Argument Evaluation Strategies in Human and Machine Reasoning: LLMs Resemble Sound Deliberate (But Not Intuitive) Thinkers

Nicolas Beauvais (nicolas1beauvais@gmail.com)[1] & Wim De Neys[1]
[1]LaPsyDÉ, CNRS, Université Paris Cité, F-75005 Paris, France

Abstract

Humans routinely evaluate arguments to decide which reasons warrant belief or action. Large language models (LLMs) are increasingly used as evaluators, yet little is known about how their argument evaluation compares to human evaluation. In this study we examined how humans and LLMs evaluate arguments supporting answers to classic reasoning problems. Human participants rated arguments either under time pressure and cognitive load or without constraints, allowing us to situate LLM evaluations relative to intuitive and deliberate human judgments. Arguments varied in response correctness and justification type. Modeling argument ratings as a function of argument validity, surface explicitness, and alignment with the evaluator's own response revealed that intuitive human evaluations were dominated by belief-consistency bias, whereas deliberate evaluations –especially among correct responders– prioritized justification validity. In contrast, LLMs generally showed stable, validity-centered evaluative strategies, closely resembling deliberate human evaluations from correct responders and diverging from intuitive human judgments.

Keywords: LLM; argument evaluation; reasoning; dual process; biases

Introduction

Humans routinely reason to evaluate arguments: from judging information and advice to assessing evidence and justifications, people constantly decide which reasons are strong enough to warrant belief or action. In recent years, advances in artificial intelligence and natural language processing have given rise to new "reasoning" agents: large language models (LLMs). Trained on massive text datasets, LLMs display impressive performance across a wide range of tasks, including natural language understanding, explanation generation, and classic reasoning benchmarks such as cognitive reflection problems (Laskar et al., 2023; Yax et al., 2024). These models are also increasingly used as evaluators in various tasks (Beucler et al., 2026; DiStefano et al., 2025; Huang et al., 2025; Le Mens et al., 2023; Ma et al., 2025; Ornstein et al., 2025). However, growing evidence indicates that they are not neutral or uniformly reliable evaluators. Recent studies show that they can exhibit systematic biases, inconsistencies, and sensitivity to superficial features when judging arguments or explanations, sometimes reproducing human-like heuristics and errors (Loru et al., 2025; Motoki et al., 2025; Stureborg et al., 2024; Wang et al., 2023; Yax et al., 2024). LLMs are often considered black boxes: we often focus on their outputs, without necessarily understanding the underlying mechanisms (Lipton, 2018; Samek et al., 2017). A central question is then to better understand the processes underlying LLM argument evaluation. What criteria do LLMs use to classify an argument as weaker or stronger? And how do their evaluative strategies compare to those of human reasoners? While most comparisons between humans and LLMs have focused on response production –whether models reach the same conclusions as humans– little work focused on evaluative strategies, that is, how different cues are weighted when judging arguments.

Human reasoning itself provides a complex benchmark. Decades of research have shown that human judgments are frequently undermined by biases. Traditional dual-process theories of reasoning attribute such biases to the tendency to rely on fast, intuitive processes rather than slower, more effortful deliberation (Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011). According to these accounts, intuitive processes (often referred to as Type 1) are rapid and effortless but prone to heuristic shortcuts that can lead to errors, whereas deliberative processes (Type 2) are slower and more cognitively demanding, but also more likely to yield normatively correct responses. Take for instance the classic bat-and-ball problem: "*A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?*" (Frederick, 2005). Most people intuitively respond "10 cents" even though the correct answer is 5 cents. According to traditional dual-process accounts, arriving at the correct solution requires deliberate reasoning to override a compelling but incorrect intuitive response (Evans & Stanovich, 2013; Kahneman, 2011). On this view, biased responses arise from intuition, whereas sound reasoning depends on deliberation correcting misleading heuristics.

Recent work, however, suggests a more nuanced picture. Studies using two-response paradigms –in which participants first give an initial, intuitive response under strict constraints and then a final, deliberate response– indicate that correct deliberative answers are often preceded by already correct intuitive responses (Bago & De Neys, 2017, 2019; Beucler et al., 2025; Raoelison et al., 2020; Thompson et al., 2011; Voudouri et al., 2024). These results suggest that intuition is not always misleading and can be normatively sound, and that correct reasoning does not always require corrective deliberation. Yet this body of work has mostly focused on problem solving, with far less known about how intuitive and deliberate processes shape argument evaluation.

