
Verify the Signals First: The Declarative–Behavioural Gap in Metacognitive Monitoring of LLM Agents

Ricky J. Sethi

Fitchburg state University
Worcester Polytechnic Institute
rickys@sethi.org

Abstract

Verification pipelines for LLM agents increasingly use signals the agent produces about itself: e.g., confidence statements, uncertainty estimates, self-assessments of competence and conflict, etc. We argue that these signals must themselves be verified before anything built on them can be trusted. In this paper, we supply a concrete validation framework for one family of such signals: a five-dimensional Metacognitive State Vector (MSV) spanning emotional response (ER), correctness evaluation (CE), experiential matching (EM), conflicting information (CI), and problem importance (PI). We assess how much evidence currently supports each MSV dimension as a measure of its stated property using behavioural probes. While all five probes are specified here, only CE is measured directly in one of its operationalizations, self-rated confidence, and evaluated herewith. We identify a central obstacle to using such signals for verification which we call the *declarative–behavioural gap*, which is the possibility that what a model reports about its own state may diverge from how it actually behaves.

We put forth four candidate explanations with distinct, testable predictions and we then measure that gap directly on 9 open-weight models. When asked prospectively how well their confidence would predict their correctness, all nine models described it as weakly or moderately predictive; measured discrimination over 1,155–1,188 binary trials per model was uniformly weak (AUC 0.488–0.543, mean 0.512, item-clustered bootstrap). A within-item order-reversal control showed that reversing option order changes accuracy by 0.375 on average, up to 0.92, while stated confidence shows almost no systematic shift with order, so confidence does not register the factor that governs correctness on this task. Two instruments intended to validate that signal failed in different ways, one becoming indeterminate for seven of nine models and the other returning plausible values with no detectable relationship to direct discrimination.

We also propose criteria for deciding when metacognitive signals are trustworthy enough to support verification as well as an accompanying audit principle: namely, that the reported state and observed behaviour should remain separately inspectable so that the declarative–behavioural gap itself can be measured.

1 Who Watches the Watchmen? Verifying the Signals that Verify Agents

Who verifies the agents, and on what evidence? Current agent-verification pipelines increasingly rely on signals produced by the agents themselves: a confidence estimate may determine whether an output is checked, stated uncertainty may trigger human review, and self-critique may stand in for external evaluation. A systems theory of agents asks how agent behaviour should be verified but we ask an even earlier question: *have the signals on which verification relies themselves been validated as measurements?* If an agent’s reported confidence, uncertainty, or conflict does not

reliably correspond to the property it purports to measure, or to the behaviour that property should predict, then a verification system built on that signal inherits a latent failure mode. The consequences are particularly important under high stakes, distribution shift, and adversarial framing, precisely where verification is intended to provide protection.

A single self-reported confidence score is the familiar case [Xiong et al., 2024, Kadavath et al., 2022], and §4 measures how little even that delivers on this cohort. Richer, multi-dimensional state estimates are now being proposed as monitoring and control signals for agentic systems, and those are what we examine here. The Metacognitive State Vector (MSV) [Sethi et al., 2025, Courchaine et al., 2026] is a five-dimensional per-agent profile $M = \langle \text{ER}, \text{CE}, \text{EM}, \text{CI}, \text{PI} \rangle \in [0, 1]^5$ over Emotional Response (ER), Correctness Evaluation (CE), Experiential Matching (EM), Conflicting Information (CI), and Problem Importance (PI). The MSV is computed through a plug-in estimator interface for each dimension, which consists of dedicated evaluation prompts by default but also allows external estimators optionally (e.g., ER is scored lexically rather than by self-report). Downstream systems use this state to decide when to deliberate and which agent takes which role.

In this paper, we ask the question that comes before those uses, namely what would make a multi-dimensional metacognitive estimate trustworthy enough to verify anything, and where the MSV currently stands against that standard. The main contribution is a validation framework and its assessment that includes: per-dimension behavioural probes with explicit construct-validity status (§2); the declarative-behavioural gap as the central threat, with four candidate mechanisms for it, each making different predictions (§3); cohort-scale evidence read through that gap (§4); and promotion criteria plus an audit-design principle for verification systems that use metacognitive signals (§5).

