Sources doubted in AI planning text

How they appear in two preserved research compilations

Richard Foster-Fletcher · MKAI · Published working paper, 11 September 2026 · Version 1.2

Research classification: descriptive coding of preserved material, undertaken as a manuscript verification pass. The amended model review ended on 10 September 2026.

Abstract

Two preserved AI research compilations contain visible planning text that questions sources cited in the completed answers. This analysis compares their presentation within each file. The primary population comprises 140 instances, each combining a source, a type of doubt and a file; these cover 98 distinct canonical sources. The retained model coding fixes 114 instances: the source is absent from every answer in 48, always qualified in 18, and unqualified at least once in 48. Another 26 remain disputed, giving a conditional range of 48–74 instances (34.3–52.9 per cent) with an unqualified appearance. The answers contain added detail in 42–68 instances. These ranges assume the retained coding and membership decisions are correct; the planned human sample check was not performed. The files do not establish which planning session produced which answer, so the comparison describes co-occurrence within the compilations. It cannot measure doubt disappearing during generation. The practical distinction is between the form of a citation and evidence that its particular claims have been checked.

A questioned date and an unqualified citation

Source A, one of the two compilations, contains a planning passage that asks, “Schrand & Zechman (2012)?” The passage also questions the paper's title and publication details. The fourth completed answer in that file states that “misreporting often begins as genuine optimism (Schrand & Zechman 2012).” The year that appeared with a question mark in the planning material is presented without a qualification in this citation. [1, lines 531 and 835]

The coding identifies this as instance PT-A-057, concerning publication venue or date. Both original coders classified the citation as an unhedged appearance at the planning text's level of detail, with the source absent from the file's other four answers. This example establishes neither that the date is wrong nor that a single generation session lost its doubt. It shows the textual relationship the comparison measures: a source is questioned in planning material and cited without qualification in an answer preserved in the same file. [3]

A qualification is visible in another publication-status example. The planning text asks whether Lopez-Lira and Tang's paper had appeared in the *Journal of Financial Economics*, adding, “I'm not certain; mark working paper.” The fourth answer labels the source a working paper and states “publication status unverified”. The retained code for this answer is A4 with a hedge: the block adds detail while leaving its publication status qualified. The supplement gives the full coding and ledger status for both examples. [1, lines 555 and 859; 3, PT-A-035]

The compilations came from an exercise about 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. Each preserves five completed answers

and four planning blocks. The accompanying research ledger was prepared on 28 August 2026, but the files do not preserve answer-level model identities or a complete session history. [1, 6]

The project began as a verification pass for Chapter Nine of *The AI Gap*, under a protocol frozen on 29 August 2026. “Doubt Attrition” remains its historical project name. The comparison measures whether the questioned source appears in the completed answers of the same file, whether any appearance lacks a qualification, and whether any appearance adds detail. [2]

What happened to the doubted sources

The main population contains 140 source-doubt instances in which the planning passage had not discarded the source. Doubt about a paper's authorship and doubt about its date count separately. Repeated references to the same source and doubt type within one file are combined. The 140 instances cover 98 distinct canonical sources, represented by 102 source-file identities because four sources occur in both files. These identities retain the original source-matching decisions. [2, 3, 7]

The retained classifications divide 114 instances as follows. The remaining 26 are kept separate because their complete coding is disputed. [7]

Presentation across the five answers in the relevant file	Instances
Source absent from every answer	48
Every appearance qualified	18
Both qualified and unqualified appearances	4
Every appearance unqualified	44
Complete coding disputed	26
Total	**140**

The four mixed instances and the 44 always-unqualified instances together produce the main count of 48. Allowing each disputed instance to meet or fail that criterion gives **48–74 of 140, or 34.3–52.9 per cent**. The same calculation produces a conditional absence count of 48–74. Those intervals happen to be equal, but describe different outcomes; an instance cannot be both absent everywhere and present without qualification.