Although the close link between reasoning and argumentation has long been emphasized (Hornikx & Hahn, 2012; Mercier & Sperber, 2011, 2017), argument evaluation has received relatively little attention within the dual-process

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

framework. Research in informal argumentation suggests that people possess at least a basic capacity to distinguish stronger from weaker arguments (Corner & Hahn, 2009; Deans-Browne & Singmann, 2026; Hoeken et al., 2014; Svedholm-Häkkinen et al., 2024), but also that argument evaluation is prone to bias. For instance, extensive evidence shows that people tend to accept at face value arguments that align with their opinions but scrutinize opposing ones (Baron, 1995; Deans-Browne & Singmann, 2026; Lord et al., 1979; Stanovich et al., 2013; Taber & Lodge, 2006). Importantly, relatively little is known about how intuitive and deliberate processes differentially contribute to argument evaluation, and existing evidence is mixed. While some studies suggest that engaging in deliberation can improve evaluative discernment (Bago et al., 2020; Beauvais & De Neys, 2026), others report mixed effects (Hornikx et al., 2022), or little to no benefit (Svedholm-Häkkinen & Kiikeri, 2022; Svedholm-Häkkinen & Hietanen, 2024). Moreover, some evidence indicates that deliberation may even reinforce reliance on prior beliefs during evaluation (Bago et al., 2023; Caddick & Feist, 2022).

Overall, the relative roles of intuition and deliberation in evaluating arguments –particularly in distinguishing strong from weak arguments and in shaping susceptibility to bias– remain poorly understood. Just as with LLMs, understanding how different features of arguments (such as logical validity, surface structure, or consistency with prior beliefs) are weighted under intuitive versus deliberate evaluation is crucial for a better characterization of reasoning processes' role in argumentation.

In the present work, we build on recent studies of human argument evaluation in classic reasoning tasks (Beauvais & De Neys, 2026) and extend this approach to a direct comparison between humans and LLMs. After solving bat-and-ball-type reasoning problems, human participants and several contemporary LLMs rated the strength of arguments supporting different answers to these problems. Human evaluations were made either under time constraint and cognitive load (fast, intuitive condition) or without constraints (slow, deliberative condition), allowing us to characterize human intuitive and deliberate evaluative strategies and to situate LLM evaluations relative to these profiles. Arguments systematically varied along three dimensions: (i) the validity of the justification (i.e., whether it included a valid calculation), (ii) the surface explicitness of the justification (whether it included an explicit calculation, regardless of its validity), and (iii) alignment between the argument's conclusion and the evaluator's own response. We modeled argument ratings as a function of these evaluative cues and compared the resulting profiles across agents and conditions, taking into account each agents' own response to the reasoning problems. By clarifying how evaluative strategies differ across intuitive and deliberate human judgments, and by situating contemporary LLMs within this space, the present work advances our understanding of both human reasoning and the extent to which LLMs approximate –or diverge from– human evaluative processes.

Methods

Human studies

Pre-registration and data availability Human data were taken from the studies of Beauvais and De Neys (2026), to which we refer for a full description of materials and procedures. These studies were preregistered on the Open Science Framework. The preregistrations, data, and materials are available at https://tinyurl.com/4mzfahjx.

Participants Participants were recruited via Prolific Academic (www.prolific.ac), provided informed consent before participation, and were compensated according to Prolific Academic's guidelines. Only native English speakers from the United Kingdom, United States, Ireland, Australia, Canada, or New Zealand were included (Study 1: N = 102, M_{age} = 39.25 years, SD = 13.26, 49% female, 75.5% had a university degree ; Study 2: N = 258, M_{age} = 37.52 years, SD = 12.92, 49.6% female, 64.7% had a university degree ; Study 3: N = 261, M_{age} = 37.93 years, SD = 12.8, 49.4% female, 62.8% had a university degree).

Material

Problems Participants were presented with six problems, adapted from Frederick's (2005) classical "bat-and-ball" problem. To minimize potential familiarity bias (Chandler et al., 2014; Haigh, 2016; Stieger & Reips, 2016), structurally similar but content-altered versions of the original problem were used (e.g., Bago & De Neys, 2019; Beauvais et al., 2025). Each problem presented the total quantity of two items and their relationship. Participants selected one of four answer options, as in the following example:

In a building residents have 370 dogs and cats in total.
There are 300 more dogs than cats.
How many cats are there?

