Risk-Aware Metacognition: A Dual-Channel Model of Human Second-Order Confidence

Qianhui Ma[1], **Guoqian Zhu**[2] **& Deqian Fu (Fudeqian@lyu.edu.cn)**[3]
[1]School of Information Science and Engineering, Linyi University
[2]School of Information Science and Engineering, Linyi University
[3]School of Information Science and Engineering, Linyi University

Abstract

A pivotal advantage of human intelligence lies in metacognitive monitoring. This ability allows individuals to detect uncertainty in their judgments and adapt their behavior accordingly, such as deferring decision-making or seeking help. However, the computational mechanism underpinning metacognitive processing still remains incompletely characterized. We propose a risk-avoidance-based computational theory of metacognition, stating that second-order confidence is formed through an asymmetric strategy: individuals conservatively downweight potential gains while overestimating potential risks inherent in their decision outputs. We formalize this principle into a dual-channel conservative confidence (DCCC) framework and instantiate it within a large language model (LLM). Evaluations across high-risk medical and legal judgment tasks demonstrate that the model's confidence signals closely align with human metacognitive profiles, including patterns in the distribution of confidence levels, patterns of error detection, and help-seeking propensities. These findings not only delineate a tractable pathway for implementing machine metacognition but also reveal a core risk-avoidance-oriented computational principle that governs human metacognition processing.

Keywords: Metacognition; Second-order confidence; Risk-avoidance; Conservative evaluation; Human–AI alignment; LLM

Introduction

Metacognitive monitoring is a core component of higher-order human cognition. It refers to the ability to reflect on and evaluate the effectiveness of one's own cognitive processes and the reliability of one's judgments. This capacity supports adaptive decision-making in uncertain environments in ways that integrate psychological, computational, and neural bases of confidence (Fleming, 2024).

Since Nelson and Narens (1990) proposed the two-process model of monitoring and control, research has mainly focused on behavioral correlations. Examples include the link between confidence judgments and decision accuracy (Fleming & Lau, 2014), the match between decision delay and task difficulty, and the positive relation (Hampton, 2001) between help-seeking behavior and subjective uncertainty. However, this line of work remains largely descriptive. It does not provide a formal account of the internal computations of metacognition. As a result, a central question remains unresolved: how do humans generate subjective confidence signals?

Most existing theories treat metacognitive signals as a linear re-encoding of first-order decision strength. For example, Bayesian models (Hangya et al., 2016) assume that subjective confidence directly reflects the posterior probability of a first-order decision. In this view, metacognition mainly serves as a form of probability readout. Recent empirical findings challenge this framework (Schulz et al., 2023). Large-sample studies show that individuals with stronger learning ability often have lower metacognitive accuracy, a pattern known as the "expert low-confidence effect" (Hu et al., 2025). This suggests that confidence judgments are not simple copies of objective performance, but show a systematic conservative bias. Work from Ning et al. (2026) using two-photon calcium imaging in macaques further shows that metacognitive judgments arise from prefrontal integration of multiple signals, including memory strength, reward history, and arousal level. They are not the output of a single probability signal.

This pattern is especially clear in high-risk settings such as medical diagnosis. Human confidence judgments show strong risk-avoidance features. Even when the probability of success is high, perceived confidence drops if potential losses are severe. This indicates that metacognition gives priority to loss weighting, rather than performing neutral probability estimation.

These studies point to a clear theoretical gap: traditional metacognition research emphasizes "monitoring accuracy" but largely ignores the role of metacognition in risk regulation under uncertainty. Based on this, the core question of our study is: do humans use an asymmetric risk-based computation when making second-order confidence judgments? We further propose a theoretical hypothesis: the core of human metacognition is not an objective estimate of success probability. Instead, it involves a conservative integration of potential gains and risks. By underestimating gains and overestimating risks, humans prioritize sensitivity to potential losses. This process generates control signals that guide behavior. This hypothesis shifts metacognition from a "probability monitoring" framework toward a "risk control" framework, offering a new perspective on the adaptive nature of human decision-making under uncertainty.

