Doubt Attrition: methods and results

Companion to *Sources doubted in AI planning text* · Published 11 September 2026 · Version 1.2

Design and available source material

This is descriptive coding of existing files. The frozen protocol, dated 29 August 2026, calls it a verification pass for Chapter Nine of *The AI Gap* and explicitly excludes a benchmark or controlled experiment. The amended model review ended on 10 September 2026. No new source answers were elicited for this verification; subsequent model calls coded the preserved material.

Source A contains 1,447 lines and Source B contains 642. Together they contain 1,293 planning lines, 768 completed-answer lines and 28 excluded header or separator lines. Each file contains five completed answers and four planning blocks. Source A came from a challenge-round compilation and Source B from a capability-round compilation concerning AI and management judgement. Model identity is absent from the completed answers. The files do not establish planning-to-answer session identity, and visible planning text is not a validated record of internal reasoning.

Frozen input	SHA-256
Source A, pasted-text.txt	a6bc3d090487c662daacdfb2ea79dcabea08b8de3362d01ac0ab393847b61699
Source B, pasted-text B.txt	6289070c381168797346288dd7c7105fc87130289f1c027b51c56d3b49b371a4
Research ledger	e733679703f0113dfec2d32226c9d86d7d466b2f63e9b18c24a183c5d9d93584
Protocol v1.3	c8f6cdfff45d6a31259d10a4d8bc62fb0821e6c295b4e3d61cabfd25832c1e1e

Only sessions that exposed planning could contribute such material. The two files are therefore a selected corpus, with no sampling basis for generalising to a provider, a model family or the full original exercise.

The original exercise concerned AI capability and its consequences for companies and their leaders. Source A challenged claims about outside analysis and management judgement; Source B considered candidate capabilities and their material consequences, including software replication. The ledger was prepared on 28 August 2026. The supplied files do not establish a complete session chronology, the number of distinct generating models, or a reliable model identity for each answer. The coding also excludes a comparison set of sources never questioned in planning, so it cannot establish whether recorded doubt altered citation practice.

Units and separate analysis populations

The whole-file unit combines file, canonical source and doubt type. Doubt types cover existence, authorship, identifier, figure or reported result, publication venue or date, and other doubts with a written reason. Repeated markers for the same key are combined, while different doubt types remain separate. The original membership and candidate-to-instance mapping are retained in evidence/analysis/instances-v2.json.

Population	Unit	Denominator	Disputed units allowed to vary in main calculation
Whole-file, non-abandoned	Source-doubt instance within file	140	26
Whole-file, abandoned	Source-doubt instance within file	21	2
Adjacency, non-abandoned	Instance-block observation	148	11
Adjacency, abandoned	Instance-block observation	23	0

Abandonment means the planning passage discards the source in that passage. An instance left open or retained is included in the primary non-abandoned set. The 122 unidentifiable candidates excluded at the membership stage are reported in their own unit. They cannot be added to any denominator in the table. These populations describe overlapping material under different comparisons and must not be pooled as independent observations.

Whole-file matching compares a source questioned anywhere in a compilation with all its completed answers. Adjacency compares a planning block with the answer immediately following it. The latter depends on an unverified assumption about preserved session order and remains secondary.

Descriptive counts added after editorial review

The following tabulations were added on 11 September 2026 in response to Claude Fable's editorial review. They use the frozen membership and codes without new classification decisions and are reproduced by reproduce-descriptive-checks.mjs against descriptive-results.json.

The primary 140 instances cover 98 distinct canonical sources and 102 source-file identities: 71 in A and 31 in B, with four canonical sources occurring in both files. The 21 abandoned instances cover 19 sources and 19 source-file identities. The two populations share four canonical sources, leaving 113 across their union, or 117 source-file identities. These counts inherit the original source-identification decisions.

The supplied routed records contain 484 distinct candidate IDs: 304 identifiable source candidates, 122 unidentifiable candidates and 58 admissions-route candidates, with no overlap between these sets. The 304 identifiable candidates form 161 source-doubt instances, including 42 instances with several candidate markers. The 143 excess candidate memberships reflect consolidation of markers, without establishing how many were duplicate extractions. The original raw extraction manifest is absent from this selected bundle, so 484 is the total represented by these routed records, not a verified count of every initial extraction. The 58 admissions-route candidates include the named-source passage later excluded from that review.