- *7*
- *35*
- *70*
- *105*

Answer options always included the correct answer (i.e., "5 cents" for the classic bat-and-ball problem, "35" in the previous example), the intuitive "heuristic" response ("10 cents" in the classic problem, "70" in the example), and two foil choices which were always the sum of the correct and heuristics options ("15 cents" in the original bat-and-ball units, "105" in the example) and their second greatest common divider ("1 cent" in original units, "7" in the example). The order of the four response options was randomized for each problem. Half of the problems (three out of six) were "conflict" problems, in which the intuitive heuristic response conflicted with the correct answer. The remaining three problems were control "no-conflict" problems, in which the logically correct answer and heuristic reasoning aligned. This alignment was achieved by omitting the critical relational statement "more than" (De Neys et al., 2013), as in the following example:

In a building, residents have 370 dogs and cats in total.

There are 300 dogs.
How many cats are there?

These no-conflict trials allowed to verify whether participants remained attentive throughout the study, as near-perfect accuracy would be expected without random guessing.

Arguments After responding to each problem, participants evaluated seven arguments supporting possible answers. Arguments varied along two dimensions: the proposed answer (correct, heuristic incorrect, or non-heuristic incorrect) and the type of justification. Justifications were either implicit (appealing to intuition) or explicit, involving a calculation. Explicit justifications were either valid, providing a normatively correct rationale for the proposed answer, or invalid, involving a syntactically plausible but irrelevant calculation. A non-heuristic incorrect "control" argument was always accompanied by an explicit invalid justification. Justifications were based on the type of justification generated by participants in Bago and De Neys (2019) and Beauvais et al. (2025), and were controlled in length and phrasing. Argument types and examples are summarized in Table 1; full materials are available on the OSF page of Beauvais & De Neys (2026).

In no-conflict problems, the justifications of correct arguments corresponded to the justifications of incorrect heuristic arguments in conflict problems. Correspondingly, the justifications of incorrect arguments in no-conflict problems perfectly mirrored the justifications of correct arguments in conflict problems.

Procedure Experiments were conducted online using Qualtrics. Problems were presented sentence by sentence, after which participants selected an answer. Participants then evaluated each of the seven arguments on a 0 ("very weak") to 10 ("very strong") scale. Arguments were presented individually in random order, with the problem text remaining visible. The order of problems was randomized. These design features minimized the risk that ratings would be influenced by sequential or comparative effects. The presence of control no-conflict problems also reduced this risk, as arguments' appropriateness changed with problem status.

In Beauvais and De Neys (2026)'s Study 1, participants could rate all arguments without constraints. Studies 2 and 3 contrasted fast (intuitive) and slow (deliberate) argument evaluation using a between-subjects design. In the present paper, we combined the data from the fast and slow conditions across studies. In the fast conditions, argument evaluations were performed under strict time limits and concurrent cognitive load (e.g., Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019; Thompson et al., 2011), whereas no constraints were imposed in the slow conditions. Cognitive load involved memorizing a briefly presented visual pattern prior to each argument. In Study 2, the pattern was a 3 x 3 grid in which 4 crosses were placed. Study 3 replicated Study 2 with stricter constraints: in Study 3 the pattern was a 4 x 4 grid with 5 crosses. In both studies, after rating the argument participants were presented with four different matrices and asked to identify the pattern presented before the argument (see Bago & De Neys, 2017 for further details). Feedback was provided to indicate whether their choice was correct or not. Time limits were 5 seconds in Study 2 and 4 seconds in Study 3. Importantly, time pressure and cognitive load applied only to argument evaluations, not to problem solving. At the end of each study, participants completed demographic questions (age, gender, education level).

Exclusion criteria In the fast condition of Studies 2 and 3, trials were excluded if participants missed the response deadline (2.99% of trials in Study 2, 5.94% in Study 3) or failed the cognitive load task (15.0% of trials in Study 2, 29.8% in Study 3). No trials were excluded in the slow condition. As preregistered, participants who reported prior familiarity with the bat-and-ball problem and solved it correctly were retained, as analyses with and without these participants yielded similar results.

Table 1: Argument types

Answer Type	Justification Type	Example
Correct (5 cents)	Implicit	"The ball costs 5 cents, because the response that readily comes to mind when considering the problem is 5."
	Explicit Valid	"The ball costs 5 cents, because if the bat costs $1 more and the total is $1.10, the bat must cost $1.05 and the ball 5 cents."
	Explicit Invalid	"The ball costs 5 cents, because 50 cents minus 45 is 5 cents, so the solution must be that the ball costs 5 cents."
Incorrect Heuristic (10 cents)	Implicit	"The ball costs 10 cents, because the response that readily comes to mind when considering the problem is 10."
	Explicit Heuristic	"The ball costs 10 cents, because subtracting $1 to the total of $1.10 amounts to a total of 10 cents."
	Explicit Invalid	"The ball costs 10 cents, because if the bat costs $1, then minus 50 is 50 cents, and 50 cents minus 40 amounts to 10 cents."
Incorrect Control (15 cents)	Explicit Invalid	"The ball costs15 cents, because $1 more minus 90 is 10 cents, and adding 5 amounts to a total of 15 cents."