2 Per-Dimension Construct Validity

A model’s ability to produce a plausible self-assessment does not establish that the assessment measures the intended construct. For each of the four non-CE dimensions, we therefore specify a behavioural probe that would test the relationship between the reported signal and the behaviour the construct is meant to capture. CE calibration against ground truth is comparatively well studied and is evaluated in companion work [Anonymous Authors, 2026]; the four dimensions below are where the multi-dimensional formulation extends beyond scalar confidence.

Emotional destabilization (ER). The probe would test sensitivity to affective framing by presenting otherwise equivalent questions with and without language that raises the emotional stakes (e.g., “this question has critical implications for patient safety”). Accuracy, stated confidence, and delegation would then be compared across the two conditions. An affective frame that adds no epistemically relevant information should not by itself impair performance; i.e., stable metacognition should not degrade under charged but epistemically irrelevant context. The probe would provide only partial validation of ER as framing is a useful manipulation but the connection between a framing effect and the specific ER construct remains to be established more carefully, so we call this *partial*.

Contradiction detection (CI). The probe would use pairs of statements in which one contains a subtle factual inconsistency and measure whether the model identifies the conflict. This would be a relatively direct behavioural test of conflict sensitivity, and it therefore gives CI the most direct construct alignment among the four probes proposed here; but the caveat is coverage, since CI also spans source agreement and temporal stability, which this probe does not touch. If successful, it would provide evidence for the interpretation of the signal rather than proof that the signal captures every form of conflicting information represented by CI, so we call it *reasonable* for the component it exercises.

Familiarity-sensitive confidence (EM). Questions would be drawn from domains that differ in estimated training-data prevalence, and the probe would ask whether confidence varies with familiarity after accounting for differences in accuracy. This is an indirect test of EM because training exposure is not observed directly; domain prevalence serves only as a proxy for familiarity. We therefore classify the proposed validation as *partial*.

Stakes-aware delegation (PI). The probe would vary the stated consequences of an otherwise comparable task (e.g., “this answer will determine a treatment decision”) and measure whether the model becomes more willing to delegate. A change in delegation would be consistent with sensitivity to problem importance, but the appropriate response to higher stakes depends on the costs and

Table 1: Construct-validity status of each MSV dimension. “Reasonable” means the probe measures an established construct by an established paradigm; “partial” means the probe is specified but has not yet been run, and its mapping onto the named dimension needs further validation. One operationalization of CE, a self-rated confidence score attached to an answer, has received a direct cohort-scale test (§4); the deployed MSV CE estimator is a separate, structured assessment and has not been tested this way. The four remaining probes are specified, not executed. No dimension is currently *validated* in the sense a verification input should require (§5).

Dim.	Behavioural probe	What it manipulates / measures	Status
CI	Contradiction detection	Statement pairs, one with a subtle factual error; detection accuracy	reasonable
ER	Emotional destabilization	Affective framing prepended; shift in accuracy, confidence, delegation	partial
EM	Familiarity-sensitive confidence	Domain prevalence varied; confidence beyond what accuracy predicts	partial
PI	Stakes-aware delegation	High-stakes framing; change in delegation rate	partial
CE (self-rated confidence)	Confidence–correctness discrimination	Direct cohort-scale test, this paper (§4)	tested; weak
CE (deployed estimator)	Structured claim assessment	Not tested here; distinct operationalization	untested

consequences of error in that particular domain. This probe would therefore provide only *partial* evidence for the PI construct.

Table 1 summarizes the resulting assessment: *reasonable* for CI on the component it tests, *partial* for ER, EM, and PI, and *mixed* for CE from the companion measurement. Among the four proposed probes, the CI probe has the most direct construct alignment, while the ER, EM, and PI probes are specified but their mapping onto the named dimensions remains to be established. None of the four has yet been run. We therefore treat the MSV as a candidate monitoring representation whose dimensions can be tested behaviourally, not as a validated measurement instrument. That distinction is especially important if these signals are later used inside verification or control mechanisms: a plausible self-assessment should not be treated as a reliable measurement until the corresponding construct has been tested directly.

