MKAI INQUIRY WORKING PAPER

The Same-Model Ceiling

A matched comparison of same-model and heterogeneous review across twelve strategic-analysis tasks

Richard Foster-Fletcher
MKAI
research@mkai.org

MKAI Inquiry Working Paper · August 2026
Published by MKAI · mkai.org/studies/same-model-ceiling

Abstract

This study tests whether three reviewers from other model families surface decision-relevant considerations that three matched reviews by the originating model do not, and whether incorporating those considerations produces analyses preferred under blind model assessment. Claude Opus 5 answered twelve strategic-analysis tasks on one constructed acquisition case; each answer was critiqued under two matched conditions, by three fresh Claude Opus 5 sessions and by GPT-5.6 Sol, Muse Spark 1.2 and DeepSeek V4 Pro. Eligible candidates, pooled under shuffled neutral identifiers, were grouped by two blinded model coders with a model adjudicator; a group containing only heterogeneous candidates counted as set-only. Claude Opus 5 rated each set-only group's decision effect and revised where a group qualified; two model scorers assessed each pair blind, with adjudication on disagreement.

The critiques produced 216 candidates; 152 passed the precommitted screen, 74 same-model and 78 heterogeneous. Blinded coding identified 37 set-only groups, at least one in every task. Claude rated 35 at Level 2, a change to terms, evidence requirements or commitment conditions; two reached Level 1 and no group reached Level 3, a route change. The final blinded preference favoured the revision in eight of twelve tasks and the original in one, with three ties. The two primary scorers agreed in five of twelve pairs, and six of the eight supported outcomes were decided by the adjudicator, which also served as reviewer H1. A post-collection audit reclassified 20 eligible candidates and retained 31 of the 37 set-only groups; four newly eligible candidates were never coded, so the audit provides no complete preference-based sensitivity result. Every judgement came from a language model, scores compressed against the top of the 0-to-2 scale, and no matched same-model revision arm was run. The result is a descriptive pilot bounded to this case, this panel and this procedure.

Keywords

Same-model review; heterogeneous critique; multi-model review; large language models; blind assessment; strategic analysis; enterprise AI; model evaluation.

Suggested citation

Foster-Fletcher, R. (2026). The Same-Model Ceiling: A matched comparison of same-model and heterogeneous review across twelve strategic-analysis tasks. MKAI Inquiry Working Paper.

1. Introduction

This study tests one narrow question about review. When a frontier model has produced a piece of strategic analysis, and three fresh sessions of the same model are asked what that analysis missed, do three reviewers drawn from other model families find decision-relevant material that the same-model reviews do not? A second question follows from the first: when the considerations only the outside panel raised are fed back to the originating model, is the revised analysis preferred over the original under blind model assessment? The study answers both questions for one originating model, one three-model panel and one constructed acquisition case.

The question bears on a deployment pattern that a single-model workflow makes possible: the model that drafts an analysis can also be asked to review it in a fresh session. Whether that arrangement has a ceiling, considerations that repeated same-model review keeps missing, is an empirical matter, and this study measures it in one tightly bounded setting rather than asserting anything about how organisations in general work. The study continues MKAI's programme of controlled work on how configuration and process shape what enterprise AI produces. The Subtraction Study (June 2026) measured what an enterprise governance prompt removes from a model's output when every other variable stays fixed, and the Separation Study (August 2026) ablated that prompt to establish which

instruction carried the effect. Both varied the instruction layer around a fixed model; this study fixes the instructions and varies the reviewing model.

The committed proposition, fixed in the study workbook before collection, reads: across twelve elicitation tasks within one constructed acquisition case, a heterogeneous three-model review may surface decision-relevant considerations that three matched reviews by the originating model do not surface. The unit of analysis is one task within the constructed case. The design is a descriptive pilot. It supports no population-level claim about multi-model review, and it cannot separate the effect of family diversity from the particular capabilities of the three reviewers tested.

The comparison has a deliberately narrow shape. The originating model answered each task once, and that answer went to two conditions of three critiques each, with reviewer identity as the intended difference between them. Considerations raised only by the heterogeneous panel were rated for decision effect by the originating model itself, folded into a revised analysis where they qualified, and the original and revised analyses were compared blind. The design therefore measures one direction of coverage, what the outside panel raised that same-model review did not, and quantifies the decision value of that direction only; the reverse direction was recorded and left unassessed, and no revision was built from the same-model critiques.

2. Related research

The same-model side of this design touches a literature on models improving their own output. Madaan et al. (2023) showed that a single model can generate an answer, critique it and refine it with its own feedback, improving results by roughly 20 per cent on average across seven tasks without further training. Shinn et al. (2023) extended the pattern to agents that keep verbal reflections in memory across attempts, reaching 91 per cent pass@1 on the HumanEval coding benchmark. Huang et al. (2024) mark the boundary of the optimistic reading: on reasoning tasks, models struggle to correct their own answers without external feedback, and performance sometimes degrades after attempted self-correction. The same-model condition here is a different arrangement: three fresh sessions of the originating model critique a completed analysis under an explicit brief, so the study tests same-model review as an external procedure rather than the self-correction setting Huang et al. examine.

A second literature varies the number of agents rather than the family. Du et al. (2024) found that multiple model instances debating over rounds improve factual accuracy and reasoning; their headline experiments debate instances of the same underlying model, so the mechanism tested is multiplicity. Chan et al. (2024) moved evaluation from a single judge to a debating panel, ChatEval, and report closer alignment with human assessment, with the panel's diversity supplied by role prompts rather than by different models. This study keeps the brief constant and varies the model family itself, which neither line of work isolates.

The assessment layer of this study depends on a third literature, models used as judges. Zheng et al. (2023) established that a strong model judge can match human preferences with over 80 per cent agreement, a level comparable with agreement between humans, while documenting position and verbosity biases; on self-enhancement bias their data were inconclusive. Panickssery et al. (2024) then showed self-preference directly: model evaluators score their own outputs above others' when human annotators rate them equal, and the effect strengthens with the model's ability to recognise its own generations. Wang et al. (2024) showed that swapping the order of two candidate answers can reverse a model judge's verdict. Chiang and Lee (2023) found model evaluation consistent with expert human rankings on two text-generation tasks, while Bavaresco et al. (2025), across twenty evaluation datasets, found agreement with humans varying widely by task and recommend validation before use; Gu et al. (2024) survey reliability practices for such judges.

These findings bear on this study's instrument rather than only its motivation. Every coder, scorer, adjudicator and decision-effect assessor here is a model; the blind pairs were allocated A and B once per pair rather than counterbalanced, so positional effects of the kind Wang et al. document are randomised across pairs rather than eliminated; and the decision-effect assessor rated considerations applied to its own earlier analysis. Panickssery et al.'s evidence concerns evaluators favouring outputs they had themselves generated, a different arrangement, so their finding marks self-preference as a related risk here rather than a demonstrated effect. The study differs from the work above in its unit and material: it compares the coverage of two review panels over long-form strategic analyses of a single constructed case, counts the considerations one panel raises that the other does not, and tests their decision value through revision and blind preference, rather than measuring answer accuracy on benchmark questions. No priority is claimed for the design.

3. Methodology

3.1 The constructed case

All twelve tasks examine one fictional transaction. Meridian Industrial Group, a precision-components manufacturer with revenue of approximately £480 million, is considering the acquisition of Calder Systems, a founder-led additive-manufacturing business with revenue of approximately £110 million, 28 per cent annual growth and an operating loss of £9 million, at a proposed price of £340 million funded through cash and new debt. The case wires ten pressure points into a single fact pattern: a board net-debt ceiling of 3.5 times EBITDA against a lending covenant of 4.0 times, synergy-dependent net debt of 3.4 or 4.1 times EBITDA, customer concentration of 34 and 26 per cent with change-of-control clauses, seven process patents, incomplete medical-device validation, a founder who owns 38 per cent and manages both key customer relationships, fourteen unretained technical staff, 7.4 per cent borrowing costs, three weeks of exclusivity and a claimed rival bidder. The full case text appears in Appendix A and was supplied verbatim with every call.

A constructed case gave every model the same supplied case record. Web access and retrieval were unavailable, although reviewers could still draw upon pretrained general knowledge. The response contract required any external proposition to be labelled external and unverified, and the precommitted screen relied upon those reviewer-supplied labels. The post-collection audit subsequently found that those labels were applied inconsistently.

3.2 The task battery

Twelve tasks, P01 to P12, each direct the model's attention to one component of analytical work on the same case: P01 memory, P02 evidence, P03 timing, P04 attention, P05 judgement, P06 framing, P07 options, P08 imagination, P09 incentives, P10 risk tolerance, P11 social validation and P12 narrative. The instrument groups P01 to P03 as analytical-reach tasks, P04 and P05 as bridge tasks and P06 to P12 as decision-conversion tasks. The exact wording of all twelve appears in Appendix A.

Every original and revised analysis was produced under a shared response contract: a maximum of 900 words, the supplied case as the factual basis, external claims identified as external and treated as unverified, facts separated from inferences and missing evidence, and four labelled closing fields covering the decision route, proposed terms or structure, evidence required before commitment, and a single switch condition. The available routes were proceed at £340 million, seek different terms, defer pending specified evidence, or leave the transaction. The word limit was breached in most outputs, a deviation recorded and discussed in sections 3.11 and 7.

Two calibration tasks, C01 and C02, ran the complete machinery on a separate constructed case before the instrument froze, and are described in section 3.10.

3.3 Model panel, interface and settings

Table 1 lists the panel. The originating model, designated F1, produced every original analysis, the three same-model critiques per task, every decision-effect assessment and every revised analysis. The three heterogeneous reviewers, H1 to H3, produced one critique each per task. Two further models coded and scored blind, and one adjudicated.