Large Language Models

Specification and prompting We evaluated a set of large language models (LLMs) on the same reasoning and argument evaluation tasks as human participants from Beauvais and De Neys (2026)'s studies. Models included GPT-4.1, GPT-4o, GPT-4o-mini (queried via OpenAI's API), LLaMA3.3-70B-Instruct, Mistral-Small-24B-Instruct-2501, Mistral-7B-Instruct-v0.3, and Qwen-2.5-7B-Instruct (queried using the Hugging Face Inference Client with the Together provider). These models were selected to span a range of architectures and capacities. To capture variability in model outputs, each model was queried multiple times (20 independent runs per model), with identical prompts but independent sampling. Each run was treated analogously to

an individual participant in the human studies. We used default settings with the temperature of each model set to 1.

Each model was presented with the same six reasoning problems used in the human studies, followed by the same seven arguments per problem. Prompts closely mirrored the human task instructions. Models were first asked to provide a single answer to each reasoning problem by selecting one of the four response options. Using the same wording as in the human studies, they were then asked to evaluate each argument's strength by returning a number on a scale from 0 ("very weak") to 10 ("very strong"), and outputs were required to be numeric. Argument order was randomized within each run. Problem solving and argument evaluation were elicited within a single API call, such that each model's own problem response was retained in the conversation history prior to argument presentation. This ensured that models had access to their own prior answer during evaluation, paralleling the human design.

Data Analyses

Argument ratings were analyzed using regression-based models to estimate the contribution of different evaluative cues. For human participants, we fitted linear mixed-effects models predicting argument ratings from argument validity (i.e., whether the argument included a valid explicit justification), surface explicitness (i.e., whether it included an explicit justification, regardless of its validity), and alignment with the participant's response (i.e., whether the argument's answer matched the participant's own answer on that problem). We fitted models separately for fast and slow evaluations following correct and incorrect responses, in order to compare their respective evaluative strategies with those of the different LLMs. Models included random intercepts for participants.

For LLMs, we estimated cue weights by fitting linear models predicting argument ratings from the same set of cues for each LLM. As with human data, analyses were conducted separately depending on whether the model's own response to the corresponding problem was correct or incorrect. All analyses focused on conflict problems.

Because raw coefficient magnitudes reflect both evaluative strategy and overall sensitivity or response variability, we normalized each coefficient using L2 normalization. For each evaluating agent j (i.e., human fast, human slow, and each LLM) and each evaluative cue k (validity, explicitness, alignment), we normalized the raw coefficient $\beta_{j,k}$ to compute the normalized cue weight coefficient $w_{j,k}$:

$$w_{j,k} = \frac{\beta_{j,k}}{||\beta_j||} \quad (1)$$

$$\text{where } ||\beta_j|| = \sqrt{\sum_{k=1}^{K} \beta_{j,k}^2} \quad (2)$$

This was done separately for evaluations following correct and incorrect responses. Without normalization, raw coefficient magnitudes would conflate relative cue weighting with overall sensitivity and noise, such that agents implementing similar evaluative strategies but differing in overall response variability would appear artificially dissimilar despite identical relative cue use. Normalization isolates the relative weighting of evaluative cues while removing differences in overall coefficient magnitude, enabling strategy-level comparisons across agents and conditions.

Similarity between evaluative strategies was quantified using cosine similarity between normalized cue-weight vectors, which reduced to the dot product of normalized profiles:

$$sim(w_i, w_j) = w_i \cdot w_j \quad (3)$$

Results

Performance on reasoning problems

Figure 1 shows the mean accuracy on conflict problems for human participants and the different LLMs. Consistent with the literature on bat-and-ball problems (Frederick, 2005; Stagnaro et al., 2018), human participants generally struggled to correctly solve conflict problems. Across studies, mean accuracy on these problems was 27.98% (SEM = 2.41). Across models, accuracy on conflict problems varied substantially. Higher-capacity models (GPT-4.1, GPT-4o, Llama70B) nearly always gave the correct response, whereas smaller models (Mistral-7B, Qwen-7B) frequently gave heuristic incorrect responses, as shown in Figure 1.

