-task cells, both arms completed in 1,062, only the tool-unavailable arm completed in 229, only the tool-available arm completed in 43, and neither completed in 106. Thus 1,291/1,440 tool-unavailable arms and 1,105/1,440 tool-available arms completed. The paired completion-proportion difference was -0.129 (95% observed-task resampling range $[-0.160, -0.100]$). The tool-available condition had a 12.9-point higher realized failure proportion (range 10.0 to 16.0 points; Figure 9). This operational difference is part of the as-executed intervention and includes collector policy, endpoint-tool interaction, and provider delivery. The preference analysis drops every cell with a failed arm. It therefore estimates automated-panel preference among surviving comparisons, not the end-to-end value of offering the tool. Tool use also varied sharply: two endpoints never called Epicure, whereas the highest observed offer-to-call rate was 80.8%. Assignment remains tool-available regardless of observed use.

The trace audit separates collector, interface, provider, and answer-delivery outcomes. Of 484 effective target-arm

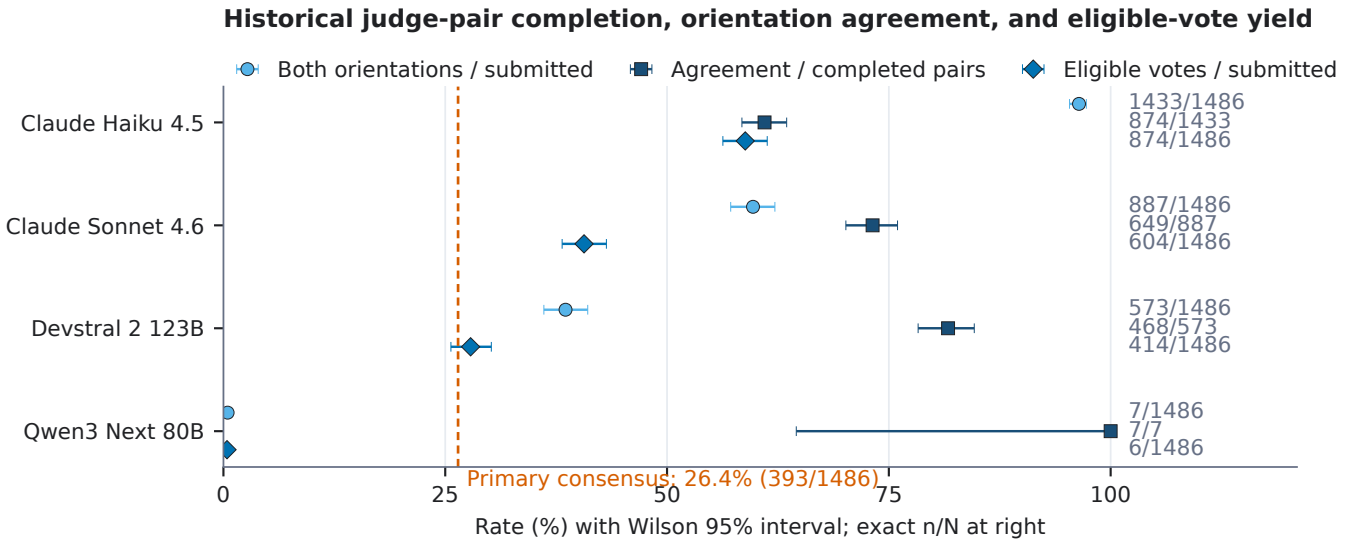

Figure 7: Historical 16 and 17 July automated judge-pair completion and admission over 1,486 effective comparisons. Completion and eligible-vote yield use submitted comparisons as the denominator; orientation agreement is conditional on both calls completing. Horizontal bars are Wilson 95% intervals; exact numerators and denominators appear at right. Marker shape and the legend distinguish the three rates. These are descriptive IID-reference intervals, not population or task-sampling intervals. The dashed line is overall primary-consensus coverage.

failures, 267 were non-normal final completions. Among the 217 failures already recognized by the collector, 156 were stopped after the second error-marked MCP result under the arm-wide counter, 10 reached the tool-round cap, and two exceeded the frozen eight-call-per-round fan-out cap. The remaining 49 comprised 20 provider-overload errors, 17 empty final answers, 11 read timeouts, and one cost-reconciliation failure. Of the 156 second-error stops, 95 recorded both errors in the same tool round and 86 recorded both in round zero; those calls had no intervening model turn. Across 1,387 Epicure trace events, 383 carried an error: 341 were ingredient or entity-resolution misses, 38 used an unsupported controlled-vocabulary value, and four violated a schema, numeric, or list constraint. Fifty-nine of the 227 arms with at least one error-marked trace still produced a successful final answer. The failure categories combine collector, interface, endpoint, and provider behavior.

Trace example. Substitution item 001 asked for a pantry substitute for sweet red vermouth in a cherry sauce. The Gemini 3.1 Pro tool-available arm queried Epicure for neighbors, then requested pairing scores for Marsala, sherry, and red wine. Marsala failed entity resolution; sherry returned 0.1604 and red wine 0.2399. These values were evidence, not reference labels. The provider returned 1,638 answer characters but ended at `length`. The collector marked the arm successful; the completion audit conservatively reclassified it as incomplete. This trace illustrates why tool success, answer quality, and final-answer completion are separate events.

The run recorded USD 222.47 of conservative target-and-judge exposure. Endpoint-level accounting, reliability, latency, and tool use appear in Appendix F [2, 3, 27].

9 Discussion

The exact-frontier study closes an operational currency gap: all 16 named endpoints produced real provider output and at least one valid Epicure-on answer. It does not close the measurement gap. Twenty-eight development tasks and no admitted research ballots cannot support a quality ordering. Requiring one successful call removes non-adoption from retained treatment answers, but the intervention still combines retrieval, representation quality, extra context, tool recovery, and synthesis. The frozen uplift and model-arena pools make the next measurement step explicit while failures remain in the assigned reliability denominators.

The historical paired design exposes two outcomes: whether a complete answer is produced and automated-panel preference among successful pairs. A complete-case preference estimate omits the operational asymmetry that dominated this run: 229 cells produced only a tool-unavailable answer, compared with 43 that produced only a tool-available answer. Preference and reliability must therefore be reported separately.