Theoretical Framework

Metacognition as Risk Regulation

Research at the intersection of behavioral ecology and neuroscience suggests that the core function of metacognition is to reduce decision risk in uncertain environments by adjusting cognitive and behavioral strategies, rather than simply

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

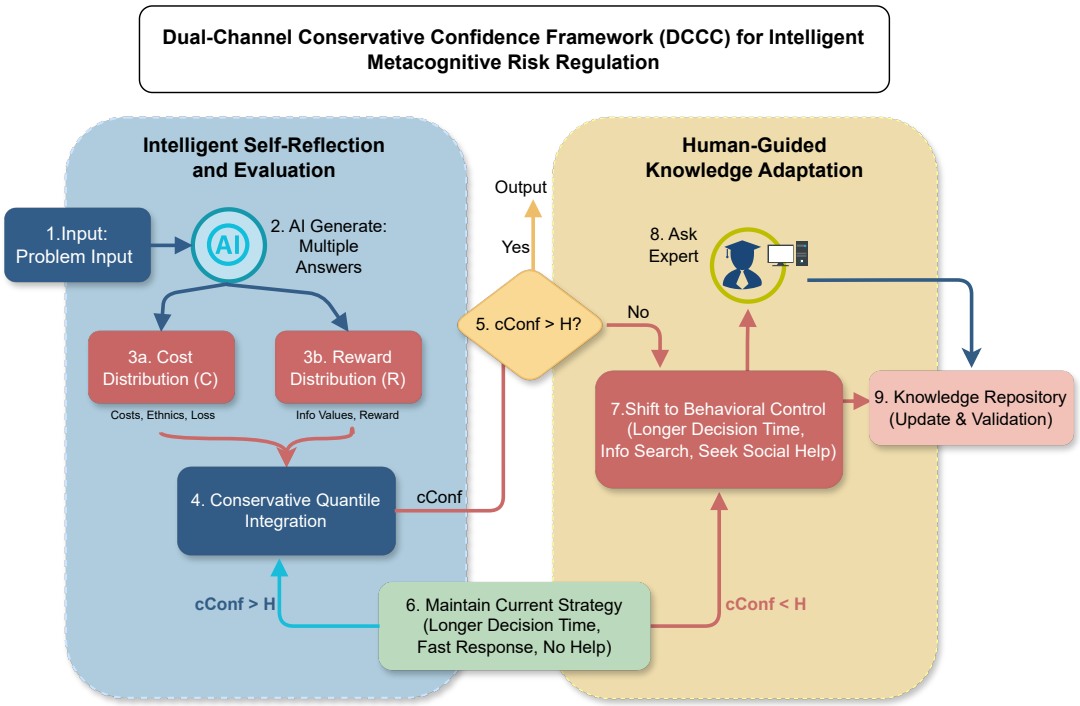

Figure 1: Dual-channel conservative evaluation model structure. After a first-order judgment, metacognitive evaluation proceeds through two parallel channels that estimate a subjective gain distribution (potential benefits) and a subjective risk distribution (potential losses and norm violations). Instead of combining these signals through neutral expectation, the model applies an asymmetric conservative evaluation that underestimates gains and amplifies risks, producing a confidence signal biased toward loss sensitivity. This conservative confidence functions as a behavioral control variable: high values sustain the current decision strategy, whereas low values trigger risk-regulatory adjustments such as increased deliberation, information search, and help-seeking.

evaluating decision accuracy. When subjective confidence decreases, individuals show three systematic behavioral responses. These behaviors are not merely the result of perceiving low probabilities; instead, they reflect an active, risk-regulating adaptation, supported by clear neural mechanisms:

Extended Decision Time When individuals perceive increased risk, they actively prolong decision-making. This shifts cognitive resources from "rapid response" to "risk assessment". Brain stimulation experiments by Kapetaniou et al. (2025) confirm this mechanism. The functional connectivity between the Frontopolar Cortex (FPC) and the Dorsolateral Prefrontal Cortex (DLPFC) directly regulates this time extension. In high-risk scenarios, coupling between these regions strengthens. This prompts more prudent information processing to avoid losses from impulsive decisions. This extension does not represent a drop in cognitive efficiency; rather, it is an optimized allocation of resources by the risk regulation system.