Table 1. Model panel and roles

Role	Study role	Provider	Model	Identifier	Reasoning setting
F1	Originating model and same-model condition	Anthropic	Claude Opus 5	claude-opus-5	High
H1	Heterogeneous reviewer 1	OpenAI	GPT-5.6 Sol	gpt-5.6-sol	High
H2	Heterogeneous reviewer 2	Meta	Muse Spark 1.2	muse-spark-1.2	Not established
H3	Heterogeneous reviewer 3	DeepSeek	DeepSeek V4 Pro	deepseek-v4-pro	No control exposed
Coder 1, Scorer 1	Blinded equivalence coding and blind scoring	Moonshot	Kimi K3	kimi-k3	Platform default
Coder 2, Scorer 2	Blinded equivalence coding and blind scoring	Mistral	Mistral Medium	mistral-medium-2604	Platform default
Adjudicator	Coding and scoring adjudication	OpenAI	GPT-5.6 Sol	gpt-5.6-sol	High

Every call ran through the TypingMind chat interface using direct provider API keys, in a fresh chat with no carried context. The account-level Web Browser and Code Sandbox tools were off, all five installed plugins were off, no knowledge base was attached, and the workspace has no persistent-memory feature. The system instruction was fixed as 'You are a helpful AI assistant.' and applied identically to every generator, reviewer, coder, scorer and adjudicator call (Deviation D009). Temperature, top-p and maximum output tokens remained at provider defaults throughout. The platform's global reasoning-effort setting was High; the model cards confirm this option for Claude Opus 5 and GPT-5.6 Sol, the effect of the setting on Muse Spark 1.2 could not be separately confirmed, and DeepSeek V4 Pro exposes no reasoning-effort control, so the setting has no effect on that model. Reasoning conditions were therefore matched in instruction and interface, and could not be matched in verified effect across providers.

Two substitutions preceded the instrument freeze, and both are recorded. H2 was specified as Google Gemini 3.1 Pro Preview, identifier gemini-3.1-pro-preview, which returned an identical provider-side quota error on every one of five attempts spanning credit and billing confirmation (Deviation D012); Muse Spark 1.2 replaced it on 13 August 2026 with the study owner's approval (Deviation D013). Coder 2 and Scorer 2 were specified as Kimi K2 Thinking, which was blocked by a model-specific access fault, and Mistral Medium replaced it the same day (Deviation D014). The adjudicator assignment follows an earlier design correction (Deviation D008): after a requirement for a separate adjudicating family was withdrawn, GPT-5.6 Sol was selected in preference to F1 so that no Claude-

family session would adjudicate disputes about Claude-family analyses. GPT-5.6 Sol therefore served both as heterogeneous reviewer H1 and as the adjudicator of coding and scoring disagreements, in fresh blinded sessions. Each adjudication session was isolated and source-blind, and the arrangement still means the adjudicator is not drawn from a family independent of the review panel. Section 7 treats this as an independence limitation.

3.4 Original analyses and the two critique conditions

For each task, F1 produced one original analysis in a fresh session from the case, the task and the response contract. The full analysis then went to six reviewers in six further fresh sessions: three more Claude Opus 5 sessions as the matched same-model condition, and H1, H2 and H3 as the heterogeneous condition. Both conditions received identical material, the case, the task, the complete original analysis and an identical critique brief, and both returned exactly three critiques per task. Reviewer identity was the intended difference between the conditions.

The critique brief asked each reviewer to identify up to three specific risks, assumptions, alternatives or decision conditions that the analysis overlooks or gives insufficient weight, restricted candidates to considerations that could change the decision route, proposed terms, evidence required before commitment or the switch condition, and prohibited repeating a consideration already present in the analysis. Each candidate carried four fields: a title of at most twelve words, a one-or-two-sentence statement, a one-sentence decision consequence, and a basis selected from supplied case fact, inference from supplied facts, or external unverified knowledge. The complete response was limited to 280 words. Calibration exposed an ambiguity when reviewers returned four candidates, so a pre-freeze correction fixed the brief to instruct explicit truncation to three, and a formal rule entered the first three candidates in returned order wherever more appeared (Deviations D019 and D020). Both brief versions appear in Appendix B. Every one of the 72 primary critiques returned three candidates, giving 216.

3.5 Eligibility screening

A precommitted rule determined which candidates entered coding, operating on the basis label each reviewer returned. Candidates labelled supplied case fact or inference from supplied facts were eligible. Candidates labelled external unverified knowledge entered coding only if the external claim was verified within the study; no external claim was verified, so every such candidate was excluded. Screening removed 64 of the 216 candidates, 34 from the same-model condition and 30 from the heterogeneous condition, leaving 152 eligible candidates: 74 same-model and 78 heterogeneous. Three reviewer responses stated compound basis lines that did not match a single committed value, and each was resolved by precedent-matching and logged (Deviations D029, D033 and D044); one of the three affected an inclusion outcome. Excluded candidates were preserved verbatim and took no further part in the primary analysis. Because the screen operated on the reviewers' own labels, a post-collection audit re-examined every basis classification; its method appears in section 3.9 and its results in section 4.8.

3.6 Neutral statements and blinded equivalence coding

Each eligible candidate entered coding as a neutral statement under a shuffled neutral identifier drawn from a single global sequence, N028 to N179 for the primary tasks. The workbook's collection sequence assigns the preparation of neutral statements to the study operator, and records no rewriting rule and no fidelity check; the post-collection audit established what was in fact done. Every one of the 152 neutral statements is character-identical to its candidate's statement field. Neutralisation therefore consisted of removing the surrounding fields, the candidate title, the decision consequence, the basis label and all run identifiers, and pooling the bare statements under shuffled identifiers. No candidate statement was rewritten. The blinding this provides is structural rather than stylistic: coders

saw statements stripped of every explicit source marker, in shuffled order, pooled across six reviewers, but each statement retains its reviewer's original wording, so stylistic cues were not removed. An audit scan found no model name and no first-person reference in any neutral statement.

Kimi K3 and Mistral Medium then received, independently and in fresh sessions, the case and the shuffled neutral statements for one task at a time, with no candidate key and no source information, and grouped the statements by underlying consideration under a fixed instruction reproduced in Appendix D. Statements describing substantially the same consideration belonged in one group; a statement with no substantive match formed a group of its own. The two coders reached full first-pass agreement on the partition in six of twelve tasks. The other six produced a coding disagreement, and each went to GPT-5.6 Sol in a fresh blinded session under the same grouping instruction (Deviations D031, D035, D037, D040, D042 and D045, covering P03, P05, P06, P07, P08 and P10). In four of the six the adjudicated partition matched Coder 1 exactly; at P03 the adjudication produced the final partition recorded in the workbook; at P07 it assigned one statement, N103, as a singleton, matching neither coder. Source labels remained concealed until every group was fixed.

3.7 The set-only rule and decision-effect assessment

After the groups froze, source identities were decoded. An equivalence group counted as set-only when it contained at least one heterogeneous-review candidate and no same-model candidate. A set-only group is the study's operational definition of a potential blind spot: a consideration the outside panel raised that no same-model review raised in any form the coders recognised as equivalent.

Every set-only group returned to Claude Opus 5 in a fresh session, with the case, the task, its own original analysis and the group's neutral statement text, and no reference to which models raised it. For two-member groups the two statements were transmitted together, separated by a blank line; an audit comparison found every transmitted statement character-identical to the coded text, apart from one dash character normalised in one transmission. The instruction, reproduced in Appendix E, asked the model to re-examine the analysis in the light of the consideration and assign exactly one level. Level 0 means no material change. Level 1 changes supporting detail or evidence discussion only. Level 2 changes proposed terms or structure, evidence required before commitment, or the switch condition, without changing the decision route. Level 3 changes the decision route itself. Under the committed design, Levels 2 and 3 qualify as decision-relevant effects.

3.8 Revision, blind pairing and blind scoring

Wherever at least one set-only group reached Level 2 or Level 3, F1 produced one revised analysis in a fresh session, incorporating all qualifying considerations for that task in ascending neutral-identifier order, under the same response contract. The revision is the model's own output; no operator edited any analysis. All twelve tasks produced a revision, because every task had at least one qualifying group.

Each original and revised analysis formed a blind pair, labelled Analysis A and Analysis B under a randomised allocation fixed in a source key kept outside every scorer-facing document; the allocation was drawn once per pair and was not counterbalanced within scorers. Kimi K3 and Mistral Medium, in fresh sessions and blind to which analysis was which, scored each analysis on three dimensions with a 0-to-2 scale, evidential discipline, decision specification and analytical reach, defined verbatim in Appendix F, and stated one overall preference, A, B or Tie. Where the two decoded preferences agreed, that preference stood as final. Where they disagreed, GPT-5.6 Sol received the identical anonymised pair in a fresh session and its decoded preference stood as final under the committed rule. The adjudicator's dimension scores were not added to the primary scorers' means, and the dimension means reported in section 4.5 use the 24 primary scorer assessments only.

A task counted as a supported outcome only when both committed conditions were met: at least one of its set-only groups reached Level 2 or Level 3, and the final blinded preference favoured the revised analysis. The primary result is the number of supported outcomes out of twelve. Figure 1 sets out the complete flow.