Outcome within the primary population	Conditional count	Percentage of 140
Source absent from every answer	48–74	34.3–52.9%
Source appears in at least one answer	66–92	47.1–65.7%
At least one appearance is unqualified	48–74	34.3–52.9%
Source appears, with every appearance unqualified	44–70	31.4–50.0%
At least one appearance contains added detail	42–68	30.0–48.6%
At least one appearance contains qualified added detail	19–45	13.6–32.1%

These outcomes overlap and their counts must not be added. Added detail can include a figure, identifier or publication detail, and can carry a qualification. It therefore measures specificity separately from qualification; it does not establish an error. [2, 3]

Source A contributes 98 instances and a conditional unqualified count of 21–46; Source B contributes 42 instances and 27–28. Twenty-five of the 26 disputed primary instances come from Source A. The supplement gives the breakdown by doubt type, so this difference can be inspected without treating it as a provider comparison. [7]

A secondary calculation compares each planning block with the answer immediately following it. This gives 40–51 unqualified appearances among 148 instance-block observations, on the unverified assumption that the preserved order reflects session order. The separately reported 21 instances whose source was discarded in its planning passage have 2–4 unqualified appearances somewhere in the file and none in the 23 adjacent-answer observations. [3]

How the coding was reviewed

The coding distinguishes absence, qualified appearance, unqualified appearance at the planning text's level of detail, and appearance with added detail. A separate hedge flag records whether the relevant uncertainty is visible to the reader. The principal whole-file result requires only one unqualified appearance; the stricter result requires every appearance to lack a hedge. [2]

Claude Opus 5 and Mistral Large initially coded 332 complete units across the whole-file and adjacency populations, agreeing on 206. Grok 4.20 received two calls for each of the other 126 units, with the coder positions exchanged. Those calls agreed on 87 units and differed on 39. An earlier adjudication run had omitted coding definitions from its instructions and was superseded by this corrected run. [3, 4]

The main calculation retains the 206 original agreements and 87 paired-call agreements. Every observation within each disputed unit remains variable. Its ranges express disagreement in those units on the assumption that the retained codes and earlier membership decisions are correct. They are not confidence intervals and do not account for errors shared by agreeing models or omissions in extraction and source identification.

Claude Opus 5 Max and DeepSeek with DeepThink subsequently reviewed the 39 disputed units through subscription and free web services. Eight met the complete agreement and evidence rule. Fixing these additionally gives a sensitivity range of 49–73 for unqualified appearances and 43–67 for added detail in

the primary population. The main analysis leaves them variable: cases selected for prior disagreement cannot provide the validation sample the protocol originally required. [3, 4]

The later panel also shared model families with earlier work. Claude Opus 5 had been an initial coder and had participated in instrument development; DeepSeek was a family represented in the original exercise. Six of the eight panel agreements match the original Claude coding across the complete unit, none match Mistral, and two match neither. This describes agreement, without establishing which coder was right. The original human sample validation remains unperformed. [4, 7]

The available material selects sessions in which interfaces exposed planning text. It also lacks a coded comparison set of sources the planning text never questioned. The results therefore cannot show whether recorded doubt changed citation practice, or estimate a rate for a wider population of AI answers. Visible planning is treated as supplied text, without assuming it is a validated account of internal reasoning.

Admissions and corrections in the ledger

General admissions about citation reliability were reviewed separately. Source A's statement, “I'll provide DOI guesses, enough”, is one such admission and is not one of the 140 source-doubt instances. Both final reviewers accepted 27 of 58 candidate passages and rejected 24, leaving seven unresolved and a conditional count of 27–34. [1, line 1249; 4, 5]

The admissions instructions initially said that none of the candidates named a source, although CAND-B-100 explicitly names arXiv:2407.17866. Both reviewers recoded all 58 candidates under corrected instructions and excluded that passage. The earlier provisional count of 30–34 was withdrawn. The passage was not moved into the source-doubt population because the original membership decisions were not reopened. [4]