Primary grouping	Instances	Fixed	Variable	At least one unhedged appearance	Added detail
Source A	98	73	25	21–46	28–53
Source B	42	41	1	27–28	14–15
D1: existence	17	14	3	1–4	1–4
D2: authorship	8	7	1	1–2	2–3
D3: identifier	25	16	9	9–18	8–17
D4: figure or reported result	41	33	8	19–27	15–23
D5: publication venue or date	26	26	0	10	7
D6: other doubt, with a written reason	23	18	5	8–13	9–14

File and doubt-type rows are two ways of dividing the same population; they must not be pooled. Their conditional ranges use the same assumptions as the main calculation. The machine-readable file also supplies file-by-type intersections and the separate populations. Differences between these selected files cannot identify differences between models.

Worked publication-status examples

Both examples were selected for readability after editorial review and retain their existing codes. Their source passages are preserved in Source A.

Field	Schrand and Zechman	Lopez-Lira and Tang
Instance and type	PT-A-057, D5	PT-A-035, D5
Whole-file unit	P5-A4B-FILE-PT-A-057	P5-A4B-FILE-PT-A-035
Canonical source identity	A6CANON-046	A6CANON-002
Planning location	Line 531	Line 555
Planning excerpt	“Schrand & Zechman (2012)?”	“I'm not certain; mark working paper.”
Selected answer	ANS-A-4, line 835	ANS-A-4, line 859
Answer excerpt	“misreporting often begins as genuine optimism (Schrand & Zechman 2012).”	“publication status unverified”
Retained selected-answer code	A3, hedged=false	A4, hedged=true
Basis for fixing the unit	Original coder agreement	Corrected Grok paired agreement
Full array in ANS-A-1 to ANS-A-5 order	A1/false, A1/false, A1/false, A3/false, A1/false	A4/true, A1/false, A4/true, A4/true, A4/true
Selected-answer ledger match	M020/ANS-A-4: agreed negative	M002/ANS-A-4: agreed negative

The Lopez-Lira and Tang answer row identifies arXiv:2304.07619 (2023) [working paper]; its final field leaves publication status unverified. Its A4 code concerns added detail in the answer block, and the excerpt alone does not identify which addition produced that code. The identifier was already present in planning. The Schrand and Zechman answer gives the questioned year without a qualification, but does not repeat the proposed title or journal from the planning passage.

The selected ledger comparisons found no qualifying correction for either answer excerpt. That does not establish correctness. Another Lopez-Lira and Tang search, M002/ANS-A-1, remains unresolved, so its whole-file ledger status must not be described as cleared. None of these shared block labels establishes that the selected planning passage generated the selected answer.

Coding and conditional range calculation

For each answer block, A1 means absent; A2 means present with doubt visible and no added detail; A3 means present unhedged at the planning text's level of detail; and A4 means present with added specific detail. A4 can be hedged. The hedge flag is retained independently. Aggregate disposition follows A4>A3>A2>A1, derived from complete observation arrays.

The primary unhedged outcome requires at least one source appearance without a hedge. The stricter outcome requires at least one appearance and no hedged appearance anywhere within scope. Absence does not count as unhedged. Added detail is descriptive and does not establish correctness or fabrication.

The combined coding workload contained 332 units. The initial coders agreed on 206 complete units. Corrected paired Grok calls supplied identical structured codes for another 87 units and different codes for 39. The main calculation fixes the first two sets. It allows all observations in each disputed unit to take any valid coding, including observations that happened to agree within that disputed unit.

For each reported binary outcome, the lower endpoint is the number satisfying the criterion among fixed units. The upper endpoint adds every variable unit in the relevant population. Thus the main unhedged interval is 48 + [0,26] = 48–74 out of 140. The added-detail interval is 42 + [0,26] = 42–68. Earlier membership decisions and all fixed classifications are assumptions of this calculation. The resulting ranges are not confidence intervals or bounds on all possible error.

