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.

Source Expertise, Argument Strength, and Multidimensional Source Sensitivity in Large Language Model Evaluation

Sujal M G

contact@sujalmg.com

Independent Researcher

Research Article

Keywords: large language models, LLM evaluation, source credibility, source attribution, AI bias, responsible AI, algorithmic fairness

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11544999/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.

Abstract

Large language models are increasingly used to evaluate arguments, responses, and decisions, yet their judgments may be influenced by contextual information about who produced the content. This study examined how source expertise relevance and argument strength affect LLM evaluation and whether source effects differ across judgment dimensions.

Twenty arguments spanning 10 topics were constructed as strong–weak pairs and presented under four source conditions: relevant expert, irrelevant expert, non-expert, and no source. Four large language models evaluated each argument on logical quality, credibility, persuasiveness, and agreement. Five repetitions were collected for each model-by-argument-by-source-by-dimension combination and averaged before inferential testing. Human validation assessed the intended argument-strength manipulation but showed negligible overall strong–weak separation in the second validation round, including two reversed topic pairs. Separate manipulation checks indicated that the models rated topic-relevant expertise and general professional expertise differently across source conditions.

Source attribution most consistently affected credibility. Topic-level robustness analyses showed significant source-condition effects on credibility across all four models, while persuasiveness remained significant only for GPT-5.6 Sol. Relevant-expert attribution did not produce a universal advantage: credibility was significantly higher for relevant than irrelevant experts for Claude Opus 5.5 and DeepSeek Flash, and Claude Opus 5.5 also showed significantly higher persuasiveness for relevant than irrelevant experts. Exploratory analyses suggested that this pattern may reflect an irrelevant-expertise penalty more than a universal bonus for relevant expertise. No robust evidence showed that relevant expertise disproportionately improved weak arguments. Credibility was descriptively more source-sensitive than logical quality, but corrected direct contrasts were not statistically significant.

Overall, LLM evaluation was neither fully source-neutral nor characterized by simple deference to experts. Source expertise produced dimension-specific and model-dependent effects, with credibility showing the clearest sensitivity.

Introduction

Large language models are increasingly used not only to generate text, but also to evaluate it. They are asked to judge the quality of arguments, compare competing responses, assess credibility, rank alternatives, and support decisions. This has made LLM-based evaluation useful in many settings, but it also raises an important question: **are these evaluations based mainly on the content being judged, or can contextual information about the source influence the result?**

Recent research suggests that LLM judgments are not fully invariant to contextual framing. Germani and Spitale (2025) showed that changing the attributed source of otherwise identical content can alter model agreement. Research on LLM-as-a-judge systems has similarly documented sensitivity to factors such as response position, verbosity, authority cues, and other contextual features (Ye et al., 2025). Other studies

have shown that wording, hierarchical roles, and characteristics associated with the speaker can also influence model decisions (Cheung et al., 2025; Hofmann et al., 2024; Oh et al., 2025).

These findings challenge the idea that an LLM evaluator can automatically be treated as a neutral judge of content. However, not every source effect should necessarily be interpreted as undesirable bias. Source information can sometimes be genuinely relevant to evaluation. For example, knowing that a cybersecurity recommendation comes from a cybersecurity researcher may reasonably affect its perceived credibility. The same source information should theoretically matter less when judging whether the internal reasoning of the argument is logically sound.

This distinction creates an important unresolved question. Existing research shows that source and authority cues can influence LLM judgments, but few studies directly distinguish topic-relevant expertise from general authority (Germani & Spitale, 2025; Oh et al., 2025; Ye et al., 2025). A professional or academic title may signal expertise, but that expertise may not be relevant to the specific topic being evaluated. Prior work also often relies on a single evaluation outcome, making it difficult to determine whether source attribution affects credibility, logical quality, persuasiveness, and agreement in the same way.

A second unresolved issue concerns **argument strength**. If an authoritative source has a larger positive effect on weak arguments than on strong arguments, this would suggest that source information can partly compensate for poor reasoning. By contrast, if strong and weak arguments are similarly affected, or if weak arguments remain weak regardless of attribution, this would provide a different picture of authority sensitivity.

The present study addresses these gaps by experimentally manipulating both **source expertise relevance** and **argument strength** while holding the underlying argument text constant across source conditions. Arguments were attributed to a relevant expert, an irrelevant expert, a non-expert, or no source and were evaluated by four large language models on logical quality, credibility, persuasiveness, and agreement.

The study asks whether source expertise attribution changes evaluations of otherwise identical arguments, whether relevant expertise produces an advantage over irrelevant expertise or non-expert attribution, whether source effects are stronger for weak arguments, and whether some judgment dimensions are more sensitive to source information than others.

By separating source relevance, argument strength, and evaluation dimension, the study aims to provide a more precise account of **source sensitivity in LLM evaluation**. Rather than treating LLM evaluators as simply biased or neutral, the goal is to identify where source information enters the judgment process and whether its influence differs depending on what the model is being asked to evaluate.

Related Work

Source Framing and Speaker Information

A growing body of research shows that LLM evaluations can change when information about the source or speaker changes, even when the substantive content remains similar or identical.

Germani and Spitale (2025) demonstrated that source framing can systematically alter model evaluations of the same statements. Their findings provide direct evidence that LLM judgments are not fully source-invariant and that information about who supposedly produced a statement can influence evaluation.

Related work on speaker characteristics reaches a similar conclusion through a different mechanism. Hofmann et al. (2024) showed that language models made systematically different judgments about speakers using African American English and Standard American English, even though race was not explicitly stated. Their results suggest that source-associated social cues can influence model behavior independently of the underlying semantic content.

Together, these studies indicate that information associated with the speaker can enter the evaluation process even when it is not part of the substantive argument itself. However, these studies do not distinguish between professional expertise that is relevant to the topic and professional status that is unrelated to it.

Bias in LLM-as-a-Judge Systems

Research on LLM-as-a-judge systems has documented a broader set of evaluation biases.

Ye et al. (2025) introduced CALM, a framework for assessing multiple forms of bias in LLM judges. Their experiments demonstrated sensitivity to factors including response position, verbosity, authority cues, self-related information, and other contextual perturbations.

Related work also shows that LLM judgments can be influenced by contextual and interpersonal cues rather than content alone. Sharma et al. (2023) found that several state-of-the-art assistants exhibited sycophancy, including a tendency to produce responses that matched users' expressed views even when doing so conflicted with truthfulness. In the LLM-as-a-judge setting, Zheng et al. (2023) likewise documented systematic evaluation biases, including sensitivity to response position, verbosity, and self-enhancement. Together, these findings reinforce the broader concern that LLM evaluations may respond to contextual cues that are not intrinsic to the quality of the argument being judged.

These findings are important because they show that strong overall judging performance does not guarantee that evaluations are based solely on the substantive quality of the material being judged.

The authority component is particularly relevant to the present study. If contextual signals associated with authority can affect evaluation, then professional expertise labels may also influence how arguments are judged. However, authority and expertise relevance are not equivalent. A source may have high professional status while lacking expertise relevant to the specific topic.

Authority and Hierarchical Influence

More direct evidence of authority sensitivity comes from multi-agent research.

Oh et al. (2025) examined hierarchical interactions between professor and student agents. When a student agent initially held the correct answer but debated a professor defending an incorrect answer, conformity to the incorrect position was substantially higher than when the same student debated another student.

This result suggests that hierarchical status can influence LLM decision-making even when following the authority figure reduces factual accuracy.

However, authority in that design was represented through a broad professor–student hierarchy. It did not test whether the authority figure's expertise was relevant to the particular question.

The present study extends this line of work by separating **relevant expertise** from **irrelevant expertise**, allowing general expert status to be distinguished from topic-specific competence.

Recent Source-Label and Expertise Studies

Recent studies have examined source and label effects in settings that are especially close to the present study. Nahar et al. (2026) compared human and LLM judgments of arguments containing logical fallacies under human, AI, mixed-authorship, and no-disclosure source labels. Human judgments were significantly influenced by the source labels, whereas LLM evaluations remained comparatively stable across conditions. This finding is particularly relevant to the present study because it suggests that source attribution may have relatively limited influence on LLM judgments of reasoning quality.

Mammen et al. (2026) examined authority bias by varying the expertise level of sources providing endorsements across mathematical, legal, medical, and scientific reasoning tasks. Across 11 models, susceptibility to incorrect or misleading endorsements increased with the perceived expertise of the source, demonstrating that LLM reasoning can be affected by graded authority cues rather than only by a simple expert-versus-non-expert distinction.

