Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Single metrics mislead when evaluating Large Language Models for evidence synthesis and decision-support

Radhika Agrawal
University of Cambridge
Sadiq Jaffer
University of Cambridge https://orcid.org/0009-0006-4120-3244
Sam Reynolds
University of Cambridge https://orcid.org/0000-0002-1255-580X
Vanessa Cutts
University of Cambridge https://orcid.org/0000-0002-4986-2934
Lynn V. Dicks
University of Cambridge https://orcid.org/0000-0002-8304-4468
Anil Madhavapeddy
University of Cambridge https://orcid.org/0000-0001-8954-2428
Rebecca K. Smith
University of Cambridge https://orcid.org/0000-0003-3294-7592
William J. Sutherland
University of Cambridge https://orcid.org/0000-0002-6498-0437
Nigel G. Taylor
University of Cambridge https://orcid.org/0000-0002-8643-826X
Ann Thornton
University of Cambridge
Alec Christie
a.christie@imperial.ac.uk

Imperial College London https://orcid.org/0000-0002-8465-8410

Research Article

Keywords: evidence synthesis, artificial intelligence, large language models, conservation evidence, biodiversity, evidence-based conservation, decision-making, AI evaluation

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11873002/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Single metrics mislead when evaluating Large Language Models for evidence synthesis and decision-support

Authors: Radhika Agrawal[1], Sadiq Jaffer[1], Sam Reynolds[2], Vanessa Cutts[2], Lynn V. Dicks[2], Anil Madhavapeddy[1], Rebecca K. Smith[2], William J. Sutherland[2], Nigel G. Taylor[2], Ann Thornton[2], Alec P. Christie[3,*]

Affiliations: [1]Department of Computer Science & Technology, University of Cambridge, Cambridge, UK. [2]Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK. [3]Centre for Environmental Policy, Imperial College London, London, UK.

*Corresponding author: a.christie@imperial.ac.uk.

Abstract

Decision-makers and scientists increasingly rely on large language models (LLMs) to find and summarise scientific evidence. However, adoption is outpacing evaluation with benchmarks typically reporting just a single headline score. Here we show that no single metric can classify how trustworthy different models are at summarising evidence and multi-metric evaluation is essential. We task four frontier LLMs to retrieve and summarise evidence on the effects of conservation actions from the Conservation Evidence database in response to 520 open-ended questions from practitioners across 19 subject areas. We decompose their evidence summaries into 23,641 statements and score them with an LLM judge validated against 820 expert-labelled statements, using metrics that measure both faithfulness to, and coverage of, relevant evidence. Three different models rank top depending on the metric: the most faithful model wrote the shortest summaries and consulted the least evidence, recalling only 49% of relevant evidence. Measuring faithfulness alone rewards omission because it is measured solely against the evidence a model chooses to retrieve. We also found that averages masked bimodal distributions in citation behaviour, and models that found evidence often failed to cite it. We propose a minimal set of jointly reported AI evaluation metrics to avoid these potential pitfalls that remain relevant even in the face of rapid advances in AI models. Failing to measure omission could reward systems that produce accurate but incomplete evidence summaries that are potentially biased and misleading.

Main text

Conservation decisions are often time-critical and context-dependent, yet the evidence needed to inform them is patchy in geographic and taxonomic coverage and also variable in quality (Christie et al., 2020; Anjan et al., 2025; González-Suárez et al., 2026). Systematic and living reviews can be robust but also too slow, high-level, and narrowly framed to be useful for rapid decision- and policy-making (Gutzat & Dormann, 2020; Webb et al., 2026). Curated resources, such as the Conservation Evidence database (Sutherland et al., 2019) and the Collaboration for Environmental Evidence Database of Evidence Reviews (CEEDER; Konno et al., 2020), broaden access but it remains difficult to flexibly answer practitioners' specific questions. Decision-makers are already using generative AI (specifically Large Language Models; LLMs) to interrogate evidence; furthermore, tools built on curated databases are being developed to close the gap between evidence and action (Iyer et al., 2025).

Whether LLMs can be trusted is increasingly questioned. Their applications for ecology and conservation science have begun to be tested through evaluations of tasks like extracting data on species names, locations (Scheepens et al., 2024), presence, threat, and range (Uryu, 2025; Dorm et al., 2026). However, evaluating the ability of LLMs to perform complex, decision-support tasks is more challenging – recent attempts have focused on retrieving relevant evidence to answer multiple-choice conservation questions (Iyer et al., 2025). For synthesising evidence in response to open-ended questions, evaluations must also consider whether LLMs might introduce bias or errors in several places at once: in what a model searches for, whether it attributes its claims, and whether those claims reliably reflect what the sources say. However, existing evaluations tend to test just one of these as a headline score – e.g. the proportion of relevant studies an LLM retrieves (Nicholson Thomas et al., 2024; Spillias et al., 2024), how many references are fabricated (Walters & Wilder, 2023), or the proportion of answers that were factually correct (Rodas-Trejo & Ocampo-González, 2024).

Scoring model outputs at scale, such as evidence summaries generated by LLMs, complicates this further. Human review is often infeasible at large scales and so an LLM is often used to 'judge' the outputs of other LLMs and validated on a single aggregate agreement figure, which can mask systematic bias for or against the very models being compared (Panickssery et al., 2024; Yang et al., 2026). This means that both the scoring mechanism and overall evaluation design of studies testing LLMs may rely on a single metric or score. The machine-learning community has previously argued for holistic, multi-metric evaluation (Liang et al., 2023), shown that leaderboard scores can mislead (Singh et al., 2025), and that metrics that become optimisation targets are gamed (Pan et al., 2022). To our knowledge, we lack tests that determine how much the choice of metric changes the conclusions drawn in an applied field that is adopting LLMs, such as ecology and conservation.

Here we produce the largest expert-grounded benchmark of agentic evidence summarisation in conservation to ask how much our conclusions on LLM performance depend on the metrics used. We tasked four frontier LLMs (Claude Sonnet 4, Gemini

2.5 Pro, GPT-5 and Kimi K2) to respond to 520 open-ended questions across 19 subject areas on the effects of conservation actions by drawing on the Conservation Evidence database – this contains expert-curated pages collating findings of studies that have tested individual actions. These models had to find relevant evidence in this database by submitting a query, receiving the best-matching action pages, before chosing which results to open, read, and use to produce an evidence summary (see Methods). Every statement in the resulting evidence summaries was scored for relevance, citation correctness and faithfulness (Box 1) by an automated 'LLM judge' selected on both its agreement with three expert reviewers and its bias across the four summariser LLMs. By pre-identifying relevant actions for each question, we were also able to measure their ability to retrieve relevant evidence. Here we show that deciding which is the 'best' model depends heavily on the metric chosen, that the most faithful model was the one that consulted the least evidence, and that averages and a single LLM judge-agreement figure each conceal patterns that would materially change our conclusions. From these findings we derive a minimal reporting standard for evaluating AI for evidence synthesis and decision-support and highlight the importance of multi-metric evaluation (i.e. ensuring multiple metrics are used that address Coverage, Attribution, and Faithfulness; Box 1). These are likely to remain relevant even as models improve because the issues they address arise from how evaluations are designed rather than from particular models.

Box 1: Metrics and analyses used in this study, what each is designed to detect and what each cannot detect on its own. We suggest evaluations of AI for evidence synthesis and decision-support use multiple metrics that cover multiple groups (i.e. at least one from each of Coverage, Attribution, and Faithfulness as a minimum).

Metrics and associated analyses		Definition in this study	Can detect	Cannot detect
Metrics				
Relevance		Proportion of statements judged relevant to the question	Off-topic content	Comprehensiveness and accuracy: a short, safe summary can be fully relevant
Coverage	**Retrieval recall**	Proportion of relevant action pages that the model opened	Omission at the search stage	Whether opened evidence was used or reported correctly
	Retrieval precision	Proportion of opened action pages that were relevant	Indiscriminate searching	Whether relevant evidence was missed in the summary
	Citation recall	Proportion of relevant action pages that each summary cites	Omission at the reporting stage: relevant evidence found or findable but absent from the summary	Whether the cited evidence is used correctly (captured by citation correctness and faithfulness)
Attribution	**Citation rate**	Proportion of statements bearing at least one citation	Unattributed claims	Whether citations are correct
	Citation correctness (when cited)	Proportion of cited statements whose citations support them	Mis-citation	Inaccurate uncited statements (captured by citation rate, or by citation correctness over all statements)
	Cite-through rate	Citation recall divided by retrieval recall: proportion of relevant action pages opened that were then cited	Failure to attribute evidence that was found	Failure to find evidence
Faithfulness		Proportion of statements fully supported by the action pages consulted to generate the summary	Misuse of consulted evidence	Omission: missing relevant evidence outside what the summary cites.
Task completion		Proportion of questions for which a summary was produced	Failure to answer certain questions	Quality of completed summaries

Analyses			
Inspect distributions	Fraction of summaries below or above quality thresholds	Tail risk hidden by means	As for the metric that is being inspected
Failure-mode profile	Analysis of controlled-vocabulary reasons for unfaithful statements	Which errors dominate	Errors of omission
LLM judge validation	Analysis of human-LLM % agreement, Cohen's κ, mean absolute bias, and spread of signed bias across summariser LLMs	An LLM judge that favours or penalises particular models	Errors shared by the LLM judge and humans

103

104

Results

Ranking models is highly metric-dependent

The four LLMs produced 1,991 evidence summaries, decomposed into 23,641 statements and with each scored by a validated LLM judge (see Methods) for relevance, citation correctness, and faithfulness, alongside retrieval and completion metrics (Figure 1; Extended Data Table 1).

Relevance narrowly separated the models (Fig.1; highest: GPT-5 at 96.5%; lowest: Gemini 2.5 Pro 88.2%), whilst citation rate separated models by the greatest degree (highest: GPT-5 at 97.8%; lowest: Kimi K2 at 42.5%). Where citations were present they were frequently correct (highest: 96.9% for both GPT-5 and Gemini 2.5 Pro; lowest: 83.3% for Kimi K2). Faithfulness spanned 18 percentage points (highest: Gemini 2.5 Pro at 91.5%; lowest: Kimi K2 at 73.7%), with the top two models separated by only a 1.2-point difference (Extended Data Table 1). In contrast, GPT-5 and Gemini 2.5 Pro were poorer at retrieving relevant action pages (retrieval recall: GPT-5 = 65.2% and Gemini 2.5 Pro = 49.2%) than Kimi K2 (68.4%) and Claude Sonnet 4 (72.9%).



Figure 1: Multi-metric profile of four frontier LLMs at summarising conservation evidence. Each point represents one evaluated summary (n = 515 Claude Sonnet 4, 520 Gemini 2.5 Pro, 467 GPT-5, 489 Kimi K2); large markers are model means with 95% bootstrapped confidence intervals (1,000 resamples at the summary level). Panels are grouped by the type of metric (see Box 1 for definitions) and are the proportion of a summary's statements meeting the criterion; retrieval recall, retrieval precision, citation recall, and cite-through are computed against the relevant action pages and those found and used by models to produce each summary.

Consequently, three different models ranked first depending on the metric chosen (Fig. 2a). GPT-5 led on relevance, citation rate and retrieval precision; Gemini 2.5 Pro on faithfulness (effectively tied with GPT-5 on citation correctness when cited); Claude Sonnet 4 on retrieval and citation recall. Kimi K2 ranked second on relevance and retrieval and citation recall but last on citation rate, citation correctness, and faithfulness. A single-metric leaderboard would therefore have produced a different 'best' model depending on which metric the evaluators selected.