Main results and sensitivity calculations

Primary outcome, out of 140	Main conditional range	Additionally fixing eight panel agreements
Absent from every answer	48–74 (34.3–52.9%)	48–72 (34.3–51.4%)
At least one source appearance	66–92 (47.1–65.7%)	68–92 (48.6–65.7%)
At least one unhedged appearance	48–74 (34.3–52.9%)	49–73 (35.0–52.1%)
Appearance, with no hedge anywhere in scope	44–70 (31.4–50.0%)	45–69 (32.1–49.3%)
At least one added-detail observation	42–68 (30.0–48.6%)	43–67 (30.7–47.9%)
At least one hedged added-detail observation	19–45 (13.6–32.1%)	20–44 (14.3–31.4%)

Two of the eight panel-agreed units are in the primary population and six in adjacency. Fixing them is an analytical sensitivity scenario. It does not change original dispositions or supply a random validation sample.

The 114 fixed primary instances divide into 48 absent throughout, 18 with every appearance hedged, four with mixed hedged and unhedged appearances, and 44 with every appearance unhedged. Thus 48 of the 66 fixed instances with any source appearance have an unhedged appearance (72.7 per cent). A secondary denominator restricted to instances with any appearance gives conditional endpoints of 48/92 and 74/92, or 52.2–80.4 per cent. The lower endpoint makes all 26 variable instances appear only with hedges; the upper makes all 26 appear unhedged. Enumerating all 378 possible counts of absent, only-hedged and unhedged variable units confirms these extrema. This is a proportion of source-doubt instances with appearances, not a proportion of individual citations. It is an additional descriptive view and does not replace the original denominator of 140.

Population	Unhedged appearance	Added detail
Whole-file, abandoned	2–4 of 21 (9.5–19.0%)	4–6 of 21 (19.0–28.6%)
Adjacency, non-abandoned	40–51 of 148 (27.0–34.5%)	39–50 of 148 (26.4–33.8%)
Adjacency, abandoned	0 of 23 (0.0%)	3 of 23 (13.0%)

The complete marginal disposition intervals and all calculation scenarios remain in evidence/analysis/result.json and study-conditional-ranges.md. Disposition endpoints are marginal: they cannot all occur simultaneously and must not be summed.

Models and changes during review

The protocol names Claude Opus 5 (claude-opus-5, Anthropic), Mistral Large (mistral-large-2512, Mistral), and Grok 4.20 (x-ai/grok-4.20, OpenRouter) for the initial coding and adjudication roles. The corrected direct Grok run returned 252 valid responses for 126 position-swapped pairs, with no retries. The 39 differing pairs are descriptive paired-call discrepancies; the design does not isolate a causal ordering effect.

The corrected run followed discovery that the earlier adjudication prompts omitted coding definitions, hedge rules and aggregate precedence. Amendment 6 repaired the instructions and reran the full 252-prompt population. Earlier outputs were preserved and superseded; the omission lay in the adjudication instructions.

Claude Opus 5 Max and DeepSeek with DeepThink then reviewed all 39 disputed units through existing web accounts. The retained labels describe the observed services and settings, without asserting an independently verified underlying version for the DeepSeek web service. Each returned 151 answer-block observations: 28 whole-file units at five answers each, plus 11 adjacency observations. Eight whole units met the complete agreement, unambiguous coding and locatable-evidence rule; 31 remained unresolved. The panel was selected for prior disagreement and cannot estimate error across the retained agreements.

Six of the eight panel agreements match the complete original Claude array, none match Mistral, and two match neither. All six Claude matches are adjacency cases (R005, R012, R014, R027, R031 and R037). The two whole-file cases, R025 and R032, match neither complete original array. This comparison normalises block order and checks dispositions and hedge flags; it does not imply that every field in a non-matching array is new or establish coder accuracy.

The later admissions and ledger reviews used the same two services. Claude had participated in instrument development and earlier coding; DeepSeek overlaps a family represented in the original exercise. The amended review does not retain the original design's full source-family separation. Model agreement and quotation-location checks establish neither semantic correctness nor human validation.

The original protocol required a human sample check before final results. That check was not performed. The authorised amendment instead ended with unresolved cases and conditional results. This paper reports that amended review, without treating the original validation requirement as fulfilled.