Saraf et al. (2026) examined label-induced bias in LLM self- and cross-evaluation using true, false, and absent model labels. They found that perceived model identity affected both overall preference judgments and fine-grained quality ratings, with informativeness showing the greatest label sensitivity while conciseness remained comparatively stable. This pattern further suggests that label-induced effects can differ substantially across evaluation dimensions.

Framing and Cognitive Bias

LLM judgments can also be influenced by the way a problem is presented.

Cheung et al. (2025) found that large language models exhibited strong cognitive biases in moral decision-making, including omission bias and sensitivity to yes–no question framing. Logically similar decisions could produce different model responses depending on how the question was phrased.

These findings reinforce the broader conclusion that LLM judgment can be sensitive to contextual information beyond the core substance of a decision problem.

They also suggest that model biases may not simply reproduce human cognitive biases and may partly arise from model-specific training or post-training processes.

Remaining Gap

Taken together, prior studies show that source attribution can alter LLM evaluation, speaker-associated characteristics can affect model judgments, authority cues can influence decision-making, and contextual framing can produce systematic evaluation differences (Cheung et al., 2025; Germani & Spitale, 2025; Hofmann et al., 2024; Mammen et al., 2026; Nahar et al., 2026; Oh et al., 2025; Saraf et al., 2026; Ye et al., 2025).

However, several gaps remain.

First, few studies directly distinguish relevant expertise from irrelevant expertise. This makes it difficult to determine whether models respond to professional authority in general or whether they evaluate whether that authority is actually appropriate to the topic.

Second, many previous studies emphasize a single primary outcome, such as agreement, correctness, conformity, or a general judge score. They therefore provide limited evidence about whether source information affects different evaluation dimensions in different ways.

Third, relatively little work has examined whether the influence of source expertise depends on **argument strength**. A source cue that substantially improves the evaluation of weak reasoning would raise different concerns from a source cue that mainly changes perceived credibility while leaving logical judgments stable.

The present study addresses these gaps by crossing source expertise relevance with argument strength and measuring four distinct outcomes: logical quality, credibility, persuasiveness, and agreement. This design makes it possible to examine not only whether source attribution matters, but also **where its influence appears and whether that influence is consistent with the type of judgment being made**.

Theoretical Framework

Source characteristics have long been known to influence human persuasion and evaluation. Early work by Hovland and Weiss (1951) showed that the credibility of a source can affect acceptance of otherwise

comparable communications. Later research reviewed across several decades likewise found substantial evidence that source credibility can influence persuasion, although the size and direction of the effect depend on characteristics of the source, message, recipient, and context (Pornpitakpan, 2004). Importantly, source expertise does not operate independently of message quality. Petty, Cacioppo, and Goldman (1981) experimentally crossed source expertise with argument quality and found that argument quality exerted greater influence under high personal relevance, whereas source expertise had greater influence under low personal relevance. These findings provide a human persuasion basis for examining whether LLM evaluations similarly respond to source expertise and argument quality as distinct but potentially interacting cues.

The hypotheses are informed by dual-process accounts of persuasion, particularly the Elaboration Likelihood Model (Petty & Cacioppo, 1986). The model distinguishes between judgments based primarily on the substantive merits of a message and judgments influenced by contextual cues such as characteristics of the source. Source expertise can therefore affect evaluation independently of argument content, although its role depends on how the information is processed. Applied to LLM evaluation, this framework motivates the expectation that relevant expertise may influence judgments beyond the argument itself (H2) and that source effects may differ across evaluative dimensions (H4). The study further tests whether this influence becomes larger when argument quality is weak (H3), treating that interaction as an empirical extension rather than a direct prediction of the original model.

Hypotheses

Based on the preceding literature and research gaps, the study tested four hypotheses:

H1: LLMs will evaluate identical arguments differently depending on source condition.

H2: Arguments attributed to a relevant expert will receive higher evaluations than the same arguments attributed to an irrelevant expert, a non-expert, or no source.

H3: The positive effect of relevant-expert attribution will be larger for weak arguments than for strong arguments.

H4: Source attribution will affect judgment dimensions differently, with credibility expected to be more source-sensitive than logical quality.

Methods

Study Design

The study used a factorial evaluation design in which otherwise identical arguments were presented under different source-attribution conditions.

The experiment included:

- 10 topics
- 2 argument-strength levels: strong and weak
- 4 source conditions: relevant expert, irrelevant expert, non-expert, and no source
- 4 judgment dimensions: logical quality, credibility, persuasiveness, and agreement
- 4 large language models
- 5 repetitions per argument × source condition × judgment dimension × model combination

This produced 20 argument stimuli in total, consisting of one strong and one weak argument for each topic.

For each model, the main experiment therefore contained:

20 arguments × 4 source conditions × 4 judgment dimensions × 5 repetitions = 1,600 evaluations

Across all four models, the main experiment contained **6,400 successful evaluations**.

Argument Stimuli

The study used 20 arguments drawn from 10 topics. Each topic included one intended strong argument and one intended weak argument.

Argument strength was manipulated by varying the quality of the reasoning while keeping the general topic and intended recommendation broadly comparable. Although the topic and intended recommendation were broadly comparable within each pair, the strong and weak arguments were not always perfectly matched in rhetorical form; in several cases, weak arguments used more explicit normative wording than the corresponding strong arguments. Strong arguments were designed to provide clearer justification and stronger support, while weak arguments relied on less well-supported or more superficial reasoning.

Human validation was conducted before the main experiment to assess whether participants perceived the intended strong–weak distinction. For each argument, the human-validation survey included three separate 1–7 ratings: logical quality, support quality, and overall argument strength. The overall-strength measure reported in Table 1 refers directly to participants' responses to the survey item, "Overall, how strong is this argument as a well-reasoned case for its conclusion?"

The first validation round included 30 participants. Inspection of the Round 1 ratings indicated that seven of the ten intended weak arguments required revision because the intended strong–weak separation was not sufficiently clear. Those seven weak arguments were revised before Round 2, while the remaining three topic pairs were carried forward without revision.

A second validation round was then conducted using 17 unique participants after exact duplicate submissions were removed. Overall differentiation between the intended strong and weak arguments was negligible in Round 2, with an average strong–weak difference of approximately +0.03 across the seven revised topic pairs, and two topic pairs, Movies/Series and Cost of Living, showed reversed strong–weak patterns.

These two topics were retained in the primary study, and an H3 sensitivity analysis was later conducted after excluding them.

Table 1. Human validation results for the seven revised argument pairs in Round 2 (N = 17)

Topic	Strong overall-strength mean	Weak overall-strength mean	Difference
Cost of Living	5.00	5.59	−0.59
Education	5.06	4.82	+0.24
Household Healthcare	5.47	5.29	+0.18
Movies/Series	4.35	4.82	−0.47
Shopping	4.76	4.59	+0.17
Sleeping Habits	5.41	4.94	+0.47
Transportation	5.41	5.18	+0.23
Overall	**5.07**	**5.03**	**+0.03**

The Round 2 export contained 21 participant codes, but four pairs produced exact duplicate response patterns submitted within seconds of one another. One response from each duplicate pair was removed, leaving 17 unique response sets for analysis.

The presence of these duplicate submissions indicates that the Round 2 survey was susceptible to repeated or duplicated entries. Removing one response from each exact duplicate pair reduced this specific source of duplication, but the possibility of other undetected non-independent submissions cannot be completely excluded.

Participants for the human-validation rounds were recruited through a combination of personal contacts, social media, and an online survey platform. Participants recruited through personal contacts and social media took part voluntarily without compensation, while participants recruited through the online survey platform received points provided by the platform.

Taken together, the human-validation results did not provide clear evidence that the intended strong–weak manipulation was consistently perceived as such across topics. Round 2 showed negligible overall separation, and two of the seven revised topic pairs showed reversed strong–weak ratings.

This independent study did not undergo formal institutional ethics review. Before participating in the human-validation survey, participants were shown a participant-information and consent statement explaining the study purpose, procedures, voluntary nature of participation, data use, and withdrawal conditions. Participants were required to indicate consent before beginning the survey. The research dataset did not collect names, email addresses, phone numbers, or other direct personal identifiers; responses were linked only through randomly generated participant codes. Survey data were stored in the research website's database, with access to the research dashboard and data export restricted to authorized website-administrator access. The planned retention period was five years after completion of the research project, after which the research data were to be securely deleted. Routine technical or server logs maintained by the website or hosting provider were separate from the research dataset.

Source Conditions

Each argument was presented under one of four source-attribution conditions.

Relevant Expert (RE):
The argument was attributed to a university researcher specializing in a field directly relevant to the topic.

Irrelevant Expert (IE):
The argument was attributed to a university researcher specializing in geology. This condition preserved general academic status while making the stated expertise unrelated to the topic.