Increased Information Search Information seeking under low confidence essentially aims to reduce cognitive uncertainty and risk exposure by supplementing data. Macaque experiments by Wang Liping's group reveal the neural mechanism. When memory certainty is low, baseline activity in the prefrontal cortex activates brain regions related to information search. This prompts animals to actively seek extra cues to verify memory accuracy. This process is closely tied to reward history: the greater the potential loss, the higher the search intensity. Human experiments show similar results. In high-risk tasks like medical diagnosis, the frequency of doctors' information searching correlates negatively with subjective confidence. Furthermore, searches prioritize risk factors that could lead to misdiagnosis.

Increased Social Help-Seeking From an evolutionary cognitive perspective, help-seeking is an adaptive strategy for risk sharing. Its core function is to reduce risk costs by introducing external decision-making resources. Research shows that even "expert" individuals with strong learning abilities actively seek help during difficult tasks. This is not due to a lack of judgment accuracy (Seo & Lee, 2012). Instead, the risk regulation system identifies the severity of potential losses and uses help-seeking to transfer risk (Desender et al., 2018). Neurally, this behavior is linked to reward expectation representations in the prefrontal cortex. Help-seeking is

activated when the expected loss of an independent decision exceeds the expected cost of asking for help.

Based on the behavioral and neural evidence above, we redefine metacognition as a second-order control system oriented toward risk regulation. Its core function is to integrate information about gains and risks. It generates subjective confidence signals that guide decision strategies. The goal is to minimize losses and stabilize gains, rather than simply "reading out probabilities".

Computational Hypothesis

To formalize the hypothesis of metacognition as a risk-regulation system, we propose the Dual-Channel Conservative Confidence (DCCC) framework, shown in Figure 1. The core idea of this computational hypothesis is that when forming second-order confidence judgments, humans compute two key variables in parallel: subjective gains and subjective risks. These are integrated using an asymmetric quantile mechanism to produce confidence signals biased toward risk avoidance (Lebreton et al., 2015). This computational formulation is not a definition of an algorithm. Rather, it is a testable formal representation of the human risk-sensitive metacognitive mechanism.

Specifically, after completing a first-order cognitive judgment, individuals form two parallel cognitive channels. These channels generate a subjective gain distribution (R) and a subjective risk distribution (C) for the judgment outcome.

Subjective gain distribution (R) This represents uncertain estimates of potential information value, task success rewards, or social recognition. For example, a doctor's gain estimate for a diagnosis includes both the therapeutic effect of a correct diagnosis and the professional recognition gained. The uncertainty of these gains forms the R distribution.

Subjective risk distribution (C) This represents uncertain estimates of negative outcomes, such as violations of ethical norms, legal rules, or the costs of errors. Its distribution depends on factors like ethical requirements, legal boundaries, and severity of potential losses. For instance (Studdert et al., 2005), a doctor's risk estimate for a diagnosis includes potential patient harm from misdiagnosis and, importantly, whether the diagnostic action complies with medical ethics or legal regulations.

Traditional computational models, such as Bayesian metacognition models (Li et al., 2020), typically integrate gain and risk information using expected values to generate confidence. However, this approach cannot explain humans' conservative bias in high-risk situations: expectation-based integration is essentially neutral and overlooks humans' heightened sensitivity to potential losses.