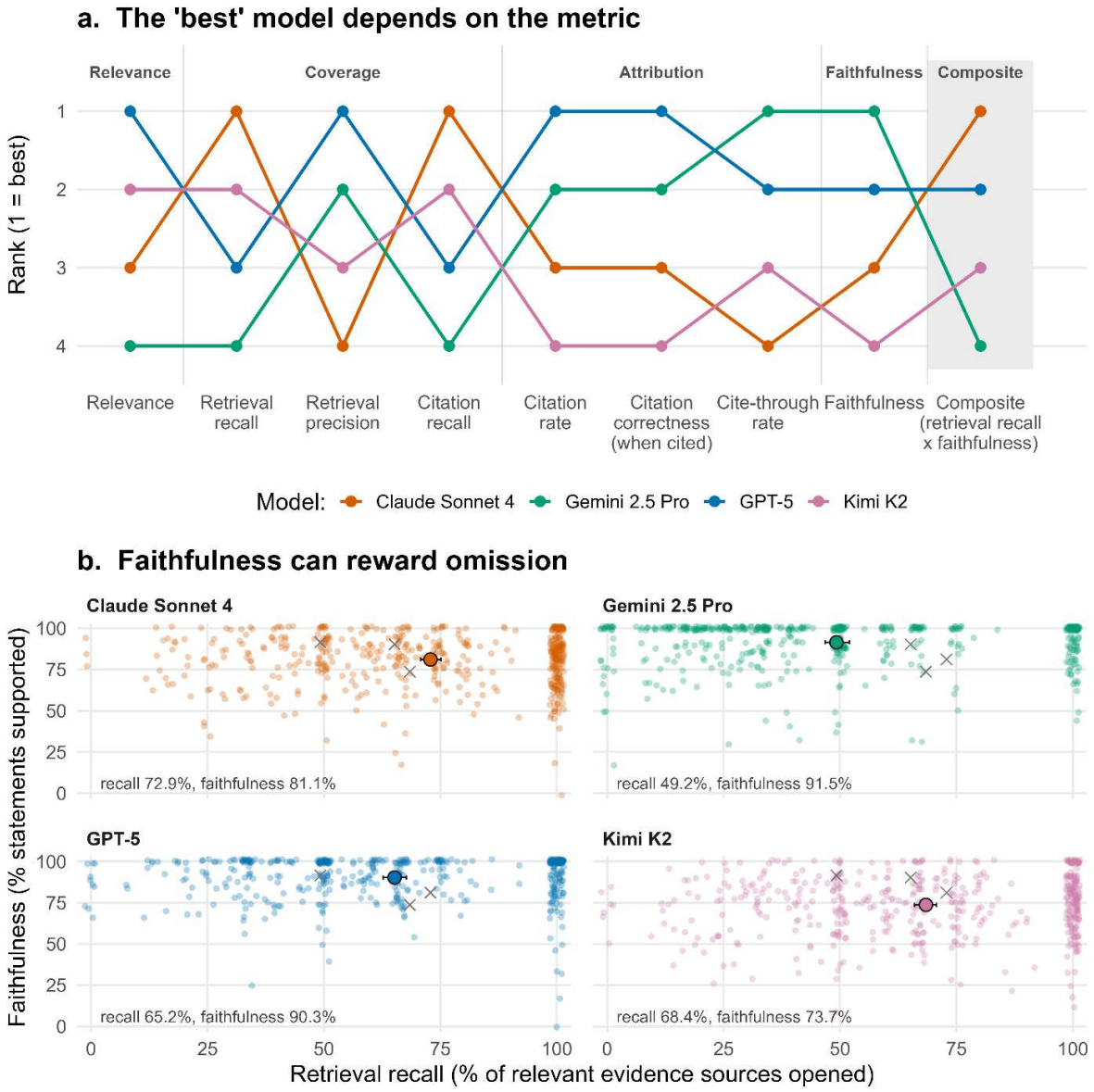


Figure 2: Rankings change depending on the metric used. Panel a. Rank of each summariser LLM (1 = best) on each metric (see Box 1 for definitions), grouped by metric as in Figure 1 and Box 1, and a composite (shaded; mean of per-summary retrieval recall × faithfulness); three different models rank first. **Panel b.** Faithfulness against retrieval recall, one panel per model. Points are individual summaries; filled markers are model means with 95% bootstrapped CIs; crosses mark the other three models' means.

Only measuring faithfulness can reward omission

Faithfulness was judged against the action pages each LLM retrieved to generate a given summary, so a model cannot be penalised for evidence it never retrieved. Gemini 2.5 Pro achieved the highest average faithfulness by searching, opening and reading the least evidence. On average per question, Gemini 2.5 Pro conducted 1.4 searches (versus 3.2–3.3 for the other models) and opened and read on average 3.1 actions per summary (versus 4.4–6.8 for the others). Furthermore, in 91% of its summaries, this model looked at just the first three search results, despite being instructed to continue until the results were no longer relevant (versus 17-20% for the other models, which either requested more results per search or paged through further results; Extended Data Table 1). As a result, Gemini 2.5 Pro found 49.2% of the actions independently verified as relevant, against 65.2–72.9% for the other models (Fig.2b), and wrote the shortest summaries (average of 9.0 statements per summary versus 14.9 for Claude Sonnet 4; Extended Data Table 1). It also had the lowest average relevance score of the four (Fig.1), consistent with answering incompletely (Fig. 2b).

An example that illustrates these problems is the question: “What are the main benefits and negative impacts of using prescribed fire as a management tool in North American forests?” (Supplementary Information Appendix D). Gemini 2.5 Pro performed a single search, opened two action pages, and cited one that reported the effects of prescribed fire on amphibians. It wrote a seven-statement summary, and every statement was scored as being faithful, relevant, and correctly cited. However, none of the seven relevant action pages were opened or cited. These action pages describe the effects of prescribed fire on mature trees, young trees, and understorey plants, including evidence that it increases tree mortality. By contrast, Claude Sonnet 4 and Kimi K2 retrieved four and three of these seven action pages, respectively. The Gemini 2.5 Pro evidence summary therefore scored perfectly on every judged metric, yet missed relevant evidence that may have materially changed the conclusions drawn by an end-user.

Finding evidence is not the same as using it

Kimi K2 recovered more of the relevant evidence than GPT-5 (68.4% versus 65.2%) yet cited fewer than half of its statements. The cite-through rate, the proportion of relevant action pages opened that were cited at least once in the summary, was high for all models (96.0% for Kimi K2 versus 94.7–98.5% for the rest). Therefore, Kimi K2 was able to retrieve relevant evidence and cite each relevant source somewhere, but left most individual statements unattributed. The two versions of citation correctness (for all statements and only cited statements) provide very different pictures of this model: 83.3% of Kimi K2 statements that contained a citation were correctly cited, but this equated to only 35.8% of all Kimi K2 statements; the problem therefore appeared to be citation quantity rather than quality. Claude Sonnet 4 illustrates a different trade-off: it opened the most action pages (6.8 per summary) and achieved the highest recall (72.9%), but almost half of what it read was not relevant (retrieval precision of 53.7%,

the lowest of the four). Using retrieval recall alone would rate Kimi K2 as second-best; citation rate alone would rate it worst (Fig.1; Fig.2a). It is only by considering retrieval recall, citation rate, and cite-through rate together that we can identify the problematic behaviour: failing to cite the evidence, rather than failing to retrieve it.

Averages conceal summary-level patterns that most affect end-users

Mean values for each metric hid important patterns between summaries. Kimi K2's citation behaviour was bimodal: in 50% of its summaries, fewer than half the statements were attributed to a specific cited action page; and in 20% of summaries it cited at least 90% of statements (Fig.3). Plotting citation rate against correctness (Extended Data Fig. 1) illustrates two main behavioural patterns, one with sparse and incorrect citations, and one with dense and accurate citations.

Similarly, Claude Sonnet 4 generated summaries containing an average of 81% faithful statements, yet in 22% of these summaries there were fewer than 70% faithful statements (Extended Data Table 2). For GPT-5 and Gemini 2.5 Pro, approximately two-thirds of summaries had ≥90% faithful statements with fewer than 8% of summaries with <70% faithful statements (Extended Data Table 2). The probability of generating a summary containing unfaithful statements is therefore a much more informative evaluation measure than average faithfulness score.

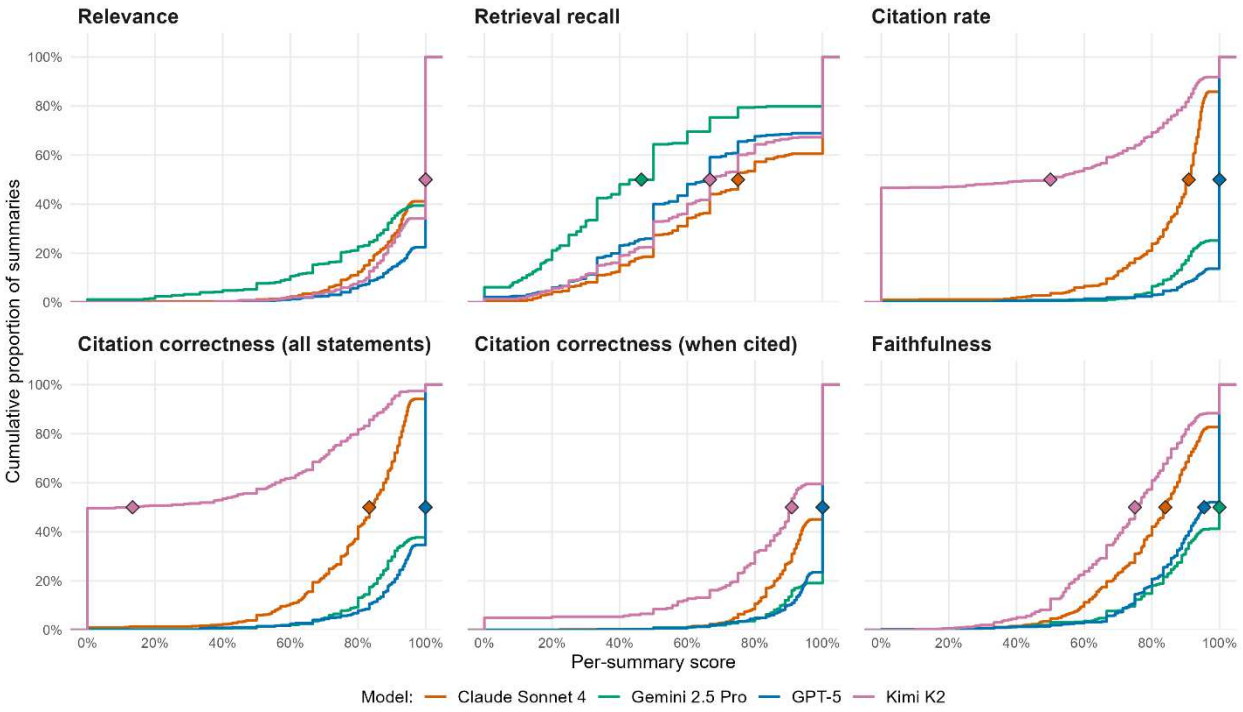


Figure 3: Averages conceal the tail risk. Empirical cumulative distributions of per-summary scores for six metrics (see Box 1 for definitions) and four **LLMs**; diamonds mark medians. Retrieval recall is computed from the model call that produced each summary.