Non-Expert (NE):
The argument was attributed to an adult member of the public.

No Source (NS):
The argument was presented without source attribution.

The argument text itself was held constant across source conditions. Only the source attribution changed.

Judgment Dimensions

Each argument was evaluated separately on four 1–7 rating dimensions:

- logical quality
- credibility
- persuasiveness
- agreement

The dimensions were analyzed separately because source information may reasonably matter more for some judgments, such as credibility, than for others, such as logical quality.

Manipulation Checks

Separate manipulation-check prompts were used to verify that the source conditions were interpreted as intended.

For each source-present condition, models rated:

1. topic-relevant expertise
2. general professional or academic expertise

The intended pattern was:

- relevant expert: high relevant expertise and high general expertise
- irrelevant expert: low relevant expertise but high general expertise
- non-expert: low relevant expertise and low general expertise

The no-source condition was excluded from the manipulation check because no source was present to evaluate. The manipulation checks were evaluated descriptively using mean ratings across the source-present conditions rather than through separate inferential hypothesis tests.

Models and Inference Settings

The study evaluated four large language models:

- GPT-5.6 Sol
- Claude Opus 5.5
- Gemini 4 Argon
- DeepSeek Flash

The models were accessed through their respective APIs. The exact API model identifiers used in the experiment were gpt-5.6-sol for GPT-5.6 Sol, claude-opus-5-5 for Claude Opus 5.5, gemini-4-argon for Gemini 4 Argon, and deepseek-flash for DeepSeek Flash.

The full model names: GPT-5.6 Sol, Claude Opus 5.5, Gemini 4 Argon, and DeepSeek Flash are used throughout the manuscript.

All model evaluations and manipulation-check runs were conducted between October 12 and October 19, 2026.

Inference settings were not identical across providers because each platform exposes different reasoning controls. GPT-5.6 Sol, Claude Opus 5.5, and Gemini 4 Argon were run using medium reasoning or thinking settings, while DeepSeek Flash used low reasoning with thinking enabled.

Provider-default temperature settings were used where applicable; temperature was not manually overridden in the API requests.

Repetitions and Data Aggregation

Each unique combination of argument, source condition, judgment dimension, and model was evaluated five times.

The repeated calls were intended to capture model-output variability rather than to increase the number of independent substantive stimuli.

To avoid pseudoreplication, the five repeated ratings for each argument-condition combination were averaged before inferential testing.

Data Integrity

Each model produced 1,600 successful and unique main-experiment evaluations.

GPT-5.6 Sol, DeepSeek Flash, and Claude Opus 5.5 completed all 1,600 main runs on the first attempt.

Nine Gemini 4 Argon runs required at least one retry. Eight of these were triggered because the model returned explanatory text rather than a valid numerical rating on the required 1–7 scale, and one followed a temporary 503 UNAVAILABLE API error. Six of the nine runs succeeded on the second attempt, while three required a third attempt, resulting in 12 retry attempts in total.

As a robustness check, the Gemini 4 Argon analyses were repeated after excluding all nine runs that required retries and retaining only the 1,591 first-attempt-success runs. The substantive conclusions were unchanged: the same planned H2 contrasts remained significant, no H3 interaction became significant after correction, and the H4 credibility-versus-logic result was unchanged.

For the manipulation checks, GPT-5.6 Sol, Gemini 4 Argon, and DeepSeek Flash each produced 600 successful unique runs.

Claude Opus 5.5 produced 620 manipulation-check response records, comprising 600 successful unique responses and 20 failed attempts.

The 20 failed Claude Opus 5.5 runs resulted from insufficient API credit balance and each exhausted the configured four-attempt retry limit. These failures were distributed across all three source conditions and across nine topics.

Only successful unique responses were included in analysis.

The failed Claude Opus 5.5 manipulation-check calls were rerun after API access was restored until a successful response was obtained for each intended experimental cell, so the final analyzed manipulation-check dataset remained complete at 600 successful unique responses.

All analyzed ratings fell within the intended 1–7 response scale.

Statistical Analysis

Where bootstrap confidence intervals are reported for H2, H3, H4, and the exploratory credibility contrasts, 95% percentile intervals were estimated using 10,000 bootstrap resamples with replacement. Resampling was performed over the independent analysis units used for each comparison: argument-level paired differences for H2, H4, and the exploratory contrasts, and topic-level interaction values for H3. The 2.5th and 97.5th percentiles of the bootstrapped mean distributions were used as the interval bounds.

H1: Overall Effect of Source Condition

H1 tested whether evaluations of identical arguments differed across the four source conditions.

For each model and judgment dimension, a Friedman test was conducted using the 20 arguments as repeated blocks.

This produced 16 omnibus tests:

4 models × 4 judgment dimensions

Holm correction was applied across all 16 H1 omnibus tests.

Because the 20 arguments were organized as strong–weak pairs within 10 topics, H1 was also repeated as a topic-level robustness analysis using 10 topic blocks. Holm correction was again applied across the same 16 model-by-dimension tests.

H2: Planned Relevant-Expert Contrasts

H2 tested whether the relevant-expert condition received higher evaluations than:

- irrelevant expert
- non-expert
- no source

For each model and judgment dimension, three paired contrasts were tested using two-sided Wilcoxon signed-rank tests:

- RE − IE

- RE − NE
- RE − NS

Holm correction was applied within each model-by-dimension family of three planned contrasts.

The multiple-comparison families differed between H1 and H2 because H1 tested a single set of 16 omnibus model-by-dimension effects, whereas H2 tested three prespecified contrasts within each model-by-dimension combination. Accordingly, adjusted p-values from H1 and H2 should be interpreted within their respective correction families rather than compared directly across analyses.

These tests were conducted at the argument level using the 20 aggregated argument observations.

H2 was considered fully supported only if relevant-expert attribution produced consistently higher evaluations than all three comparison conditions. Evidence limited to one or two contrasts, or to only some models or judgment dimensions, was interpreted as partial support.

H3: Source × Argument-Strength Interaction

H3 tested whether the effect of relevant-expert attribution was larger for weak arguments than for strong arguments.

The interaction was defined as:

(RE − NS) for weak arguments − (RE − NS) for strong arguments

Positive values represented the direction predicted by H3.

Because strong and weak arguments were paired within topic, the analysis was conducted at the topic level using 10 topic pairs.

For each model and judgment dimension, the topic-level interaction estimates were tested against zero using a two-sided Wilcoxon signed-rank test.

Because H3 constituted a within-model hypothesis across the four judgment dimensions, Holm correction was applied across the four H3 dimension tests within each model.

A sensitivity analysis was also conducted after excluding Movies/Series and Cost of Living, the two topic pairs that showed reversed strong–weak patterns during human validation.

Because the human validation did not establish a consistently strong distinction between the intended strong and weak arguments across topics, H3 should be interpreted as a hypothesis test conducted under a partially validated manipulation rather than as a fully validated test of argument strength.

H4: Differences in Source Sensitivity Across Judgment Dimensions

H4 tested whether credibility was more sensitive to source attribution than logical quality.

For each argument and judgment dimension, source sensitivity was operationalized as the range between the highest and lowest mean rating across the four source conditions.

The primary comparison was:

credibility source-sensitivity range − logic source-sensitivity range

This comparison was calculated separately for each model.

For each model, credibility and logical-quality source-sensitivity ranges were compared across the 20 arguments using a two-sided Wilcoxon signed-rank test. The four resulting p-values, one for each model, were adjusted together using the Holm procedure.

Bootstrap 95% confidence intervals were generated for the credibility-minus-logic differences.

The direct H4 significance tests were corrected across models using Holm correction.

The bootstrap confidence intervals were not adjusted for multiple comparisons and were treated as descriptive uncertainty estimates rather than as substitutes for the corrected hypothesis tests.

Exploratory Analyses

Post hoc exploratory credibility analyses were conducted to clarify the pattern observed under H2.

Four comparisons were examined within each model:

- RE − NS
- IE − NS
- NE − NS
- IE − NE

These four comparisons were Holm-corrected together within each model.

The exploratory analyses were conducted at the argument level and were not repeated at the topic level.

Comparable post hoc no-source analyses were not conducted for logic or persuasiveness.

Results

Data Integrity and Manipulation Checks

The final analyzed dataset contained the complete intended set of successful main-experiment observations: 1,600 evaluations per model. Manipulation-check analyses included 600 successful unique responses per model. All analyzed ratings fell within the intended 1–7 scale. Details of retries, failed calls, and data-quality handling are reported in the Methods section.