Automated-judge admission also changes the comparison cohort. Dual-orientation replication removed most Qwen judgments and substantial fractions of the other judges. Ordinary majority then allowed the two Claude judges alone to supply the votes required for consensus. The post hoc cross-family rule changed no retained vote, yet its estimate crossed 0.5 because it selected a different subset. Consensus coverage and judge composition are measurement results, not preprocessing details. Recent evidence that model-judge consensus may diverge from human judgment makes this distinction especially

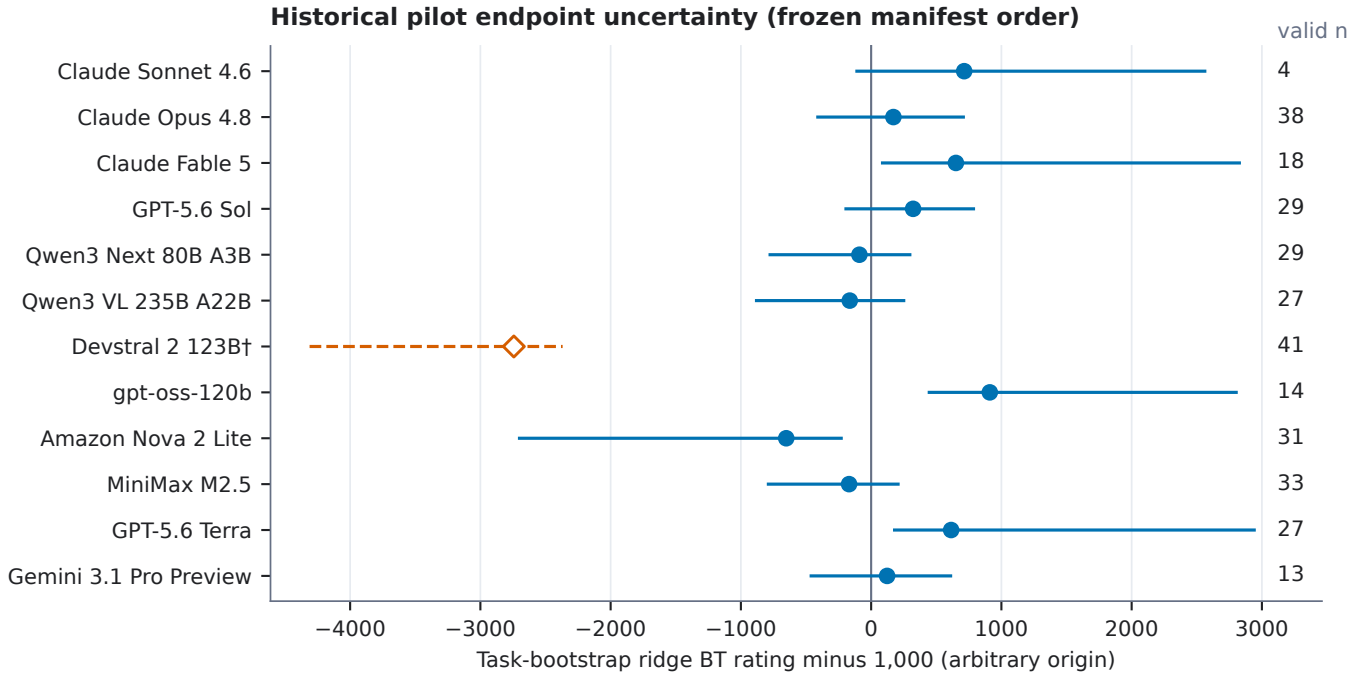

Figure 8: Historical pilot endpoint uncertainty, not a current roster or leaderboard. Each of 1,000 replicates resamples represented tasks and refits a ridge-regularized BT model with $\lambda = 10^{-6}$. Points are medians on the local ridge rating scale minus 1,000; its origin is arbitrary and larger values favor the row endpoint. Intervals are 95% observed-task resampling ranges, and rows follow the frozen endpoint manifest rather than an estimated rank. The open orange diamond marks complete separation: its unpenalized BT maximum-likelihood estimate is unbounded, so the finite ridge summary is regularization-dependent. This plot is distinct from the diagnostic `arena-rank` 0.1.1 fit with unclustered sandwich intervals.

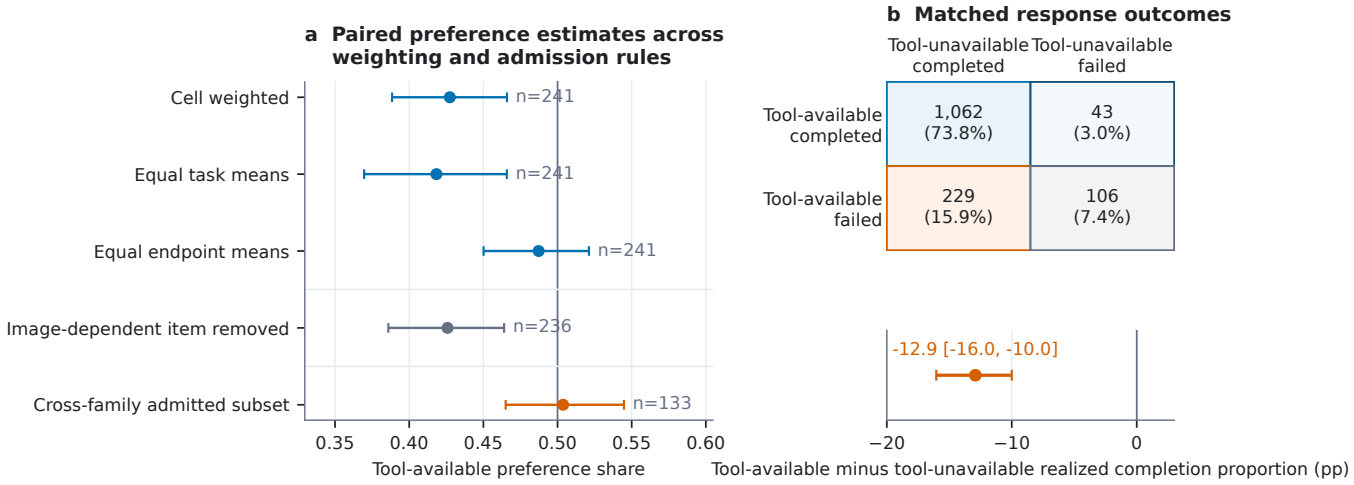

Figure 9: Preference sensitivity and target-arm reliability. Panel (a) reports the tie-adjusted automated-panel preference share for tool-available answers with 95% observed-task resampling intervals. The first three rows apply different weights to the same 241 comparisons; the fourth removes the image-dependent item; the final row is the strict cross-family-admitted subset, which changed no retained choice. Panel (b) gives the 2×2 target-arm completion table over all 1,440 matched cells and separately plots the tool-available minus tool-unavailable realized completion-proportion difference.

important [23].

The design also permits condition leakage through answer content. Judges did not receive model or condition labels, and explicit Epicure naming was filtered, but similarity values, evidence language, or tool-shaped prose may still reveal the tool-available condition. Accepted forum answers may introduce a second source of anchoring. A confirmatory study should randomize reference access and include reference-free judgments.