Variance decomposition confirms where the major risks lie (Extended Data Table 3). Model identity explained 49% of variance in citation rate and 54% in citation correctness over all statements, but only 18% of faithfulness and 5% of relevance. Subject area explained at most 5%; the remaining 74–88% of variance in faithfulness, relevance and citation correctness when cited lay between individual summaries. Performance between subject areas did differ somewhat (Extended Data Fig. 2; Extended Data Table 4): the subject areas Mediterranean Farmland and Soil Fertility were hardest to summarise accurately (76-77% faithfulness), whereas Management of Captive Animals and Sustainable Aquaculture were the easiest (90% faithfulness). However, the spread between models within a subject area was as high as 31 percentage points (Mediterranean Farmland) and no model was uniformly poor on a subject that the others handled well. Most variation that end-users would encounter in an evidence summary therefore seem to lie between individual summaries rather than between subject areas – however, it is important for evaluations to consider variation at different scales.

Scores are conditional on completing the task

Model calls that exhausted the ten-iteration search budget produced no summary, and did so unevenly (Extended Data Table 5). GPT-5 failed to complete 38 of 520 questions

(7%) (it hit this cap on at least one attempt for 105 questions, of which 67 were completed on re-execution); Kimi K2 failed to complete 13, Claude Sonnet 4 two, and Gemini 2.5 Pro none. Evaluation scores are conditional on completion, and the uncompleted questions may be the hardest to answer. Completing searches and searching comprehensively are also two different properties: the model with perfect completion (Gemini 2.5 Pro) searched the least.

Unfaithful statements arise mainly from absent facts and overgeneralisation

The LLM judge's reasons for marking statements as unfaithful followed a similar profile across models but differed two- to five-fold in frequency (Fig. 4; see Extended Data Table 6 for definitions). The most common was that the cited evidence did not contain, or contradicted, the claim: at least one such statement appeared in 80% of Kimi K2 summaries, 69% of Claude Sonnet 4, 40% of GPT-5, and 33% of Gemini 2.5 Pro summaries (2.3, 1.9, 0.9 and 0.6 such statements per summary on average). For example, Kimi K2 generated an unfaithful statement suggesting "One study showed adding lime to soil *after* planting decreased pine seedling survival…." but the cited evidence (action 1259) stated the study looked at adding lime "…*before* restoration planting…"

The most differentiating was overgeneralisation (extending or transferring evidence beyond its stated scope) which affected 54% and 46% of Kimi K2 and Claude Sonnet 4 summaries against 19% and 11% of GPT-5 and Gemini 2.5 Pro summaries. An example of this from Claude Sonnet 4 was to state "Fertilizing typically increased plant species richness in bogs." when the evidence was drawn from a single study of a bog from New Zealand (action 1812).

Wrong quantities, unsupported comparisons, mismerged studies (combining findings across studies that match no single source), wrong effect direction, and unsupported evidence summaries (unsupported claims about overall state of evidence base) each affected fewer than a third of summaries for every model, and fewer than one in ten GPT-5 or Gemini 2.5 Pro summaries (except for wrong quantities and mismerged studies). Overall, the most common reasons for unfaithful statements suggest that models were providing confidently worded statements whose errors could not be detected without returning to the source.

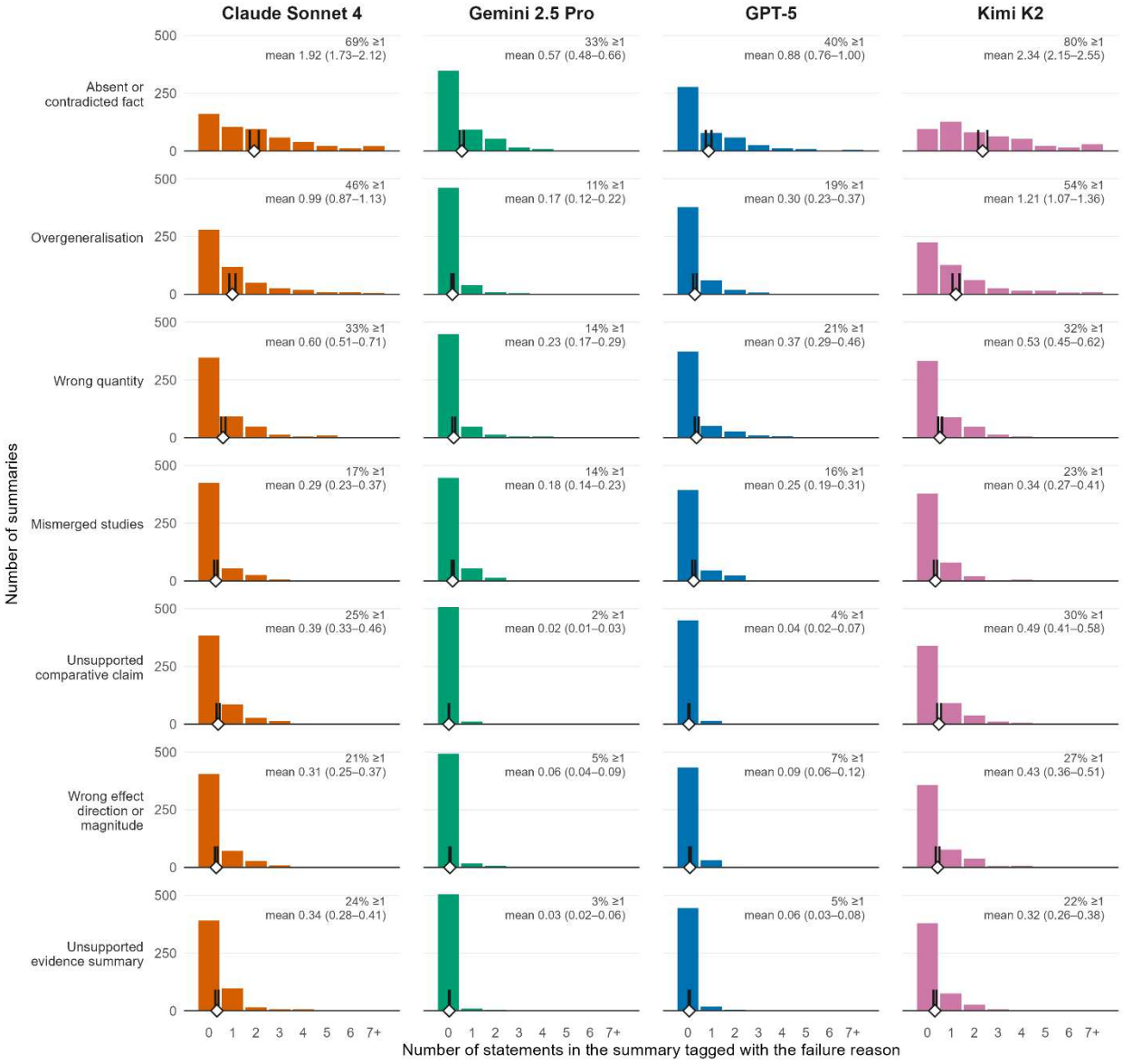


Figure 4: Why statements were judged as unfaithful. Sea stack plots (Stuart et al., 2024): for each failure reason assigned by the LLM judge (rows, ordered by overall frequency) and each summariser LLM (columns), a histogram of the number of statements per evaluated summary tagged with that reason (bin width 1; seven or more pooled). On each baseline, the diamond marks the mean number of tagged statements per summary and the ticks its 95% bootstrapped CI (1,000 resamples at the summary level); the cell text gives the percentage of summaries containing at least one such statement. Definitions and full results, including rates per 100 statements, are in Extended Data Table 6.

The LLM judge can introduce bias and needs thorough evaluation

In our study, every evaluation score for the models generating summaries (hereon ‘summariser models’) was generated by an LLM judge selected based on its similarity to human expert scores for 820 statements (see Methods for details). We compared three candidate LLM judges (Gemini 3.1 Pro Preview, Claude Opus 4.7 and GPT-5.4) and four prompts (manually curated, human–AI refined, and two prompts optimised automatically; see Methods). We selected a judge based on inter-rater reliability between the LLM judge and human experts (Cohen’s κ) and the bias for/against each summariser model (both the range and mean level of bias). We found that using only a single criterion would have led to a different LLM judge being selected. When judging faithfulness, Claude Opus 4.7 had the highest Cohen’s κ marginally but was more biased against some summariser LLMs than others (i.e. had a wide bias range) so we selected Gemini 3.1 Pro Preview (Extended Data Table 7). GPT-5.4 matched the other LLM judges when scoring relevance, yet when judging faithfulness penalised Claude Sonnet 4 and Kimi K2 summaries by 20.5 and 24.6 percentage points. An aggregate score for agreement between the LLM judge and experts would have missed this behaviour, introducing bias into the evaluation. Automatically optimised prompts reached competitive κ on some metrics but were systematically strict, under-crediting summariser models by 5-21 points when judging relevance – this is because optimisation was focused on overall agreement rather than uniformity of error. The selected LLM judge showed only modest favour towards its own model family (+2.5, +1.1 and +5.1 points when judging relevance, citation correctness and faithfulness for Gemini 2.5 Pro summaries), within the range observed for other model families. Expert reviewers themselves agreed on 84–90% of faithfulness labels (pairwise κ 0.00–0.46; Extended Data Table 8), so the LLM judge’s κ of 0.40 sits near the human ceiling. Judging faithfulness is difficult even among human experts and so binary scores carry uncertainty that evaluations should report explicitly.

Discussion

Our aim is not to compare LLMs, especially as they are continually evolving, but to examine the challenges of estimating the quality of their outputs. Across 520 open-ended conservation questions, the answer to “which LLM summarises evidence the best?” changed depending on the metric used. Three models ranked first on at least one metric, for example, the most faithful model consulted the least evidence, while a model that retrieved more evidence failed to cite it. Averages also concealed bimodal and heavy-tailed distributions, and the LLM judge used for scoring would have been biased had it been chosen on agreement with human experts alone. None of these findings implies we need a new metric, but rather calls for reporting more than one, considering them in combination, and looking beyond simple averages to uncover how well LLMs perform on a given task (Burnell et al., 2023). Future engineering of frontier LLMs may improve absolute performance for certain metrics, but methodologies, such as we present here, for highlighting the differences between metrics will always matter for meaningful interpretation and evaluation of model outputs.

Our most consequential finding is that a faithfulness metric can reward omission. This is analogous to an ecological survey method that never records a false positive but detects half the species present, or a systematic review that faithfully reports half of all relevant studies. The evidence synthesis standards adopted in ecology and conservation for human synthesis exist because selective, non-comprehensive searching can bias conclusions (Cooke et al., 2023). Faithfulness and related source-consistency metrics (e.g. groundedness, hallucination rate) are conditional on the evidence a system consulted (Maynez et al., 2020; Es et al., 2024). A model can therefore score highly on faithfulness by consulting few sources and making few claims. The most faithful model (Gemini 2.5 Pro) wrote summaries with fewer statements based on the fewest action pages from the Conservation Evidence database. At best this means its summaries lacked comprehensiveness, and at worst could be considered cherry-picking – if conducted selectively. Consulting less evidence makes faithful summarisation easier as the potential to find conflicting evidence and misinterpret multiple findings is reduced. However, the evidence a narrow search omits could also be the evidence that an action failed or was only effective in certain contexts. Therefore, omission is likely to make a summary misleading in a way an evidence-user cannot detect because it is far harder to check what is missing than whether what is present is correct. This model behaviour probably reflects a default search policy under a low reasoning budget rather than a pre-determined strategy. This makes it all the more important that we use evaluations to detect this kind of default behaviour as it is likely to be present in typical uses of these models. Faithfulness and related source-consistency metrics are already prominent training and product targets for retrieval-augmented systems (Es et al., 2024), and as agents become more capable optimisers of whatever they are scored on (Kalai et al., 2025), evaluations that do not account for omission could select for systems that are not comprehensive (Amodei et al., 2016; Pan et al., 2022).