The manipulation-check pattern was broadly consistent with the intended source-condition distinctions. Across the four models, relevant experts received high ratings for both relevant and general expertise, whereas irrelevant experts were rated as having high general expertise but low topic-relevant expertise. Non-experts received comparatively low ratings on both dimensions. The manipulation-check means for all four models are presented in Table 2.

Table 2 presents the descriptive manipulation-check means for the three source-present conditions. These values are reported to show whether the source descriptions produced the intended pattern of perceived topic relevance and general expertise; they were not subjected to separate inferential significance testing.

Table 2. Mean manipulation-check ratings by model and source condition

Model	RE relevant	RE general	IE relevant	IE general	NE relevant	NE general
GPT-5.6 Sol	6.90	6.89	2.10	6.16	2.59	2.99
Claude Opus 5.5	6.13	6.01	2.00	5.91	2.27	2.79
Gemini 4 Argon	6.83	6.66	1.55	6.01	3.35	2.78
DeepSeek Flash	6.61	6.77	1.04	5.28	1.08	1.19

Note. Ratings were made on 1–7 scales. RE = relevant expert; IE = irrelevant expert; NE = non-expert. "Relevant" refers to perceived topic-relevant expertise, whereas "General" refers to perceived general professional or academic expertise.

These results indicate that the models rated topic-relevant expertise and general professional or academic expertise differently across the source conditions, supporting the intended interpretation of the source manipulation.

As an additional check on the intended argument-strength manipulation, logical-quality ratings in the no-source condition were compared between the intended strong and weak arguments. Strong arguments received higher mean logical-quality ratings than weak arguments in all four models, with strong-minus-weak differences of +1.92 for GPT-5.6 Sol, +2.46 for Claude Opus 5.5, +2.60 for Gemini 4 Argon, and +2.82 for DeepSeek Flash. These model-based differences provide supplementary evidence that the intended strong and weak stimuli were distinguishable to the evaluated models, although they do not resolve the weak separation observed in the human-validation data.

H1: Effect of Source Condition on Argument Evaluation

H1 predicted that evaluations of identical arguments would differ as a function of source condition. To avoid treating repeated API calls as independent observations, the five repetitions for each argument-condition were averaged before inferential testing. Friedman tests were then conducted at the argument level, using 20 arguments as the independent blocks, and Holm correction was applied across the 16 model-by-dimension omnibus tests.

In the primary argument-level analysis, significant source-attribution effects remained for GPT-5.6 Sol credibility and persuasiveness, Claude Opus 5.5 credibility, Gemini 4 Argon credibility and persuasiveness, and DeepSeek Flash credibility. A topic-level robustness analysis using 10 topic blocks, with Holm correction applied across the same 16 model-by-dimension tests, produced the same credibility pattern across all four models, while persuasiveness remained significant only for GPT-5.6 Sol. Gemini 4 Argon persuasiveness did not remain significant after topic-level aggregation.

Two Gemini 4 Argon results fell on opposite sides of the conventional .05 threshold after correction (persuasiveness, adjusted $p = .0487$; logic, adjusted $p = .0503$), so this distinction should be interpreted cautiously rather than as evidence of a sharp substantive difference.

The H1 omnibus tests and H2 planned contrasts used different Holm-correction families. The 16 H1 omnibus tests were corrected together across all model-by-dimension combinations, whereas the H2 planned contrasts were corrected within each model-by-dimension contrast family. Accordingly, although no omnibus logic effect survived H1 correction, planned H2 contrasts showed lower logic ratings for the relevant expert than the non-expert for GPT-5.6 Sol (−0.24) and Gemini 4 Argon (−0.27).

H1 was partially supported. The most robust source-attribution effect was on credibility, which remained significant across all four models at both the argument and topic levels. Persuasiveness showed a robust effect for GPT-5.6 Sol, while Gemini 4 Argon's persuasiveness effect was not robust to topic-level aggregation. There was little corrected evidence of systematic effects on logical quality or agreement.

H2: Advantage of Relevant-Expert Attribution

H2 predicted that arguments attributed to a relevant expert would receive higher evaluations than the same arguments attributed to an irrelevant expert, a non-expert, or no source.

Relevant-expert attribution produced higher mean credibility ratings than irrelevant-expert attribution in all four models (+0.31 for GPT-5.6 Sol, +0.34 for Claude Opus 5.5, +0.19 for Gemini 4 Argon, and +0.45 for DeepSeek Flash), but this planned contrast remained statistically significant after Holm correction within its model-by-dimension contrast family only for Claude Opus 5.5 and DeepSeek Flash.

Relevant-expert attribution did not produce a general advantage over the other source conditions. After Holm correction within each model-by-dimension contrast family, GPT-5.6 Sol rated the relevant expert

lower than the non-expert on logic (−0.24) and persuasiveness (−0.26). Gemini 4 Argon rated the relevant expert lower than the non-expert on logic (−0.27) and credibility (−0.25), while DeepSeek Flash rated the relevant expert lower than the non-expert on credibility (−0.29). Other relevant-expert versus non-expert or no-source contrasts did not remain significant after correction.

Claude Opus 5.5 showed a somewhat different pattern. Relevant-expert attribution produced significantly higher credibility than irrelevant-expert attribution and significantly higher persuasiveness than irrelevant-expert attribution after correction. Differences relative to the non-expert and no-source conditions did not remain statistically significant.

Agreement showed little evidence of a general relevant-expert advantage. In most comparisons, differences were small or nonsignificant, and Gemini 4 Argon showed descriptively lower agreement for the relevant expert than the irrelevant expert, although this contrast was not statistically significant after Holm correction.

Taken together, H2 was **partially supported**. Relevant-expert attribution produced higher mean credibility than irrelevant-expert attribution across all four models, but corrected statistical significance remained only for Claude Opus 5.5 and DeepSeek Flash. Exploratory comparisons with the no-source condition suggest that the credibility pattern may reflect an irrelevant-expertise penalty rather than a universal relevant-expertise bonus. However, this asymmetry should be interpreted cautiously because ceiling effects, particularly for GPT-5.6 Sol and Gemini 4 Argon strong-argument ratings, restricted the amount of possible upward movement in the relevant-expert condition while still allowing ratings to decrease under irrelevant-expert attribution.

Descriptively, the no-source comparisons also differed by argument strength. For GPT-5.6 Sol, the relevant-expert minus no-source credibility difference was 0.00 for strong arguments and −0.04 for weak arguments, whereas the irrelevant-expert minus no-source difference was −0.06 for strong arguments and −0.60 for weak arguments. For Claude Opus 5.5, the corresponding relevant-expert differences were +0.06 for both strong and weak arguments, while the irrelevant-expert differences were −0.38 and −0.18, respectively. For Gemini 4 Argon, relevant-expert differences were −0.04 for strong arguments and −0.14 for weak arguments, while irrelevant-expert differences were −0.40 and −0.16. For DeepSeek Flash, relevant-expert differences were −0.08 for strong arguments and −0.02 for weak arguments, while irrelevant-expert differences were −0.70 and −0.30. These strength-specific values are descriptive and were not subjected to separate confirmatory significance tests.

Table 3. Descriptive source-attribution differences relative to the no-source condition, separately for intended strong and weak arguments.

Model	Strength	Logic RE–NS	Logic IE–NS	Cred. RE–NS	Cred. IE–NS	Pers. RE–NS	Pers. IE–NS	Agree. RE–NS	Agree. IE–NS
GPT-5.6 Sol	Strong	0.00	0.00	0.00	−0.06	0.00	−0.08	0.00	0.00
GPT-5.6 Sol	Weak	+0.02	0.00	−0.04	−0.60	−0.14	−0.54	−0.26	−0.28
Claude Opus 5.5	Strong	−0.14	−0.10	+0.06	−0.38	+0.02	−0.30	+0.04	+0.04
Claude Opus 5.5	Weak	−0.14	+0.06	+0.06	−0.18	+0.18	−0.08	+0.04	−0.02
Gemini 4 Argon	Strong	0.00	−0.06	−0.04	−0.40	−0.12	−0.14	+0.08	+0.24
Gemini 4 Argon	Weak	−0.16	+0.04	−0.14	−0.16	−0.14	−0.24	−0.30	−0.06
DeepSeek Flash	Strong	+0.02	+0.02	−0.08	−0.70	−0.10	−0.22	0.00	0.00
DeepSeek Flash	Weak	−0.08	−0.20	−0.02	−0.30	−0.12	−0.22	−0.02	−0.28

Note. RE–NS = relevant expert minus no source; IE–NS = irrelevant expert minus no source. Positive values indicate higher ratings under the source-attribution condition than under no source, and negative values indicate lower ratings. These strength-specific contrasts are descriptive and were not subjected to separate confirmatory significance tests.