The core innovation of our model is to replace expected-value integration with conservative quantile estimation. By underestimating the lower bound of gains and overestimating the upper bound of risks, the model prioritizes sensitivity to potential losses. Its formal representation is as follows:

$$cConf = Q_\alpha(R) - Q_\beta(C) \tag{1}$$

Here, $Q_\alpha(R)$ represents the lower quantile of the gain distribution ($\alpha < 0.5$). This means that a proportion α of gain estimates falls below this quantile, reflecting a conservative underestimation of gains. $Q_\beta(C)$ represents the upper quantile of the risk distribution ($\beta > 0.5$). It means that a proportion β of risk estimates falls below this quantile, reflecting an amplified perception of risk.

Since the raw confidence signal $cConf$ inherits the scale of the underlying gain and risk scores, we apply a monotonic normalization to map it into the range [0,1] for comparability across conditions. Specifically, we use a logistic transformation:

$$\tilde{cConf} = \sigma(cConf/s) = \frac{1}{1 + \exp(-cConf/s)} \tag{2}$$

where s is a scaling factor defined based on the nominal ranges of gain and risk scores. This transformation preserves the relative ordering of confidence values while producing a bounded control signal.

Based on our pilot data and existing literature, we set $\alpha = 0.25$ (lower quartile of gains) and $\beta = 0.75$ (upper quartile of risks). This parameter choice aligns with the principle of "distribution-independent conservative evaluation" in conformal prediction theory (Angelopoulos & Bates, 2021). It also matches the "expert low-confidence effect". Lower-quartile gain estimates prevent individuals from ignoring potential risks due to overconfidence. Upper-quartile risk estimates enhance sensitivity to possible losses.

It is important to emphasize that conservative confidence ($cConf$) in this model is not an estimate of success probability, but a control signal for risk regulation. When $cConf$ exceeds an internal threshold H, the system maintains its current decision strategy, favoring rapid responses without external intervention. When $cConf$ falls below H, it triggers a shift to a controlled behavioral state, characterized by prolonged decision time, increased information search, and potential help-seeking.

This threshold-based transition captures how low-confidence states activate adaptive regulatory responses under uncertainty. It also distinguishes the proposed model from traditional probability-based metacognition frameworks, where confidence serves primarily as a passive estimate rather than an active control variable.

Methods

Dual Cognitive Channels

We constructed a trainable neural system to instantiate the metacognitive theory of this study. In this research, the Large Language Model (LLM) assumes only two core roles. Its purpose is strictly to verify cognitive mechanisms:

Uncertainty Distribution Generator This component generates multiple candidate answers. It then feeds this content into the reward and cost evaluation models within the Safe-RLHF (Dai et al., 2023) framework.

Confidence Judgment Observation System This component receives scores from the dual evaluation models. It integrates these scores to calculate conservative confidence and decides on the appropriate method of behavioral regulation.

The neural system employs a three-stage architecture: "Main Model Generation - Dual Model Evaluation - Quantile Integration." We use Apache-7B as the core model and integrate the reward and cost models from the Safe-RLHF framework. The specific design is as follows: In the first stage, Apache-7B processes inputs (such as Medical or Legal tasks) and generates diverse first-order judgments. It extracts core semantics and task difficulty features, simulating human initial information processing. In the second stage, the Safe-RLHF dual models quantify scores: the reward model evaluates positive value, while the cost model evaluates potential risk. These scores are normalized to construct the Reward Distribution (R) and Risk Distribution (C). In the third stage, we use quantile calculation to extract and integrate $Q_\alpha(R)$ and $Q_\beta(C)$. This generates the $cConf$, simulating the risk asymmetry mechanism found in human metacognition (Sun et al., 2023).

Assistance-Seeking as Metacognitive Behavior

Based on the conservative confidence signal, we implement a human–AI collaboration mechanism that translates metacognitive states into adaptive behavior. When $cConf$ falls below a threshold (H), the system triggers a Human-in-the-Loop assistance process, mirroring the human tendency to seek help under low confidence.

This process consists of three stages. First, the system defers autonomous output to avoid error propagation under uncertainty (Mozannar & Sontag, 2020). Second, it submits the query, preliminary response, and confidence signal to a human expert for evaluation, enabling context-aware intervention (Kompa et al., 2021). Finally, the system incorporates expert feedback, either confirming the response or correcting errors for future updates.