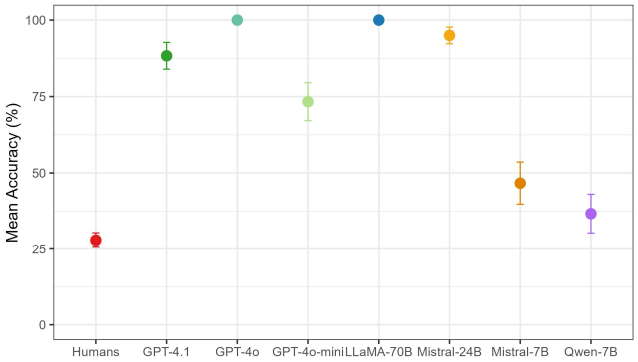

Figure 1: Mean accuracy on conflict problems for human participants and large language models. Error bars are standard errors of the mean.

Argument evaluations

Figure 2A shows the normalized cue-weight profiles for humans (fast and slow) and LLMs evaluations after correct and incorrect responses. In humans, fast evaluations following both incorrect and correct responses were characterized by a strong reliance on alignment and comparatively weak sensitivity to validity. Slow evaluations placed proportionally greater weight on validity and reduced weight on alignment. This was especially the case for slow evaluations that followed correct responses. In these evaluations, validity became the dominant predictor of ratings, exceeding the relative weight of alignment. Surface

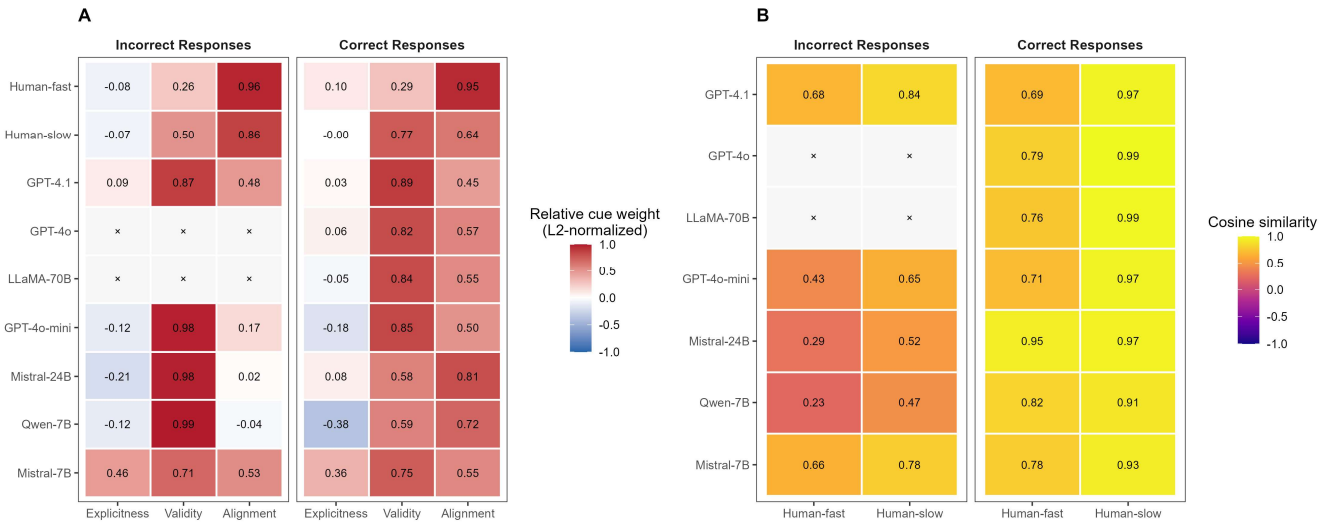

Figure 2: **A** Relative cue-weight profiles for human and LLMs argument evaluations. Tiles represent L2-normalized regression coefficients for argument' surface explicitness, validity, and alignment, shown separately for evaluations following incorrect (left panel) and correct (right panel) responses. **B** Cosine similarity between LLM and human evaluative profiles, separately for evaluations following incorrect (left panel) and correct (right panel) responses. Values closer to 1 indicate greater similarity in the relative weighting of evaluative cues. In both figures, grey cells marked with × indicate model–condition combinations with no available data (i.e., models that never produced incorrect responses).

explicitness had only minimal influence across all human evaluation conditions.