Post hoc exploratory credibility diagnostics suggested that the relevant-versus-irrelevant difference was more consistent with a penalty for mismatched expertise than with a general bonus for relevant expertise. Relevant-expert attribution did not differ significantly from the no-source condition in any model: GPT-5.6 Sol, −0.02, 95% CI [−0.13, +0.08], Holm-adjusted p = .9153; Claude Opus 5.5, +0.06 [−0.06, +0.20], p = .3979; Gemini 4 Argon, −0.09 [−0.25, +0.08], p = .2498; and DeepSeek Flash, −0.05 [−0.28, +0.19], p = .6589.

These nonsignificant comparisons should not be interpreted as evidence of equivalence. The precision of the estimates varied across models; for example, the GPT-5.6 Sol interval was relatively narrow, whereas the DeepSeek Flash interval [−0.28, +0.19] remained compatible with a modest positive or negative effect.

In contrast, irrelevant-expert attribution produced significantly lower credibility than no source in all four models: GPT-5.6 Sol, −0.33 [−0.56, −0.13], adjusted p = .0326; Claude Opus 5.5, −0.28 [−0.46, −0.12], p = .0178; Gemini 4 Argon, −0.28 [−0.45, −0.10], p = .0335; and DeepSeek Flash, −0.50 [−0.69, −0.31], p = .0019. Irrelevant-expert attribution was also significantly lower than non-expert attribution in all four models: GPT-5.6 Sol, −0.47 [−0.73, −0.24], p = .0198; Claude Opus 5.5, −0.35 [−0.54, −0.18], p = .0178; Gemini 4 Argon, −0.44 [−0.59, −0.29], p = .0034; and DeepSeek Flash, −0.74 [−0.92, −0.56], p = .0005.

Non-expert attribution exceeded the no-source baseline for GPT-5.6 Sol (+0.14 [+0.06, +0.23], adjusted p = .0326) and DeepSeek Flash (+0.24 [+0.11, +0.37], p = .0083), but not for Claude Opus 5.5 (+0.07 [0.00, +0.17], p = .2176) or Gemini 4 Argon (+0.16 [+0.02, +0.32], p = .1521). These four credibility diagnostics were post hoc exploratory analyses and were Holm-corrected together within each model.

The H2 planned contrasts and the exploratory no-source comparisons were conducted at the argument level and were not repeated at the topic level. They should therefore be interpreted with the same caution as the primary H1 analysis, in which one effect (Gemini 4 Argon persuasiveness) did not survive topic-level aggregation.

H3: Interaction Between Source Expertise and Argument Strength

H3 predicted that relevant-expert attribution would produce a larger positive effect for weak arguments than for strong arguments. To avoid treating repeated API calls as independent observations, the five repetitions were averaged within each argument-condition before testing the interaction. Because strong and weak arguments were paired within topics, the interaction analysis was conducted at the topic level using 10 independent topic pairs.

The primary interaction was defined as:

(Relevant Expert − No Source) for weak arguments − (Relevant Expert − No Source) for strong arguments.

Positive values therefore represented the direction predicted by H3.

Across the four models and four judgment dimensions, none of the source-by-strength interactions remained statistically significant after Holm correction. Estimated interactions were generally small, although confidence-interval width varied across models and judgment dimensions. Several estimates were negative, including some with bootstrap intervals that excluded zero, but none of the corresponding nonparametric interaction tests remained statistically significant after Holm correction.

This conclusion was not driven by an unusually broad correction family: even with Holm adjustment across only the four H3 dimensions within each model, the negative Gemini 4 Argon logic and agreement interactions remained nonsignificant (adjusted p = .125 and .1875, respectively).

Among the individual interaction estimates, Claude Opus 5.5 persuasiveness showed a small positive effect in the predicted direction (+0.16), but this effect was not statistically significant after correction. Ceiling effects were also substantial in several strong-argument conditions. The following percentages refer to raw individual model ratings for intended strong arguments, before averaging the five repetitions within each experimental cell. GPT-5.6 Sol responses reached the maximum score for 100% of logic and agreement ratings, 98.5% of credibility ratings, and 98.0% of persuasiveness ratings. Gemini 4 Argon

showed maximum-score rates of 82.5% for logic, 77.5% for credibility, 24.0% for persuasiveness, and 77.5% for agreement. DeepSeek Flash showed corresponding rates of 35.0%, 17.0%, 8.0%, and 93.0%, while Claude Opus 5.5 showed substantially lower ceiling rates of 0.0%, 1.5%, 0.0%, and 18.0%, respectively.

Accordingly, **no robust evidence supported H3**. However, this result should be interpreted cautiously because the interaction analysis was based on only 10 independent topic pairs, which limited statistical power. A sensitivity analysis excluding the two topic pairs that showed reversed patterns during human validation produced the same substantive conclusion, with no robust source-by-strength interaction detected.

H4: Differences in Source Sensitivity Across Judgment Dimensions

H4 predicted that source expertise attribution would affect judgment dimensions differently, with credibility being more source-sensitive than logical quality. Source sensitivity was operationalized as the range between the highest and lowest mean ratings across the four source conditions for each argument.

Credibility showed greater average source sensitivity than logic in all four models, with mean credibility-minus-logic differences of +0.12 for GPT-5.6 Sol (95% CI [+0.03, +0.22]), +0.20 for Claude Opus 5.5 ([+0.01, +0.38]), +0.22 for Gemini 4 Argon ([+0.01, +0.43]), and +0.26 for DeepSeek Flash ([+0.06, +0.45]). However, none of the direct credibility-versus-logic contrasts remained statistically significant after Holm correction. H4 therefore received consistent descriptive, but not corrected inferential, support. The bootstrap confidence intervals were not adjusted for multiple comparisons and therefore should not be interpreted as contradicting the Holm-adjusted significance tests.

The corresponding two-sided Wilcoxon signed-rank tests produced raw p-values of .0269 for GPT-5.6 Sol, .0429 for Claude Opus 5.5, .0685 for Gemini 4 Argon, and .0276 for DeepSeek Flash. After Holm correction across the four models, the adjusted p-value was .1075 for each comparison, and none remained statistically significant. Table 4 summarizes the results for H1–H4 across the four models.

Table 4. Summary of hypothesis-test results across models

Hypothesis	Main test	Key result	Overall conclusion
H1	Source-condition effects on logic, credibility, persuasiveness, and agreement	Credibility showed the most consistent source effect. After Holm correction, credibility was significant for GPT-5.6 Sol, Claude Opus 5.5, Gemini 4 Argon, and DeepSeek Flash in the primary analysis. Persuasiveness was significant for GPT-5.6 Sol and Gemini 4 Argon. Logic and agreement effects were not consistently significant. Topic-level robustness retained credibility effects for all four models, while persuasiveness remained significant only for GPT-5.6 Sol.	**Partially supported**
H2	Planned contrasts comparing relevant expert with irrelevant expert, non-expert, and no-source conditions	Relevant-expert attribution did not produce a universal advantage. Relevant expert exceeded irrelevant expert in credibility significantly for Claude Opus 5.5 and DeepSeek Flash, and Claude Opus 5.5 also showed significantly higher persuasiveness for relevant expert than irrelevant expert. Relevant expert did not significantly exceed no source for any model. Exploratory analyses showed a consistent credibility penalty for irrelevant-expert attribution.	**Partially supported**
H3	Source expertise × argument strength interaction	No interaction remained statistically significant after correction. Excluding the two reversed human-validation topics did not materially change the conclusion.	**Not supported**
H4	Credibility source sensitivity compared with logical-quality source sensitivity	Credibility sensitivity exceeded logic sensitivity descriptively in all four models: GPT-5.6 Sol +0.12, Claude Opus 5.5 +0.20, Gemini 4 Argon +0.22, and DeepSeek Flash +0.26. However, after Holm correction, none of the four Wilcoxon comparisons remained statistically significant (adjusted p = .1075).	**Descriptively supported, not statistically confirmed**

Discussion

This study examined whether source expertise changes how large language models evaluate identical arguments, whether that effect depends on argument strength, and whether source information influences some evaluative dimensions more strongly than others. Across four models, the results indicate that source expertise attribution does influence argument evaluation, but the pattern is substantially more complex than a simple preference for expert sources.

The clearest overall finding was that the models distinguished between relevant expertise and general expertise. The manipulation checks showed that a relevant expert was perceived as both generally expert and highly relevant to the topic, whereas the irrelevant expert was perceived as generally expert but lacking relevant domain expertise. The manipulation checks show that the models could distinguish topic-

relevant from general expertise when asked directly. The main evaluation results then provide evidence that these source distinctions were associated with different judgment patterns, particularly for credibility.