Rather than a simple correction mechanism, this process reflects a metacognitively driven strategy: low confidence triggers the recruitment of external cognitive resources to mitigate risk. The model thus operationalizes assistance-seeking as a control behavior, serving as a computational analogue of human help-seeking under uncertainty.

Experiments

Research Question

We employ a two-stage design: "Human Behavior Measurement + Computational Model Instantiation + Structural Alignment Analysis." We treat the model's output as the computational implementation of cognitive theory, human data serves as the empirical reference for these metacognitive phenomena. To ensure this problem is testable, we decompose it into two interlinked operational cognitive hypotheses. Together, they address our core objective: determining if the model can serve as a working theoretical instance of human metacognitive risk regulation mechanisms.

H1 (Calibration Structure Hypothesis) The model's generated conservative confidence should show a significant positive correlation with human subjective confidence scores. This would indicate that both signals reflect a similar "dimension of uncertainty perception".

H2 (Behavioral Regulation Hypothesis) Human help-seeking or delayed decision-making should increase significantly within low-confidence intervals. Similarly, the model should trigger its assistance mechanism when its $cConf$ falls below a threshold that aligns closely with these human behavioral thresholds.

Table 1: Summary of Participant Information.

Measure	Description	Value
Number	Completed experiment	40
Age	Mean ± SD	23.5 ± 1.7
Gender	Male / Female	18 / 22
Total trials	Per participant	60
Task Type	Question / Judgment	30 / 30
Confidence	0–100	68.2 ± 15.4

Human Study

To obtain reliable data on human metacognitive behavior, we recruited 40 graduate students from medical and law schools as participants. None had a history of cognitive disorders. All were proficient in computer use and had basic medical and legal knowledge. Each participant signed an informed consent form before the experiment. The tasks covered two high-risk domains: medical question answering MedQA (Jin et al., 2020) and legal judgment Legal QA (Zhong et al., 2020). These tasks involve substantial potential gains and losses, making them well suited for studying metacognitive risk regulation. They also allow controlled manipulation of uncertainty while preserving ecological validity.

Each participant answered the same set of mixed questions. For each item, they completed three steps: providing an answer as the first-order decision, rating subjective confidence on a 0–100 scale, and choosing a follow-up action (submit directly, search for information, or seek help). These measures captured cognitive output, confidence evaluation, and behavioral regulation.

Computational Instantiation

To operationalize the risk-asymmetric metacognitive theoretical model, we built a complete computational system using the Apache-7B Large Language Model as the core "cognitive

agent". We integrated this with the dual evaluation models (reward and cost) from the Safe-RLHF framework. The system processes the exact same set of questions used for human subjects and generates corresponding responses. It then utilizes a dual-channel assessment mechanism to calculate two key metrics: estimated reward (R) and risk distribution (C). Based on these metrics, we calculate the $cConf$ using the quantile integration mechanism described previously. Besides, we additionally implement two baseline variants for comparison: a reward-only proxy (R-only) and a symmetric integration model ($\alpha = \beta = 0.5$), which serve as reference points for evaluating the effect of asymmetric risk integration.

Model–Human Structural Alignment

This study examines structural similarity between human and model metacognition. All reported alignment analyses are conducted at the item level by pairing each question item's human confidence (averaged across participants) with the model's $cConf$ on the same item; for completeness, participant-level correlations are also computed and summarized. We focus on two analyses based on our cognitive hypotheses.

First, the confidence-error relationship: we compare human confidence scores and model $cConf$ against error rates. By examining curve shape, slope, and correlation, we test H1: the Calibration Structure Hypothesis, assessing whether model and human confidence reflect the same uncertainty dimension. We further include baseline comparisons to evaluate whether asymmetric integration provides a better fit to human calibration patterns than simpler alternatives.