3 The Declarative–Behavioural Gap

Calibration alone does not establish that a reported metacognitive state is reflected in behaviour. A model may give a sensible account of its uncertainty or conflict yet fail to use that information when choosing whether to answer, defer, or revise; the converse is also possible, with behaviour discriminating among cases that the reported state does not. Declared and behavioural metacognition can therefore come apart [Ackerman, 2026, Anonymous Authors, 2026]; Ackerman [2026] reaches the same conclusion from behavioural paradigms that avoid self-reports entirely, scoring whether a model strategically delegates questions it would answer incorrectly. For verification, this distinction is especially relevant because a declared signal provides evidence about behaviour only to the extent that the relationship between the two has been established and several mechanisms could produce such a dissociation:

Training-objective separation. Models receive abundant examples of hedging, qualification, and uncertainty language, but far less direct supervision linking uncertainty to decisions such as abstention or delegation. The ability to express uncertainty may therefore improve without a corresponding change in how uncertainty governs action. If this account is correct, the gap should vary with post-training strategy and should narrow when uncertainty is tied explicitly to action-level supervision.

Representational separation. Reported confidence and uncertainty-sensitive action may depend on different internal representations. Verbal confidence may be recoverable from features that are not sufficient for deciding whether to answer or defer, so changes to one need not affect the other. This hypothesis predicts that the two should be separable across layers or representational subspaces, and that an intervention affecting one need not affect the other.

Table 2: Four candidate mechanisms for the declarative–behavioural gap; each row gives the observation that would tell it apart from the others. All four imply the same operational caution but they differ in what would fix the gap.

Mechanism	Prediction	Evidence that would distinguish it
Training-objective separation	Gap size tracks the post-training recipe and narrows under action-level supervision	Compare families by abstention/deferral training exposure; fine-tune on action-conditioned data
Representational separation	Declared confidence and action policy are separable in different layers or subspaces	Probing classifiers on hidden states for each readout
Format compliance without uncertainty use	Perturbing report format changes declared state but not behaviour	Format-perturbation sweep at fixed item and context
Model-family differences	Gap size clusters by model family rather than by scale	Family $\times$ scale factorial over the probe battery

Format compliance without uncertainty use. Structured self-reporting may reflect instruction-following rather than access to a state that also governs behaviour. A model could emit the requested confidence or conflict field while relying on a different signal when making the subsequent decision. Under this idea, changing the reporting format should alter the declaration more readily than the behaviour.

Model-family differences. Training for abstention, calibration, and confidence expression is unlikely to be uniform across model families. This could allow improvements in declaration and behaviour to vary independently, producing systematic differences in declarative–behavioural agreement even at similar model scale. This view predicts stronger clustering by training lineage or post-training regime than by parameter count.

Table 2 pairs each mechanism with an observation that would help distinguish it from the alternatives. These are hypotheses, not findings, and several could hold at once; discriminating among them is future work that the probes of §2 make possible. The implication for verification is narrower and does not depend on which mechanism dominates: a declared metacognitive state should not be treated as a measurement of behavioural disposition until the dissociation rate between the two has been measured and bounded in the intended deployment setting.

4 Measuring the Gap: Declared Self-Knowledge versus Measured Discrimination

We measure the gap directly on 9 open-weight models (0.5B–9B parameters), using a prospective design: models first declare their expected calibration, and the declaration is then tested against behaviour on 1,155–1,188 binary-choice trials per model (198 source questions $\times$ 3 answer pairs $\times$ 2 presentation orders), drawn from graduate-level GPQA items [Rein et al., 2023, Anonymous Authors, 2026]. Discrimination and calibration are distinct properties of such a signal and can vary independently [Steyvers et al., 2024]. In this work, we measure discrimination, because the probe asks a discrimination question, whether confidence predicts correctness, and because calibration is not identifiable at this task’s difficulty: with accuracy between 0.502 and 0.589, the empirical correctness rate is close to 0.5 in every confidence bin, leaving a reliability curve with almost no vertical extent to fit.