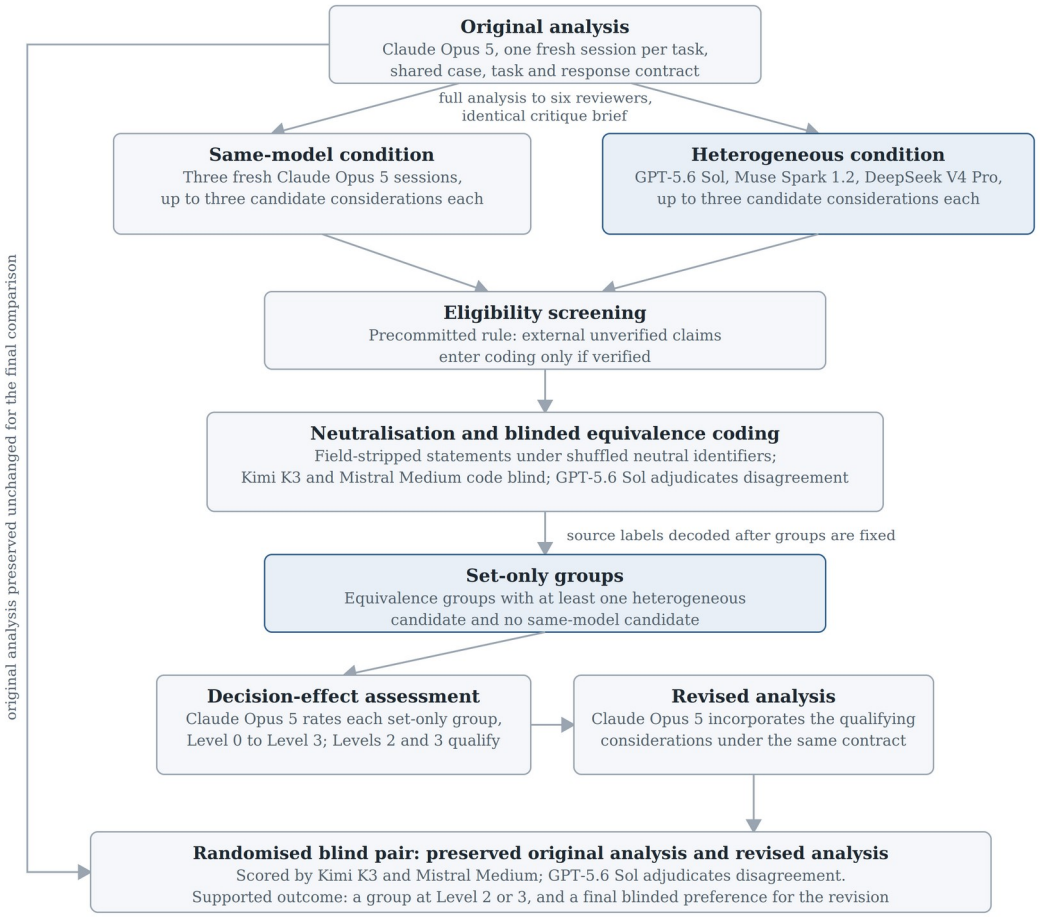

Figure 1. The study flow for one task. The preserved original analysis and the revised analysis form the final blind pair; shaded stages mark the heterogeneous condition and the set-only path it feeds.

3.9 Post-collection audits

Two post-collection checks were performed on the completed record during the preparation of this paper, with no new study interactions. Both were exploratory checks conducted through one model pass by the AI reviser session that prepared this publication package, on 16 August 2026, working from the workbook and the case text with full access to the study record, including source identities, groups and outcomes. Neither check was blinded, and neither is equivalent to the study's precommitted blinded coding procedure. The study record does not capture the reviser's exact model identity, provider or reasoning settings, so those details are not reported; the classification rule stated below served as the audit instruction, and no separate fixed prompt was preserved. The neutralisation audit compared all 152 eligible candidate statements with their neutral statements character by character, scanned for source-identifying content, and traced the decision-effect transmissions back to the coded texts; its findings are incorporated in sections 3.6 and 3.7. The eligibility audit reclassified all 216 candidates from the supplied case alone, treating a statement as an inference only where its reasoning completes from the case without relying on an external legal, regulatory, technical, accounting or market proposition. Under that rule, a statement asserting an outside practice or the scope of a legal instrument, phrases of the form most credit agreements, lenders typically,

qualifications usually attach, or a freedom-to-operate opinion does not establish title, was classed external; a statement flagging a possibility or an open question arising from the supplied facts, or proposing an alternative structure without asserting an outside fact, was classed inference. The audit classification was then compared with each reviewer's original basis label, and every disagreement was traced through its equivalence group, set-only status, decision-effect result and prompt outcome, with exclusions applied mechanically to the coded groups and group composition recalculated. The precommitted screen remains the primary analysis; the audit is reported in section 4.8 as an exploratory sensitivity check, and the full 216-row audit file accompanies this paper.

3.10 Calibration, freeze and collection window

Two calibration tasks, C01 and C02, ran the complete machinery on a separate constructed case, a licensing decision between Northbank Health Devices and Vectra Materials, before any primary output was generated. Their committed purpose was mechanical: to verify every stage the study depends on and to correct procedural faults without contributing data to the findings. Calibration performed exactly that function. C01 absorbed the study's early technical failures, including prompt-transmission faults that forced logged retries, and C02 ran the full pipeline end to end, exercising both adjudication paths. Both calibration tasks returned a substantive null, with the final blinded preference favouring the original analysis in each, and both results were preserved as recorded and excluded from every primary count. The instrument froze on 14 August 2026 after C02 completed. Primary collection ran inside the committed 72-hour window: the first primary output is logged at 01:15 BST on 14 August 2026 and the final primary generation call at 16:05 BST on 15 August 2026.

3.11 Interaction counts and deviations

The primary battery comprised 96 generation calls, eight per task: one original analysis, three same-model critiques, three heterogeneous critiques and one revised analysis. Assessment added 37 decision-effect sessions, 24 coding sessions with six coding adjudications, and 24 scoring sessions with seven scoring adjudications, giving 194 fresh-session model interactions across the primary study. Calibration added its own logged interactions, including six generation retries recorded against C01 and C02.

Fifty-three deviations, D001 to D053, are logged in the study workbook, and Appendix H summarises the register. Most record pre-freeze setup faults, calibration corrections or routine adjudications. Five bear directly on the primary evidence. The 900-word analysis limit was breached in 22 of the 24 primary analyses; only P03's original at 880 words and P05's revision at 863 complied, the longest output was P07's revision at 1,079 words, and no output was truncated after collection (Deviation D053, with full counts in Appendix G). The direction of bias is uncertain in both arms, since longer originals may leave reviewers less omission space while longer revisions may attract scoring credit for added material. The eight P02 session identifiers were never recorded, a gap discovered at final verification and preserved rather than reconstructed (Deviation D052). P12's revision was blocked twice by an exhausted provider credit balance before succeeding on a third attempt after the balance was restored; the two failed attempts are preserved in the deviation record rather than as invented log rows (Deviation D050). A source-attribution tag left on P12's revision file by a non-blind file convention was caught and removed before any scorer transmission, with a programmatic check confirming removal (Deviation D051). The assessor instructions for coding and decision-effect checks were drafted by the study operator to complete the instrument before calibration, within the committed structure but without separate owner approval of the wording (Deviation D016). Post-collection corrections to the workbook are documented on its Publication changes sheet and in the published change log; every raw output remains unchanged.

4. Results

4.1 Candidate production and eligibility

The 72 primary critiques returned 216 candidate considerations, three per critique, 108 from the same-model condition and 108 from the heterogeneous condition. Screening on the reviewers' basis labels excluded 64, every one for resting on external unverified knowledge that the study did not verify: 34 same-model candidates and 30 heterogeneous candidates. The 152 eligible candidates divide into 74 same-model and 78 heterogeneous, so the two conditions entered coding at nearly equal strength, 74 of 108 against 78 of 108. Per task, eligible candidates ranged from nine at P04 to seventeen at P12. Of the 152, the reviewers labelled 144 as inference from supplied facts and eight as a supplied case fact.

4.2 Equivalence groups, overlap and set-only groups

Blinded coding organised the 152 eligible candidates into 80 equivalence groups across the twelve tasks. Twenty-five groups contained candidates from both conditions, which is the overlap a matched design expects: the same-model and heterogeneous reviews frequently converged on the same underlying considerations. Eighteen groups contained only same-model candidates. Thirty-seven groups contained at least one heterogeneous candidate and no same-model candidate, and these are the set-only groups the design is built to isolate. Every task produced at least one, with a spread from one set-only group at P11 to five at P03: four at P01, three at P02, five at P03, two at P04, four at P05, three at P06, four at P07, four at P08, two at P09, two at P10, one at P11 and three at P12. Thirty-three of the 37 are singletons, one heterogeneous reviewer raising a consideration alone, and four contain two candidates from different heterogeneous reviewers. The 41 candidates inside the 37 groups distribute unevenly across the panel: 18 from GPT-5.6 Sol, 13 from DeepSeek V4 Pro and 10 from Muse Spark 1.2.

The eighteen same-model-only groups are the mirror image of the design's target, considerations the three Claude sessions raised that no heterogeneous reviewer raised. Under the committed design they received no decision-effect assessment, so their decision relevance is unmeasured; the design quantified the decision value of one direction of coverage only.

4.3 Decision effects

Claude Opus 5 assessed all 37 set-only groups in fresh sessions. Thirty-five reached Level 2, a change to proposed terms or structure, evidence required before commitment, or the switch condition, without a change of decision route. Two reached Level 1, supporting detail only: group E at P04 and group G at P12. No group received Level 0, and no group received Level 3, a change of decision route. When assessed individually, therefore, 35 of the 37 groups changed proposed terms, evidence requirements or commitment conditions, and no individual group changed the decision route; the combined revisions subsequently changed the stated route at P06, P07 and P10, as section 4.7 describes.

4.4 Blind preferences and supported outcomes

All twelve tasks had at least one qualifying group, so all twelve produced a revised analysis and a blind pair. Table 2 sets out the prompt-level outcomes. The final blinded preference favoured the revised analysis in eight tasks, favoured the original in one, P06, and recorded a tie in three, P02, P08 and P12. Eight of twelve tasks therefore met the committed definition of a supported outcome: at least one set-only group at Level 2 or above and a final preference for the revision.

Table 2. Prompt-level outcomes

Prompt	Component	Set-only groups	Level 2 or 3 groups	Final preference	Supported outcome
P01	Memory	4	4	Revised	Yes
P02	Evidence	3	3	Tie	No
P03	Timing	5	5	Revised	Yes
P04	Attention	2	1	Revised	Yes
P05	Judgement	4	4	Revised	Yes
P06	Framing	3	3	Original	No
P07	Options	4	4	Revised	Yes
P08	Imagination	4	4	Tie	No
P09	Incentives	2	2	Revised	Yes
P10	Risk tolerance	2	2	Revised	Yes
P11	Social validation	1	1	Revised	Yes
P12	Narrative	3	2	Tie	No
Total		37	35	8 revised, 1 original, 3 ties	8 of 12

The two unsupported ties at P02 and P08 occurred despite three and four Level 2 groups respectively, and P06's three Level 2 groups preceded a final preference for the original analysis. Considerations the originating model itself rated as changing terms, evidence or conditions did not reliably convert into a preferred revision. Figure 2 traces the counts through the pipeline.