These findings imply a minimal reporting standard is needed for AI evidence synthesis (i.e. selecting multiple metrics that span different criteria with at least one from each of

Coverage, Attribution, and Faithfulness from Box 1), to complement existing guidance on AI for evidence synthesis (e.g. Thomas et al., 2025). The distribution of scores for these metrics should also be considered, for example, as the fraction of summaries below decision-relevant thresholds, not only means. Any use of automated scoring, such as an LLM judge, should consider not only the agreement between human experts and the LLM judge (e.g. Cohen's κ) but also mean absolute bias and the range in bias across the models or systems being scored (Zheng et al., 2023; Panickssery et al., 2024; Yang et al., 2026). This should be combined with other important disclosures, such as the configuration of the models or systems (e.g. reasoning effort, provider and quantisation, temperature, prompts; Hochlehnert et al., 2025; Marchisio et al., 2024). Overall, multi-metric evaluations should report individual model performance as a profile rather than collapsing findings into averages or rankings.

Our results also suggest a potential design principle for improving evidence synthesis with LLMs. If no single agent can be trusted to report its own omissions, pipelines could separate incentives rather than pool them: one component rewarded purely for recall of relevant evidence, another for attribution, another for faithfulness, with human effort concentrated on cross-checking where statements and sources disagree. Requiring a supporting quotation for every citation (Menick et al., 2022), or a second pass that re-checks each statement against its source, are testable variants of the same idea (Dhuliawala et al., 2024). Future evaluations like ours could test whether such cross-checking systems outperform a single well-configured agent.

For those applying LLMs for conservation, our evaluation further suggests that the latest LLMs at a given point in time are unlikely to represent a single tier of reliability. We found an 18 percentage-point gap in faithfulness and a difference of 7% to 37% in the chance of generating a summary with poor levels of faithfulness. Such differences matter when determining whether model outputs are sufficiently trustworthy to inform decisions. Model choice is a critical design decision for tools built on curated evidence repositories (Berger-Tal et al., 2024) – such as the Conservation Evidence database (Sutherland et al., 2019) we used in this study – which will need revisiting with every model release (Chen et al., 2024). Since most of the variance we found lay between individual summaries rather than between subject areas, deployed tools should also report per-summary provenance – i.e. what was searched, what was opened, and which source supports each statement.

Several limitations apply to our study. Questions were generated by an LLM from the background and key-message text of Conservation Evidence actions and filtered by two authors, with recall measured against a reference set of relevant action pages derived from two independent LLM passes; both could potentially miss other relevant evidence. All summarisers ran at low reasoning effort and Kimi K2 was called via a quantised endpoint, which may contribute to its weaker performance, illustrating why disclosing configuration is crucial. Furthermore, our benchmark study only covers one curated database and neglected questions where evidence is absent or unstructured. Future research could design and apply evaluation methods to test how different LLMs respond to questions for which no evidence exists and their ability to abstain and identify

important evidence gaps. Future work could also extend our evaluation study to alternative retrieval pipelines, models with higher reasoning budgets, multi-turn interactions, and the cross-checked, multi-agent designs proposed above.

In conclusion, the most faithful model in our study was the one that looked at the least evidence. This matters because AI evaluations that focus on accuracy and faithfulness may lead us to develop AI systems that are optimised to lack comprehensiveness, whilst users typically expect a thorough synthesis of the available evidence. For rigorous evidence synthesis applications, this means missing a crucially important objective, which is to reduce the risk of bias in evidence syntheses. In ecology and conservation, if important evidence is overlooked this could lead to poorly informed decisions with significant environmental consequences. Ranking which model is ‘best’ should not be the goal of evaluations, but rather reporting a wide range of metrics accounting for coverage, attribution, and faithfulness together, whilst also examining the distributions of these metrics rather than simply their averages. Making this a routine part of AI evaluations in applied fields, like ecology and conservation, will help ensure that we select and develop appropriate models and systems that accelerate the translation of evidence into environmental decisions whilst minimising the risk of introducing misinformation or bias.

Methods

Source database

We used the Conservation Evidence database (www.conservationevidence.com), a freely available, curated repository of summarised studies (primary studies, reviews and meta-analyses) that test the effects of conservation actions – this is compiled through systematic discipline-wide searches of the literature (Sutherland et al., 2019). The database organises evidence into thematic synopses, referred to here as 'subject areas' (e.g. Amphibian Conservation, Bee Conservation, Peatland Conservation), each comprising a set of actions with individual study summaries, narrative key messages summarising the findings of studies for that action, and background information defining the action and its underlying logic. Actions (specifically 'action pages' containing study summaries and key messages) are the unit of retrieval and citation throughout.

Question generation

Questions for our evaluation were generated synthetically using Gemini 2.5 Pro (thinking budget 8,192 tokens; outputs constrained to a JSON schema; Supplementary Information Appendix A). For each subject area, the model received the background information and key messages for every action, each prefixed with its Conservation Evidence action ID, and was asked to produce seven short, open-ended questions that could only be answered by drawing on the evidence under multiple actions.

The generation prompt enforced diversity along three dimensions. Each of the seven questions had to be of a different type from a controlled list of eleven (e.g. available actions for a stated objective or threat, effects of an action, geographical variation, expected timeframes; Supplementary Information Appendix A). These were derived from elicited example questions obtained from conservation practitioners via informal online surveys sent to representatives in conservation organisations that are part of the Conservation Evidence project's Evidence Champions scheme (see Ethics Approval).

Half of the questions in each batch were general and half were scoped by a specific source of impact, geography or habitat type. To limit redundancy, up to 15 previously generated questions for that subject area were supplied with an instruction to avoid overlap. To ensure the evaluation focused on the ability of LLMs to summarise evidence in response to questions on the effects of conservation actions, we disallowed questions that were answerable solely from the background sections of an action page as these did not represent a meaningful test of a model's ability to synthesise evidence on the effects of actions (i.e. contained in the individual study summaries and key messages). Similarly, to ensure questions were realistic, we also disallowed comparisons between actions that had very different goals and objectives – for example, creating ponds versus engaging volunteers.

Of the 24 synopses available at the time of analysis, five (Bird Conservation, Marsh and Swamp Conservation, Reptile Conservation, Subtidal Benthic Invertebrate

Conservation, and Terrestrial Mammal Conservation) were excluded because their documents exceeded the 100,000-token input limit of the generation model.

Question filtering

Generated questions passed through an independent completeness filter. Using a separate Gemini 2.5 Pro call on the same documents, in batches of ten questions and blind to the generator's output, we asked the model to identify all actions relevant to each question. A question was retained only where this independently derived set was a subset of the pool recorded at generation; questions for which the filter identified a relevant action absent from the pool were discarded. This ensured the reference set of relevant action pages was complete with respect to two independent passes, so that retrieval recall is not penalised by an under-specified reference.

Two authors (APC, SR) then independently checked each question for answerability given the database's purpose and scope (within the remit of Conservation Evidence) and with at least one study permitting a defensible synthesis. They met to discuss questions flagged as inappropriate or unanswerable and discarded them where necessary; further questions were generated until each subject area had at least 20.

The final corpus comprised 520 questions across 19 subject areas, drawing on 1,243 unique actions. The reference set contained 1 to 35 relevant action pages per question (median 4). Questions were distributed unevenly across subject areas (range 20-43, median 25), reflecting the number of generation batches completed and the filtering pass rate in each. The 19 subject areas were: Amphibian Conservation, Bat Conservation, Bee Conservation, Biodiversity of Marine Artificial Structures, Butterfly and Moth Conservation, Control of Freshwater Invasive Species, Farmland Conservation, Forest Conservation, Grassland Conservation, Management of Captive Animals, Marine Fish Conservation, Marine and Freshwater Mammal Conservation, Mediterranean Farmland, Natural Pest Control, Peatland Conservation, Primate Conservation, Shrubland and Heathland Conservation, Soil Fertility, and Sustainable Aquaculture.

Summary generation

We evaluated four frontier LLMs, at the time of analysis, as candidate summariser LLMs: Anthropic Claude Sonnet 4 (anthropic/claude-sonnet-4), Google Gemini 2.5 Pro (google/gemini-2.5-pro), OpenAI GPT-5 (openai/gpt-5) and Moonshot Kimi K2 (moonshotai/kimi-k2-0905), each accessed via the OpenRouter API with reasoning_effort = "low" and default temperature 1.0. Provider routing was pinned for Kimi K2 to a quantised provider (fireworks/fp8) for reproducibility.

Critically, models were not told which action pages were relevant or not. Each model received only the question and was required to locate its own evidence from the Conservation Evidence database through an agentic retrieval loop, enabling us to evaluate retrieval and summarisation together and to build on the work of Iyer et al. (2025). Three tools were available via function calling:

- search_actions(query_string, k, offset): hybrid retrieval over the full Conservation Evidence corpus, returning the top k actions (default 3, maximum 10) with offset pagination. Retrieval fused sparse lexical matching (BM25) and dense embedding similarity, reranked by a cross-encoder, following the best-performing method in Iyer et al. (2025). The index was built over action ID, title, key messages and background information.

- get_action_details(action_id): returned action ID, title and key messages for a specified action. Background information was indexed for retrieval but not returned by this tool. Calls were logged and provide the basis for retrieval recall.

- get_formatted_result(query, summary, action_ids): produced the final summary together with the list of action IDs used, as structured JSON.

The system prompt instructed models to search iteratively, increasing the offset until returned actions ceased to be relevant; to expand candidate actions through parallel get_action_details calls; to use no knowledge beyond the retrieved documents; to cite an action ID for every statement; and to summarise rather than answer or qualify the evidence (Supplementary Information Appendix B). The loop terminated successfully on a get_formatted_result call and was capped at ten iterations (parallel tool calls counted as one iteration).

Runs that reached the iteration cap terminated without producing a summary (reported in the Results). Further summaries were lost to transient API errors and refusals, giving a final corpus of 1,991 summaries (467 GPT-5, 489 Kimi K2, 515 Claude Sonnet 4, 520 Gemini 2.5 Pro). Attrition is reported in Extended Data Table 5.

Statement-level decomposition

Each summary was decomposed into claim statements by a single standardised extraction model (Gemini 2.5 Pro), applied identically to all four summariser LLMs' outputs to avoid confounding statement-boundary differences with summary quality. The same extractor identified, for each statement, any action IDs cited in support; multi-citation statements were preserved as a list of IDs. This yielded 23,641 statements (mean 11.9 per summary, range 1-47), which became the units for evaluation using an automated LLM judge.