Declared Self-Knowledge. Before any trial data was collected, each model answered a structured self-knowledge probe; its first question asks, verbatim, how well the model expects its confidence to predict whether it is actually correct, on a four-point scale from “essentially random relative to my accuracy” to “highly predictive”. All nine declared their confidence weakly (two models) or moderately (seven models) predictive. The declaration is a single unreplicated response per model, and it is generic rather than task-specific; both limitations bound the claim below.

Measured Discrimination. The gold-standard discrimination measure, the Mann–Whitney AUC of stated confidence against correctness, ranges from 0.488 to 0.543 across the nine models (mean 0.512). Because the six trials per source question are not independent, intervals are computed by item-clustered bootstrap (resampling source questions); empirical intraclass correlations are low

(−0.18 to 0.15), and two per-model clustered intervals exclude 0.5 (gemma2:9b [0.501, 0.557]; qwen2.5:7b [0.514, 0.572]), without correction for the nine model-level comparisons, while none reaches a 0.60 practical reference level for useful discrimination (a threshold we adopt as a reference, not a preregistered criterion). Figure 1 (left) shows the nine intervals. The seven declarations of moderately predictive confidence are difficult to reconcile with discrimination this close to chance; we do not attempt to map the four-point verbal scale onto an AUC, so we take no position on whether the two weakly-predictive declarations are compatible with effects this small.

Mechanism: positional response bias with confidence invariant to it. Each answer pair was presented in both orders, giving a within-item control. Accuracy swings by 0.375 on average when the order is reversed, up to 0.92 for the most biased model, which answers the first-listed option on 96% of trials; the deviation of each model’s accuracy from what a pure position-responder would score averages 0.039, and the actual content signal (Youden’s J) ranges from 0.003 to 0.178 (mean 0.077). Stated confidence, meanwhile, shifts systematically with order by only 0.019 on average, and is flat to two decimals for several models; it varies from trial to trial, but not with the manipulation. Performance on this task is dominated by positional response bias for most models while genuine content discrimination is small, and confidence is essentially invariant to the single factor that governs correctness (Figure 1, right). The fact that LLMs exhibit such selection bias is established in literature; e.g., Zheng et al. [2024] document a preference for particular option identifiers across twenty models and three benchmarks. Our observation is the second half of the pair; i.e., that stated confidence does not register the bias, moving by 0.019 while accuracy moves by 0.375. A paired analysis of the two presentations of each answer pair confirms the mechanism and identifies one exception, in which repeating the same *letter* across reversal (cohort mean 0.613) is uninformative on its own: independent responding at each model’s observed letter preference already predicts 0.598, and the observed excess is +0.015. The informative quantity is semantic consistency, the rate at which a model repeats its *underlying choice*. Under an independent, order-invariant content-following model with per-presentation accuracy a , semantic consistency is $a^2 + (1 - a)^2 \geq 0.5$; the observed sub-chance consistency is therefore inconsistent with that model, and in this reversal design it is the pattern expected from position-sensitive responding. Seven of nine models fall strictly below (item-clustered bootstrap intervals entirely under 0.5, as low as 0.054 [0.035, 0.074]), one is indistinguishable from 0.5, and one, gemma2:9b, lies strictly above at 0.582 [0.534, 0.630]. That model is also the one with the largest first-order sensitivity and the largest content signal net of position, so the analysis separates a content-follower from a position-dominated cohort rather than flattening both. Stated confidence varied across reversals (mean absolute change 0.276) but not *systematically* with order (mean shift 0.019), so confidence fluctuates without responding to the manipulation that governs correctness. The reversed-order design is what makes the bias measurable and separable, so the task instrument is doing its job; what requires reinterpretation is raw accuracy, along with any measure built on it that does not separate bias from sensitivity.

