THINKING WITH METACOGNITION: REBALANCED EXPLORATION-EXPLOITATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent advances in reasoning capabilities, Large Language Models (LLMs) often exhibit strikingly greedy behavior, exploring far less than humans in sequential decision-making tasks. We investigate this discrepancy across three bandit settings with human reference data. To characterize the underlying choice dynamics, we adapt a computational model of human exploration-exploitation (EE) behavior. Across ten LLMs, we find that their native reasoning (*i.e.*, Thinking mode) amplifies sensitivity to estimated value, causing value differences to dominate choice while limiting the strength and persistence of uncertainty-directed exploration. Analyses of reasoning traces reveal that LLMs tend to compare value estimates and translate provisional judgments directly into choices, with little assessment of their reliability or evidential support. Such assessment and regulation of one's own judgments constitute *Metacognitive* monitoring and control, whose limited engagement gives rise to a "value-to-action" shortcut. Motivated by this diagnosis, we scaffold metacognitive reasoning in LLMs by prompting them to assess value, uncertainty, confidence, and whether the current state calls for exploration or exploitation, and to use these assessments to regulate choice. Across LLMs and experimental settings, metacognitive reasoning rebalances sensitivity to value and uncertainty, promoting early uncertainty-directed exploration and improving subsequent exploitation. These gains persist in more challenging long-horizon four- and five-arm tasks, where improved optimal-arm selection and cumulative reward show that metacognitive reasoning can support adaptive exploration-exploitation over extended decision horizons. An overview of the study, visualizations, and interactive results are available at this *link*.

1 INTRODUCTION

LLMs are increasingly deployed as agents in sequential decision-making settings, where effective behavior requires balancing immediate reward with information acquisition that can improve future choices (Wilson et al., 2021). Yet recent studies show that LLMs often behave greedily, exploring little and committing prematurely based on limited evidence (Krishnamurthy et al., 2024, Schmied et al., 2026, Chen et al., 2026). Such behavior can restrict information gathering, reduce option coverage, and lead to premature lock-in to suboptimal choices (Figure 1, top).

Why LLMs underexplore remains poorly understood. Greedy behavior alone does not identify its underlying mechanism. An agent may underexplore because it is insufficiently responsive to uncertainty, places excessive weight on current value estimates, or fails to regulate how these signals influence choice as evidence accumulates. Prior work shows LLM exploration can be altered through prompting, fine-tuning, or explicit algorithmic support (Nie et al., 2025, Schmied et al., 2026). Without identifying the source of this imbalance, however, it remains unclear what these interventions correct or whether the same mechanism holds across models and decision settings.

To distinguish these possibilities, we evaluate ten LLMs across three two-armed bandit settings with human reference data. We adapt a computational model of human exploration and exploitation (EE) behavior that separates value-related and uncertainty-related influences on trial-level choice (Gershman, 2018). This decomposition allows us to distinguish whether limited exploration arises from weak sensitivity to uncertainty or excessive reliance on estimated value. Our evaluation spans multiple model families and both open- and closed-source systems, yielding over 200,000 sequential choices in each experimental condition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