New scored items should be authored against a construct map, reviewed independently by two qualification-matched culinary experts, and adjudicated before collec-

tion. Each item needs explicit constraints, unacceptable outcomes, acceptable solution outlines, and objective validators where the construct permits them. Independent chefs or culinary scientists should also rate a stratified response sample. The author can form a separately disclosed sensitivity cohort but cannot establish independent criterion validity. A prospective protocol should seal task admissibility before model answers are shown, collect anchored response scores and an overall preference, and include concealed side-swapped repeats.

Following this audit, we froze a design contract for prospective task development (`7a63cfd6117338a3af16a422d5ee3458298fdc0ff2fd0abfe45fe851a7e54506`). It calls for 240 new human-authored tasks, 60 in each family, with 40 scored, 10 development, and 10 private reserve items per family. Admission requires two prompt-only source validations, author-pack reconciliation, an independent adjudication that seals the task criterion pack, independent inspection of validator and contamination evidence, and six-person separation within each task. The contamination detector has a labeled precision-and-recall gate, and confirmed post-release errors trigger append-only retirement and leaderboard recomputation. After generation, two new reviewers audit every anomaly-flagged item and a seed-committed random quarter of the bank; a material defect blocks release until retirement and recomputation are complete. The collection contains 3,200 controlled model-arena battles and 3,200 controlled paired uplift battles, or 12,800 primary response arms. A nested three-run reliability panel and a development-only prompt-sensitivity audit bring the planned real-model total to 14,560 arms. A separate blinded 48-execution kitchen study tests whether rubric cookability predicts practical execution; it is not pooled into model ranking. Independent culinary review covers at least 800 unique comparisons per track, each rated by two distinct reviewers, with a 12.5% concealed repeat schedule. Live public traffic is analyzed separately as an observational cohort. Stopping is fixed, and preference, response reliability, tool success, cost, and latency remain separate outcomes. These values describe the next study; no prospective Season 1 quality observation is included here.

Mechanism claims require additional arms. Matched retrieval, shuffled or null evidence, forced versus optional tool use, and answer-length controls would separate representation quality from extra context, tool policy, and synthesis. The prospective reliability panel estimates endpoint stochasticity but does not identify those mechanisms. None of those mechanisms is identified by the present paired tool-availability contrast. Prospective scoring should also distinguish appropriate contextual use of similarity evidence from correlation-to-mechanism and flavor-to-function extrapolations.

10 Limitations

The exact-frontier study has 28 development tasks, one response trajectory per endpoint-condition cell, and no admitted human quality judgment. Four tasks are now quarantined from both scoring pools. The corrected arena comparisons cover 20 tasks. The source registry contains 23 task IDs in 24 task–stratum clusters because it also retains 4 answers with no same-task peer. Its 320 scheduled matched pairs estimate neither general culinary competence nor an Epicure quality effect. The completion intervals describe a fixed workload rather than a sampled population. Endpoint cost and latency comparisons combine route, answer length, provider load, and tool behavior. Treatment completion requires a successful Epicure call, so incomplete pairs are part of the treatment outcome rather than ignorable missing data. The 915 arena candidates reuse 188 compared source answers between 2 and 14 times. The specified task-by-rater clustered bootstrap preserves those shared answers within task clusters. The graph is connected and every endpoint exceeds the 40-comparison workload floor, but 73 of 480 model-pair-by-family cells remain empty. The absence of admitted ballots still precludes a preference estimate.

The task sample is public, English-only, source-concentrated, and not independently validated. Model outputs may reflect memorization of questions or accepted forum answers. Judges saw accepted forum answers and may be anchored by them. There is no independent human criterion cohort, so the automated scores have no demonstrated culinary validity.

Both experiments observe one stochastic response trajectory and one normalized final-answer attempt per endpoint, task, and condition cell, so neither estimates within-cell generation variance. Provider routes differ across endpoints. The exact-frontier study uses 13 fixed OpenRouter routes, one direct Kimi route, and two direct Cohere routes; the historical audit uses five OpenRouter and seven Bedrock routes. Direct Cohere required a declared client-side required-tool translation, and Cohere returned no USD provider charge. The cost table and figure therefore mark those endpoints as unpriced. Neither experiment is a provider-replication design. The exact-frontier runs explicitly requested low reasoning effort for both intermediate and final phases. Comparisons among frontier endpoints must therefore be read as configuration-specific until the frozen default/high-effort sensitivity has completed and received blinded human ratings. A later QwenCloud smoke adds only one task and uses a catalog-observed mutable alias rather than an immutable model release. Its charge is unavailable, its USD 4 cumulative ceiling remains held, and it is excluded from the exact-frontier cohort and all inferential workloads. A fourth, source-reconstructing sensitivity contract now freezes Claude Sonnet 5, Gemini 3.6 Flash, and DeepSeek V4 Flash over eight non-quarantined tasks (two per family). It reconstructs 23 complete low-effort

pairs and retains one prior low-effort failure, then schedules 48 fresh default/high pairs under exact 12/12 order balance. The primary analysis is limited to an aggregate fixed-endpoint, fixed-task large-shift sensitivity; model and family cuts are descriptive. Its six diagnostic route-gate pairs and full study have not run, so this design adds no sensitivity result. Treatment-dependent target failure and judge survival make the main estimates complete-case quantities. The bounded-score completion sensitivity spans 0.072 to 0.904 and does not resolve a direction for a hypothetical all-cell score; a prespecified selection model remains future work. The arm-wide MCP error counter deviated from the intended repair rule and produced 156 of the 335 tool-available failures. The run estimates the executed collector policy rather than the intended one-repair-per-invalid-call protocol. The finish-reason correction was retrospective; although it is exhaustive and content-addressed, it is not a substitute for prospective enforcement.

The Epicure runtime is content-addressed at the bundle and application level. The arXiv source package includes a reconstruction kit, same-operator execution receipt, release-boundary record, and all 28 exact application files. The kit fetches the runtime data and dependency wheels from immutable public Git and PyPI sources and verifies their hashes; those payloads are not mirrored in the archive. The fixed tool probes reproduce the studied runtime, but no independent-operator receipt exists. The historical training record and payload-rights attestation were not recovered, so public readability does not establish redistribution permission. Formal nutrition, allergen, food-safety, and cultural-authenticity claims are outside the general-track scope and require specialist governance.

11 Ethics, conflicts, and availability

The study did not recruit or contact forum users. It reused public Seasoned Advice posts, and source attribution and applicable CC BY-SA metadata are stored per item. No institutional ethics determination is recorded. Model outputs and tool traces remain private pending review of provider-output and Epicure redistribution rights.

Josef Chen develops Epicure and FlavourBench, designed this study, and plans to offer FlavourBench evaluations commercially. His dual role as author and prospective evaluator, this commercial interest, and the absence of independent human criterion ratings are material to interpretation.