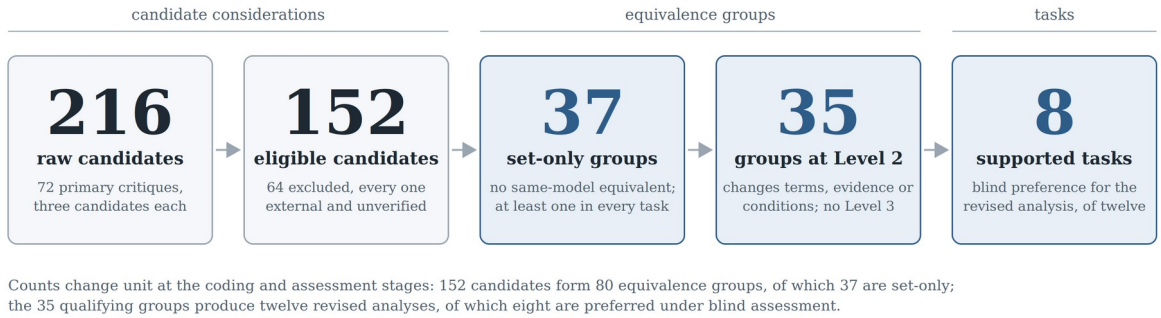

Figure 2. The numeric flow from 216 raw candidates to eight supported tasks. Shaded stages follow the decoding of source labels.

4.5 Dimension scores

Table 3 reports the mean dimension scores across the 24 primary scorer assessments, twelve pairs scored by two scorers, with the adjudicator's scores excluded. All three dimensions moved in the revision's favour, and the largest movement, 0.2083 on a 0-to-2 scale, concerns analytical reach, the dimension closest to what the added considerations supply.

Table 3. Mean dimension scores across the 24 primary scorer assessments

Dimension	Original mean	Revised mean	Mean change
Evidential discipline	1.8750	1.9167	+0.0417
Decision specification	1.9167	2.0000	+0.0833
Analytical reach	1.7917	2.0000	+0.2083
Total score out of six	5.5833	5.9167	+0.3333

These means come from a severely compressed scale. Of the 144 dimension scores the two primary scorers assigned, 132 were the maximum of 2 and twelve were 1; no score fell below 1. Thirteen of the 24 assessments awarded 2 on all six dimensions, leaving the overall preference line as the only discriminating signal in more than half the assessments. Mean changes this small, on a scale this compressed, order the two sets of analyses; they do not measure the size of any quality difference, and they support no claim that the revisions were better in any sense beyond preference under this procedure.

4.6 Scorer agreement and adjudication

The two primary scorers agreed on a decoded overall preference in five of twelve pairs: the ties at P02, P08 and P12 and the revised preferences at P03 and P09. The other seven pairs went to adjudication. Table 4 sets out every assessment. Six of the eight supported outcomes, P01, P04, P05, P07, P10 and P11, were decided by the adjudicator after scorer disagreement; only P03 and P09 received a preference for the revision from both primary scorers. The one preference for an original, P06, was also adjudicated. P01 produced a full three-way split, the first scorer preferring the original, the second a tie, and the adjudicator the revision.

The scorers' individual profiles differ visibly. Kimi K3 committed to a definite preference in eight of twelve pairs, decoding to six revised and two original. Mistral Medium returned eight ties and four revised preferences, and never preferred an original. The adjudicator, GPT-5.6 Sol, returned a definite preference in all seven adjudications, six for the revision and one for the original, and never a tie. The final tally of eight revised, one original and three ties therefore rests heavily on the behaviour of the adjudicating model, which is also heterogeneous reviewer H1.

Table 4. Scorer preferences and adjudication, decoded

Prompt	Scorer 1 (Kimi K3)	Scorer 2 (Mistral Medium)	Agreement	Adjudicator (GPT-5.6 Sol)	Final preference
P01	Original	Tie	No	Revised	Revised
P02	Tie	Tie	Yes	Not required	Tie
P03	Revised	Revised	Yes	Not required	Revised
P04	Tie	Revised	No	Revised	Revised
P05	Original	Revised	No	Revised	Revised
P06	Revised	Tie	No	Original	Original
P07	Revised	Tie	No	Revised	Revised
P08	Tie	Tie	Yes	Not required	Tie
P09	Revised	Revised	Yes	Not required	Revised
P10	Revised	Tie	No	Revised	Revised

Prompt	Scorer 1 (Kimi K3)	Scorer 2 (Mistral Medium)	Agreement	Adjudicator (GPT-5.6 Sol)	Final preference
P11	Revised	Tie	No	Revised	Revised
P12	Tie	Tie	Yes	Not required	Tie

4.7 Decision routes in the combined revisions

No individual set-only group was rated as changing the decision route, yet three revised analyses state a different route from their originals. At P06, P07 and P10 the original analyses ended on seek different terms and the revisions ended on defer pending specified evidence. In each case the revision was produced from several Level 2 considerations at once, and the shift in route wording is the model's own output under the standard revision instruction; no operator edited a route, and no instruction asked for one. The other nine revisions kept their original routes, seek different terms in four cases and defer pending specified evidence in five.

The two observations describe different phenomena. Assessed one at a time, every qualifying consideration changed terms, evidence requirements or conditions and left the route intact; combined into a single rewrite, the accumulated changes shifted the stated route in three of twelve tasks, from renegotiation towards deferral pending evidence. An individually assessed Level 3 effect and a route change emerging from several Level 2 considerations incorporated together are distinct events, and only the second occurred. P06 also shows that a route shift is no guarantee of preference: its revision moved to deferral and the final blinded preference went to the original.

4.8 Sensitivity: the post-collection eligibility audit

The exploratory audit described in section 3.9 reclassified all 216 candidates from the case alone and compared the result with the reviewers' basis labels. The two classifications agree on 184 of 216 candidates. The 32 disagreements divide three ways. Twenty eligible candidates, 13 same-model and 7 heterogeneous, reclassify as external unverified knowledge, most because their statements assert an outside practice or the scope of a legal instrument, in phrases such as most credit agreements, lenders typically, or a freedom-to-operate opinion does not establish title. Four excluded heterogeneous candidates reclassify in the other direction, as eligible inference. Eight candidates the reviewers labelled supplied case fact reclassify as inference, which leaves eligibility unchanged. The audit also documents labelling inconsistency inside the original screen: materially identical considerations carry different basis labels from different reviewers, and in one case from the same reviewer at different tasks.

The stricter audit therefore identifies 136 eligible candidates, 61 same-model and 75 heterogeneous. Only 132 of the 136 already possess blind coding, because the four newly eligible candidates were excluded before the coding stage and were never entered into it. Applying the twenty exclusions to the coded groups changes the inventory in three ways. Six of the 37 set-only groups lose their member or members entirely, P01-B, P01-D, P02-D, P03-E, P05-F and P07-F, every one rated Level 2 in the primary analysis. Thirty-one set-only groups remain, 29 previously rated Level 2 and the two Level 1 groups. Four originally mixed groups lose their same-model members and become heterogeneous-only, P01-F, P03-C, P09-C and P09-D; these would qualify as set-only under the committed definition and carry no decision-effect assessment. Three same-model-only groups empty entirely.

The audit does not deliver a preference-based sensitivity result. It retained 31 of the original 37 set-only groups, including at least one previously rated Level 2 group in every task supported under the primary analysis. Four previously excluded candidates became eligible but were never coded, four mixed groups became heterogeneous-only without decision-effect assessment, and four preferred revisions, at P01, P03, P05 and P07, included material later excluded by the audit. The existing runs

therefore do not provide a complete preference-based sensitivity result, and the eight-of-twelve outcome stands solely as the precommitted primary analysis. Appendix J gives the descriptive account, and the full 216-row audit accompanies this paper as a data file.

5. Interpretation

Three fresh sessions of the originating model, given an identical brief against the same analysis, did not surface 37 consideration groups that three other models surfaced, and at least one such group appeared in every task. Assessed individually, the distinctive additions changed the specification of decisions, with 35 of 37 groups rated Level 2 by the originating model, no individual group rated Level 3, and the largest scoring movement in analytical reach. Additional considerations did not guarantee a preferred answer: 35 qualifying groups produced twelve revisions of which eight were preferred blind, three tied and one lost to its original.

These are facts about six specific critiques per task under one brief. They do not establish that the originating model could never surface the missing considerations under further sampling, different settings or a differently worded brief; three same-model draws sample a small part of the model's output distribution, and the eighteen same-model-only groups show that fresh sessions of one model do produce material distinctive to a run. Several mechanisms are compatible with the pattern and the design does not separate them: the heterogeneous reviewers may attend to different considerations because their training and tuning emphasise different material, or they may simply be strong reviewers whose output any three strong reviewers might have matched, a reading kept live by the concentration of set-only contributions in GPT-5.6 Sol, 18 of the 41 candidates. Nor does the preference count isolate what the additions contributed. Each revision was a complete fresh response, so the comparison shows how revisions incorporating heterogeneous additions were assessed against their originals under this procedure; it does not isolate the contribution of those additions from redrafting or ordinary sampling variation.

The preference results are model judgements about model outputs. The deciding voice in six of the eight supported outcomes was the adjudicator, a model that also served as heterogeneous reviewer H1, and the eligibility screen that defined the candidate pool ran on reviewer-assigned labels that the post-collection audit found noisy in both directions. The prompt-level preference count was precommitted as the study's primary result. The later concentration of dimension scores at the top of the scale limits the interpretive value of those secondary measures; the mean change of +0.3333 on a six-point scale records an ordering under this procedure and measures neither the size nor the objectivity of any improvement.

6. What the study establishes and what it does not

Table 5 sets out the claims against the evidence.

Table 5. Assessment of claims

Claim	Assessment	Basis
Three heterogeneous reviewers surfaced decision-relevant considerations that three matched same-model reviews did not surface, within this case, panel and brief.	Supported	37 set-only groups across twelve tasks, at least one per task; 35 rated Level 2 by the originating model.
Assessed individually, the distinctive additions changed decision specification, and no individual group changed a decision route.	Supported	35 of 37 groups at Level 2 and no group at Level 3; the combined revisions changed the stated route at P06, P07 and P10 (section 4.7).