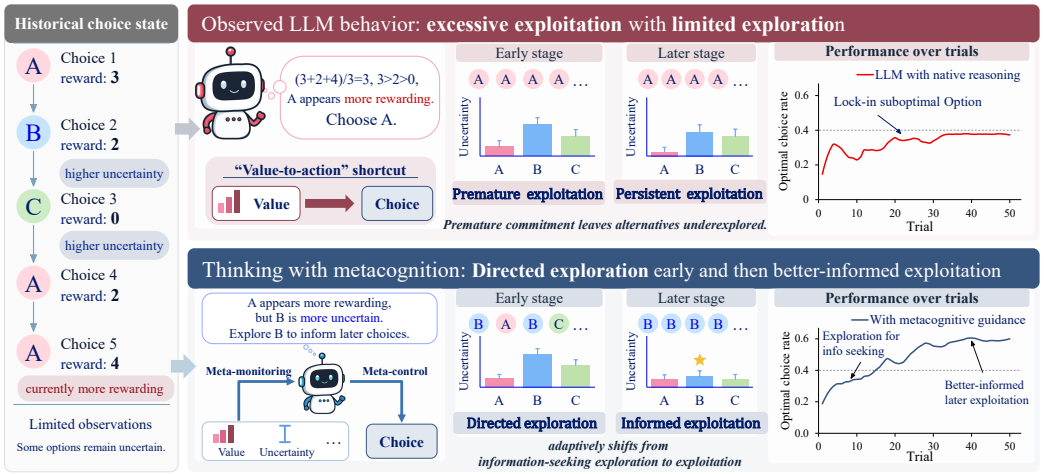


Figure 1: Rebalancing LLM exploration and exploitation through metacognition. Native reasoning can become value-dominated and commit prematurely under unresolved uncertainty. Metacognitive reasoning promotes early uncertainty-directed exploration and improves later exploitation.

Our analysis reveals a consistent value and uncertainty imbalance across LLMs. Relative uncertainty influences choice when uncertainty differences are pronounced, but its influence is quickly overwhelmed as value evidence emerges. Within-model comparisons show that native reasoning strengthens both signals asymmetrically: sensitivity to uncertainty increases, but sensitivity to estimated value increases more, shifting control of choice toward value. As a result, uncertainty-directed exploration is confined to an early window before value-dominated choice takes over.

This behavioral pattern is mirrored in the reasoning process itself. Reasoning traces show that LLMs frequently focus on forming and comparing value estimates, then translate provisional value judgments directly into choices with little assessment of their reliability or evidential support. We refer to this pattern as a "value-to-action" shortcut. Viewed through the framework of metacognition, it suggests limited engagement of monitoring and control: the processes by which an agent evaluates its own judgments and regulates how strongly those judgments guide action (Nelson, 1990). This perspective connects the observed choice imbalance to a specific limitation in how native reasoning is regulated, rather than to an absence of uncertainty representation itself.

Motivated by this diagnosis, we scaffold metacognitive reasoning with minimal, task- and model-independent prompts that require LLMs to assess value, uncertainty, confidence, and whether the current state calls for exploration or exploitation, and to use these assessments to regulate choice. Across models and settings, this guidance rebalances sensitivity to value and uncertainty, promoting early uncertainty-directed exploration that reduces unresolved uncertainty and improves subsequent exploitation (Figure 1, bottom). We further evaluate this intervention in more challenging long-horizon settings, including a four-arm, 50-trial task and an established five-arm, 100-trial exploration benchmark. The same adaptive pattern persists: metacognitive reasoning promotes earlier information acquisition, reduces premature lock-in, and supports stronger later exploitation, leading to improved optimal-arm identification and cumulative reward. Gains appear across nearly all tested models in the four-arm task and relative to strong prompting baselines in the five-arm setting.

2 Related Work

Exploration-exploitation in LLMs. As LLMs are increasingly used to empower decision-making agents that adapt through sequential interaction and reward feedback (Monea et al., 2024, Song et al., 2026), their ability to balance information gathering with reward seeking has become a central challenge. Early work found that GPT-3 (Binz and Schulz, 2023) demonstrated little directed exploration, and systematic bandit evaluations later showed that robust in-context exploration was difficult to elicit across models and prompting configurations (Krishnamurthy et al., 2024). Beyond insufficient directed exploration, LLMs also exhibit limitations in when and how they commit to exploitation. In dueling bandits, strong LLMs can identify promising actions early yet fail to converge

to consistent exploitation (Xia et al., 2025). These findings point to limitations in both directed exploration for information acquisition and the subsequent commitment to exploitation.