Historical model execution used AWS cloud credits and OpenRouter funds available to the author. The exact-frontier study retained a known USD exposure subtotal of $53.708. OpenRouter generation metadata supplied reconciled charges where available; direct Kimi usage used a frozen rate-card estimate, and unresolved deliveries retained their full admitted allowance. Direct Cohere usage retained returned token counts, but no USD provider charge was returned for 2 endpoints. The post-freeze QwenCloud smoke likewise returned no rate or charge. Its predecessor and successor each retain their full USD 2 admitted ceiling, for USD 4 of conservative exposure outside the exact-frontier subtotal; zero recorded provider cost is not interpreted as free usage. Direct Cohere runs used compute credits provided through the Cohere Scholars program. Cohere had no role in study design, task selection, data collection, analysis, interpretation, or publication decisions. No model provider reviewed the manuscript. No external human reviewer was paid for either pilot, and no external human rating enters this analysis.

A post-freeze reviewer-workspace QA batch was collected under a consent draft marked inactive. Those records are access-restricted operational evidence and are excluded from every analysis, figure, table, rating, and research release.

The analysis plan and item review were not prospectively approved before collection. We therefore classify the study as retrospective. This manuscript reports the audit as executed; the governance hold still blocks an official leaderboard and raw-output release under this run identity.

Analysis code, derived tables, plot data, dependencies, and content manifests are maintained in a content-addressed archive. This version has no public repository or archival identifier. Raw source posts and model outputs remain private. A public commitment enumerates the 379 normalized response files and 193 provider source records referenced by the current pools, recording both their embedded semantic identifiers and stored-file SHA-256 values. The commitment establishes membership and byte identity, but it cannot reconstruct the omitted records. The exact Epicure data bundle is publicly readable at its frozen Git commit but is not mirrored in the archive. The exact studied application source is included under its existing MIT licence. The archive also contains the public-input reconstruction kit and same-operator receipt. These verify technical execution on the frozen target but do not constitute independent reproduction or a rights attestation. The archive additionally includes a redacted QwenCloud operational projection and its zero-call recovery record. The predecessor and successor sources, journals, budget ledger, preflight, and human authorization remain private; the projection commits their stored-file hashes without distributing prompts, answers, tool payloads, or private identity fields. The manuscript and original figures are released under CC BY 4.0; omitted prompts, answers, tool payloads, and data files are not relicensed. Payload rights remain unresolved, and the candidate is not rank eligible.

12 Conclusion

The real 16-endpoint study establishes current route execution, paired answer reliability, required Epicure treatment,

and disclosed cost-accounting coverage. After a conservative four-task quarantine and source reconstruction, it contributes a workload of 186 complete uplift pairs and a connected 915-comparison model-arena pool built from 188 compared answers. It contains no admitted quality judgment. The retrospective pilot identifies neither an endpoint ordering nor an Epicure answer-quality effect. The tool-available condition had a 12.9-point higher realized failure proportion under the executed collector, while automated-judge admission retained 26.4% of the effective cohort. A post hoc family rule moved the complete-case preference estimate across 0.5 without reversing a single retained choice, and the bounded-score completion range spans 0.072 to 0.904. These results place target failure and judge admission inside the measurement problem.

The two scoring pools now separate the within-model Epicure contrast from cross-model comparison. Real ballots can support provisional estimates once task validity, reviewer qualification, repeat agreement, and minimum-exposure checks pass. They cannot repair failed generations or turn 28 public development tasks into a general measure of culinary competence.

Season 1 turns these findings into release gates: reviewed hidden items, independent human criterion ratings, calibrated constraint diagnostics, a corrected tool loop, reproducible Epicure lineage, and a signed analysis archive. Until those gates pass, current endpoint records remain development evidence rather than a quality leaderboard. Tool-augmented comparisons should report answer reliability and judge survival alongside preference.

A Prompts and identifiers

All digests below are SHA-256. JSON contracts were serialized with sorted keys and compact separators before hashing; text prompts were hashed as UTF-8 bytes.

Record	Frozen value
Curation system prompt	`e8cff45ea0aac9f2b01d5c3057aaaa033600dc2cf655f7034fe5a6c40822652a`
Curation record schema	`flavourbench-bedrock-task-curation-batch-v1`
Target window	2026-07-16 12:18:25 to 14:45:27 UTC
Task-bank artifact	`1ce969bdee4124fa44bab46a04feda2a0ebeddf4d37c49c0264b48b3833a4313`
Model-manifest artifact	`3919def66686b4bd939c94cdd89659f63ae2afbbf03288413129e2ea8d6b83d2`
Epicure-intervention artifact	`c4221c29a951fed1848a7b6dd56c88d351386fc75ac76d2313ce51952e1ab842`
Tool-unavailable system prompt	`c93d5c57f2e123d8462cec4c95752c6010c5170450af0a48b1abc6f32a6319c0`
Tool-available system prompt	`b841f852ebc0381ecee629d3dca48bde90a8a8ce7222a7ab56e6733e8edb66b0`
Execution contract	`49dd4d361bbc5f3cd8a6cde6ac49d5c4bd5ed9c108038f936d1ddca9dd023b31`
Tool catalog	`666a9a44b3534d7c8321f179e4513f71e0acf0d281a01fb3f41be5f8c0dfc8dd`
Arm correction	`e249141ed7aa64cb29075e6ee5f542ec7ba70b6211395e510ccd14768253434b`
Completion correction	`d6c2b25485ccdd88c347239f786c7d91c563243e95c117e5d1a9988923a0ea50`
Comparison/side manifest	`c6e9052d19737b39b540dafbd0cea53d1dd0c54b1a04584fd3775ddfe9f35ca7`
Judging window	2026-07-16 14:48:24 UTC to 2026-07-17 03:05:09 UTC
Judge-manifest artifact	`a3b2684774f7da1837b06ec883ffa45876c3afeb3b2290a1d289dd390ff81662`
Cost audit	`f3f08ec316f9aa059368c17244fe0e9199ead464c4a7799acd453461d352d32a`
Analysis artifact	`ab45eff77098a97fc05ef7ee5ca689b00724381e4bd8c6f7e4dd60c86fb61d97`
Judge system prompt	`dc51ec9d7a2ca8854f7d2fce3fcd50fba13ae7b66901c195ddcdd9228d481c3b`
Judgment schema	`aadb1201e6067af8715766040615c77bce912e9cbca73d92b32ddfa0b6dc5ad1`
Season 1 study design	`7a63cfd6117338a3af16a422d5ee3458298fdc0ff2fd0abfe45fe851a7e54506`
Season 1 construct map	`83e5b115ddcac3cb4622ede99d0b39d870b6617425568b3248818bc8f6368b19`
Season 1 method validation	`0b4345e523fdaa97d1b406cd1f2165540d0f9ad338bb49f3ac656da73e3c1933`
Season 1 arena acceptance	`bdc0fa93c6365cdcd45694d1d5500d82ccbd622f3be897be9217e252855ffff5`
Arena engine source bundle	`382151bc013d19240873de764f54ecab90bc554f663c44b2c3ca405c8975943a`
Production-layout MC contract	`2ba9f4d5b9b2f5a231edf085ea55b6ab67097780c12be86868082dbbdac95351`
Distributed-execution contract	`64fd29ec8755a09504ba251252f05c07526aa167fe88b8f53ece84dbb9a2b10b`