Claim	Assessment	Basis
Revisions incorporating the qualifying additions were preferred under this model-based blind assessment in most tasks.	Qualified	8 of 12 supported outcomes; 6 of the 8 decided by adjudication; both scorers preferred the revision only at P03 and P09.
The primary supported outcomes would survive a stricter eligibility screen.	Not established	The exploratory audit retains 31 of 37 set-only groups, with at least one previously rated Level 2 group in every supported task, but four newly eligible candidates were never coded and four newly heterogeneous-only groups carry no decision-effect assessment; no complete preference-based sensitivity result exists.
The originating model could not have surfaced these considerations under any conditions.	Not established	Three same-model draws per task under one brief; further sampling, other settings and other briefs were untested.
Heterogeneous review is generally superior to same-model review.	Out of scope	One constructed case, one originating model, one panel, twelve dependent tasks; descriptive pilot by design.
The advantage came from model-family diversity as such.	Not established	Family diversity and individual reviewer capability are confounded; 18 of 41 set-only candidates came from one reviewer.
Every distinctive addition improved the final answer.	Not supported	35 qualifying groups against 8 supported tasks; one revision lost to its original and three tied; preference under this procedure is the only quality measure available.
The additions frequently reversed top-level decisions.	Not supported	No individual group rated Level 3; route wording shifted in three combined revisions, one of which lost the blind comparison.
The three same-model reviews raised no distinctive considerations of their own.	Not supported	18 same-model-only groups existed; their decision relevance was not assessed under the committed design.
The study compared a heterogeneous revision against a matched same-model revision.	Out of scope	No same-model revision arm existed; the same-model critiques served to define overlap and set-only status.
The blind scores measure analytical quality as an expert human panel would.	Not inferable	All coding, scoring, adjudication and decision-effect assessment came from language models; 132 of 144 dimension scores at the scale maximum.

7. Limitations and deviations

The case is one constructed transaction, divided into twelve tasks that share its facts. The tasks are dependent draws on a single fact pattern, so the twelve outcomes cannot be read as twelve independent samples, and no rate for other cases, sectors or task types follows from them.

The panel is fixed and named, and its capability is confounded with its diversity. The originating model was Claude Opus 5 and the heterogeneous panel was GPT-5.6 Sol, Muse Spark 1.2 and DeepSeek V4 Pro; a different originating model or panel could produce a different set-only count in either direction, and the design cannot attribute the observed groups to family diversity rather than to

the particular strengths of these three reviewers. Two panel substitutions preceded the freeze, Muse Spark 1.2 for Gemini 3.1 Pro Preview and Mistral Medium for Kimi K2 Thinking (Deviations D012 to D014), so the roster tested differs from the roster first specified. Reasoning settings could not be matched in verified effect across providers, and sampling stayed at provider defaults.

The design has no matched revision arm, and each revised analysis was a complete fresh response. The comparison shows how revisions incorporating heterogeneous additions were assessed against their originals under this procedure; it does not isolate the contribution of those additions from redrafting or ordinary sampling variation, and whether revisions built from the same-model critiques would have been preferred as often was not tested. The mirror inventory is likewise unassessed: the eighteen same-model-only groups, and the four groups the audit finds newly heterogeneous-only, carry no decision-effect ratings.

Every judgement in the pipeline came from a language model, under blinding and precommitted rules. The decision-effect assessor was the analysis's own author, and the direction of any self-assessment bias is unclear: a model may resist findings that it missed something or credit them too readily, and the recorded concentration at Level 2, with no group at Level 0 or Level 3, is consistent with either tendency operating at the margins. The adjudicator was not independent of the review panel, since GPT-5.6 Sol served as reviewer H1 and as adjudicator in separate blinded sessions; seven of twelve final preferences required adjudication and six of the eight supported outcomes rest on the adjudicated preference, so a different adjudicator could change the headline count without any change in the analyses. No claim about expert human agreement is available for any of these judgements.

The measurement instruments limit what the numbers can carry. The 0-to-2 scale compressed, with 132 of 144 dimension scores at the maximum, so the small positive means in Table 3 record ordering rather than measured size. The A and B allocation was randomised once per pair and not counterbalanced, so positional effects documented in the judge literature are randomised across pairs rather than removed. The eligibility screen ran on reviewer-assigned basis labels that the post-collection audit found inconsistent in both directions, and the audit that documents this is itself a single exploratory, non-blinded pass applying a stricter rule; its reclassifications are judgements, and the four excluded candidates it classes as eligible cannot be restored to the coded record without rerunning collection.

The neutral statements were not rewritten. Neutralisation removed titles, consequences, basis labels and identifiers and pooled the bare statements under shuffled identifiers, and every neutral statement is character-identical to its candidate statement, so the blinding is structural and stylistic cues remain in the text. The record does not name the agent who prepared the neutral statements and documents no fidelity check; the character-identity finding is the post-collection audit's, and it also confirms that the blinded partition was made over the reviewers' own sentences, so no regrouping is required.

The 900-word output limit failed in practice, with 22 of 24 primary analyses exceeding it (Deviation D053); the breach was discovered after collection and left uncorrected because truncation would have destroyed collected evidence, and its bias direction is uncertain in both arms. Some session metadata is incomplete: the eight P02 session identifiers were never recorded (Deviation D052), and the two failed P12 revision attempts exist only inside Deviation D050. The blind-pair texts in the workbook are restorations from the verbatim raw outputs and the frozen source key, verified character-exact. Each recorded output is one stochastic response under the recorded settings, and re-running any stage could move individual cells.

8. What remains unresolved

The most direct extension is the arm this study did not run: folding the same-model critiques into a matched revision under identical machinery and comparing the two revisions blind. A matched same-model revision arm would permit direct comparison between revisions incorporating same-model and

heterogeneous critiques, and would control the general effect of redrafting more directly than the original-versus-revision comparison reported here. It would not on its own separate critique effects from ordinary sampling variation, because each arm would contain a single stochastic revision draw; repeated revision draws, or a fresh-redraft condition receiving no critique material, would be needed for that. The symmetric assessment stands open for a related reason: the eighteen same-model-only groups, and the four newly heterogeneous-only groups the sensitivity audit identifies, have no decision-effect ratings. Similar qualification rates in both directions would demonstrate decision-relevant coverage in both directions; they would not establish equal value without matched revisions and matched scoring.

Whether the effect comes from diversity or from capability remains untested; swapping single panel members, running three strong same-family reviewers at varied settings against three families of matched strength, or repeating the design with a different originating model would begin to separate the readings. The assessment instrument invites the same treatment: a finer scale, counterbalanced pair orders, an adjudicator from outside the review panel and a human expert pass on a sample of pairs would each remove a dependence this study discloses. The drift the combined revisions showed towards deferral pending evidence, in three tasks where no single consideration justified a route change, remains unexplained, and whether a second heterogeneous panel reviewing the revised analyses would find another 37 set-only groups or almost no new ones is unknown.

9. Data availability and reproducibility

The complete study record is a single workbook of fifteen sheets: the committed design, the model and assessor register, the scenario and instruments, a run log of all 118 planned and retry generation calls with session identifiers and statuses, verbatim raw outputs for every call that produced output, the candidate source key, the blinded equivalence coding, the decision-effect record, the blind pairs, the source key, the blind assessments, the calculated findings, the deviation register D001 to D053, and a publication-changes sheet documenting every post-collection correction. A publication copy, its change log and a verification note listing every headline number against its source rows accompany this paper, together with the 216-row neutralisation and eligibility audit file, a post-collection verification script reconstructed during publication revision, and the figure-generation script with its data files and dependency list; the verification note contains the run instructions. The script's 21 checks cover nine areas, the primary candidate and eligibility totals, the neutral-statement identity, the equivalence-group totals, the decision-effect level totals, the aggregate final blind preferences, the scorer agreement totals, the dimension means, the word-limit breaches and the blind-pair text correspondence; values outside that coverage are traced to cited workbook cells in the verification note, and the sensitivity figures derive from the audit file. Every count in the paper traces to identified rows, and every primary generation call except the eight P02 rows carries a recorded session identifier.

The workbook contains the full verbatim outputs of a commercial frontier-model panel and the source keys that unblind them, so it is not posted publicly. It is available from the author on request to bona fide researchers for peer review, replication and academic citation. Requests should be sent to research@mkai.org.

10. Disclosures

Conflict of interest. The author is the founder of MKAI and of Reality & Reason. This research was conducted independently and received no third-party funding.

Use of AI. Language models are the instruments of this study by design: the coders, scorers, adjudicator and decision-effect assessor named in Table 1 are all models, and their roles, blinding and

limits are reported in full. AI assistance was also used in study operation, including prompt transmission, logging, verification passes, workbook assembly and drafting. The post-collection neutralisation and eligibility audits reported in sections 3.9 and 4.8 were performed by the AI reviser session that prepared this publication package, in a single exploratory, non-blinded pass from the workbook and case text; the reviser's exact model identity and settings were not captured in the study record. The related-research references in section 2 were verified against primary sources with AI assistance. The author retains full responsibility for the research design, the findings and the final text.

Scope. This study documents the observable behaviour of one originating model, one three-model review panel and one model-based assessment pipeline on one constructed acquisition case. It does not evaluate the overall quality of any named model or provider, it does not claim that any organisation operates the review workflow tested here, and it makes no claim about how organisations in general use frontier models.

11. References

Foster-Fletcher, R. (2026). The Subtraction Study: Measuring Capability Reduction in Enterprise AI Configurations. MKAI Inquiry Working Paper.

Foster-Fletcher, R. (2026). The Separation Study: Whether Decision Policy Can Be Separated From Action Constraints in Enterprise System Prompts. MKAI Inquiry Working Paper.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., et al. (2025). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 238-255. https://doi.org/10.18653/v1/2025.acl-short.20.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. (2024). ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. International Conference on Learning Representations 2024. arXiv:2308.07201.

Chiang, C.-H., and Lee, H.-y. (2023). Can Large Language Models Be an Alternative to Human Evaluations? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 15607-15631. https://doi.org/10.18653/v1/2023.acl-long.870.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2024). Improving Factuality and Reasoning in Language Models through Multiagent Debate. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 11733-11763. arXiv:2305.14325.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., et al. (2024). A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations 2024. arXiv:2310.01798.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. Advances in Neural Information Processing Systems 36. arXiv:2303.17651.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. Advances in Neural Information Processing Systems 37. arXiv:2404.13076.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems 36. arXiv:2303.11366.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. (2024). Large Language Models are not Fair Evaluators. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 9440-9450. https://doi.org/10.18653/v1/2024.acl-long.511.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track. arXiv:2306.05685.