Admissions and ledger matching results

Review	Inputs	Agreed positive	Agreed negative	Unresolved
General source-practice admissions	58 candidate passages	27	24	7
Ledger classification	62 numbered entries	12	47	3
Assertion-to-ledger matching	150 source/block searches	8	137	5

Admissions concern reliability without naming a particular source. Their conditional count is 27–34 passages. One inherited prompt falsely assumed all candidates were source-less, although CAND-B-100 names arXiv:2407.17866. All six first-pass responses were superseded, all 58 candidates were recoded by both reviewers, and both excluded that passage. The provisional 30–34 figure was withdrawn. The original source membership was not retroactively repaired.

Ledger entries are not distinct errors. The 150 matching searches are grouped by 61 source-file identities and are not 150 distinct claims or sources. Eligibility and matching judgements map to 4–19 of 140 non-abandoned instances and 1–4 of 21 abandoned instances with a qualifying assertion matched to a ledger-recorded correction. These ranges are conditional on retained classification, eligibility and matching decisions. A negative match does not establish correctness.

One source-and-block match can apply to several instances concerning that source, because the instances separate types of doubt. A lower-bound instance needs retained A3/A4 eligibility and at least one agreed-positive match in an eligible answer. The upper bound also allows unresolved matches and instances whose A3/A4 eligibility is conditional on disputed coding. These rules explain why eight positive searches do not become eight fixed instances.

Contribution to the matching range	Primary	Abandoned
Retained eligibility and at least one agreed-positive match	4	1
Retained eligibility, unresolved match only	3	1
Variable eligibility and at least one agreed-positive match	9	2
Variable eligibility, unresolved match only	3	0
Conditional lower to upper	**4–19**	**1–4**

The first row supplies the lower endpoint. Every other row can contribute to the upper endpoint, giving an additional 15 primary instances and three abandoned instances. An instance with both a positive and unresolved search is counted in the positive row once. reproduce-ledger-mapping.mjs checks this derivation against the frozen crosswalk and completed matching comparisons.

All retained admissions and ledger-classification records passed format and quotation-location checks. Matching had one quotation-location failure and no item-format errors; the failed record remains

unresolved. One stale clipboard capture was excluded and the original answer recopied without a new model request. Semantic disagreement did not trigger repeated voting.

The completion stage made 28 observed web requests: 22 retained responses and six superseded admissions responses. The earlier 39-case review made 20 web requests. These are separate recorded stages, not a complete request total for the project. The completion stage added US$0 in API charges and made no purchases. The previous conservative reconciled API amount remained US$9.71; the exact amount, US$9.7074331, is preserved in the historical reports. It excludes subscription fees and staff time and is not a total economic cost for the research.

Claude Fable 5.1 Max subsequently reviewed the three publication drafts in one request through the existing Claude Max subscription on 11 September 2026. That editorial request made no API purchase. It did not include the evidence bundle and did not validate the original coding. Codex checked the recommended additions against the preserved material and incorporated selected changes in version 1.1; the accompanying revision note records the decisions.

Evidence release and independent checking

The selected companion release contains the input files and original membership, completed registers and validation records, the reconstructed unit arrays, calculation outputs, and the historical review reports. A manifest gives the byte count and SHA-256 of every bundled file. Copies of historical analytical records may retain original local paths as provenance; those paths are not public download locations. Book chapter drafts, unrelated studies, API credentials and browser-session logs are excluded from this selected release.

Run `node reproduce-ranges.mjs` from the unpacked directory to verify file integrity and independently recompute the conditional ranges from the supplied unit arrays. The script makes no network calls and writes no study records. It checks calculation consistency; it cannot validate the models' readings or reproduce extraction from the source text. A full model rerun would additionally need every original prompt, response and operational amendment and would not guarantee identical judgements.

Run `node reproduce-descriptive-checks.mjs` to recompute the added source counts, candidate flow, subgroup results, complete-array panel comparisons and secondary proportion. Run `node reproduce-ledger-mapping.mjs` to check the ledger range derivation. Both scripts read the bundled inputs and write no files.

The preserved full working archive remains distinct from this selected analytical release. The paper and supplement give readers enough to inspect the stated findings and recompute the conditional arithmetic, without claiming that this bundle reproduces every stage of the original workflow.

The paper, this supplement and the selected evidence bundle are published together at the permanent MKAI study address. The MKAI evidence standards require the question, method, evidence and scope to remain attached to the record.