Per-summary metrics

From the statement-level judgements we computed five per-summary metrics: (i) relevance, the proportion of statements judged relevant to the question; (ii) citation rate, the proportion of statements bearing at least one citation; (iii) citation correctness over all statements, the proportion of statements bearing a correct citation, with uncited statements scored as failures; (iv) citation correctness when cited, the proportion of cited statements whose citations correctly supported the claim; and (v) faithfulness, the proportion of statements fully supported by the action pages the summary consulted. The two citation-correctness metrics separated the quantity of citation behaviour from its quality.

Retrieval and completion metrics

As the summariser models were not given information on whether an action page was relevant or not to answer a given question, we could measure retrieval performance directly. We computed two recall measures based on the proportion of relevant actions that (i) the model opened via get_action_details ('retrieval recall') and (ii) the model cited in its summary ('citation recall'). Their ratio gave a cite-through rate, separating failure to locate relevant evidence from failure to attribute evidence that was located. Retrieval precision was the proportion of opened actions that were relevant.

We additionally extracted agentic behaviour from the logs of every conversation: loop iterations, search_actions calls, distinct actions opened, deepest result rank reached (k + offset), and the rate of malformed or failed tool calls. Behavioural measures were computed over the completed run that produced each scored summary; error rates were computed over all logged attempts, since a failed tool call may itself end a run. Task completion was the proportion of questions for which any attempt produced a summary.

LLM judge: prompts, validation and selection

To score 23,641 statements at scale for relevance, citation correctness (when cited), and faithfulness we used an LLM judge validated against human expert labels. Each metric was operationalised as an independent prompt returning a binary verdict with reasoning under a JSON-schema-constrained response format (Supplementary Information Appendix C). For faithfulness, statements judged not fully supported were also assigned one or more failure tags from a controlled vocabulary of seven categories (Extended Data Table 6). Faithfulness was judged in a single call with all cited action pages concatenated.

LLM judge prompts were developed in three ways. First, we wrote initial manual prompts grounded in each metric's criteria. Second, we iteratively refined these prompts using Perplexity Pro (with Claude Sonnet 4.6) against criteria and feedback from three independent expert reviewers (R1, R2, R3; authors AT, NGT, VC who have each curated at least one Conservation Evidence synopsis and are highly familiar with the database). These three reviewers also produced a labelled dataset of approximately 820 statements (617 for citation correctness when cited) drawn proportionally from the four summariser LLMs and 19 subject areas. Third, we generated two automatically optimised prompts per metric using GEPA (Genetic-Pareto prompt optimiser; optimize_anything; Agrawal et al., 2026) under two random seeds, initialised from the manual prompts and the labelled dataset. This gave four candidate prompts per metric.

We compared these prompts across three candidate LLM judges (Google Gemini 3.1 Pro Preview (google/gemini-3.1-pro-preview), Claude Opus 4.7 (anthropic/claude-opus-4.7) and GPT-5.4 (openai/gpt-5.4); temperature 0, strict JSON output schema, 4,096-token output limit, up to three retries; reasoning effort at provider default) against the expert labels for each metric. LLM judges were assessed independently per metric (relevance, citation correctness (when cited), faithfulness) using two criteria: Cohen's κ

against the consensus expert label and levels of bias for and against each of the four summariser LLMs. Full results are in Extended Data Table 7.

Human inter-rater reliability

To provide a reference level for achievable agreement, inter-rater reliability was assessed pairwise between the three expert reviewers across 57 statements (19 per reviewer pair), drawn proportionally from the four summariser LLMs and 19 subject areas as for the wider labelled dataset. Pairwise raw agreement and Cohen's κ are reported in Extended Data Table 8.

Application of the validated LLM judge to the full corpus

We applied the selected configuration (Gemini 3.1 Pro Preview with settings as described earlier) to all 23,641 statements. Each statement was scored independently for relevance, citation correctness (when cited) and faithfulness. For multi-citation statements, citation correctness was evaluated against each cited action page independently and aggregated disjunctively: the statement was scored incorrect if any cited document did not fully support the claim. Statements with no citation were assigned an automatic 'no' for citation correctness over all statements, on the principle that an uncited verifiable claim is a citation failure; we also recorded whether each statement was cited (citation rate). Faithfulness was evaluated in a single LLM judge call against the concatenated text of the actions cited by that summary. The pipeline produced one flat JSON record per statement containing the verdict, reasoning, claim decomposition and failure tags.

Statistical analysis

Statement- and summary-level score tables, failure-reason tags and retrieval metrics from the tool-call logs were computed in Python 3; all statistical analyses and figures were produced in R Statistical Software version 4.4. Per-model metrics are reported as the unweighted mean of per-summary scores with 95% bootstrapped confidence intervals computed by resampling at the summary level (1,000 resamples), rather than the statement level, to account for non-independence of statements within a summary. Pairwise model differences are reported as differences of summary-level means with 95% bootstrapped CIs. Retrieval recall, citation recall, precision and cite-through rates were summarised in the same way. Model ranks (Fig. 2a) were assigned from model-level means, with ties sharing a rank; the composite is the mean over summaries of each summary's retrieval recall multiplied by its faithfulness, with a 95% bootstrapped CI. To quantify how much variation in each metric reflected model identity versus subject area, we fitted Type II two-way ANOVAs on each summary-level metric with model, subject area, and their interaction as factors, reporting η^2 for each term. Analysis of reasons for summaries being classified as unfaithful used the reasoning tags returned by the LLM judge (Extended Data Table 6). For each summary we computed a per-reason rate (occurrences of each failure reason divided by statements scored); per-model rates are the mean of per-summary rates with 95% bootstrapped CIs, so that each summary contributes equally regardless of length. As a single unfaithful statement

could be due to multiple reasons, counts of failure reasons exceed counts of statements; we therefore also report statement-level failure rates.

Ethics approval

Ethical approval was granted by Imperial College London's Science Engineering Technology Research Ethics Committee (SETREC; ID 7602196; approved 24/03/2025) for eliciting example questions from conservation practitioners and expert validation methods for the LLM judge.

Data availability

The data supporting this study are available on Zenodo at https://doi.org/10.5281/zenodo.24202230. Conservation Evidence content is reproduced under a CC BY-NC 4.0 licence.

Code availability

All code is available at https://github.com/alecchristie888/llm_evidence_summary_analysis and archived on Zenodo (release v1.0.0) at https://doi.org/10.5281/zenodo.24372257.

References

Agrawal, L. A. et al. (2026). optimize_anything: A Universal API for Optimizing any Text Parameter. Preprint at arXiv:2605.19633. https://doi.org/10.48550/arXiv.2605.19633

Amodei, D. et al. (2016). Concrete Problems in AI Safety. Preprint at arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Anjan, A., Geldmann, J., Harfoot, M. & Christie, A. P. (2025). The spatial distribution of tests of conservation interventions does not align with global conservation needs. Biological Conservation 309, 111313. https://doi.org/10.1016/j.biocon.2025.111313

Berger-Tal, O. et al. (2024). Leveraging AI to improve evidence synthesis in conservation. Trends in Ecology & Evolution 39(6), 548–557. https://doi.org/10.1016/j.tree.2024.04.007

Burnell, R. et al. (2023). Rethink reporting of evaluation results in AI. Science 380(6641), 136–138. https://doi.org/10.1126/science.adf6369

Chen, L., Zaharia, M. & Zou, J. (2024). How Is ChatGPT's Behavior Changing Over Time? Harvard Data Science Review 6(2). https://doi.org/10.1162/99608f92.5317da47

Christie, A. P. et al. (2020). The challenge of biased evidence in conservation. Conservation Biology 35(1), 249–262. https://doi.org/10.1111/cobi.13577

Cooke, S. J. et al. (2023). Environmental evidence in action: on the science and practice of evidence synthesis and evidence-based decision-making. Environmental Evidence 12(1), 10. https://doi.org/10.1186/s13750-023-00302-5

Dhuliawala, S. et al. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. Findings of the Association for Computational Linguistics ACL 2024, 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212

Dorm, F., Millard, J., Purves, D., Harfoot, M. & Mac Aodha, O. (2026). Large language models possess some ecological knowledge, but how much? Ecological Informatics 95, 103699. https://doi.org/10.1016/j.ecoinf.2026.103699

Es, S., James, J., Espinosa Anke, L. & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

González-Suárez, M. et al. (2026). Conservation evidence is biased but can support decision-making for prevalent and severe threats in tetrapods. Journal of Applied Ecology 63(3), e70337. https://doi.org/10.1111/1365-2664.70337

Gutzat, F. & Dormann, C. F. (2020). Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice. Environmental Management 66(3), 435–449. https://doi.org/10.1007/s00267-020-01312-6

Hochlehnert, A. et al. (2025). A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility. Preprint at arXiv:2504.07086. https://doi.org/10.48550/arXiv.2504.07086

Iyer, R. et al. (2025). Careful design of Large Language Model pipelines enables expert-level retrieval of evidence-based information from syntheses and databases. PLOS One 20(5), e0323563. https://doi.org/10.1371/journal.pone.0323563

Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. (2025). Why Language Models Hallucinate. Preprint at arXiv:2509.04664. https://doi.org/10.48550/arXiv.2509.04664

Konno, K. et al. (2020). The CEEDER database of evidence reviews: An open-access evidence service for researchers and decision-makers. Environmental Science & Policy 114, 256–262. https://doi.org/10.1016/j.envsci.2020.08.021

Liang, P. et al. (2023). Holistic Evaluation of Language Models. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2211.09110

Marchisio, K. et al. (2024). How Does Quantization Affect Multilingual LLMs? Findings of the Association for Computational Linguistics: EMNLP 2024, 15928–15947. https://doi.org/10.18653/v1/2024.findings-emnlp.935

Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Menick, J. et al. (2022). Teaching language models to support answers with verified quotes. Preprint at arXiv:2203.11147. https://doi.org/10.48550/arXiv.2203.11147

Nicholson Thomas, I., Roche, P. & Grêt-Regamey, A. (2024). Harnessing artificial intelligence for efficient systematic reviews: A case study in ecosystem condition indicators. Ecological Informatics 83, 102819. https://doi.org/10.1016/j.ecoinf.2024.102819

Pan, A., Bhatia, K. & Steinhardt, J. (2022). The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. Preprint at arXiv:2201.03544. https://doi.org/10.48550/arXiv.2201.03544

Panickssery, A., Bowman, S. & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. Advances in Neural Information Processing Systems 37, 68772–68802. https://doi.org/10.52202/079017-2197

Rodas-Trejo, J. & Ocampo-González, P. (2024). Assessment of ChatGPT's potential as an innovative tool in searching for information on wild mammals. Ecological Informatics 83, 102810. https://doi.org/10.1016/j.ecoinf.2024.102810

Scheepens, D., Millard, J., Farrell, M. & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential pest controllers. Methods in Ecology and Evolution 15(7), 1261–1273. https://doi.org/10.1111/2041-210x.14341

Singh, S. et al. (2025). The Leaderboard Illusion. Advances in Neural Information Processing Systems 38, 86910–86964. https://doi.org/10.52202/085713-2620

Spillias, S. et al. (2024). Human-AI collaboration to identify literature for evidence synthesis. Cell Reports Sustainability 1(7), 100132. https://doi.org/10.1016/j.crsus.2024.100132

Stuart, A. D., Ilić, M., Simmons, B. I. & Sutherland, W. J. (2024). Sea stack plots: Replacing bar charts with histograms. Ecology and Evolution 14(4), e11237. https://doi.org/10.1002/ece3.11237