Appendix A. The constructed case and task battery

A.1 Study scenario, supplied verbatim with every call

Meridian Industrial Group manufactures precision components for aerospace and medical-device companies. Its annual revenue is approximately £480 million, it is profitable, and its net debt is 1.9 times EBITDA. The board has set an internal net-debt ceiling of 3.5 times EBITDA, while Meridian's principal lending covenant is 4.0 times EBITDA. Meridian is considering the acquisition of Calder Systems, a founder-led manufacturer with annual revenue of approximately £110 million. Calder grew by 28 per cent during the previous year and made an operating loss of £9 million. Its proprietary additive-manufacturing process is protected by seven patents with between eleven and fifteen years remaining. Two aerospace components using the process have completed customer qualification, while medical-device validation remains incomplete. Calder's two largest customers provide 34 per cent and 26 per cent of its revenue. Their contracts expire in eleven and seventeen months, and both contracts contain change-of-control provisions that permit renegotiation or termination. Calder's founder-CEO owns 38 per cent of the company, directs the technical programme and manages both major customer relationships. He has offered an eighteen-month transition after completion. Fourteen employees form Calder's principal technical team, and no retention agreements have been signed. The proposed price is £340 million, funded through Meridian's cash and new debt. New borrowing would carry an expected interest rate of 7.4 per cent. Meridian's advisers calculate a post-acquisition net-debt ratio of 3.4 times EBITDA if £18 million of forecast annual synergies are achieved within two years, and 4.1 times EBITDA if those synergies do not appear. Calder has granted Meridian three weeks of exclusivity and says another strategic buyer remains interested. The executive committee must decide whether to proceed at £340 million, seek different terms or leave the transaction.

A.2 The twelve analytical tasks

P01, Memory. Drawing on common patterns from comparable acquisitions, where an established manufacturer buys a smaller founder-led company for its process and patents, identify the precedents or recurring deal patterns that should inform Meridian's decision. Distinguish any named external example from a general pattern, and explain the decision consequence of each pattern.

P02, Evidence. Identify the evidence that would most strengthen or weaken the case for proceeding at £340 million. Separate evidence Meridian probably possesses from evidence it still needs, and explain how each material absence should affect the decision.

P03, Timing. Assess whether Meridian should act during the three-week exclusivity period. Compare the exposure created by acting now with the exposure created by waiting, using Calder's growth, customer renewals, founder dependence, financing cost and the claimed competing interest.

P04, Attention. Identify the considerations that Meridian's committee is most likely to underweight because the patents, process and growth rate dominate the proposed transaction. Explain the decision consequence of each displaced consideration.

P05, Judgement. Weigh the complete case and state Meridian's strongest current decision route. Identify the reasoning that determines that route and the single change that would alter it.

P06, Framing. Examine whether the stated choice among proceeding, seeking different terms or leaving the transaction frames Meridian's decision adequately. Identify concealed assumptions or excluded choices and provide a stronger frame where the evidence supports one.

P07, Options. Identify the available structures beyond an immediate acquisition at £340 million. Consider staged ownership, licensing, commercial partnerships, options, earn-outs and other arrangements that might secure Meridian's objective. Explain what each serious option secures and gives up.

P08, Imagination. Describe two or three specific ways the transaction could develop during its first three years, including an unexpected source of success and a failure mechanism absent from the obvious risk list. Connect each scenario to facts supplied in the case.

P09, Incentives. Examine how the interests of Calder's founder, its two major customers, both companies' advisers and the people within Meridian supporting the deal could affect the price, urgency, evidence and proposed terms.

P10, Risk tolerance. Assess the risk appetite implied by Meridian using cash and new debt to acquire a loss-making, founder-dependent and customer-concentrated company at a premium. Identify where the downside concentrates and whether the proposed terms fit Meridian's stated financial limits.

P11, Social validation. Assess how the claimed competing bidder, adviser support, board expectations and possible competitor reaction could reassure Meridian without improving the underlying economics. Identify any decision that would change if those signals disappeared.

P12, Narrative. Examine the account of a proven manufacturer acquiring a fast-growing innovator with protected technology. Identify which case facts that account elevates, which facts it suppresses and how a more sceptical account would alter Meridian's decision.

The instrument groups P01 to P03 under analytical reach, P04 and P05 as bridge tasks and P06 to P12 under decision conversion.

A.3 The excluded calibration case and tasks

Northbank Health Devices is considering a ten-year exclusive licence for a new antimicrobial coating developed by Vectra Materials. Northbank would pay £60 million upfront and a five per cent royalty. The coating has completed one hospital trial, while a second trial remains unfinished. One competitor has challenged a related Vectra patent. Vectra produces the coating at a single plant and has offered no alternative supply arrangement. Northbank's board meets in ten days, and Vectra says another licensee is waiting.

C01, calibration judgement task: assess whether Northbank should accept the proposed licence, seek different terms, defer pending evidence or leave the opportunity, stating the decision route and the evidence that determines it. C02, calibration framing task: examine whether a ten-year exclusive licence is the right frame for Northbank's objective and identify a stronger structure if the supplied facts support one.

Appendix B. Response contracts and critique briefs

B.1 Shared response contract for original and revised analyses

Write no more than 900 words. Use the supplied case as the factual basis for your analysis. You may draw on general knowledge, although any external example or factual claim must be identified as external and treated as unverified within this study. Separate supplied facts, inferences and missing evidence. End with four labelled fields: Decision route; Proposed terms or structure; Evidence required before commitment; Single switch condition. The available routes are proceed at £340 million, seek different terms, defer pending specified evidence, or leave the transaction.

B.2 Critique brief, corrected version in force from C02 onward and used for all primary tasks

Below is a strategic analysis of the Meridian-Calder acquisition. Read the case and analysis closely. Identify up to three specific risks, assumptions, alternatives or decision conditions that the analysis overlooks or gives insufficient weight. Include only considerations that could change the decision route, proposed terms, evidence required before commitment or switch condition. Do not repeat a consideration already present in the analysis. Do not provide general commentary. Return no more than three candidates. If more occur to you, rank them and omit everything after the third. Return each consideration separately using this format: Candidate title, maximum twelve words; Candidate statement, one or two sentences; Decision consequence, one sentence; Basis, selected from supplied case fact, inference from supplied facts, or external unverified knowledge. The complete response must remain within 280 words.

The original brief, used only for the seven C01 calibration calls, was identical except that it lacked the two truncation sentences, 'Return no more than three candidates. If more occur to you, rank them and omit everything after the third.'; two calibration critiques returned four candidates, and a pre-freeze correction added those sentences (Deviations D019 and D020). Under the accompanying rule, wherever a critique returned more than three candidates, only the first three in returned order entered the candidate key; the full response was preserved verbatim. Every primary critique returned exactly three.

For C02 only, whose framing task fits no four-route ending, the response contract ended with three labelled fields, recommended structure, evidence required before commitment, and the condition under which the originally proposed ten-year exclusive licence would remain the right frame (Deviation D023).

Appendix C. Model register

Generator and reviewer models. F1: Anthropic Claude Opus 5, identifier claude-opus-5, TypingMind catalogue release date 2026-07-24, reasoning effort High. H1: OpenAI GPT-5.6 Sol, identifier gpt-5.6-sol, catalogue release date 2026-07-09, reasoning effort High, accessed through the OpenAI response API type; the register records the verification that Sol is the frontier variant of the GPT-5.6 family. H2: Meta Muse Spark 1.2, identifier muse-spark-1.2, catalogue release date 2026-08-05, confirmed as a direct Meta endpoint in public preview; the model card lists a thinking mode, and the effect of the platform's High reasoning setting on this model was not separately confirmed. H3: DeepSeek V4 Pro, identifier deepseek-v4-pro, catalogue release date 2026-04-24; the model card lists no reasoning-effort control, so the High setting has no effect on this model.

Assessor models. Coder 1 and Scorer 1: Kimi K3, identifier kimi-k3, Moonshot, direct endpoint. Coder 2 and Scorer 2: Mistral Medium, identifier mistral-medium-2604, Mistral, direct endpoint; the platform displays this pinned snapshot under the label mistral-medium-latest, a labelling discrepancy

investigated and resolved before the freeze (Deviation D022) and re-verified when the catalogue later showed duplicate display labels (Deviation D048). Adjudicator for coding and scoring: GPT-5.6 Sol, identifier gpt-5.6-sol, in fresh blinded sessions, selected under the fallback rule of Deviation D008.

Common settings for every call: TypingMind chat interface with direct provider API keys; a fresh chat per call; system instruction fixed as 'You are a helpful AI assistant.'; temperature, top-p and maximum output tokens at provider defaults; account-level Web Browser and Code Sandbox tools off; all five installed plugins off; no knowledge base attached; no persistent-memory feature present.

Substitution history. H2 was specified as Google Gemini 3.1 Pro Preview, identifier gemini-3.1-pro-preview, and was replaced by Muse Spark 1.2 on 13 August 2026 after five attempts returned an identical provider-side quota error (Deviations D012 and D013). Coder 2 and Scorer 2 were specified as Kimi K2 Thinking, identifier kimi-k2-thinking, and were replaced by Mistral Medium on 13 August 2026 after a model-specific access fault (Deviation D014). Both replacements preceded the instrument freeze of 14 August 2026.

Appendix D. Coding rules

D.1 Eligibility rule

Each candidate carried one basis value assigned by its reviewer: supplied case fact, inference from supplied facts, or external unverified knowledge. Candidates on either of the first two bases entered coding. A candidate on the external basis entered coding only if its external claim was verified within the study; no external claim was verified, so all 64 such candidates were excluded and preserved. Three responses carried compound basis lines naming two values; each was resolved to a single value by precedent-matching and logged (Deviations D029, D033 and D044).

D.2 Equivalence-coding instruction, sent identically to both coders in fresh blinded sessions

CASE: [calibration or primary case text as applicable]. Below is a set of independently produced statements, each identifying a specific risk, assumption, alternative or decision condition relevant to the case above. Each statement has a neutral ID. Group the statements by the underlying consideration each one describes: statements that describe substantially the same underlying consideration belong in the same group, even where the wording, emphasis or proposed remedy differs. Statements addressing genuinely different considerations belong in different groups; a statement with no substantive match to any other belongs in a group of its own. Base each grouping decision only on the substance of the consideration described, not on wording, length, phrasing or style. Return your answer as a list of groups. Label each group with a single capital letter starting at A. For each group, list every neutral ID belonging to it. Every neutral ID supplied must appear in exactly one group. Do not merge or split a statement's content — group by ID only. Do not add commentary, explanation or a summary before or after the list.