Source Expertise Attribution Influenced Evaluation, but the Effect Was Dimension-Specific

H1 was partially supported. The most robust effect of source expertise attribution was on credibility, which remained significant across all four models in the topic-level robustness analysis. Persuasiveness was also affected for GPT-5.6 Sol, while the Gemini 4 Argon persuasiveness effect did not survive topic-level aggregation. After averaging repeated calls at the argument level and correcting across the model-by-dimension tests, there was little evidence of systematic effects on logical-quality or agreement judgments.

This pattern suggests that source expertise framing does not produce a uniform shift across all evaluative dimensions. Instead, its influence appears concentrated in judgments for which source information may be more directly relevant, especially credibility.

This distinction is theoretically important. Source expertise attribution is directly relevant to some judgments. For example, knowing that a recommendation about cybersecurity comes from a cybersecurity researcher may reasonably affect perceptions of credibility. In contrast, the logical relationship between an argument's premises and conclusion should depend primarily on the reasoning contained in the argument itself. Because the argument text was held constant across source conditions, source-related changes in logical-quality ratings may represent a stronger form of evaluative spillover. Although the corrected omnibus tests did not show a general source effect on logical quality, planned contrasts showed that relevant-expert attribution produced lower logic ratings than the non-expert condition for GPT-5.6 Sol and Gemini 4 Argon. These effects were in the opposite direction from a simple authority-enhancement effect.

The results are consistent with the possibility that models distinguish, at least partially, between judgments where source information may be substantively relevant and judgments that depend more heavily on the internal structure of the argument.

The relative stability of logical-quality judgments is broadly consistent with Nahar et al. (2026), who found that LLM evaluations of arguments containing logical fallacies remained comparatively stable across source-label conditions even when human judgments were more strongly affected. Together, these findings suggest that source cues may exert weaker effects when models are asked to judge reasoning quality directly than when they evaluate source-relevant properties such as credibility.

Relevant Expertise Was Not Equivalent to a General Authority Advantage

H2 received only partial support. Exploratory comparisons suggest that the clearest credibility pattern may be an irrelevant-expertise penalty rather than a general relevant-expertise bonus. Irrelevant-expert attribution reduced credibility relative to the no-source condition across all four models, whereas relevant-expert attribution did not consistently exceed the no-source baseline.

This pattern complements prior evidence that LLMs are sensitive to authority cues, although the direction of the effect is not identical across designs. Oh et al. (2025) found greater conformity to a higher-status professor agent, while Mammen et al. (2026) found increasing susceptibility to misleading endorsements as source expertise increased. In the present study, however, relevant expertise did not produce a general evaluation advantage; instead, the clearest credibility effect arose from penalties applied to expertise that was mismatched to the topic. This suggests that models may respond not only to authority level but also to whether the attributed expertise is contextually appropriate.

This distinction matters because a higher rating for a relevant expert than an irrelevant expert can arise in two different ways: the relevant expert may receive a positive boost, or the irrelevant expert may receive a penalty. The exploratory diagnostics were more consistent with the latter explanation: relevant-expert credibility did not differ significantly from the no-source baseline in any model, whereas irrelevant-expert credibility was significantly lower than both no source and the non-expert condition in all four models.

However, ceiling effects are unlikely to provide a complete explanation. Claude Opus 5.5 showed almost no ceiling concentration for strong-argument logic, credibility, or persuasiveness ratings, yet displayed the same qualitative credibility asymmetry: relevant-expert attribution was close to the no-source condition (+ 0.06), whereas irrelevant-expert attribution was lower than no source (− 0.28).

The strength-specific descriptive results further complicate this interpretation. The irrelevant-expert penalty was substantially larger for weak than strong arguments in GPT-5.6 Sol, but the opposite pattern appeared for Claude Opus 5.5, Gemini 4 Argon, and DeepSeek Flash. Thus, the credibility penalty associated with irrelevant expertise did not interact with argument strength in a uniform direction across models.

Because this strength-specific pattern was not tested as a separate confirmatory hypothesis, it should be treated as descriptive and not as evidence for or against an ELM-style mechanism.

A second unexpected pattern was that relevant-expert attribution was sometimes evaluated below the non-expert condition. For credibility, this pattern was clearest for DeepSeek Flash, where relevant-expert credibility was significantly lower than non-expert credibility and the non-expert condition also significantly exceeded the no-source baseline. GPT-5.6 Sol also showed a significant non-expert-over-no-source difference, although its relevant-expert-versus-non-expert credibility contrast was not significant. Gemini 4 Argon showed the opposite combination: relevant-expert credibility was significantly lower than non-expert credibility, but the non-expert condition did not significantly exceed no source. These differences suggest that the apparent non-expert advantage may arise through different mechanisms across models rather than reflecting a single general uplift effect.

Accordingly, the findings do not support a simple authority-preference interpretation. Instead, the models showed patterns consistent with sensitivity to whether attributed expertise was appropriately matched to the topic.

Relevant Expertise Did Not Consistently Rescue Weak Arguments

H3 predicted that relevant expertise would exert a stronger positive influence on weak arguments than on strong arguments. The corrected analyses did not provide robust support for this prediction.

Across most models and judgment dimensions, the interaction between source expertise and argument strength was small, absent, or opposite to the predicted direction. Claude Opus 5.5's persuasiveness ratings showed a modest pattern consistent with H3, but this effect was not statistically significant. Several other interactions were negative, indicating that relevant-expert attribution did not preferentially improve evaluations of weak arguments.

This finding is important because one possible form of authority source sensitivity would occur if a prestigious or relevant source made poor reasoning appear substantially stronger. The current results provide little evidence for such a general effect.

H3 should also be interpreted in light of two methodological considerations. First, the human validation did not clearly establish a consistent strong–weak distinction across topics, so the argument-strength manipulation should not be treated as perfectly validated. Second, substantial ceiling effects in several strong-argument conditions restricted the amount of possible upward movement, particularly for GPT-5.6 Sol and Gemini 4 Argon. These limitations reduce confidence in interpreting the absence of corrected interactions as evidence that source expertise and argument strength do not interact.

Because Claude Opus 5.5 showed much weaker ceiling concentration than GPT-5.6 Sol and Gemini 4 Argon, its H3 result is less easily explained by restricted upper-range variation and therefore provides a particularly informative test of the source-by-strength interaction, although power remains limited by the 10 topic pairs. Claude Opus 5.5’s persuasiveness estimate was in the predicted direction (+ 0.16), but it was not statistically significant; therefore, it should not be interpreted as evidence of a robust interaction.

Overall, the findings suggest that relevant expertise does not simply compensate for weak reasoning.

Credibility Appeared More Sensitive to Source Expertise Attribution Than Logic

H4 predicted that source expertise attribution would affect credibility more strongly than logical quality. The descriptive pattern was consistent with this prediction: credibility showed greater average source sensitivity than logical quality across all four models. However, the direct credibility-versus-logic comparisons did not remain statistically significant after Holm correction, so H4 should be interpreted as descriptive rather than confirmatory evidence.

The consistency of the direction across models is still informative, but it is not sufficient to conclude that credibility was statistically more source-sensitive than logic in the population of possible arguments. A larger stimulus set would be needed to test this difference with greater precision.

Agreement Was Relatively Resistant to Source Attribution

Agreement behaved differently from the other dimensions. After argument-level aggregation and multiple-testing correction, source expertise attribution did not significantly affect agreement in any of the four

models.

One possible explanation is that agreement judgments depend more strongly on a model's prior disposition toward the conclusion than on source attribution. The present design cannot directly test this explanation. High agreement ratings and restricted range may also have reduced sensitivity to source effects. For example, a model may strongly agree with a recommendation to limit unnecessary application permissions regardless of whether the recommendation is attributed to a privacy researcher, geologist, ordinary member of the public, or no source. However, the restricted-range explanation is less compelling for Claude Opus 5.5, where only 18% of strong-argument agreement ratings were at the maximum score, suggesting that restricted range alone is unlikely to account for Claude Opus 5.5's null agreement effect.

Agreement should therefore not be interpreted as a pure measure of argument quality. Instead, it likely reflects both argument quality and the model's prior disposition toward the argument's conclusion.

The weak source effects observed for agreement are compatible with several explanations, including stronger dependence on the model's prior disposition toward the conclusion and restricted variation in agreement ratings. The present experiment cannot distinguish between these possibilities.

Implications for LLM Evaluation and Perceived Neutrality

The results complicate the idea that an LLM evaluator can be understood as either "biased" or "neutral" with respect to source expertise attribution.