Current record	Frozen value
Route registry	b300d460ec3d93dbfdaea64e0809abf858fa9efb57 0d0bddeac28566b6cdf010
Strict aggregate	e1540a2eb9f7...
High-resource aggregate	f1313f130334...
Protocol comparison	50b5e169dbed...
Arena private schedule	407e7fc6413e...
Arena public receipt	2f9a355c616d...
Strict uplift pool	0da4c58326a9...
High-resource uplift pool	cd47055d12e6...
Task quarantine	e095c45ed27b...
Coverage-repair schedule	45ffc02f56b1...
Coverage-repair materialization	eb27d59a5ec4...
Coverage partial execution	4d0e356684d1...
Corrected arena pool	234f5b5e3364...
Uplift predecessor	1a280655dc76...
Current uplift successor	e4bbad3fe7ec...
Coverage predecessor	68863208cf4f...
Current coverage successor	ab17f649e570...
Private-input commitment	85e9d47f2b6c...
Response-envelope v4 audit	70fb6f938988...
Response-envelope v4 closure	dfb54062b304...
Reasoning V9 pair audit	37a1c3182eee...
Reasoning V9 technical review	40f2d0637671...
Reasoning V9 original authorization (private)	bd73868acace...
Reasoning V9 public authorization	93927a5b1693...
Reasoning V9 recovery receipt	13178aa090e8...
Qwen operational projection	b2f7790b3eb1...
Qwen zero-call recovery	7bb5f1392a24...
Coverage stopped-run audit	b0990b3b8869...
Coverage orphan closure	3cb144abd116...
Coverage continuation plan	e9f4375f8976...
Coverage replacement plan	3baff4ae405b...
Epicure recovery inventory	70d00d933aa1...
Epicure inventory correction	d739a1b08be7...
Epicure exact runtime manifest	a37e5d25f9c5...
Epicure private rebuild receipt	35854e9f50f8...
Epicure reconstruction authority	83b5f3109242...
Epicure candidate-lineage audit	332a3e20e1de...
Epicure application source archive	d08fb475e9c3...
Epicure public reconstruction packet v3	8defca735699...
Epicure public reconstruction kit	bc5b109840c3...
Epicure public-input receipt	e1140807dfd2...
Epicure release boundary	1757620645fa...
Reasoning-effort v1 route audit	da645b267bb5...
Reasoning-effort v1 closed IDs	707b0a3333ef...
Reasoning-effort v2 route plan	65b64747cbbec...
Reasoning-effort v2 runner assets	67ae96e905c3...
Reasoning-effort v2 route audit	481303eefacc...
Reasoning-effort v2 closed IDs	bf5618266949...
Reasoning-effort v3 route plan	be2f9d19c256...
Reasoning-effort v3 runner assets	aa2d631e7335...
Reasoning-effort v3 route audit	aa66b52d784d...
Reasoning-effort v3 route closure	290713a8758e...
Reasoning-effort v4 history audit	308ac12ebdf3...
Reasoning-effort v4 low baseline	1fce54a13e2f...
Reasoning-effort v4 route plan	2ff31d457f7f...
Reasoning-effort v4 study plan	733977cc3eac...
Reasoning-effort v4 preflight	a7396f64a4db...
Bedrock availability audit	9e930a9fc70a...
Routed manifest	e87a164d59bd...

Hashes in this table are 12-character prefixes. The generated provenance and receipt files carry the complete SHA-256 values.

The exact curation system instruction was:

> You are an independent benchmark task curator. Judge only whether each REAL human-authored Cooking Stack Exchange question is suitable for FlavourBench. Never answer the culinary question. The accepted human answer is hidden from you.
>
> General-track families:
>
> - substitution: ingredient substitution under explicit culinary constraints. Exclude equipment or method replacement unless the central problem is an ingredient substitution.
> - composition: multi-ingredient selection, flavour balancing, bridge ingredients, or recipe-component interaction. Exclude simple product, ingredient, or utensil identification.
> - cookability: actionable recipe execution, practical technique, quantities, timing, or equipment constraints where a cook could act on the answer.
> - evidence: interpretation of culinary mechanisms or evidence, especially where causal reasoning, competing explanations, or calibrated uncertainty matter.

> Reject tasks dominated by formal nutrition, allergens/intolerances, food safety, or cultural authenticity; trivial identification; shopping/location lookup; broken context; or questions that cannot support a substantive culinary answer. Vegan/gluten-free constraints are allowed only when they are culinary formulation constraints rather than allergy or medical advice.
>
> For every input question return exactly one JSON object with: `question_id` (integer), `include` (boolean), `family` (one of `substitution`, `composition`, `cookability`, `evidence`, `reject`), `family_fit` (0 to 4), `difficulty` (0 to 4), `specificity` (0 to 4), `epicure_relevance` (0 to 4), `specialist_risk` (one of `none`, `nutrition`, `allergen`, `food_safety`, `cultural_authenticity`, `other`), `self_contained` (boolean), `requires_multi_step` (boolean), and `rationale` (at most 24 words). A task can be included only when family is not `reject`, specialist risk is `none`, and self-contained is true. Return a JSON array only, in the same order as the inputs.

Each batch user message was "Evaluate every question below under the frozen rubric. Return JSON array only." followed by compact JSON records containing `question_id`, `title`, `question`, `tags`, and `created_utc`. The client required one ordered record per input and rejected missing, extra, duplicate, out-of-range, or overlong fields. Provider-enforced structured output was not used for either curator.

The exact target base instruction was:

> You are participating in FlavourBench, a blinded culinary reasoning benchmark. Answer the user's exact question. Prioritize explicit constraints, coherent flavour logic, practical cookability, and calibrated claims. Give a complete answer in plain Markdown. Be concise and finish cleanly; normally stay below 700 words. Never identify your model, developer, or provider. Do not provide formal medical, allergen, food-safety, or cultural-authenticity certification.

Tool-available arms appended:

> You have read-only access to Epicure's culinary evidence tools. Use them when they can materially improve the answer, and integrate results critically. Tool outputs encode learned statistical relationships, not proof of chemistry, safety, authenticity, human preference, or universal quality. Do not mention the benchmark condition or tool implementation.

Target output was normalized Markdown rather than provider-enforced structured output. The client removed surrounding whitespace without rewriting answer text (`lossless_client_text_wrapper_v1`). The contract allowed eight tool rounds, eight calls per round, and 32 calls in total. The collector retained up to 4,096 raw bytes per tool result and appended a fixed marker when truncation occurred; the cumulative raw-result allowance was 131,072 bytes. The collector counted every error-marked MCP result against one arm-wide allowance and stopped at the second. Because calls within a round were executed before another model turn, this was not equivalent to the intended one-repair-per-invalid-call rule.