As shown in Figure 2A, most LLMs weighted validity most heavily across conditions, indicating that LLM argument evaluation was primarily driven by the normative quality of justifications. Alignment with the model's own response contributed less to LLM evaluations than to human evaluations, and its influence varied as a function of models' response correctness. Following incorrect responses, all models showed strongly validity-dominated profiles, with moderate-to-low alignment weights –even near zero for some models (GPT-4o-mini, Mistral-24B, Qwen-7B). This sharply contrasts with the alignment-bias observed in human evaluations: contrary to humans, models commonly rated correct-response arguments higher even though they themselves gave the incorrect heuristic response to the problem. Following correct responses, validity also largely dominated cue-weighting profiles, and alignment played a secondary role, except for two lower capacity models (Mistral-24B and Qwen-7B). Surface explicitness played only a minor role in LLM evaluations: across models, normalized weights associated with explicitness were small relative to those associated with validity and alignment, and in several cases near zero, indicating limited sensitivity to the mere presence of explicit computations once validity and conclusion alignment were accounted for. Only one model, Mistral-7B, showed a more balanced profile with a moderate contribution from surface explicitness (though still smaller than other cues).

To quantify the similarity between evaluative strategies across agents and conditions, we computed the cosine similarity between L2-normalized cue-weight vectors. Cosine similarity captures the extent to which two agents rely on evaluative cues in the same relative proportions, independently of overall sensitivity or response variability. Values close to 1 indicate highly similar weighting profiles, values close to 0 indicate little similarity, and negative values indicate opposing cue-weighting patterns. Figure 2B shows the cosine similarity between each LLM and the human fast and slow evaluative profiles, separately for correct and incorrect responses. Across models, LLM evaluative strategies were consistently closer to human slow evaluations than to fast evaluations. After correct responses, all models showed a very high similarity with slow human evaluations (range: 0.91-0.99), and a moderate-to-high similarity with fast human evaluations (range: 0.69-0.95).

After incorrect responses, similarity with slow human evaluations was moderate-to-high (range: 0.47-0.84), and similarity with fast human evaluations was lower but positive, with substantial heterogeneity across models: the few models that put moderate weight on alignment had the highest cosine similarities with fast human evaluations, (range: 0.66-0.68) while other models had low similarities (0.23-0.43) with them, reflecting divergent weighting of evaluative cues. As the previous analysis indicated, this is likely because during fast evaluations, humans relied largely on whether arguments aligned with their own response, especially when they gave the incorrect heuristic answer. On the contrary, models showed no such alignment-bias and relied mostly on justification validity in evaluations, often rating correct-response arguments higher even when it contradicted their response.

Discussion

In this work, we compared how humans and large language models evaluate arguments supporting answers to classic

reasoning problems. By modeling argument ratings as a function of argument validity, surface explicitness, and alignment with the evaluator's own response, we were able to characterize evaluative strategies rather than mere rating outcomes, and compare these across agents and conditions. We discuss below the main findings that emerge from our results, along with their implications for theories of human reasoning and for the interpretation of LLM behavior.

First, human argument evaluation showed a clear dissociation between fast and slow judgments. Fast evaluations –whether following correct or incorrect responses– were dominated by whether the argument aligned with the evaluator's own answer, reflecting a strong belief-consistency (or myside) bias. Argument validity played a comparatively minor role in these evaluations, but it is important to note that participants still moderately relied on this cue to assess arguments' strength, suggesting some intuitive sensitivity to argument validity (Beauvais & De Neys, 2026). In slow evaluations, human participants placed substantially greater weight on validity and less weight on alignment, particularly when they had themselves solved the problem correctly. In such cases, validity became the dominant cue, outweighing alignment. These findings refine dual-process accounts by showing how intuition and deliberation differ in the structure of cue integration during argument evaluation. Fast evaluations appear primarily driven by arguments' coherence with one's current belief state, whereas slow evaluations prioritize normative justification quality. These results align with previous work showing that deliberation improves discrimination of argument quality (Bago et al., 2020; Beauvais & De Neys, 2026). However, note that even in slow evaluations, participants who had initially given an incorrect heuristic response continued to weight alignment more heavily than validity. Thus, even though deliberation increases sensitivity to argument quality and attenuates myside bias, it is not always sufficient to override it.

Second, LLMs exhibited relatively homogeneous evaluative strategies across models and conditions. They relied primarily on justification validity, with minimal influence of surface explicitness and moderate reliance on alignment with their own responses. Cosine similarity analyses revealed that these LLM profiles were structurally closest to that of human slow evaluations following correct responses, and most distinct from human fast evaluations following incorrect responses. These results suggest that LLMs evaluate arguments in a manner that resembles a specific subset of human reasoning, namely, deliberative evaluation by sound reasoners, rather than incorrect responders' intuitive, belief-driven judgment.