Coders received the case and the shuffled neutral statements for one task at a time, with no candidate key and no indication of source model, family or review arm. Where the two partitions disagreed, the adjudicator received the identical material in a fresh blinded session under the same instruction.

Appendix E. Decision-effect rubric

Instruction to F1, one fresh blinded session per set-only equivalence group, with no reference to which models raised the consideration:

CASE: [calibration or primary case text as applicable]. TASK: [original task text as applicable]. ORIGINAL ANALYSIS: [F1's original analysis for this prompt, verbatim]. ADDITIONAL CONSIDERATION: [the neutral statement(s) making up this set-only equivalence group]. The additional consideration above was not present in your original analysis. Re-examine your original analysis in light of it only. State whether, and how, this consideration would change your analysis, using exactly one of the following four levels. Level 0: no material change. Level 1: changes supporting detail or evidence discussion only, with no change to terms, evidence required before commitment, or the decision route. Level 2: changes proposed terms or structure, evidence required before commitment, or the switch condition, without changing the decision route itself. Level 3: changes the decision route stated in your original analysis. State the level first, on its own line as 'Level: [0/1/2/3]', then explain your reasoning in no more than 150 words.

Levels 2 and 3 qualified as decision-relevant under the committed design. For C02 only, the level definitions were re-mapped onto that task's three-field contract before the freeze (Deviation D026).

Appendix F. Blind scoring rubric

Instruction to both scorers, sent identically and independently in fresh blinded sessions; the adjudicator received the same anonymised pair only where the two primary preferences disagreed:

CASE: [case text as applicable]. TASK: [task text as applicable]. Below are two independently produced analyses responding to the same case and task, labelled Analysis A and Analysis B. You do not know, and should not attempt to infer, which model produced either analysis, whether either has been revised, or anything about the process that produced them. Score each analysis separately on three dimensions, each on a scale of 0 to 2. Evidential discipline (0-2): does the analysis cleanly separate supplied facts, inferences and missing evidence, and treat unverified claims as unverified? Decision specification (0-2): is the stated decision route specific, actionable and directly supported by the analysis, with proposed terms, evidence required and a switch condition that are concrete rather than generic? Analytical reach (0-2): does the analysis identify risks, dependencies or structuring options beyond the most obvious reading of the case, without inventing facts not in evidence? For each analysis, give three scores (0, 1 or 2) in the form 'A evidential discipline: [0/1/2]', 'A decision specification: [0/1/2]', 'A analytical reach: [0/1/2]', then the same three lines for B, each with a brief one-line justification. Then, on its own final line, state your overall preference as exactly one of: 'Overall: A', 'Overall: B', or 'Overall: Tie'. Do not explain the overall preference beyond the six dimension scores already given, and do not add any other commentary.

Appendix G. Prompt-level results matrix

Table G1. Candidates and equivalence groups by task

Prompt	Raw candidates	Eligible (same-model / heterogeneous)	Equivalence groups	Mixed	Same-model-only	Set-only
P01	18	11 (5 / 6)	7	1	2	4
P02	18	12 (6 / 6)	6	2	1	3
P03	18	13 (6 / 7)	8	1	2	5
P04	18	9 (4 / 5)	5	2	1	2
P05	18	13 (6 / 7)	8	3	1	4
P06	18	11 (6 / 5)	6	1	2	3
P07	18	12 (6 / 6)	8	2	2	4

Prompt	Raw candidates	Eligible (same-model / heterogeneous)	Equivalence groups	Mixed	Same-model-only	Set-only
P08	18	14 (7 / 7)	9	2	3	4
P09	18	13 (5 / 8)	5	3	0	2
P10	18	14 (8 / 6)	7	3	2	2
P11	18	13 (7 / 6)	4	2	1	1
P12	18	17 (8 / 9)	7	3	1	3
Total	216	152 (74 / 78)	80	25	18	37

Table G2. The 37 set-only groups, their sources and decision-effect levels

Prompt	Group	Effect ID	Neutral ID(s)	Source candidate(s)	Level
P01	B	E004	N029	P01-H1-E2	2
P01	C	E005	N030	P01-H3-E1	2
P01	D	E006	N031	P01-H3-E3	2
P01	G	E007	N035	P01-H1-E3	2
P02	D	E008	N046	P02-H1-E2	2
P02	E	E009	N048, N050	P02-H2-E1, P02-H3-E3	2
P02	F	E010	N049	P02-H2-E3	2
P03	D	E011	N058	P03-H1-E2	2
P03	E	E012	N059	P03-H1-E3	2
P03	F	E013	N060, N063	P03-H2-E1, P03-H3-E2	2
P03	G	E014	N061	P03-H2-E3	2
P03	H	E015	N062	P03-H3-E1	2
P04	D	E016	N068	P04-H1-E2	2
P04	E	E017	N071	P04-H3-E1	1
P05	E	E018	N079	P05-H1-E1	2
P05	F	E019	N081	P05-H2-E2	2
P05	G	E020	N082	P05-H2-E3	2
P05	H	E021	N083	P05-H3-E1	2
P06	D	E022	N092	P06-H1-E1	2
P06	E	E023	N095	P06-H3-E1	2
P06	F	E024	N096	P06-H3-E3	2
P07	E	E025	N103	P07-H1-E1	2
P07	F	E026	N104	P07-H1-E3	2
P07	G	E027	N106	P07-H2-E2	2
P07	H	E028	N108	P07-H3-E2	2
P08	F	E029	N116	P08-H1-E1	2

Prompt	Group	Effect ID	Neutral ID(s)	Source candidate(s)	Level
P08	G	E030	N117	P08-H1-E2	2
P08	H	E031	N118	P08-H1-E3	2
P08	I	E032	N121	P08-H3-E1	2
P09	B	E033	N128, N133	P09-H1-E1, P09-H3-E1	2
P09	E	E034	N132	P09-H2-E3	2
P10	F	E035	N145	P10-H1-E2	2
P10	G	E036	N147	P10-H2-E3	2
P11	D	E037	N159	P11-H1-E3	2
P12	E	E038	N171	P12-H1-E1	2
P12	F	E039	N173, N179	P12-H1-E3, P12-H3-E3	2
P12	G	E040	N175	P12-H2-E2	1

Table G3. Word counts of the 24 primary analyses against the 900-word limit

Prompt	Original words	Revised words	Within limit
P01	916	938	Neither
P02	989	978	Neither
P03	880	932	Original only
P04	923	958	Neither
P05	939	863	Revised only
P06	912	937	Neither
P07	917	1,079	Neither
P08	1,001	1,036	Neither
P09	948	1,037	Neither
P10	994	966	Neither
P11	930	1,067	Neither
P12	954	1,012	Neither

Appendix H. Deviation register

The workbook logs 53 deviations. Every entry preserves its full description, reason, possible effect and approval trail; the register below gives one line each. Entries D001 to D027 precede the instrument freeze and concern setup, calibration and pre-freeze corrections; D028 onwards arise during and after primary collection.

ID	Date	Task	Summary
D001	2026-08-11	Setup	Account defaults found contrary to protocol: system instruction set, web and sandbox tools on; flagged for correction.

ID	Date	Task	Summary
D002	2026-08-11	C01	First C01 original-analysis call never transmitted; platform requested an API key.
D003	2026-08-11	Setup	H2 and H3 present in catalogue and uncallable for want of API keys; stop condition invoked.
D004	2026-08-13	Setup	Google and DeepSeek keys added; D003 resolved; H2 and H3 confirmed accessible.
D005	2026-08-13	Setup	Specified assessor models Qwen 3.8 Max and MiniMax M2.7 unavailable as direct endpoints; stop condition invoked.
D006	2026-08-13	Setup	Account preparation: tools and plugins switched off; system instruction still set.
D007	2026-08-13	Setup	Two inert duplicate catalogue entries created by misclick; cosmetic only.
D008	2026-08-13	Design	Assessor register corrected by owner: Qwen and MiniMax requirements withdrawn; GPT-5.6 Sol assigned adjudicator; dual-role limitation logged.
D009	2026-08-13	Design	System instruction fixed as 'You are a helpful AI assistant.' for every call, in place of blanking.
D010	2026-08-13	C01	Automation sent an incomplete prompt when a blank line acted as Enter; retry with single-line construction.
D011	2026-08-13	C01	Second transmission fault on long critique prompts; two invalid calls excluded and re-run.
D012	2026-08-13	C01	Gemini 3.1 Pro Preview blocked by an identical provider quota error across five attempts.
D013	2026-08-13	C01 on	H2 substituted: Muse Spark 1.2 replaces Gemini 3.1 Pro Preview, approved by the study owner.
D014	2026-08-13	Design	Coder 2 and Scorer 2 substituted: Mistral Medium replaces blocked Kimi K2 Thinking, approved by the study owner.
D015	2026-08-13	C01	A calibration critique returned four candidates; first three retained under the mechanical rule.
D016	2026-08-13	Instrument	Coding and decision-effect instruction wording drafted by the operator to complete the instrument; not separately owner-approved.
D017	2026-08-13	C01	Platform unreachable during decision-effect stage; calls completed after connectivity returned.
D018	2026-08-13	C01	Transcription error found on re-verification: a fourth candidate had been mis-handled; corrected mechanically.
D019	2026-08-13	C02 on	Critique brief corrected pre-freeze to instruct ranking and truncation to three candidates.
D020	2026-08-13	All	Formal truncation rule: first three candidates in returned order enter coding; responses preserved verbatim.
D021	2026-08-13	C01	Stale coder-column values found by post-write verification and corrected before any downstream use.
D022	2026-08-13	Register	Display label mistral-medium-latest resolved to pinned identifier mistral-medium-2604; notes corrected.