Sutherland, W. J. et al. (2019). Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Thomas, J., Flemyng, E. & Noel-Storr, A. (2025). Responsible use of AI in evidence SynthEsis (RAISE): recommendations and guidance. OSF. https://doi.org/10.17605/OSF.IO/FWAUD

Uryu, S. (2025). Evaluating Large Language Models for IUCN Red List Species Information. Preprint at arXiv:2510.02830. https://doi.org/10.48550/arXiv.2510.02830

Walters, W. H. & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports 13(1), 14045. https://doi.org/10.1038/s41598-023-41032-5

Webb, J. A. et al. (2026). A Standardized Definition of Rapid Evidence Assessment for Environmental Applications. Conservation Letters 19(1), e70005. https://doi.org/10.1111/con4.70005

Yang, Z., Hou, Y. & Yang, X. (2026). When the Judge Changes, So Does the Measurement: Auditing LLM-as-Judge Reliability. Preprint at arXiv:2607.08535. https://doi.org/10.48550/arXiv.2607.08535

Zheng, L. et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36, 46595–46623. https://doi.org/10.52202/075280-2020

Extended Data

Extended Data Table 1: Per-model metrics with 95% bootstrap confidence intervals. Values are means across evaluated summaries, in percent unless stated, with 95% CIs in brackets (1,000 resamples at the summary level, percentile method). Judge metrics are the proportion of each summary's statements meeting the criterion; citation correctness (CC) over all statements treats uncited statements as failed, whereas CC (when cited) conditions on cited statements and so excludes summaries with no cited statement (n = 510, 519, 467, 259). Coverage and attribution metrics are computed from the single agentic run that produced each evaluated summary, against the withheld set of relevant action pages. Search behaviour is per evaluated summary and the depth of searches considers both paging (offset) and requesting more results per search (k, default 3, maximum 10); 'Saw only the first three results' is the share of summaries in which no search returned anything beyond the top three results, and 'deepest result rank reached' is the mean of the largest k + offset across a summary's searches. Composite = mean of per-summary retrieval recall × faithfulness. n evaluated summaries: Claude Sonnet 4 515, Gemini 2.5 Pro 520, GPT-5 467, Kimi K2 489 (Gemini 2.5 Pro faithfulness n = 519).

Group	Metric	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
Judge metrics	Relevance	93.1 [92.1, 94.1]	88.2 [86.4, 90.0]	96.5 [95.7, 97.2]	94.4 [93.5, 95.3]
	Citation rate	85.8 [84.5, 87.1]	96.2 [95.4, 96.9]	97.8 [97.0, 98.5]	42.5 [38.8, 46.1]
	CC (all statements)	80.0 [78.6, 81.4]	93.2 [92.2, 94.1]	94.7 [93.7, 95.7]	35.8 [32.2, 39.2]
	CC (when cited)	93.3 [92.5, 94.2]	96.9 [96.2, 97.5]	96.9 [96.2, 97.6]	83.3 [80.5, 86.2]
	Faithfulness	81.1 [79.7, 82.5]	91.5 [90.3, 92.6]	90.3 [89.0, 91.4]	73.7 [72.1, 75.3]
Coverage and attribution	Retrieval recall	72.9 [70.6, 75.2]	49.2 [46.6, 51.9]	65.2 [62.5, 67.8]	68.4 [65.9, 70.8]
	Retrieval precision	53.7 [51.3, 56.2]	67.6 [64.7, 70.5]	70.3 [67.7, 73.0]	57.6 [55.1, 60.0]
	Citation recall	69.0 [66.5, 71.4]	48.6 [45.9, 51.5]	62.9 [60.2, 65.7]	65.7 [63.1, 68.3]
	Cite-through rate	94.7 [93.4, 95.9]	98.5 [97.7, 99.2]	96.6 [95.6, 97.7]	96.0 [94.8, 97.1]
Search behaviour	Action pages opened per summary	6.82 [6.60, 7.06]	3.06 [2.95, 3.18]	4.42 [4.21, 4.65]	6.04 [5.81, 6.29]
	Searches per summary	3.22 [3.11, 3.33]	1.40 [1.33, 1.46]	3.26 [3.13, 3.39]	3.28 [3.16, 3.39]
	Deepest result rank reached per summary (k + offset)	8.61 [8.31, 8.95]	3.36 [3.25, 3.48]	7.94 [7.59, 8.31]	10.85 [10.21, 11.49]
	Saw only the first three results (% of summaries)	18.4 [15.1, 21.6]	91.3 [89.0, 93.7]	16.9 [13.3, 20.3]	20.2 [17.0, 23.7]
Summary length	Statements per summary	14.9 [14.4, 15.3]	9.0 [8.6, 9.4]	11.6 [10.9, 12.1]	12.1 [11.8, 12.4]
Completion	Task completion (% of 520 questions)	99.6	100.0	92.7	97.5
Composite	Recall × faithfulness (0–1)	0.591 [0.571, 0.612]	0.453 [0.429, 0.479]	0.589 [0.565, 0.614]	0.508 [0.486, 0.531]

Extended Data Table 2: Per-model distribution of summary-level scores. Statistics are computed across evaluated summaries for each model–metric cell (scores in percent). %≥90 and %≥80 are right-tail fractions; %<70 and %<50 are left-tail fractions. Retrieval recall is from the run that produced each summary.

Model	Metric	n	Mean	Median	SD	IQR	%≥90	%≥80	%<70	%<50
Claude Sonnet 4	Relevance	515	93.1	100.0	11.5	10.3	75.0	88.9	4.5	0.8
Claude Sonnet 4	Citation rate	515	85.8	90.9	15.3	12.6	55.9	79.0	11.1	2.7
Claude Sonnet 4	CC (all statements)	515	80.0	83.3	16.5	19.4	34.2	62.9	21.2	3.9
Claude Sonnet 4	CC (when cited)	510	93.3	100.0	9.9	11.1	72.9	91.6	2.8	0.4
Claude Sonnet 4	Faithfulness	515	81.1	84.0	15.8	21.1	34.0	61.6	22.3	3.5
Claude Sonnet 4	Retrieval recall	515	72.9	75.0	26.8	50.0	40.0	46.6	44.3	18.6
Gemini 2.5 Pro	Relevance	520	88.2	100.0	20.4	15.4	67.5	78.8	15.6	5.2
Gemini 2.5 Pro	Citation rate	520	96.2	100.0	8.8	3.3	84.2	96.2	1.4	0.4
Gemini 2.5 Pro	CC (all statements)	520	93.2	100.0	11.6	11.1	71.7	90.8	4.6	0.8
Gemini 2.5 Pro	CC (when cited)	519	96.9	100.0	8.0	0.0	87.9	96.3	1.9	0.4
Gemini 2.5 Pro	Faithfulness	519	91.5	100.0	13.7	13.5	69.2	85.2	7.7	1.9
Gemini 2.5 Pro	Retrieval recall	520	49.2	46.4	31.6	41.7	20.2	20.6	75.4	50.0
GPT-5	Relevance	467	96.5	100.0	8.5	0.0	86.5	94.2	2.6	0.2
GPT-5	Citation rate	467	97.8	100.0	7.9	0.0	92.5	97.6	1.9	0.9
GPT-5	CC (all statements)	467	94.7	100.0	10.6	6.7	81.4	93.2	4.1	1.1
GPT-5	CC (when cited)	467	96.9	100.0	7.9	0.0	89.9	95.5	2.1	0.2
GPT-5	Faithfulness	467	90.3	95.5	13.6	16.7	63.0	82.2	6.9	1.5
GPT-5	Retrieval recall	467	65.2	66.7	28.5	54.5	31.5	34.0	59.1	25.9
Kimi K2	Relevance	489	94.4	100.0	10.1	9.1	77.1	92.6	3.7	0.8
Kimi K2	Citation rate	489	42.5	50.0	41.9	85.7	20.4	32.5	59.9	49.7
Kimi K2	CC (all statements)	489	35.8	13.3	38.0	73.3	8.0	20.2	69.9	55.6
Kimi K2	CC (when cited)	259	83.3	90.9	24.1	23.1	55.6	73.0	16.6	6.6
Kimi K2	Faithfulness	489	73.7	75.0	18.2	26.0	20.2	42.7	37.2	8.2
Kimi K2	Retrieval recall	489	68.4	66.7	27.9	50.0	33.5	39.3	51.3	22.5

Extended Data Table 3: Variance decomposition of summary-level scores. Two-way ANOVA with model, synopsis and their interaction as fixed effects. SS = type-II sum of squares; η^2 = SS divided by total SS. Residual = variance between individual summaries within model–synopsis cells.

Metric	Term	SS	η^2 (%)
Relevance	model	1.84	4.8
Relevance	synopsis	1.47	3.8
Relevance	model × synopsis	1.18	3.0
Relevance	Residual (between summaries)	34.21	88.4
Citation rate	model	98.91	48.6
Citation rate	synopsis	1.72	0.9
Citation rate	model × synopsis	5.46	2.7
Citation rate	Residual (between summaries)	97.59	47.9
CC (all statements)	model	111.64	53.6
CC (all statements)	synopsis	1.80	0.9
CC (all statements)	model × synopsis	4.74	2.3
CC (all statements)	Residual (between summaries)	90.24	43.3
CC (when cited)	model	3.81	12.7
CC (when cited)	synopsis	0.86	2.9
CC (when cited)	model × synopsis	1.73	5.8
CC (when cited)	Residual (between summaries)	23.56	78.6
Faithfulness	model	10.37	17.9
Faithfulness	synopsis	3.13	5.4
Faithfulness	model × synopsis	1.25	2.2
Faithfulness	Residual (between summaries)	43.04	74.5

Extended Data Table 4: Mean summary-level score (%) by subject area (synopsis) and model. Cells are means across evaluated summaries; n is the number of evaluated summaries per synopsis summed over models. Rows are ordered by mean faithfulness across models.