The DOI admission should not be read as evidence that a guessed identifier reached an answer. The ledger specifically locates the Hutton DOI variants ending 00457.x in planning text, including Source A lines 1091 and 1240. Their presence in planning is separate from the answer's issue-number error described below. [6, entries 3.4 and 6.8]

The frozen research ledger supplied a separate basis for examining factual corrections. The review accepted 12 of its 62 numbered entries as corrections, rejected 47 and left three unresolved. Some entries repeat a correction, so the count does not represent 12 distinct errors. Subsequent matching produced eight positive, 137 negative and five unresolved source-and-answer-block searches. Mapping these to eligible source-doubt instances gives a conditional count of 4–19 of the primary 140 with an assertion matched to a correction in the ledger. The supplement explains the mapping and its uncertainty. [4, 6]

One example concerns Hutton, Lee and Shu's *Do Managers Always Know Better? The Relative Accuracy of Management and Analyst Forecasts* (2012). The fifth answer in Source A gives *Journal of Accounting Research* 50(2); ledger entry 3.4 gives 50(5), while entry 6.9 describes the answer's account of the finding as accurate in direction. Source A's first answer also attributes a withdrawal date and a cause involving look-ahead leakage to arXiv:2407.17866. The ledger states that the fetched notices did not establish those details. These are the frozen ledger's findings, without a fresh external check in this analysis. [1, lines 42, 1346 and 1379; 6]

A negative ledger match means that no qualifying correction was matched to the eligible assertion. It does not certify correctness. The ledger covers a limited set of checks, so the matching count cannot serve as an error rate for the answers.

What this means for a reader

A precise citation can accompany an answer whose description of a finding is sound while a publication detail is wrong, as the Hutton example illustrates. Citation form alone does not establish that the particular source and claim have been checked. Evidence of those checks would have to accompany an assessment of the answer's reliability.

Materials and citation details

The companion **Methods and results** document provides the full populations, sensitivity results, review history and reproduction instructions. The selected research bundle preserves inputs and analytical records, with separate scripts for the original conditional arithmetic and the descriptive tabulations added after editorial review.

1. Frozen Source A (pasted-text.txt) and Source B (pasted-text B.txt). Bundle: evidence/source/. Line references in the body use these files.

2. *The planning-to-answer tally*, frozen protocol v1.3, 29 August 2026, especially sections 2, 4–7, 9 and 11. Bundle: evidence/source/planning-to-answer-tally-protocol-v1.3.md.

3. *Doubt Attrition verification: conditional ranges*, calculation plan, instances-v2.json, units.json and result.json. Bundle: evidence/analysis/.

4. *Doubt Attrition verification: final amended review*, additional LLM review and completion method, 10 September 2026. Bundle: evidence/reports/ and evidence/review/.

5. Complete admissions register and validation records. Bundle: evidence/review/.

6. *Research ledger: When current models assessed future capability*, prepared 28 August 2026; subsequent ledger classification and assertion-matching registers. Bundle: evidence/source/chapter-nine-research-ledger.md and evidence/review/.

7. Descriptive tabulations added on 11 September 2026 using unchanged membership and codes: descriptive-results.json and reproduce-descriptive-checks.mjs.

Suggested citation: Foster-Fletcher, R. (2026). *Sources doubted in AI planning text: How they appear in two preserved research compilations*. MKAI, published working paper, version 1.2, 11 September 2026. https://mkai.org/studies/sources-doubted-in-ai-planning-text.

AI assistance disclosure: AI models performed extraction, classification and review in the roles described above. Codex assisted with analysis code, evidence checks and drafting. Claude Fable 5.1 Max reviewed the three publication drafts; Codex checked and incorporated selected recommendations. That editorial review did not inspect the raw evidence bundle or add substantive validation of the research coding.