ID	Date	Task	Summary
D023	2026-08-13	C02	Response contract adapted for the calibration framing task, which fits no four-route ending.
D024	2026-08-13	C02	Fresh-chat isolation failure contaminated one critique; excluded and re-run in a verified fresh chat.
D025	2026-08-13	C02	Neutral-ID namespace collision caught; single global sequence imposed.
D026	2026-08-13	C02	Decision-effect level wording re-mapped to the C02 contract fields.
D027	2026-08-14	C02	Mis-dated calibration scoring rows corrected before any reference to them.
D028	2026-08-14	P01	First three-way scoring split: original, tie and revised across the three assessments; adjudicated preference final under the committed rule.
D029	2026-08-14	P02	Compound basis line from H3 resolved by precedent; candidate excluded as external unverified.
D030	2026-08-14	P02	Interrupted chunked transmission discarded; verified single-call transmission used before any send.
D031	2026-08-14	P03	First coding disagreement; adjudicated in a fresh blinded session.
D032	2026-08-14	P03	Decision-effect prompt-construction bug caught before any transmission.
D033	2026-08-14	P04	Two compound basis lines from H3 resolved by precedent; one affected an inclusion outcome.
D034	2026-08-14	P04	Second scoring disagreement; adjudicated.
D035	2026-08-14	P05	Second coding disagreement; adjudication matched Coder 1 exactly.
D036	2026-08-14	P05	Third scoring disagreement; adjudicated.
D037	2026-08-14	P06	Third coding disagreement; adjudication matched Coder 1 exactly.
D038	2026-08-14	P06	API-key lapse interrupted the decision-effect stage; no output lost.
D039	2026-08-14	P06	Fourth scoring disagreement; adjudicated preference favoured the original analysis.
D040	2026-08-14	P07	Fourth coding disagreement; adjudication placed N103 as a singleton, matching neither coder.
D041	2026-08-14	P07	Fifth scoring disagreement; adjudicated.
D042	2026-08-14	P08	Fifth coding disagreement; adjudication matched Coder 1 exactly.
D043	2026-08-15	P08	Two transient connectivity failures on Scorer 2's first attempt; successful verified retry.
D044	2026-08-15	P10	Two compound basis lines from H3 resolved by precedent; two candidate-key rows added late.
D045	2026-08-15	P10	Sixth coding disagreement on N147; adjudication matched Coder 1 exactly.

ID	Date	Task	Summary
D046	2026-08-15	P10	Sixth scoring disagreement; adjudicated.
D047	2026-08-15	P10	Contemporaneous-logging shortfall: outputs saved locally before Run log entry; reconstructed and noted.
D048	2026-08-15	P11	Duplicate catalogue display labels re-verified; pinned Mistral snapshot confirmed in use.
D049	2026-08-15	P11	Seventh scoring disagreement; adjudicated.
D050	2026-08-15	P12	Revision blocked twice by provider credit exhaustion; completed on a third attempt after restoration; failed attempts preserved here.
D051	2026-08-15	P12	Source-attribution tag on the revision file caught and removed before any scorer transmission; programmatic guard confirmed removal.
D052	2026-08-15	P02	All eight P02 session identifiers found unrecorded at final verification; gap preserved, no identifiers reconstructed.
D053	2026-08-16	P01-P12	900-word analysis limit breached in 22 of 24 primary analyses; full counts recorded; no output truncated.

Appendix I. Data map

The study workbook contains fifteen sheets. Read me and status records the workbook guide, the stage-completion register and the collection sequence. Committed design fixes the proposition, the two review conditions, the candidate and word limits, the set-only rule, the decision-effect levels, the supported-instance definition, the freeze status and date, the fixed system instruction and the assessor register. Model register contains the generator rows F1 and H1 to H3 and the assessor rows, with identifiers, access surfaces, settings, tool states and substitution history. Scenario and prompts contains the case, the twelve tasks, the calibration case and tasks, both critique briefs, the response contracts and the assessor instructions. Run log records one row per planned call, rows 5 to 116 covering C01, C02 and P01 to P12, with retries at rows 117 to 122, each with date, session identifier and status. Raw outputs preserves the verbatim text of every logged call that produced output. Candidate key lists all 252 extracted candidates with source run, review arm, basis, neutral statement, external-claim status, inclusion flag and neutral identifier. Equivalence coding records the blinded partitions of both coders, agreement flags, adjudicated groups and final groups for N001 to N179. Decision effects records one row per verified set-only group, E001 to E040, with sources, levels, qualifying flags and verbatim responses. Blind pairs preserves the fourteen A and B texts as transmitted. Score key records the concealed A and B mapping. Blind assessments contains all 42 assessment rows with dimension scores, overall preferences, decoded values, dates, session notes and agreement flags. Findings calculates the twelve prompt outcomes, the supporting counts and the mean dimension scores from the source sheets. Deviations contains the register D001 to D053. Publication changes documents every post-collection correction to the workbook, with the word-count audit and the reconciled totals.

A publication copy of the workbook, its change log and a verification note accompany this paper, together with the post-collection audit file and the verification and figure scripts. The verification note lists every headline number beside the sheet and rows that produce it.

Appendix J. Post-collection audit summary

The full audit is the 216-row data file accompanying this paper. The tables below summarise the entries that change anything.

Table J1. Eligible candidates reclassified as external unverified knowledge by the post-collection audit

Candidate	Neutral ID	Condition	Group	Ground for reclassification
P01-F1C1-E1	N037	Same-model	P01-F	Asserts what most credit agreements cap or exclude
P01-H1-E2	N029	Heterogeneous	P01-B	Asserts the scope of a freedom-to-operate opinion
P01-H2-E1	N038	Heterogeneous	P01-F	Asserts what lenders typically disallow
P01-H3-E3	N031	Heterogeneous	P01-D	Asserts legal consequences of share and asset purchase structures
P02-H1-E2	N046	Heterogeneous	P02-D	Imports product-liability, warranty and recall exposure classes
P03-F1C2-E3	N054	Same-model	P03-C	Asserts trailing-twelve-month covenant testing
P03-H1-E3	N059	Heterogeneous	P03-E	Asserts the scope of validity and freedom-to-operate opinions
P05-H2-E2	N081	Heterogeneous	P05-F	Imports minimum-liquidity and working-capital covenant types
P06-F1C1-E2	N086	Same-model	P06-A	Asserts that written confirmation is rarely given
P06-F1C3-E2	N090	Same-model	P06-A	Asserts that customers rarely give pre-completion commitments
P07-F1C2-E1	N098	Same-model	P07-B	Asserts that facility definitions often permit add-backs and cures
P07-F1C3-E1	N101	Same-model	P07-B	Asserts that lenders commonly measure adjusted EBITDA
P07-H1-E3	N104	Heterogeneous	P07-F	Imports latent defect, warranty and recall exposure classes
P08-F1C2-E2	N111	Same-model	P08-C	Asserts that covenant terms are negotiated pre-signing
P09-F1C1-E2	N123	Same-model	P09-C	Asserts that dominant customers frequently hold embedded IP rights
P09-F1C2-E1	N125	Same-model	P09-D	Asserts that covenants are typically tested on group EBITDA
P10-F1C1-E1	N136	Same-model	P10-A	Asserts that standard documentation often permits add-backs and cures
P10-F1C3-E1	N141	Same-model	P10-E	Asserts likely national-security and export-control screening
P11-F1C1-E2	N151	Same-model	P11-B	Asserts that facilities of this type usually carry further covenants
P11-F1C2-E3	N154	Same-model	P11-C	Asserts that founder-led programmes are a common source of defective assignment

Four excluded candidates reclassify in the other direction, as inference from supplied facts: P01-H2-E3, P03-H3-E3, P04-H1-E1 and P07-H3-E1. They were excluded at screening and never coded, so they cannot be restored mechanically. Eight candidates labelled supplied case fact reclassify as inference with no effect on eligibility: P03-H3-E1, P04-H3-E2, P05-H3-E2, P05-H3-E3, P06-H3-E1, P07-H2-E2, P07-H2-E3 and P09-H3-E2.

Table J2. Group-level and candidate-level consequences of the audit

Change	Detail
Set-only groups removed entirely (all previously rated Level 2)	P01-B (N029), P01-D (N031), P02-D (N046), P03-E (N059), P05-F (N081), P07-F (N104)
Set-only groups retained	31 of 37: 29 previously rated Level 2 and the two Level 1 groups (P04-E, P12-G)
Mixed groups becoming heterogeneous-only, no decision-effect assessment available	P01-F (N034), P03-C (N057), P09-C (N135), P09-D (N129, N130, N134)
Same-model-only groups emptied	P07-B, P08-C, P10-E
Newly eligible candidates, never entered into blind coding	P01-H2-E3, P03-H3-E3, P04-H1-E1, P07-H3-E1
Preferred revisions containing material excluded by the audit	P01-REV, P03-REV, P05-REV, P07-REV

Table J3. Descriptive account of the audit by task

Prompt	Coded candidates retained (same-model / heterogeneous)	Set-only groups retained	Retained groups previously rated Level 2	Newly heterogeneous-only, unassessed	Newly eligible, never coded
P01	7 (4 / 3)	2 of 4	2	1	1
P02	11 (6 / 5)	2 of 3	2	0	0
P03	11 (5 / 6)	4 of 5	4	1	1
P04	9 (4 / 5)	2 of 2	1	0	1
P05	12 (6 / 6)	3 of 4	3	0	0
P06	9 (4 / 5)	3 of 3	3	0	0
P07	9 (4 / 5)	3 of 4	3	0	1
P08	13 (6 / 7)	4 of 4	4	0	0
P09	11 (3 / 8)	2 of 2	2	2	0
P10	12 (6 / 6)	2 of 2	2	0	0
P11	11 (5 / 6)	1 of 1	1	0	0
P12	17 (8 / 9)	3 of 3	2	0	0
Total	132 (61 / 71)	31 of 37	29	4	4

The post-collection audit retained 31 of the original 37 set-only groups, including at least one previously rated Level 2 group in every task supported under the primary analysis. Four previously excluded candidates became eligible but were never coded, four mixed groups became heterogeneous-only without decision-effect assessment, and four preferred revisions included material later excluded by the audit. The existing runs therefore do not provide a complete preference-based sensitivity result; the eight-of-twelve outcome is the precommitted primary result and carries no sensitivity confirmation. The blinded preferences and revision texts predate the audit and are unchanged.