The validators failed informatively. Two instruments intended to score this signal were applied to the same data. A ratio-based metacognitive coefficient (meta-level sensitivity over first-order sensitivity, in the signal-detection formulation of Maniscalco and Lau [2012] as adapted to LLMs by Wang et al. [2025]) was indeterminate for seven of nine models: five fits produced ratios outside the implementation’s admissible range (up to 6.09) and two degenerated to zero numerators, consistent with instability under the small first-order sensitivity (0.04–0.45) that position-driven responding produces; the two “successful” values are ratios of near-zero quantities and no more trustworthy. An entropy-based fallback assigned discrimination scores spanning its full range, yet those scores show no detectable relationship to the gold-standard AUC across the cohort; structurally, the proxy measures within-condition concentration of confidence while discrimination requires between-condition separation, so a model emitting one confidence level everywhere would receive a perfect score while carrying no information. The failures differ in kind as the primary estimator could not return a stable measurement in this regime, which is the responsible outcome when its assumptions fail. More concerningly, the fallback returned apparently meaningful values that do not track direct discrimination, and a verification pipeline consuming those would have received plausible numbers unrelated to the construct. Together, these failures show why metacognitive validators must themselves be validated before their outputs are used in verification as failure to validate the validator can lead either to unusable measurements or, more seriously, to plausible-looking scores that do not measure the intended construct.

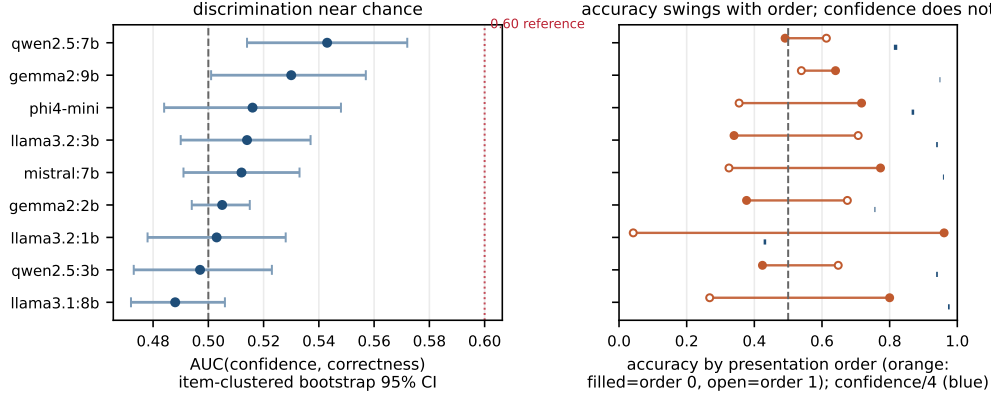


Figure 1: The measured gap on the 9-model cohort. Left: item-clustered bootstrap 95% intervals for AUC(confidence, correctness); two per-model intervals exclude 0.5 and none reaches the 0.60 reference level (dotted). Right: per-model accuracy under the two presentation orders (orange, filled = order 0, open = order 1) against mean stated confidence rescaled to $[0, 1]$ (blue): accuracy swings with order by 0.375 on average while the systematic shift in confidence is only 0.019, showing that stated confidence largely fails to reflect the strong effect of option order on performance.

Scope of the claim. The self-assessment probe was generic, whereas the behavioural measurement was conducted on a single task, so the two are not perfectly matched in scope. A model could show a different relationship between confidence and correctness on other tasks. The cohort is also limited to small open-weight models answering graduate-level science questions near their difficulty ceiling, and the position-dominance finding is not uniform: one model (gemma2:9b) shows content-following and is also one of the two models whose per-model confidence-discrimination interval excludes 0.5. Finally, each model contributed only a single declarative self-assessment. These caveats limit the scope of the result but do not remove the verification problem it demonstrates. On this cohort and task, prospective self-assessment, direct behavioural measurement, and the two automated validators gave different accounts, and only the bias-separating behavioural analysis revealed the strong effect of presentation order.