Second, behavioral regulation thresholds: we track how the probabilities of human help-seeking behaviors—checking materials or seeking assistance—change as confidence decreases. We compare this with the model's $cConf$ trigger region for assistance. This tests H2: the Behavioral Regulation Hypothesis, checking whether model and human thresholds align.

All correlations are computed using the normalized confidence signal ($c\tilde{Conf}$) and human confidence rescaled to [0,1].

Result

The experimental results support both cognitive hypotheses and confirm structural alignment between human and model metacognition. Our findings show that the risk-asymmetric confidence model effectively captures key features of human metacognitive monitoring and behavioral regulation.

For H1, model confidence ($cConf$) correlates positively with human subjective confidence ($r = 0.71$, $p < 0.001$, $N = 60$). Participant-level correlations show consistent patterns. As shown in Figure 2(a), both humans ($r = -0.62$) and the model ($r = -0.68$) exhibit consistent negative relationships between confidence and error. Notably, while all model variants (including R-only and symmetric integration) follow this general trend, the proposed asymmetric formulation more closely tracks the human curve, particularly in intermediate confidence regions. Sensitivity analysis further confirms that this alignment is stable across $\alpha \in [0.15, 0.4]$ and $\beta \in [0.6, 0.85]$.

For H2, help-seeking behavior increases as confidence decreases in both humans and the model. As illustrated in Figure 2(b), human consultation frequency rises progressively with uncertainty, reaching 0.98 in the lowest confidence bin. In contrast, the model adopts a deterministic threshold policy, where requests are triggered when $cConf < 0.5$, leading to a saturated request rate in the low-confidence regime. Despite this difference in mechanism, both exhibit a consistent directional trend: intervention becomes more likely under lower-confidence conditions.

Stratified analyses showed no significant differences between medical and legal tasks (MedQA: $r = 0.70$; Legal QA: $r = 0.68$), indicating that the model's representation of human metacognitive structure generalizes across domains. Overall, these results demonstrate that the proposed model provides a valid instantiation of human-like metacognitive risk regulation, reproducing key structural patterns under uncertainty.

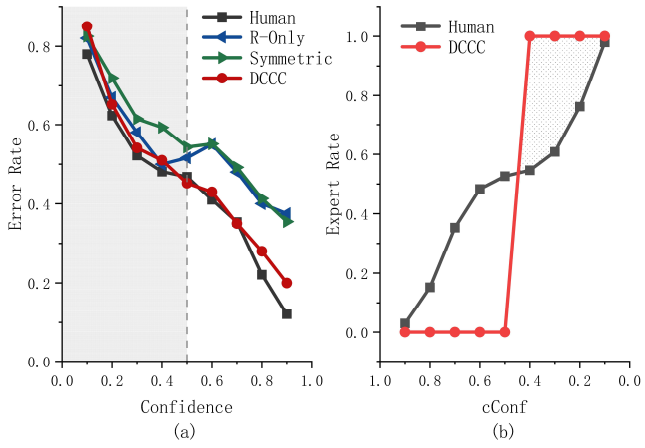

Figure 2: Empirical validation of metacognitive alignment hypotheses. (a) Confidence–error relationship. Confidence is binned and mean error rates are computed. We include two baselines (R-only and symmetric $\alpha = \beta = 0.5$). All methods show a negative trend, while DCCC more closely matches human behavior. (b) Confidence vs. expert consultation frequency. The y-axis denotes request probability. Human behavior increases gradually as confidence decreases, whereas the model applies a threshold policy ($cConf < 0.5$), resulting in saturated requests in low-confidence regions.

Discussion

This study presents a computational account of metacognition as risk regulation. In this view, confidence is not a direct estimate of correctness, but a control signal that guides behavior under uncertainty. The results show that human second-order confidence is better characterized by asymmetric integration

of gains and risks, rather than neutral or reward-dominant formulations.