Synopsis	n	Metric	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
Mediterranean Farmland	117	Relevance	89.3	76.2	95.0	95.0
		Citation rate	87.2	93.8	97.4	62.1
		CC (when cited)	90.5	94.0	97.6	73.1
		Faithfulness	75.0	80.6	90.1	59.3
Soil Fertility	86	Relevance	91.8	85.4	96.3	93.2
		Citation rate	87.8	96.4	100.0	59.1
		CC (when cited)	87.9	96.1	96.4	88.4
		Faithfulness	67.7	92.0	80.6	65.3
Butterflyand Moth Conservation	132	Relevance	96.2	91.9	94.9	98.0
		Citation rate	83.2	97.3	99.0	29.7
		CC (when cited)	93.0	95.3	94.7	80.2
		Faithfulness	79.5	90.2	83.6	67.2
Marine Fish Conservation	81	Relevance	94.5	94.7	96.9	92.6
		Citation rate	83.2	96.3	97.1	27.8
		CC (when cited)	91.0	97.5	98.2	62.1
		Faithfulness	75.0	87.4	87.5	73.8
Grassland Conservation	88	Relevance	89.9	84.0	98.7	95.6
		Citation rate	85.2	90.4	99.0	52.7
		CC (when cited)	93.8	92.4	91.3	83.5
		Faithfulness	79.9	88.1	85.7	71.0
Biodiversityof Marine Artificial Structures	169	Relevance	97.0	93.7	98.2	97.5
		Citation rate	90.6	96.6	96.8	43.7
		CC (when cited)	92.0	96.5	96.0	76.5
		Faithfulness	80.2	87.1	88.7	72.2
Marineand Freshwater Mammal Conservation	101	Relevance	94.4	91.4	97.6	97.1
		Citation rate	81.0	98.0	96.4	38.0
		CC (when cited)	87.0	96.9	93.2	87.1
		Faithfulness	76.9	91.9	89.5	71.6

Synopsis	n	Metric	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
Amphibian Conservation	124	Relevance	95.2	93.9	97.4	95.2
		Citation rate	89.2	94.6	99.2	22.4
		CC (when cited)	91.8	95.3	93.9	85.3
		Faithfulness	81.4	85.7	87.8	76.8
Bee Conservation	98	Relevance	91.8	82.5	97.8	90.7
		Citation rate	82.1	97.0	99.7	41.3
		CC (when cited)	95.9	98.9	99.8	71.6
		Faithfulness	81.1	91.6	89.6	71.7
Peatland Conservation	91	Relevance	94.9	91.5	95.8	97.0
		Citation rate	85.0	94.9	96.2	55.7
		CC (when cited)	93.9	95.9	98.3	84.9
		Faithfulness	84.9	91.6	93.9	73.4
Primate Conservation	85	Relevance	90.8	73.3	94.0	90.9
		Citation rate	80.8	97.3	93.3	25.2
		CC (when cited)	89.3	98.5	96.7	90.2
		Faithfulness	81.5	95.3	93.8	74.2
Bat Conservation	90	Relevance	96.5	93.5	98.2	95.8
		Citation rate	85.4	96.6	95.7	61.6
		CC (when cited)	94.9	97.5	96.0	90.4
		Faithfulness	81.4	92.6	92.7	79.6
Natural Pest Control	74	Relevance	93.6	90.6	97.0	93.5
		Citation rate	89.4	95.8	99.0	37.9
		CC (when cited)	96.6	97.7	98.8	83.0
		Faithfulness	85.1	93.4	92.9	76.1
Farmland Conservation	100	Relevance	96.4	89.6	99.1	97.5
		Citation rate	89.3	97.7	99.4	45.0
		CC (when cited)	96.0	98.2	97.3	71.4
		Faithfulness	87.3	94.8	90.3	75.2
Shrublandand Heathland Conservation	87	Relevance	87.0	88.5	94.2	95.9
		Citation rate	82.3	98.5	99.4	41.0
		CC (when cited)	97.5	98.3	98.7	95.2
		Faithfulness	85.3	95.3	89.2	78.1
Controlof Freshwater Invasive Species	159	Relevance	94.4	87.1	95.0	93.0

Synopsis	n	Metric	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
		Citation rate	86.9	97.1	96.0	49.5
		CC (when cited)	93.4	99.2	99.4	87.0
		Faithfulness	81.6	97.7	94.7	74.8
Forest Conservation	92	Relevance	87.4	84.1	91.7	90.9
		Citation rate	82.3	97.8	98.8	42.8
		CC (when cited)	97.0	96.6	98.5	95.3
		Faithfulness	84.9	96.3	92.3	80.2
Sustainable Aquaculture	110	Relevance	90.3	90.3	96.9	92.5
		Citation rate	89.4	92.9	98.3	34.8
		CC (when cited)	96.9	97.6	98.0	88.9
		Faithfulness	87.8	94.5	94.4	82.3
Managementof Captive Animals	107	Relevance	92.2	90.7	96.9	88.6
		Citation rate	84.8	98.0	98.6	39.0
		CC (when cited)	96.4	98.9	99.5	94.7
		Faithfulness	85.7	96.6	96.7	80.8

Extended Data Table 5: Loss of summaries to the ten-iteration search cap, by model. A run that reached the cap produced no summary. 'Questions never completed' counts questions for which no attempt produced a summary; 'hit cap on ≥1 attempt' counts questions that reached the cap on any attempt, including those later recovered on re-execution; 'summaries missing' is 520 minus the number of evaluated summaries. All quality metrics are conditional on completion.

Model	Questions never completed (n)	Questions never completed (%)	Hit cap on ≥1 attempt	Recovered on re-execution	Summaries missing
Claude Sonnet 4	2	0.4	5	3	5
Gemini 2.5 Pro	0	0.0	1	1	0
GPT-5	38	7.3	105	67	53
Kimi K2	13	2.5	45	32	31

Extended Data Table 6: Failure-reason categories for statements judged as not fully faithful: definitions and prevalence by model. Reasons are drawn from a controlled vocabulary; a statement may be tagged with more than one reason for failure. For each model: % of evaluated summaries containing at least one statement with the failure reason; mean number of tagged statements per summary with 95% bootstrap CI; and failure reason tags per 100 statements scored.

Failure reason	Definition	Model	Summaries with ≥1 (%)	Tagged statements per summary	Tags per 100 statements
absent or contradicted fact	A specific factual claim (effects, sample sizes, locations, species, contexts, study design) that is absent from or directly contradicted by the cited action pages.	Claude Sonnet 4	68.7	1.92 [1.74, 2.09]	12.5
		Gemini 2.5 Pro	33.1	0.57 [0.49, 0.66]	6.2
		GPT-5	40.5	0.88 [0.76, 1.00]	6.7
		Kimi K2	80.4	2.34 [2.16, 2.54]	19.1
overgeneralisation	Extends the cited evidence beyond its stated scope, e.g. 'most', 'all', 'typically' or 'generally' when the evidence is mixed, limited or from few studies.	Claude Sonnet 4	45.6	0.99 [0.86, 1.13]	5.9
		Gemini 2.5 Pro	11.0	0.17 [0.12, 0.22]	1.5
		GPT-5	19.1	0.30 [0.23, 0.37]	2.2
		Kimi K2	54.0	1.21 [1.07, 1.37]	9.6
wrong quantity	Reports a number, count or quantifier that is wrong or not clearly supported (e.g. 'several studies' when only one or two exist).	Claude Sonnet 4	32.8	0.60 [0.51, 0.70]	4.0
		Gemini 2.5 Pro	13.7	0.23 [0.17, 0.29]	2.2
		GPT-5	20.6	0.37 [0.29, 0.46]	2.9
		Kimi K2	31.9	0.53 [0.45, 0.62]	4.2
mismerged studies	Combines findings from different studies into a single claim that matches no individual study.	Claude Sonnet 4	17.5	0.29 [0.23, 0.37]	2.0
		Gemini 2.5 Pro	13.9	0.18 [0.14, 0.23]	2.0
		GPT-5	15.8	0.25 [0.19, 0.31]	2.1
		Kimi K2	22.7	0.34 [0.27, 0.41]	2.8
unsupported comparative claim	A comparative claim ('X is more	Claude Sonnet 4	25.4	0.39 [0.33, 0.46]	2.7

Failure reason	Definition	Model	Summaries with ≥1 (%)	Tagged statements per summary	Tags per 100 statements
	effective than Y', 'no difference from Z') not explicitly supported by at least one cited action page.				
		Gemini 2.5 Pro	2.1	0.02 [0.01, 0.04]	0.2
		GPT-5	3.6	0.04 [0.02, 0.07]	0.3
		Kimi K2	30.5	0.49 [0.41, 0.57]	4.0
wrong effect direction or magnitude	Misrepresents the direction, size or certainty of an effect, e.g. 'increased' when the action page reports no significant effect.	Claude Sonnet 4	21.2	0.31 [0.25, 0.37]	1.9
		Gemini 2.5 Pro	5.0	0.06 [0.04, 0.09]	0.7
		GPT-5	7.3	0.09 [0.06, 0.12]	0.7
		Kimi K2	26.8	0.43 [0.36, 0.52]	3.4
unsupported evidence summary	A claim about the overall state of the evidence base ('no studies', 'only one study', 'many studies') not supported by the cited action pages.	Claude Sonnet 4	24.3	0.34 [0.28, 0.41]	2.3
		Gemini 2.5 Pro	2.5	0.04 [0.01, 0.06]	0.3
		GPT-5	4.7	0.06 [0.03, 0.08]	0.4
		Kimi K2	22.3	0.32 [0.26, 0.38]	2.7

Extended Data Table 7: LLM-judge validation by candidate judge, prompt and metric. Three candidate judges were compared under four prompt regimes (seed manual prompt v7; human–AI refined prompt; two prompts optimised by GEPA with different seeds) against the expert-labelled subset (n statements). Agreement = % of statements matching the expert label; κ = Cohen's kappa; |bias| = mean absolute difference between judge and expert Yes-rates across the four summariser models (percentage points); bias by summariser = judge − expert Yes-rate for each summariser's statements (positive = judge more lenient). Rows marked ✓ are the selected judge–prompt combination for each metric (Gemini 3.1 Pro Preview, refined prompt).

Metric	Judge	Prompt	Selected	n	Agreement (%)	κ	Mean |bias| (pp)	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
Relevance	Gemini 3.1 Pro Preview	refined (v16)	✓	820	94.1	0.40	2.7	+2.7	+2.5	+3.1	+2.4
Relevance	Claude Opus 4.7	refined (v16)		819	93.8	0.38	2.3	+2.3	+1.0	+2.6	+3.3
Relevance	GPT-5.4	refined (v16)		820	93.8	0.38	2.3	+1.8	+2.0	+3.7	+1.9
Relevance	Gemini 3.1 Pro Preview	GEPA seed 1234		820	84.4	0.34	13.4	-16.8	-16.2	-10.5	-10.0
Relevance	Gemini 3.1 Pro Preview	GEPA seed 5678		820	85.6	0.35	11.6	-15.0	-11.1	-8.4	-11.8
Relevance	Gemini 3.1 Pro Preview	seed prompt (v7)		820	92.3	0.37	1.6	-1.4	-2.0	+2.1	+0.9
Relevance	Claude Opus 4.7	GEPA seed 1234		818	88.1	0.40	8.6	-11.4	-11.6	-5.3	-6.2
Relevance	Claude Opus 4.7	GEPA seed 5678		818	86.7	0.36	9.9	-10.5	-12.6	-8.9	-7.6
Relevance	Claude Opus 4.7	seed prompt (v7)		811	91.9	0.33	1.7	-2.8	-1.0	+2.1	+1.0
Relevance	GPT-5.4	GEPA seed 1234		820	82.2	0.31	15.8	-16.8	-20.2	-13.1	-13.3
Relevance	GPT-5.4	GEPA seed 5678		820	81.0	0.28	16.5	-21.4	-19.7	-14.1	-10.9
Relevance	GPT-5.4	seed prompt (v7)		820	93.3	0.33	2.4	+0.9	+1.0	+4.7	+2.8
Citation correctness (when cited)	Gemini 3.1 Pro Preview	refined (v17)	✓	617	94.3	0.38	2.6	-3.3	+1.1	+1.1	-4.9
Citation correctness (when cited)	Claude Opus 4.7	refined (v17)		616	94.0	0.27	1.2	-1.1	-0.5	+1.6	+1.6
Citation correctness (when cited)	GPT-5.4	refined (v17)		617	93.8	0.26	1.1	-1.7	-0.5	+2.1	+0.0
Citation correctness (when cited)	Gemini 3.1 Pro Preview	GEPA seed 1234		617	92.4	0.41	5.1	-13.3	-1.1	-1.1	-4.9