5 Implications for verification design

Promotion criteria. A metacognitive signal should be promoted to verification use only when it meets, at minimum: (1) *predictive validity*: the declared dimension predicts the behaviour it names on held-out items, not merely on the items used to fit any calibration; (2) *nuisance robustness*: the signal is unchanged by perturbations carrying no epistemic content (the ER probe is such a manipulation), together with the complementary requirement of *interventional sensitivity*, that the signal does move when the construct it names is genuinely manipulated (the PI probe is of this second kind); (3) *bounded dissociation*: the degree of declarative-behavioural disagreement, measured as the rate at which the declared signal and the behaviour it names diverge on the same items, is quantified and below a stated threshold for the deployment; and (4) *estimator-source invariance*: self-produced and external estimates of the same dimension agree within tolerance, which the plug-in estimator interface makes directly testable by swapping estimators on identical inputs. Signals failing these remain useful as *monitoring* telemetry but they should not be automatically promoted into *verification* inputs. The failures in §4 illustrate why these criteria are complementary: direct predictive validation would have exposed the entropy proxy, while explicit measurement of declarative-behavioural disagreement would have exposed the mismatch between prospective self-description and subsequent behaviour.

An audit-design principle. Because dissociation is the failure mode, expose the declared state and the behaviour conditioned on it as *separately inspectable surfaces*: an auditor shown only the self-report cannot detect that behaviour ignored it, and an auditor shown only behaviour cannot detect that the report was confabulated. This follows from the gap itself rather than from any preference about

presentation, and it costs little, since both surfaces can be logged and rendered and the dissociation rate can be audited alongside them.

Falsifiable next steps. The proposed probe battery can be run across a model cohort to estimate the validity of each MSV dimension. Runs that record declared CE and routing decisions on the same items would allow the declarative-behavioural disagreement to be measured directly but the existing logs do not support that comparison. Estimator robustness can be tested by swapping self-produced and external estimators on identical inputs, one dimension at a time. Finally, if the state estimates are genuinely useful for control, state-conditioned routing or assignment should outperform otherwise matched state-blind policies downstream. These tests are designed to separate failures of measurement, estimator choice, and control, rather than treating all negative results as evidence against the framework as a whole. Different failure patterns would also provide evidence for or against the candidate mechanisms in §3.

5.1 Position

Verification systems depend on measurements and, when those measurements are produced by the agents being verified, their validity must also be established. Whatever verifies the agents will utilize signals, and signals an agent produces about itself are instruments that require validation like any other, through behavioural probes, adversarial manipulation, and measured dissociation. We have offered one such framework together with its current scorecard. We would welcome the scrutiny on the two points where the program is most exposed: whether the per-dimension probes measure the constructs they name, and whether the promotion criteria of §5 are sufficient for signals that will sit inside safety-relevant verification loops.

References

- Christopher Ackerman. Evidence for limited metacognition in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2509.21545. URL <https://arxiv.org/abs/2509.21545>. arXiv:2509.21545v2 explicitly identifies this as the ICLR 2026 camera-ready version; introduces the Delegate Game and Second-Chance Game behavioral paradigms for LLM metacognition.
- Anonymous Authors. Declarative vs. behavioral metacognition in large language models: A benchmark on expert questions. NeurIPS 2026 Evaluations & Datasets Track (under review), 2026. Companion submission; details available upon de-anonymization.
- Charles Courchaine, Ricky J. Sethi, and Hefei Qiu. Implementation of a metacognition framework for self-awareness and self-regulation in ensembles of llms. In *Proceedings of The Web Conference (TheWebConf)*. ACM, 2026. Both authors Charles Courchaine and Ricky J. Sethi contributed equally.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint*, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Brian Maniscalco and Hakwan Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430, 2012. doi: 10.1016/j.concog.2011.09.021.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- Ricky J. Sethi, Hefei Qiu, Charles Courchaine, and Joshua Iacoboni. Do llms dream of electric emotions? towards quantifying metacognition and generalizing the teacher-student model using ensembles of llms. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2025.

- Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 2024. URL <https://www.nature.com/articles/s42256-024-00976-7>.
- Guannan Wang, Wenzhe Wu, Guanxiong Ye, Zhaoxuan Cheng, Xuan-Yi Chen, and Hang Zheng. Decoupling metacognition from cognition: A framework for quantifying metacognitive ability in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25353–25361, 2025.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2306.13063>.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=shr9PXz7T0>. arXiv:2309.03882; ICLR 2024 Spotlight.