Empirically, the proposed DCCC framework reproduces two key structural properties of human metacognition. First, confidence serves as a calibrated uncertainty signal: both human judgments and model-derived $cConf$ exhibit consistent negative relationships with error rates, with the asymmetric formulation more closely matching human patterns than baseline alternatives. Second, confidence functions as a control signal for behavior: decreasing confidence is associated with increased help-seeking in both humans and the model, despite differences in implementation.

These findings suggest that metacognition operates as a second-order control system oriented toward risk management. Rather than estimating objective success probabilities, the system biases confidence toward loss sensitivity, enabling adaptive regulation under uncertainty. This perspective also offers a computational explanation for conservative phenomena observed in human cognition, such as reduced confidence in high-stakes contexts.

At a broader level, the alignment between human and model behavior indicates that risk-asymmetric confidence may serve as a useful design principle for machine metacognition. By explicitly encoding sensitivity to potential losses, such models can better approximate human-like uncertainty awareness and support safer decision-making in high-risk domains.

Conclusion

This paper introduces a risk-asymmetric computational framework for modeling human metacognition and instantiates it within a large language model. By integrating gain and risk through asymmetric quantile estimation, the proposed DCCC model produces a confidence signal that aligns with human judgments in both calibration structure and behavioral regulation.

Experimental results demonstrate that the model captures key features of human metacognitive behavior, including the relationship between confidence and error, as well as uncertainty-driven help-seeking. Compared to simpler baselines, the asymmetric formulation provides a closer approximation to human patterns, supporting the hypothesis that metacognition is fundamentally shaped by risk-sensitive evaluation.

Overall, this work reframes metacognition as a risk-control process and provides a tractable computational instantiation of this principle. It offers both theoretical insight into human cognition and practical implications for developing more reliable and self-aware AI systems in high-stakes environments.

References

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., & Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback [arXiv preprint arXiv:2310.12773].

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences*, *98*(9), 5359–5362.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

Hu, M., Zhao, W., Li, A., Shanks, D. R., Yu, Y., Tian, X., Liu, M., Hu, X., Luo, L., & Yang, C. (2025). Good learners are poor monitors: A negative relation between learning ability and monitoring accuracy. *Psychological Science*, *36*(9), 746–764. https://doi.org/10.1177/09567976251358838

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2020). What disease does this patient have? a large-scale open domain question answering dataset from medical exams [arXiv preprint arXiv:2009.13081].

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., Clark, J., Amodei, D., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kapetaniou, G. E., Moisa, M., Ruff, C. C., Tobler, P. N., & Soutschek, A. (2025). Frontopolar cortex interacts with dorsolateral prefrontal cortex to causally guide metacognition. *Human Brain Mapping*, *46*(2), e70146. https://doi.org/10.1002/hbm.70146

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 4.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167.

Li, H. H., et al. (2020). Confidence reports in decision-making with multiple alternatives. *Nature Communications*, *11*, 4469. https://doi.org/10.1038/s41467-020-15581-6

Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. *International Conference on Machine Learning*, 7076–7087.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 125–173, Vol. 26). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Ning, C., Fu, G., Zhang, Y. Y., Meyniel, F., & Wang, L. (2026). Macaque prefrontal cortex integrates multiple components for metacognitive judgments of working memory [Advance online publication]. *Neuron*. https://doi.org/10.1016/j.neuron.2025.11.014

Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, *130*(3), 604–639. https://doi.org/10.1037/rev0000401

Seo, H., & Lee, D. (2012). Neural basis of learning and preference during social decision-making. *Current Opinion in Neurobiology*, *22*(6), 990–995. https://doi.org/10.1016/j.conb.2012.05.010

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*, 8634–8652.

Studdert, D. M., Mello, M. M., Sage, W. M., DesRoches, C. M., Peugh, J., Zapert, K., & Brennan, T. A. (2005). Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *JAMA*, *293*(21), 2609–2617.

Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., & Gan, C. (2023). Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, *36*, 2511–2565.

Xue, K. (2024). Challenging the bayesian confidence hypothesis in perceptual decision making. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2410487121

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). Jec-qa: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5). https://doi.org/10.1609/aaai.v34i05.6519