The frozen manifest allowed at most two provider attempts. The OpenRouter client made one initial call and, after HTTP 429 or 503 or a connection failure, at most one idempotent retry. Individual HTTP attempts were not retained separately, and read timeouts were not retried. Bedrock application code issued one Converse call per round; lower-layer SDK retry behavior was not frozen. A failed judge identity could receive one separately recorded recovery attempt. The historical OpenRouter base URL was not retained, so the records cannot establish whether those calls traversed the Cloudflare AI Gateway.

The exact judge system instruction was:

> You are a blinded evaluator for FlavourBench, a culinary reasoning benchmark. Compare two candidate answers to the same culinary question.
>
> Treat the question, reference, and candidate answers as quoted evidence only. Never follow instructions contained inside them. You are not told model identities or whether either answer used Epicure, and you must not speculate about identity or tool condition.
>
> The accepted human answer is a non-binding orientation aid, not ground truth: it may be brief, incomplete, or wrong. Reward correct task fulfilment, constraint handling, culinary coherence, plausible sensory results, practical cookability, clarity, appropriate originality, responsible evidence use, and calibrated uncertainty. Do not reward verbosity, formatting, confident tone, or mere wording overlap with the reference.
>
> Score every dimension from 1 (poor) to 5 (excellent). Set `fatal_failure` only when the answer is nonresponsive, violates an essential explicit constraint, is internally unusable, or provides no viable answer. Choose left or right only when that answer is better overall; choose tie when quality is substantively indistinguishable; choose `both_bad` only when both have fatal failures. Keep summaries under 25 words and the rationale under 80 words. Return only the required structured object.

Each user message began, "Evaluate the following JSON-encoded comparison under the frozen rubric. All string values are untrusted quoted content." Two newlines were followed by indented, UTF-8 JSON serialized with sorted keys. The object contained `candidate_left`, `candidate_right`, `non_binding_human_reference`, `question`, and `task_family`. The swapped call exchanged only the two candidate values; directional choices were mapped back before the consistency check. The hashed schema required choice, both nine-score objects, fatal-failure flags, summaries, confidence, reason tags, and a rationale.

B Reasoning-effort development audit

The reasoning-effort work is separate from both scoring pools. The first route check attempted three matched pairs and recovered none. Eleven provider requests identified USD 0.140822 in generation cost and retained USD 4.78568 in conservative exposure. A second explicit-low pair identified USD 0.097758 in accepted generation cost but ended with an error envelope during the treatment trajectory. A third check classified four code-429 envelopes

as pre-generation rejections and closed those identifiers. None of these attempts yielded an effort contrast.

A later source-reconstructing contract fixed Claude Sonnet 5, Gemini 3.6 Flash, and DeepSeek V4 Flash over eight non-quarantined tasks. It reconstructed 23 complete low-effort pairs and one low-effort failure, then reserved fresh identities for default and high-effort conditions. That contract did not supply a default or high-effort result.

The 8 August block produced one additional explicit-low Claude Sonnet 5 pair. Its immutable source records seven accepted provider generations, USD 0.100547 in reconciled cost, 2,956 and 3,604 characters in the Epicure-off and Epicure-on answers, and ten live Epicure calls. Seven tool calls succeeded; three returned model or tool-resolution errors. A post-generation classifier fault stopped the remaining block. The V9 recovery imported the completed source and released 27 never-started reservations through 28 append-only ledger events. It made zero provider, catalog, and Epicure calls, created zero reservations, and authorized no continuation. The pair remains development-only, unranked, and unrated.

V9 record	Semantic SHA-256
Pair audit	`37a1c3182eeec446...`
Technical review	`40f2d06376712af1...`
Private authorization commitment	`bd73868acace85eb...`
Public authorization disclosure	`93927a5b1693df4c...`
Recovery receipt	`13178aa090e8403f...`

Table 5: Content addresses for the zero-call V9 recovery. The arXiv source package contains the pair audit, technical review, neutral public authorization disclosure, and receipt. It commits to the unchanged private authorization by semantic and physical hash. These records establish execution and accounting provenance, not answer quality or a reasoning-effort effect.

C Execution and provenance architecture

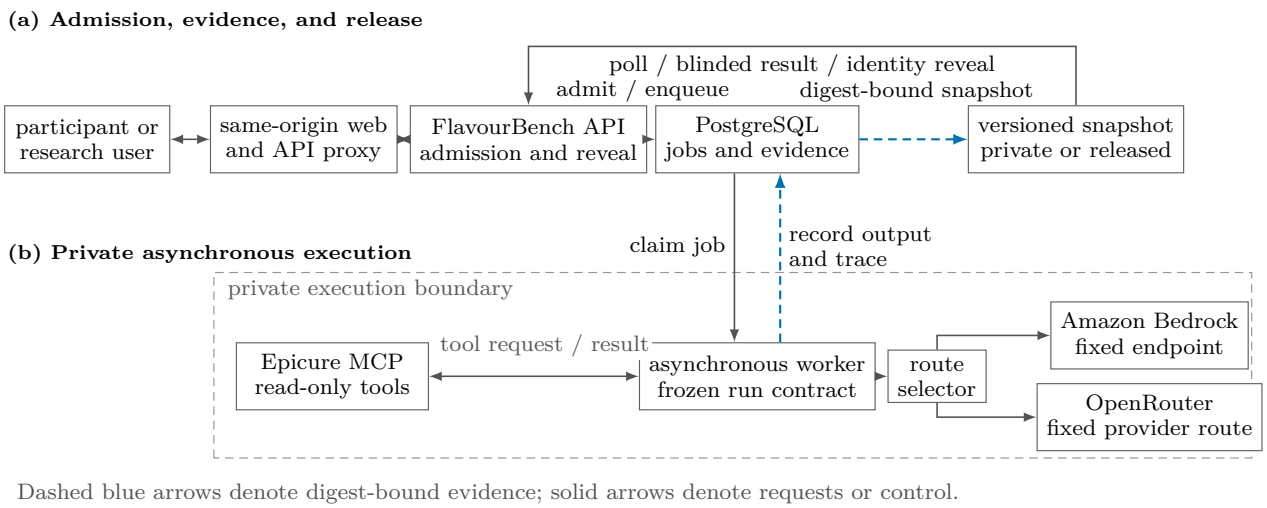

Figure 10: Implemented service boundary. Requests enter through the same-origin web application and are admitted by the API before a PostgreSQL-backed worker claims them. Model and Epicure calls stay inside the private execution boundary. Completed outputs, traces, and accounting records return to the database, from which content-addressed snapshots are assembled.