Crucially, a key divergence between humans and models was that LLMs showed markedly reduced alignment (or myside) bias. Humans' fast evaluations were tightly coupled to their own responses, especially following incorrect answers, whereas LLMs showed no such coupling – remaining largely anchored to justification validity regardless of their own response. As a result, cosine similarity between LLMs and human fast evaluations was lower than with slow human profiles, particularly after incorrect responses. This contrast highlights a fundamental difference between human and model evaluation: humans' intuitive evaluations are tightly coupled to prior commitments, whereas LLMs' evaluations remain largely anchored to justification validity even when their own responses are wrong (with variations depending on model architecture and capacity).

These findings caution against interpreting surface similarities in task performance as evidence of shared reasoning processes. Although LLMs often match or exceed human accuracy on reasoning problems (Laskar et al., 2023; Yax et al., 2024), their evaluative strategies resemble a specific subset of human judgments (slow evaluations from sound reasoners). In particular, LLMs appear weakly sensitive to myside bias –a central feature of human reasoning (Deans-Browne & Singmann, 2026; Mercier, 2022; Stanovich et al., 2013). While this may appear normatively desirable and "more rational", a narrower evaluative regime that prioritizes explicit normative cues over coherence with prior commitments might be advantageous in the present task but maladaptive in other contexts. For instance, by undermining consistency across responses or fostering excessive sensitivity to external cues, as observed in sycophantic behavior (Fanous et al., 2025; Kumaran et al., 2026).

Several limitations and open questions remain. Our analyses focused on a specific class of reasoning problems and a constrained set of argument features. Whether similar patterns would generalize to richer, real-world arguments, or in domains involving moral or causal reasoning, remains to be tested. In addition, we tested LLMs under a standard, unconstrained prompting regime because our goal was to ask how ordinary LLM evaluations compare to human intuitive and deliberate evaluations, not to simulate human cognitive constraints in models. Consequently, the present findings cannot determine whether LLMs' cue-weight profiles would change under resource-constrained, speeded, or distraction-like conditions. Future work could directly manipulate these conditions to test whether model evaluations can be shifted toward more intuitive-like profiles. Similarly, inducing explicit chain-of-thought reasoning (Kojima et al., 2022; Wei et al., 2023) might alter cue-weight profiles. More broadly, as LLM evaluations were obtained under a single prompting regime and temperature setting, future work should examine how evaluative strategies vary with prompting, instruction, or fine-tuning.

In sum, this study shows that argument evaluation provides a powerful lens for comparing human and machine reasoning. Humans flexibly shift evaluative strategies depending on cognitive constraints, whereas contemporary LLMs implement a more stable, validity-centered strategy that resembles deliberate human evaluation from sound reasoners. Understanding these similarities and divergences is essential both for theories of reasoning and for the responsible use of LLMs as evaluators in scientific and applied settings.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback, as well as Jérémie Beucler for his helpful comments in the preparation of the manuscript. Preparation of this manuscript benefitted from support from the European Union under Horizon Europe Programme Grant Agreement no.101120763 – TANGO. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking and Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bago, B., Rand, D. G., & Pennycook, G. (2023). Reasoning about climate change. *PNAS Nexus*, *2*(5), pgad100. https://doi.org/10.1093/pnasnexus/pgad100

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235. https://doi.org/10.1080/13546789508256909

Beauvais, N., & De Neys, W. (2026). *Argument evaluation, Fast and Slow: Deliberation boosts argument strength discrimination in reasoning problems*. https://osf.io/qz7ph

Beauvais, N., Voudouri, A., Boissin, E., & De Neys, W. (2025). System 2 and cognitive transparency: Deliberation helps to justify sound intuitions during reasoning. *Thinking & Reasoning*, *0*(0), 1–26. https://doi.org/10.1080/13546783.2025.2532653

Beucler, J., Purcell, Z., Charles, L., & De Neys, W. (2026). Using large language models to estimate belief strength in reasoning. *Behavior Research Methods*, *58*(2), 43. https://doi.org/10.3758/s13428-025-02923-9

Beucler, J., Voudouri, A., & De Neys, W. (2025). Moses illusions, fast and slow. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001530

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Caddick, Z. A., & Feist, G. J. (2022). When beliefs and evidence collide: Psychological and ideological predictors of motivated reasoning about climate change. *Thinking & Reasoning*, *28*(3), 428–464. https://doi.org/10.1080/13546783.2021.1994009

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199–212. https://doi.org/10.1037/a0016533

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273. https://doi.org/10.3758/s13423-013-0384-5

Deans-Browne, C., & Singmann, H. (2026). For everyday arguments prior beliefs play a larger role on perceived argument quality than argument quality itself. *Cognition*, *266*, 106257. https://doi.org/10.1016/j.cognition.2025.106257

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2025). Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*, *37*(4), 555–569. https://doi.org/10.1080/10400419.2024.2326343