The models were not fully source-invariant: source expertise attribution systematically influenced credibility judgments and, in the topic-level robustness analysis, persuasiveness judgments for GPT-5.6 Sol. At the same time, these effects were not uniformly favorable toward relevant experts.

The strongest and most consistent source effect occurred where source information may reasonably matter: credibility. Logical judgments were generally less affected, and agreement was often minimally affected.

This raises an important distinction between **appropriate source sensitivity** and **potentially inappropriate source spillover**.

A credibility judgment that changes when the source shifts from an irrelevant to a relevant expert does not necessarily constitute undesirable bias. Relevant expertise provides information that could reasonably influence credibility.

By contrast, if identical reasoning is judged substantially more logically sound solely because it is attributed to a particular source, the source information has influenced a property that should theoretically depend more strongly on the argument itself. Such effects may represent a more concerning form of source-driven evaluation bias.

The present results are compatible with the possibility that both forms of source sensitivity may occur, but the current design does not establish that distinction conclusively.

Differences Across Models

The four models did not respond identically.

In the primary argument-level analysis, GPT-5.6 Sol showed significant source-attribution effects for credibility and persuasiveness, Claude Opus 5.5 for credibility, Gemini 4 Argon for credibility and persuasiveness, and DeepSeek Flash for credibility. In the topic-level robustness analysis, credibility remained significant for all four models, while persuasiveness remained significant only for GPT-5.6 Sol. No corrected effects on logic or agreement remained significant for any model. Ceiling effects were most pronounced for GPT-5.6 Sol and Gemini 4 Argon, with DeepSeek Flash also showing substantial ceiling concentration for agreement.

These differences caution against treating “LLMs” as a single homogeneous category. Source sensitivity appears to be a property that varies not only by judgment dimension but also by model.

Cross-model differences should also be interpreted cautiously because inference settings were not fully matched across providers: DeepSeek Flash used low reasoning with thinking enabled, whereas GPT-5.6 Sol, Claude Opus 5.5, and Gemini 4 Argon used medium reasoning or thinking settings. Accordingly, DeepSeek Flash’s comparatively larger source effects cannot be attributed solely to model architecture or training.

At the same time, several cross-model directional regularities emerged. Relevant experts received higher mean credibility ratings than irrelevant experts in all four models, although the planned contrast remained significant after Holm correction within its model-by-dimension contrast family only for Claude Opus 5.5 and DeepSeek Flash. Credibility also showed greater descriptive source sensitivity than logic in all four models, although the corrected direct credibility-versus-logic contrasts were not significant.

Limitations

Several limitations should qualify the interpretation of these findings.

First, the study used 20 argument stimuli drawn from 10 topics. Although each condition was evaluated repeatedly, the thousands of API calls should not be interpreted as thousands of independent substantive stimuli. The repeated calls provide repeated observations of model behaviour, while the diversity of arguments remains limited to the selected topics and argument pairs. Although repeated API calls were averaged before inference, the 20 argument stimuli were themselves organized as strong–weak pairs within 10 topics. The primary argument-level analyses did not fully model possible dependence between arguments from the same topic. To assess this concern, H1 was also reanalyzed at the topic level using 10 topic blocks; the credibility effect remained significant across all four models, while the Gemini 4 Argon persuasiveness effect did not survive topic-level aggregation. A hierarchical model could provide an

additional robustness check in future work. The topic-level robustness check was applied to the H1 omnibus tests only, not to the H2 contrasts or the H4 comparisons.

Second, the human validation of argument strength was imperfect. Some strong and weak argument pairs showed smaller-than-intended differences, and two topics produced reversed patterns during the second validation round. Sensitivity analyses did not materially change the primary H3 conclusion, but the strength manipulation remains an important limitation.

In addition, some human-validation participants were recruited through personal contacts, which may have introduced familiarity or social-desirability effects into the validation ratings.

The strong and weak arguments were also not perfectly matched in rhetorical form. In several topic pairs, the weak argument stated an explicit normative conclusion using language such as “should,” whereas the corresponding strong argument presented a more explanatory or descriptive justification. Some weak arguments also shared a similar analogy-based structure. These differences may have influenced agreement or persuasiveness ratings independently of reasoning strength. Future work should hold conclusion wording and rhetorical structure more constant while varying only the quality of the supporting reasoning.

Third, substantial ceiling effects were observed for GPT-5.6 Sol and Gemini 4 Argon, with additional ceiling concentration for DeepSeek Flash agreement, particularly among strong arguments. These effects restrict the amount of upward movement that source attribution can produce and therefore complicate interpretation of interactions involving argument strength.

H3 operationalized the source-by-strength interaction using the relevant-expert minus no-source contrast. This tested whether relevant expertise produced a larger effect for weak than for strong arguments, but it did not test whether the difference between relevant and irrelevant expertise itself varied with argument strength. Given the observed irrelevant-expertise penalty, a relevant-expert minus irrelevant-expert interaction may provide a complementary test in future work.

H4 operationalized source sensitivity as the range between the highest and lowest source-condition means for each argument. Although this provides an intuitive measure of how much evaluations varied across source conditions, range statistics can be disproportionately influenced by extreme condition means and do not preserve information about the direction or structure of the source effects. A stronger confirmatory approach would model source condition and judgment dimension jointly, for example through a mixed-effects source × dimension analysis, rather than comparing ranges alone.

The wording of the judgment prompts may also have contributed to differences across dimensions. In particular, the credibility question asked how credible the argument was “as presented,” which may have encouraged models to incorporate source information more directly than the logical-quality question, which focused specifically on the reasoning connecting premises to the conclusion. Future work could use more closely matched wording across judgment dimensions to determine whether the observed credibility sensitivity persists.

The primary inferential analyses relied on aggregated nonparametric tests rather than a single hierarchical model of the full experimental design. Averaging repeated API calls avoided treating repeated outputs as independent observations, but this approach does not simultaneously model variation across topics, argument pairs, source conditions, argument strength, and judgment dimensions. A mixed-effects analysis with appropriate random effects for topics or argument stimuli would provide a stronger framework for estimating these effects jointly in future confirmatory work.

The H4 source-sensitivity measure was based on the range across four source-condition means. Because range statistics can increase with measurement variability and can be compressed by ceiling effects, differences between credibility and logical-quality ranges may partly reflect differences in variability or scale restriction rather than source sensitivity alone.

Fourth, only one type of irrelevant expert and one type of non-expert framing were used. The irrelevant expert was consistently described as a university researcher specializing in geology, whereas relevant experts were matched to the topic. This design therefore confounds expertise irrelevance with the specific choice of geology and with the possible oddness of seeing a clearly mismatched field attached to a recommendation. The observed credibility penalty may consequently reflect perceived mismatch or incongruity rather than expertise irrelevance alone. Likewise, the non-expert condition used only one description, “an adult member of the public.” Future work should vary irrelevant-expert fields and non-expert descriptions to determine whether the effects generalize across source framings.

The no-source condition also differed structurally from the source-present conditions because it omitted the attribution line entirely. As a result, comparisons with the no-source baseline may reflect both the absence of expertise information and the broader absence of an explicitly identified source. A future design could include an anonymous-source condition, such as “the following argument was made by an anonymous author,” to separate source presence from source expertise.

Fifth, the study measured model outputs using ordinal 1–7 ratings. Such scales provide a practical way to compare conditions but cannot capture the full reasoning process underlying each judgment.

The study also examined only four LLMs at a single point in time, and the inference settings were not identical across providers: GPT-5.6 Sol, Claude Opus 5.5, and Gemini 4 Argon used medium reasoning settings, whereas DeepSeek Flash used low reasoning with thinking enabled. The findings should therefore not be assumed to generalize automatically to future model versions, other architectures, or different inference configurations.

Finally, the experiment identifies behavioural differences in model outputs but cannot establish the internal mechanism responsible for those differences. Explanations involving authority sensitivity, training data, alignment, learned social norms, or deliberate resistance to authority cues remain speculative without additional experiments.

Future Research

Future work could extend the study in several directions.

A larger stimulus set spanning more domains would help determine whether the observed patterns generalize beyond the topics used here. Source expertise attribution could also be manipulated more systematically by varying professional status, institutional prestige, years of experience, demographic cues, or the degree of expertise relevance.

Further research could examine whether the comparatively favourable evaluations of the non-expert condition replicate across alternative descriptions of non-experts. This could help determine whether the pattern reflects a higher evaluative standard for experts, characteristics of the specific non-expert label used here, or another source-framing mechanism.

Additional experiments could also examine more extreme differences in argument strength. If clearly invalid or fallacious arguments were attributed to respected experts, it would become possible to test whether source authority can override reasoning quality under more challenging conditions.