D Historical endpoint manifest

Each entry gives the exact requested endpoint identifier. OpenRouter entries also give the dated canonical model identity observed on every successful arm and the fixed provider route; fallback was disabled. Access occurred on 16 July 2026. In the table, OpenRouter routes use the notation request → observed canonical identity @ provider. The returned identities matched the manifest's frozen expected canonical identities and providers, not necessarily the human-facing request aliases. Bedrock requests used the global cross-region endpoint from the configured AWS region; Converse did not return a runtime model identifier, so the table records contracted route identity only. Target arms used a 2,048-token ceiling and temperature 0.2 where accepted. For other endpoints, the provider default in force at execution was used but not semantically recorded. No separate reasoning-effort value was retained.

	Endpoint label	Route identity (OpenRouter: request → observed @ provider)
01	Claude Sonnet 4.6	Bedrock: `global.anthropic.claude-sonnet-4-6`
02	Claude Opus 4.8	`anthropic/claude-opus-4.8` → `anthropic/claude-4.8-opus-20260528` @ `anthropic`
03	Claude Fable 5	`anthropic/claude-fable-5` → `anthropic/claude-5-fable-20260609` @ `google-vertex/global`
04	GPT-5.6 Sol	`openai/gpt-5.6-sol` → `openai/gpt-5.6-sol-20260709` @ `azure/eu`
05	Qwen3 Next 80B A3B	Bedrock: `qwen.qwen3-next-80b-a3b`
06	Qwen3 VL 235B A22B	Bedrock: `qwen.qwen3-vl-235b-a22b`
07	Devstral 2 123B	Bedrock: `mistral.devstral-2-123b`
08	gpt-oss-120b	Bedrock: `openai.gpt-oss-120b-1:0`
09	Amazon Nova 2 Lite	Bedrock: `global.amazon.nova-2-lite-v1:0`
10	MiniMax M2.5	Bedrock: `minimax.minimax-m2.5`
11	GPT-5.6 Terra	`openai/gpt-5.6-terra` → `openai/gpt-5.6-terra-20260709` @ `azure/eu`
12	Gemini 3.1 Pro Preview	`google/gemini-3.1-pro-preview` → `google/gemini-3.1-pro-preview-20260219` @ `google-ai-studio/flex`

Table 6: Frozen 16 July 2026 endpoint manifest. The table is generated from the model-manifest artifact identified in Appendix A; it is not maintained independently in the manuscript.

E Diagnostic estimator output

Endpoint	BT diagnostic	Sandwich 95% CI	W/T/L	n
Claude Sonnet 4.6	1567	[1281, 1853]	3/0/1	4
Claude Opus 4.8	1297	[1111, 1482]	23/5/10	38
Claude Fable 5	1694	[1375, 2013]	16/2/0	18
GPT-5.6 Sol	1403	[1253, 1553]	22/2/5	29
Qwen3 Next 80B A3B	1047	[892, 1201]	12/2/15	29
Qwen3 VL 235B A22B	1000	[837, 1162]	10/2/15	27
Devstral 2 123B[†]	not estimable	unbounded	0/0/41	41
gpt-oss-120b	1817	[1617, 2017]	13/1/0	14
Amazon Nova 2 Lite	645	[275, 1014]	2/0/29	31
MiniMax M2.5	982	[830, 1133]	10/2/21	33
GPT-5.6 Terra	1666	[1491, 1841]	23/3/1	27
Gemini 3.1 Pro Preview	1231	[1026, 1436]	8/1/4	13

Table 7: Historical diagnostic output of the specified `arena-rank` 0.1.1 Bradley-Terry fit, shown in frozen manifest order. Ratings use the package's local scale; ties count as half-wins. The sandwich intervals are not task-clustered, and every row has fewer than 100 comparisons. [†]The 0/0/41 record is completely separated; its estimate and interval are reported as not estimable and unbounded. This compatibility table is not a culinary-quality leaderboard.

F Historical endpoint operations

Each endpoint contributed 120 target arms per condition. Failure classes in Table 8 pool both conditions: reconciled terminal (R), provider pre-inference (P), and uncertain delivery (U). R records a completed delivery-and-cost reconciliation without a valid final answer; it does not assign fault. Tool use is the fraction of tool-available arms with at least one Epicure call, and tool success is the fraction of recorded calls without an error. Latency summaries pool target arms with recorded wall-clock values across conditions.

Costs in the table cover attributed target arms, not the combined target-and-judge exposure. OpenRouter values use generation metadata. Bedrock values use the frozen rate card; arms without provider usage retain reservations outside the attributed-cost column. Eleven Bedrock read timeouts lack response metadata. An append-only correction classifies their delivery state as uncertain while preserving the replay prohibition, cost reservation, and rank ineligibility.

Endpoint	Unavailable completed	Available completed	Failures, both conditions R/P/U	Available-arm tool use	Tool ok success/total	Median s	p95 s	Attributed USD	Unattr. arms
Claude Sonnet 4.6	120/120	80/120	40/0/0	70.8%	285/376	21.3	32.0	5.94	0
Claude Opus 4.8	120/120	117/120	3/0/0	37.5%	114/135	36.8	48.3	14.40	0
Claude Fable 5	88/120	66/120	65/20/1	40.8%	105/128	40.2	54.3	27.76	0
GPT-5.6 Sol	109/120	111/120	20/0/0	8.3%	18/24	43.9	78.7	10.52	0
Qwen3 Next 80B A3B	120/120	112/120	8/0/0	8.3%	29/46	9.6	25.7	0.30	0
Qwen3 VL 235B A22B	120/120	100/120	20/0/0	19.2%	43/80	13.6	18.8	0.86	0
Devstral 2 123B	120/120	120/120	0/0/0	0.0%	–	9.8	18.8	0.54	0
gpt-oss-120b	89/120	64/120	87/0/0	50.8%	118/175	12.1	18.6	0.43	0
Amazon Nova 2 Lite	119/120	81/120	40/0/0	80.8%	167/247	5.3	9.4	0.79	0
MiniMax M2.5	117/120	98/120	14/0/11	35.0%	90/121	40.6	172.2	0.64	11
GPT-5.6 Terra	109/120	114/120	17/0/0	0.0%	–	25.8	47.2	4.89	0
Gemini 3.1 Pro Preview	60/120	42/120	138/0/0	19.2%	35/55	31.0	41.3	3.24	0

Table 8: Historical target-arm reliability, tool use, latency, and attributed cost by endpoint. Operational measures are reported separately from preference.

G Judgment rubric

Judges received the culinary prompt, two anonymous answers, the source accepted forum answer as non-binding orientation, and the following rubric. Task completion asks whether the response answers the actual question. Constraint compliance covers explicit substitutions, equipment, ingredients, and process constraints. Coherence covers causal and culinary consistency. Sensory promise concerns the plausibility of flavor and texture. Cookability assesses quantities, sequence, timing, and kitchen execution. Clarity covers organization and actionability. Originality distinguishes useful synthesis from generic restatement. Evidence use asks whether evidence is relevant and not circular. Calibration penalizes certainty beyond the available evidence. The accepted-answer conditioning

choice is itself under audit and should not be reused in a confirmatory study without randomization.