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, *49*(8), 709–724.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM Sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(1), 893–900. https://doi.org/10.1609/aies.v8i1.36598

Franiatte, N., Boissin, E., Delmas, A., & De Neys, W. (2024). Boosting debiasing: Impact of repeated training on reasoning. *Learning and Instruction*, *89*, 101845. https://doi.org/10.1016/j.learninstruc.2023.101845

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Haigh, M. (2016). Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in Cognitive Psychology*, *12*(3), 145–149. https://doi.org/10.5709/acp-0193-5

Hoeken, H., Šorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking & Reasoning*, *20*(1), 77–98. https://doi.org/10.1080/13546783.2013.807303

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking & Reasoning*, *18*(3), 225–243. https://doi.org/10.1080/13546783.2012.674715

Hornikx, J., Weerman, A., & Hoeken, H. (2022). An exploratory test of an intuitive evaluation method of perceived argument strength. *Studies in Communication Sciences*, *22*(2), Article 2. https://doi.org/10.24434/j.scoms.2022.02.003

Huang, H., Bu, X., Zhou, H., Qu, Y., Liu, J., Yang, M., Xu, B., & Zhao, T. (2025). *An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4* (arXiv:2403.02839). arXiv. https://doi.org/10.48550/arXiv.2403.02839

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kumaran, D., Fleming, S. M., Markeeva, L., Heyward, J., Banino, A., Mathur, M., Pascanu, R., Osindero, S., De Martino, B., Veličković, P., & Patraucean, V. (2026). Competing Biases underlie Overconfidence and Underconfidence in LLMs. *Nature Machine Intelligence*, *8*(4), 614–627. https://doi.org/10.1038/s42256-026-01217-9

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. (2023). A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 431–469). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.29

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, *120*(49), e2309350120. https://doi.org/10.1073/pnas.2309350120

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lord, C., Ross, L., & Lepper, M. (1979). Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, *122*(42), e2518443122. https://doi.org/10.1073/pnas.2518443122

Ma, J., Hu, L., Li, R., & Fu, W. (2025). LoCal: Logical and Causal Fact-Checking with LLM-Based Multi-Agents. *Proceedings of the ACM on Web Conference 2025*, 1614–1625. https://doi.org/10.1145/3696410.3714748

Mercier, H. (2022). Confirmation bias – myside bias. In *Cognitive Illusions* (3rd ed.). Routledge.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press. https://www.jstor.org/stable/j.ctv2sp3dd8

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Motoki, F. Y. S., Pinho Neto, V., & Rangel, V. (2025). Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, *234*, 106904. https://doi.org/10.1016/j.jebo.2025.106904

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2025). How to train your stochastic parrot: Large language models for political texts. *Political Science Research and Methods*, *13*(2), 264–281. https://doi.org/10.1017/psrm.2024.64

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*(2), 170–178. https://doi.org/10.1017/S1930297500003405

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (arXiv:1708.08296). arXiv. https://doi.org/10.48550/arXiv.1708.08296

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, *13*(3), 260–267. https://doi.org/10.1017/S1930297500007695

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside Bias, Rational Thinking, and Intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: Familiarity. *PeerJ*, *4*, e2395. https://doi.org/10.7717/peerj.2395

Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). *Large Language Models are Inconsistent and Biased Evaluators* (arXiv:2405.01724). arXiv. https://doi.org/10.48550/arXiv.2405.01724

Svedholm-Häkkinen, A. M., & Hietanen, M. (2024). On being drawn to different types of arguments a mouse-tracking study. *Thinking & Reasoning*, *0*(0), 1–26. https://doi.org/10.1080/13546783.2024.2370074

Svedholm-Häkkinen, A. M., Hietanen, M., & Baron, J. (2024). Individual Differences in Argument Strength Discrimination. *Argumentation*, *38*(2), 141–167. https://doi.org/10.1007/s10503-023-09620-x

Svedholm-Häkkinen, A. M., & Kiikeri, M. (2022). Cognitive miserliness in argument literacy? Effects of intuitive and analytic thinking on recognizing fallacies. *Judgment and Decision Making*, *17*(2), 331–361. https://doi.org/10.1017/S193029750000913X

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Voudouri, A., Białek, M., & De Neys, W. (2024). Fast & slow decisions under risk: Intuition rather than deliberation drives advantageous choices. *Cognition*, *250*, 105837. https://doi.org/10.1016/j.cognition.2024.105837

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). *Large Language Models are not Fair Evaluators* (arXiv:2305.17926). arXiv. https://doi.org/10.48550/arXiv.2305.17926

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8