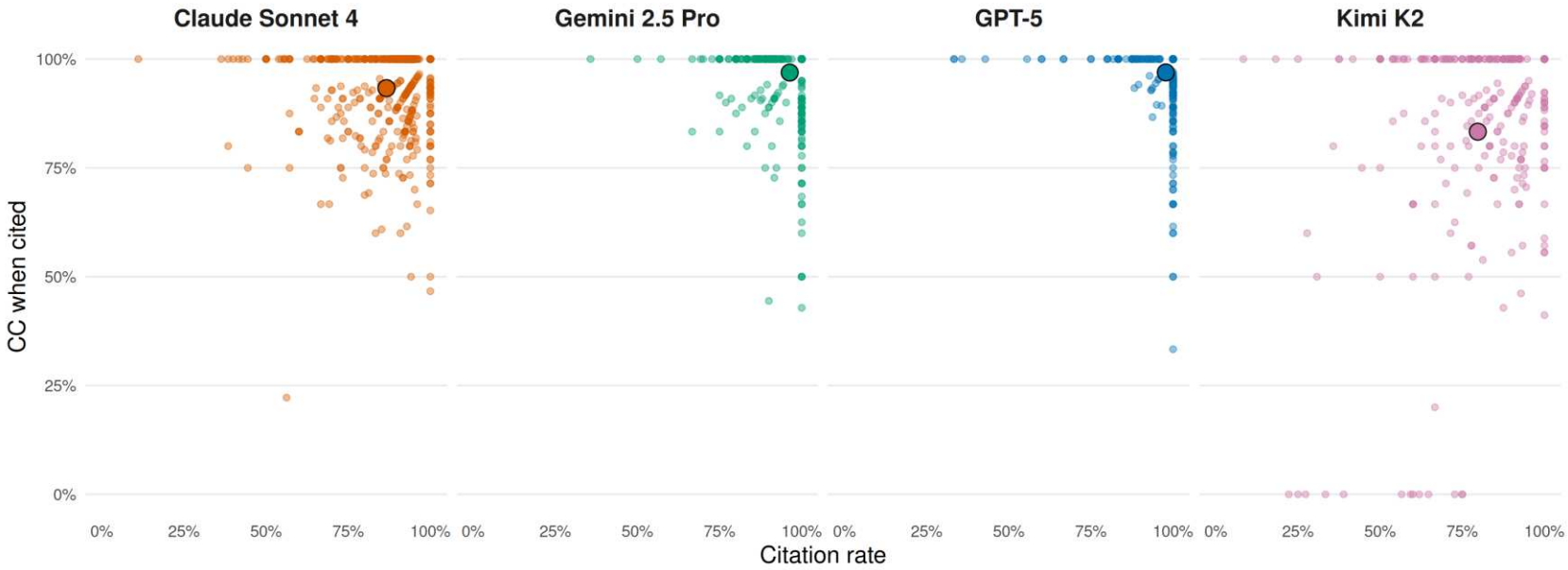
Metric	Judge	Prompt	Selected	n	Agreement (%)	κ	Mean |bias| (pp)	Claude Sonnet 4	Gemini 2.5 Pro	GPT-5	Kimi K2
Citation correctness (when cited)	Gemini 3.1 Pro Preview	GEPA seed 5678		617	91.7	0.35	4.8	-11.7	-0.5	-2.1	-4.9
Citation correctness (when cited)	Gemini 3.1 Pro Preview	seed prompt (v7)		617	93.0	0.28	2.6	-5.6	+0.5	+1.1	-3.3
Citation correctness (when cited)	Claude Opus 4.7	GEPA seed 1234		616	93.3	0.42	3.2	-8.4	-1.6	-1.1	-1.6
Citation correctness (when cited)	Claude Opus 4.7	GEPA seed 5678		616	93.3	0.35	2.1	-6.1	+0.5	-1.6	+0.0
Citation correctness (when cited)	Claude Opus 4.7	seed prompt (v7)		612	93.5	0.28	1.1	-2.2	+0.0	+0.5	-1.6
Citation correctness (when cited)	GPT-5.4	GEPA seed 1234		617	93.5	0.40	2.8	-8.9	+0.0	+0.5	-1.6
Citation correctness (when cited)	GPT-5.4	GEPA seed 5678		617	90.0	0.23	5.2	-13.3	-2.1	-0.5	-4.9
Citation correctness (when cited)	GPT-5.4	seed prompt (v7)		617	93.4	0.25	1.5	-3.9	+0.0	+2.1	+0.0
Faithfulness	Gemini 3.1 Pro Preview	refined (v14)	✓	820	83.5	0.40	4.3	-4.5	+5.1	+4.2	-3.3
Faithfulness	Claude Opus 4.7	refined (v14)		817	83.6	0.43	5.1	-3.7	+5.6	+1.6	-9.5
Faithfulness	GPT-5.4	refined (v14)		820	72.7	0.28	15.3	-20.5	-4.0	-12.0	-24.6
Faithfulness	Gemini 3.1 Pro Preview	GEPA seed 1234		820	82.0	0.35	5.8	-5.5	+6.1	+4.7	-7.1
Faithfulness	Gemini 3.1 Pro Preview	GEPA seed 5678		820	81.0	0.38	4.4	-8.6	+0.0	-1.0	-8.1
Faithfulness	Gemini 3.1 Pro Preview	seed prompt (v7)		820	84.4	0.34	5.7	+4.1	+6.6	+7.3	+4.7
Faithfulness	Claude Opus 4.7	GEPA seed 1234		819	84.2	0.37	4.4	+3.2	+6.6	+7.3	-0.5
Faithfulness	Claude Opus 4.7	GEPA seed 5678		819	82.9	0.36	3.7	+0.5	+5.1	+4.7	-4.7
Faithfulness	Claude Opus 4.7	seed prompt (v7)		818	84.6	0.29	8.8	+9.6	+7.1	+10.5	+8.1
Faithfulness	GPT-5.4	GEPA seed 1234		820	81.5	0.34	3.1	-4.1	+2.0	+2.6	-3.8
Faithfulness	GPT-5.4	GEPA seed 5678		820	81.1	0.35	3.7	-5.9	+1.5	+2.1	-5.2
Faithfulness	GPT-5.4	seed prompt (v7)		820	82.9	0.30	4.5	+0.0	+5.1	+7.3	+5.7

Extended Data Table 8: Inter-rater reliability between the three expert reviewers. Nineteen statements were dual-scored within each reviewer pair; citation correctness (when cited) is computed over the cited statements among them. Raw agreement and Cohen's κ; κ is 0 or undefined when one reviewer's labels have no variance, which is expected on small, highly skewed subsets, so raw agreement is the more informative figure here.

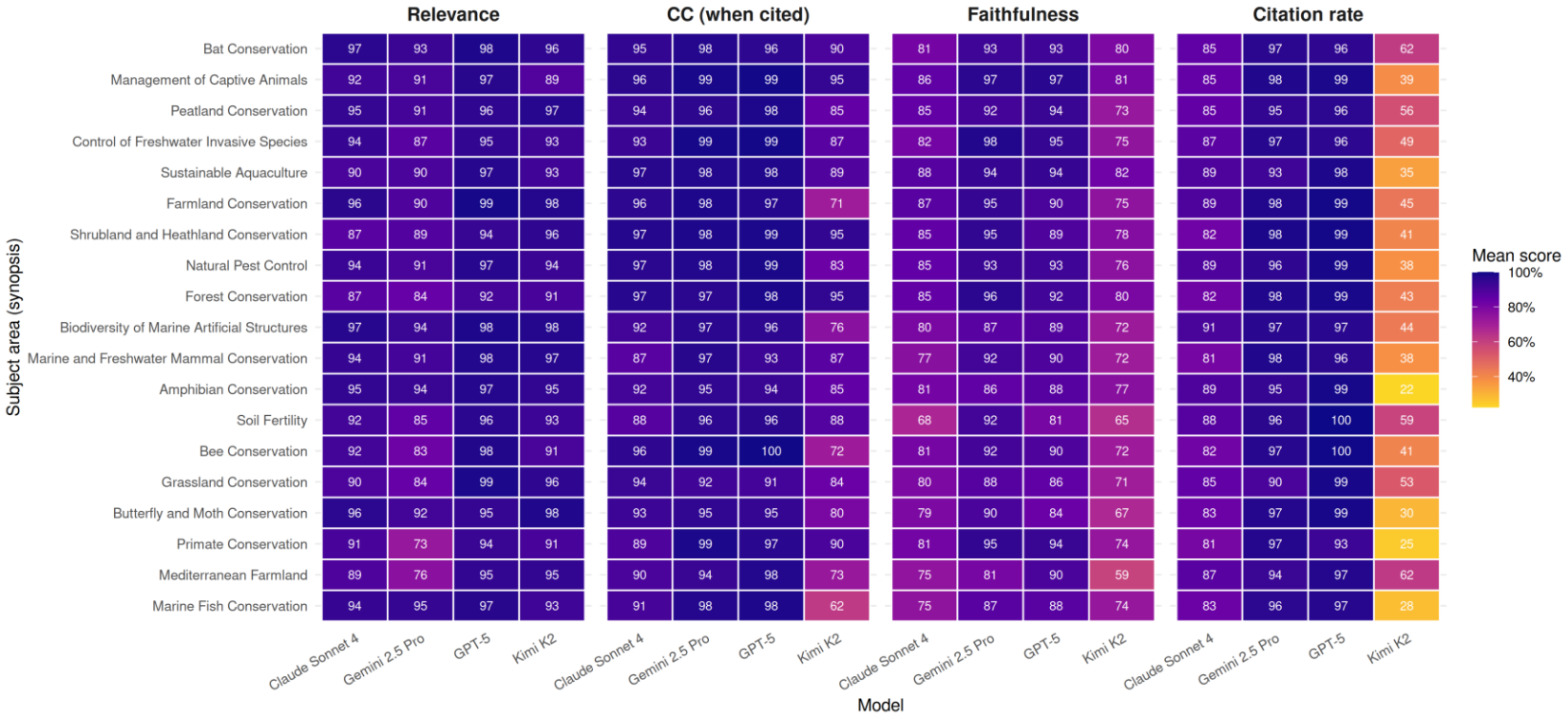
Metric	Comparison	n	Raw agreement (%)	κ
Relevance	R1 × R2	19	100.0	undefined†
Relevance	R1 × R3	19	84.2	-0.075
Relevance	R2 × R3	19	94.7	0.000
Citation correctness (when cited)	R1 × R2	15	100.0	1.000
Citation correctness (when cited)	R1 × R3	15	93.3	0.000
Citation correctness (when cited)	R2 × R3	15	100.0	1.000
Faithfulness	R1 × R2	19	89.5	0.457
Faithfulness	R1 × R3	19	89.5	0.457
Faithfulness	R2 × R3	19	84.2	0.000

†All 19 statements were labelled Yes by both reviewers; Cohen's κ is undefined when a rater's labels have zero variance.

Extended Data Fig. 1: Citation rate against citation correctness (when cited) per summary. One point per evaluated summary, one panel per model. Kimi K2 shows two regimes: sparsely and poorly cited summaries and densely and accurately cited ones.



Extended Data Fig. 2: Mean evaluation metric scores by subject area and model. Heatmaps of mean summary-level relevance, citation correctness (when cited), faithfulness and citation rate (%) for each of the 19 subject areas (rows) and four summariser LLMs (columns). Subject areas are ordered by their mean score across all four metrics and models. Values as in Extended Data Table 4.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationupdated.docx