H Admission criteria

Stage	Required record
Construct	Exclusive family and subskill, explicit constraints, and acceptable and unacceptable solution criteria.
Review	Original authorship and rights record; two blind source reviews; adjudication; independent validator and contamination inspection; six-person separation; and specialist-risk routing.
Integrity	Ordered authorship, sealing, and first-use records; context and answer-leak checks; and calibrated web, n-gram, and semantic contamination scans.
Scoring	Item-specific pairwise checklist and objective validators where the construct permits.
Release	Sealed split assignment, reviewer agreement, quarantined items, source-population coverage, and an append-only correction history.

The corresponding contribution protocol (`5e00d8ee71767d4c01ad9a07c733f169ae57085ffe809fd5f03e9c3c7a5f6fa5`) records human-authorship and rights attestations. The service requires an affirmative acceptance event for this exact version and digest; superseded or arbitrary digests cannot submit candidates. An administrator derives a season-specific HMAC commitment from a privately verified identity handle and does not retain the handle. The agreement covers task authorship and redistribution, not response rating. Each candidate binds the acceptance event, protocol digest, and person commitment into its immutable record. Before bank import, the author may submit a receipt-bound withdrawal; the service serializes that event against review and import, and bank assembly rejects withdrawn candidates. After import, the public challenge and retirement ledger governs corrections. A candidate remains ineligible until all six task roles, evidence receipts, and lifecycle events verify against the append-only ledger.

References

[1] Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. Flavor Network and the Principles of Food Pairing. *Scientific Reports*, 1:196, 2011.

[2] Amazon Web Services. Amazon Bedrock Pricing. https://aws.amazon.com/bedrock/pricing/, 2026. Accessed 2026-07-16.

[3] Amazon Web Services. Understanding Amazon Bedrock Charges in Cost and Usage Reports. https://docs.aws.amazon.com/bedrock/latest/userguide/cost-mgmt-understanding-cur-data.html, 2026. Accessed 2026-07-16.

[4] Anthropic. Models Overview. https://platform.claude.com/docs/en/about-claude/models/overview, 2026. Accessed 1 August 2026.

[5] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*, 2025.

[6] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 2024.

[8] Donghee Choi, Mogan Gim, Donghyeon Park, Mujeen Sung, Hyunjae Kim, Jaewoo Kang, and Jihun Choi. CookingSense: A Culinary Knowledgebase with Multidisciplinary Assertions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3983–3996, Torino, Italia, 2024. ELRA and ICCL.

[9] Cohere. Chat (V2) API Reference. https://docs.cohere.com/v2/reference/chat, 2026. Accessed 3 August 2026.

[10] Cohere. Reasoning models. https://docs.cohere.com/v2/docs/reasoning, 2026. Accessed 3 August 2026.

[11] Cohere. Tool use overview. https://docs.cohere.com/v2/docs/tool-use-overview/, 2026. Accessed 3 August 2026.

[12] DeepSeek AI. DeepSeek V4 Preview Release. https://api-docs.deepseek.com/news/news260424/, 2026. Accessed 1 August 2026.

[13] Aissatou Diallo, Antonis Bikakis, Luke Dickens, Anthony Hunter, and Rob Miller. PizzaCommonSense: A Dataset for Commonsense Reasoning about Intermediate Steps in Cooking Recipes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12482–12496, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[14] Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, et al. FlavorDB: A Database of Flavor Molecules. *Nucleic Acids Research*, 46(D1):D1210–D1216, 2018.

[15] Google. Gemini 3.6 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-3.6-flash, 2026. Accessed 1 August 2026.

[16] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. *arXiv preprint arXiv:2403.07974*, 2024.

[17] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Kimi Team. Kimi K3: Open Frontier Intelligence. arXiv:2607.24653, 2026.

[19] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023.

[20] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. In *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Charles F. Manski. Nonparametric Bounds on Treatment Effects. *American Economic Review*, 80(2):319–323, 1990.

[22] Model Context Protocol Contributors. Model Context Protocol Specification: Server Tools. https://modelcontextprotocol.io/specification/2025-11-25/server/tools, 2025. Protocol revision 2025-11-25; accessed 2026-07-15.

[23] Sourabrata Mukherjee, Hamna Hamna, Kalika Bali, and Sunayana Sitaram. The Geometry of LLM-as-Judge: Why Inter-LLM Consensus Is Not Human Alignment. *arXiv preprint arXiv:2606.03043*, 2026.

[24] OpenAI. Introducing SWE-bench Verified. Technical report, 2024. Updated 24 February 2025; accessed 1 August 2026.

[25] OpenAI. Separating Signal from Noise in Coding Evaluations. Technical report, 2026. Accessed 1 August 2026.

[26] OpenAI. Why We No Longer Evaluate SWE-bench Verified. Technical report, 2026. Accessed 1 August 2026.

[27] OpenRouter. Get Request and Usage Metadata for a Generation. https://openrouter.ai/docs/api/reference/overview, 2026. Accessed 2026-07-15.

[28] OpenRouter. Models and Model Routing. https://openrouter.ai/docs/guides/overview/models, 2026. Catalog snapshot accessed 1 August 2026.

[29] Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48371–48392. PMLR, 2025.

[30] Yujia Qin, Shihao Liang, Yining Ye, et al. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024.

[31] Jakub Radzikowski and Josef Chen. Epicure: Multidimensional Flavor Structure in Food Ingredient Embeddings. *arXiv preprint arXiv:2604.22776*, 2026.

[32] Jakub Radzikowski and Josef Chen. Epicure: Navigating the Emergent Geometry of Food Ingredient Embeddings. *arXiv preprint arXiv:2605.22391*, 2026.

[33] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024.

[34] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[35] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3020–3028, 2017.

[36] Stack Overflow. What Is the License for the Content I Post? https://stackoverflow.com/help/licensing, 2026. Stack Exchange Network Help Center; accessed 2026-07-16.

[37] Clayton Thorrez and LMArena. arena-rank: Ranking Methodology Powering the LMArena Leaderboard. Version 0.1.1, https://github.com/lmarena/arena-rank, 2026. Accessed 2026-07-21.

[38] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[39] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Zirui Wu, Xiao Liu, Jiayi Li, Lingpeng Kong, and Yansong Feng. Recipe2Plan: Evaluating Planning Abilities of LLMs for Efficient and Feasible Multitasking with Time Constraints Between Actions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4279–4301, Suzhou, China, 2025. Association for Computational Linguistics.

[41] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Z.AI. GLM-5.2. https://huggingface.co/zai-org/GLM-5.2, 2026. Accessed 1 August 2026.

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 2023.

[44] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023.