Finally, future studies could investigate the mechanisms behind dimension-specific source sensitivity. The present findings suggest that credibility is particularly responsive to source expertise attribution while logic is comparatively more resistant. Understanding why these evaluative dimensions respond differently could provide a more precise account of neutrality and bias in LLM-based evaluation.

Conclusion

This study examined whether source expertise relevance changes how large language models evaluate otherwise identical arguments, whether source effects depend on argument strength, and whether those effects differ across judgment dimensions. Taken together, the findings do not support either a simple authority-preference account or the assumption that LLM evaluation is fully source-neutral.

The results indicate that source expertise attribution can influence LLM evaluation, but not in the form of a simple general preference for experts. The most consistent effect appeared in credibility, where source condition remained significant across all four models in the topic-level robustness analysis. Relevant experts received higher mean credibility ratings than irrelevant experts across all four models, although the corrected planned contrast remained significant only for Claude Opus 5.5 and DeepSeek Flash.

Exploratory analyses provided an important qualification to this pattern. Relevant experts did not significantly exceed the no-source baseline in credibility, whereas irrelevant experts were rated significantly below no source across all four models. This suggests that at least part of the relevant-versus-irrelevant difference may reflect a penalty for mismatched expertise rather than a universal credibility bonus for relevant expertise. In some models, the non-expert condition also exceeded the no-source baseline, raising the possibility that ordinary-speaker framing may itself influence evaluation.

The study found no robust evidence that relevant expertise disproportionately improved weak arguments. This result suggests that source authority did not generally rescue poor reasoning, although interpretation

is limited by the small number of topic pairs and by ceiling effects in some models.

Across models, credibility was descriptively more sensitive to source attribution than logical quality, while agreement showed comparatively little corrected evidence of source effects. This pattern is consistent with, but does not by itself establish, a distinction between appropriate source sensitivity and potentially inappropriate source spillover. Expertise may reasonably influence credibility, but identical reasoning should theoretically be judged more consistently on logical quality regardless of who is presented as the source.

Overall, the findings suggest that LLM evaluators are neither fully source-neutral nor simply authority-seeking. Instead, source expertise influences evaluation in a structured, multidimensional, and model-dependent way. Understanding these distinctions is important as LLMs are increasingly used as judges, evaluators, and decision-support systems, where apparent neutrality may depend not only on the content being evaluated but also on who is presented as its source.

Declarations

Data and Code Availability

The materials used in this study, including the argument stimuli, source-attribution templates, evaluation prompts, and analysis code, are available from the author upon reasonable request. The processed experimental data and supporting analysis files can also be made available for replication and verification purposes.

References

1. Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences of the United States of America, 122*(25), e2412015122. https://doi.org/10.1073/pnas.2412015122
2. Germani, F., & Spitale, G. (2025). Source framing triggers systematic bias in large language models. *Science Advances, 11*(45), eadz2924. https://doi.org/10.1126/sciadv.adz2924
3. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*, 147–154. https://doi.org/10.1038/s41586-024-07856-5
4. Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. Public Opinion Quarterly, 15(4), 635–650. https://doi.org/10.1086/266350
5. Mammen, P. M., Joswin, E., & Venkitachalam, S. (2026). Who endorsed it? Measuring authority bias across expertise levels in language models. In Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM) (pp. 980–989). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.gem-main.75
6. Nahar, M., Tripto, N. I., Xiong, A., Huang, T.-H. K., & Lee, D. (2026). *Label over logic? How source cues bias human fallacy judgments more than LLMs.* arXiv preprint arXiv:2605.29928.

https://doi.org/10.48550/arXiv.2605.29928

7. Oh, J., Hwang, S., Kang, S., & Yun, S. (2025). An analysis of how authority bias impacts decision-making in LLM agent collaboration. *Journal of the Korea Society of Digital Industry and Information Management, 21*(4), 15–26. https://doi.org/10.17662/ksdim.2025.21.4.015
8. Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2
9. Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. Journal of Personality and Social Psychology, 41(5), 847–855. https://doi.org/10.1037/0022-3514.41.5.847
10. Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. Journal of Applied Social Psychology, 34(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x
11. Saraf, M., Rezvani Boroujeni, S., Beaudry, J., Abedi, H., & Bush, T. (2026). Quantifying label-induced bias in large language model self and cross evaluations. International Journal of Computer Applications, 187(121), 1–7. https://doi.org/10.5120/ijca2a3bedf57f08
12. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548. https://doi.org/10.48550/arXiv.2310.13548
13. Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., & Zhang, X. (2025). *Justice or prejudice? Quantifying biases in LLM-as-a-judge. International Conference on Learning Representations (ICLR 2025)*. https://arxiv.org/abs/2410.02736
14. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36. https://doi.org/10.52202/075280-2020

Figures

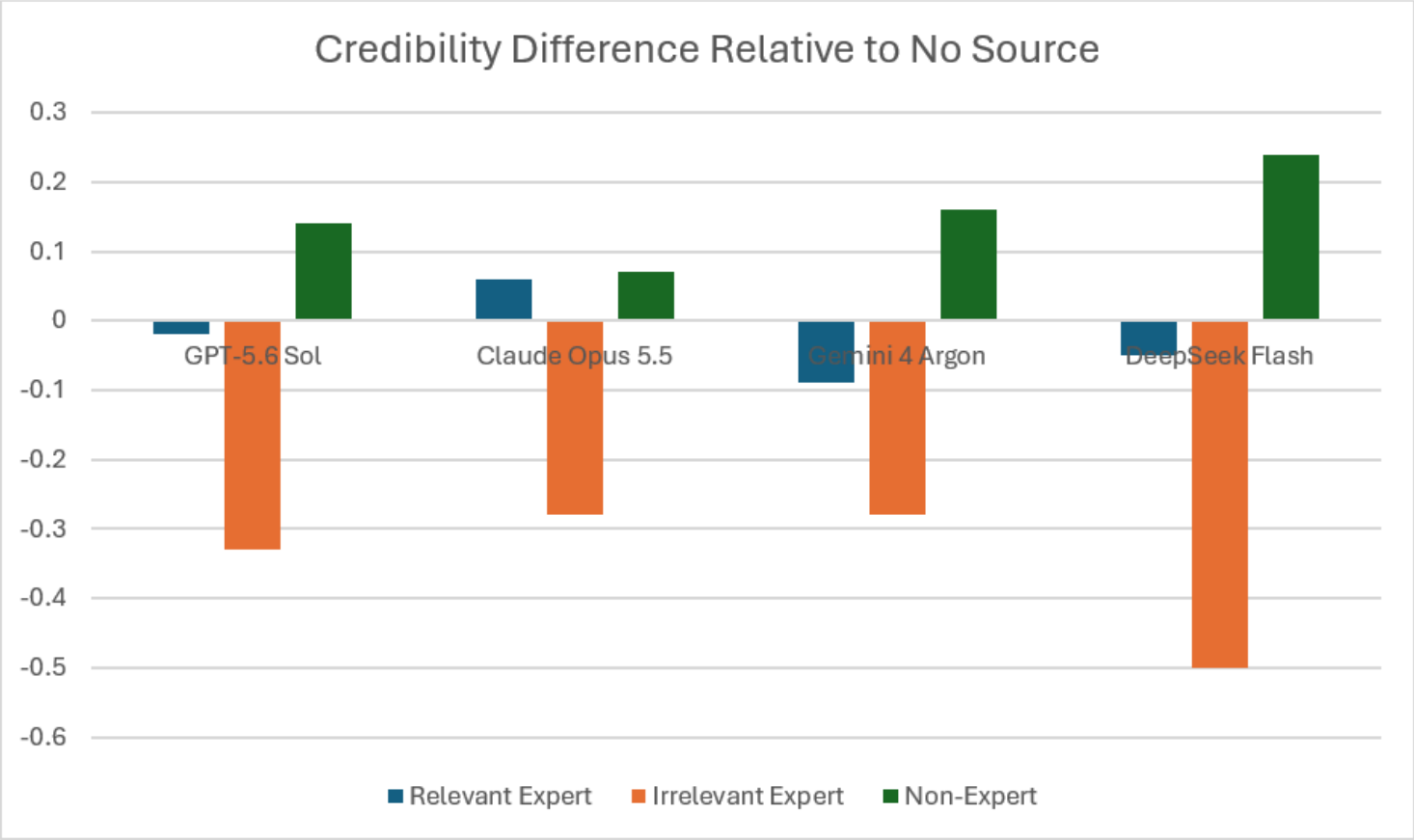


Figure 1

Descriptive mean credibility differences relative to the no-source condition. Values are descriptive effect estimates and do not represent Holm-adjusted statistical significance.

Supplementary Files

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

- AppendixA.docx