Recent studies have reported seemingly contradictory effects of reasoning on LLM EE behavior across decision settings. In sequential bandits, Zhang et al. (2026) find that reasoning shifts LLMs toward more human-like exploration in stationary settings. However, this apparent benefit does not extend uniformly to information acquisition. Guiomar et al. (2026) show that, when evidence acquisition and integration are separated, reasoning yields substantially larger gains in inference than in active sampling. Consistent with this asymmetry, Xiong et al. (2026) find under controlled histories that inference-time reasoning strengthens value-guided choice without increasing measured exploration signatures at the first free choice. These findings leave the role of reasoning in LLM EE behavior unclear, with seemingly contradictory effects across settings warranting further investigation.

Intervention studies show that exploration can be improved through algorithm-guided inference and distillation (Nie et al., 2025), explicit posterior-sampling implementations (Arumugam and Griffiths, 2026), and RL fine-tuning (Schmied et al., 2026). Yet better post-training performance can coexist with greedier policies and premature abandonment of exploration (Chen et al., 2026), showing that aggregate outcomes need not reveal the underlying EE strategies. Complementary evaluations explicitly separate exploration from exploitation (Grams et al., 2025, Park et al., 2026). These results highlight a remaining gap between improving decision outcomes and understanding how the underlying exploration-exploitation process is altered.

Metacognitive theory and guidance. Metacognitive monitoring evaluates ongoing cognition, whereas metacognitive control uses these assessments to regulate subsequent behavior (Nelson, 1990). In human decision making, confidence and metacognitive monitoring have been linked to exploration and information search (Boldt et al., 2019, Schulz et al., 2023). Recent work distinguishes metacognitive sensitivity and calibration from first-order task performance (Steyvers and Peters, 2026), while frameworks such as DMC seek to disentangle metacognitive ability from cognitive performance (Wang et al., 2025). Existing approaches instantiate different aspects of metacognition through skill-aware demonstration selection (Didolkar et al., 2024), structured self-assessment (Wang and Zhao, 2024), verbal reflection over feedback (Shinn et al., 2023), learned strategic oversight (Wan et al., 2025), and explicit planning-monitoring-evaluation cycles (Elenjical et al., 2026). Against this background, we examine how metacognitive guidance within native reasoning affects sequential exploration-exploitation. We compare monitoring alone with monitoring plus control to test their effects on early information acquisition and subsequent exploitation.

3 Cognitive Modeling and Experimental Design

Computational Model and Experimental Design. We repurpose the computational model Gershman (2018), developed to characterize how humans resolve the EE dilemma, for probing LLM EE strategies. The model formalizes human exploration strategies through distinct effects of uncertainty on value-based choice. We leverage this structure as an interpretable decomposition of value (an estimate of expected reward)- and uncertainty-related influences on choice.

Three model-matched bandit settings are used for analysis and evaluation, corresponding respectively to Experiments 1 and 2 of Gershman (2018) and the task of Gershman (2019). The settings differ in their reward-uncertainty structure: Setting 1 pairs a stochastic arm with a fixed-zero arm, Setting 2 uses two stochastic arms, and Setting 3 combines risky (R) and safe (S) arms across RS, SR, RR, and SS configurations, enabling controlled variation in relative and total uncertainty (Appendix A.1).

Following Gershman (2018), a Bayesian observer reconstructs, before choice t, the posterior mean $Q_t(k)$ and posterior standard deviation $\sigma_t(k)$ for arm $k \in \{1, 2\}$ (Appendix B.1). Here, $Q_t(k)$ represents the arm's current value estimate. These quantities yield

$$V_t = Q_t(1) - Q_t(2), \qquad \mathrm{RU}_t = \sigma_t(1) - \sigma_t(2),$$
$$\mathrm{TU}_t = \sqrt{\sigma_t^2(1) + \sigma_t^2(2)}, \qquad \mathrm{VTU}_t = \frac{V_t}{\mathrm{TU}_t}, \tag{1}$$

where V_t, RU_t, TU_t, and VTU_t denote value difference, relative uncertainty, total uncertainty, and uncertainty-normalized value difference, respectively.

Trial-level choice $a_t \in \{1, 2\}$ is modeled as

$$P(a_t = 1) = \begin{cases} \Phi\left(\beta_{\mathrm{RU}}\widetilde{\mathrm{RU}}_t + \beta_{\mathrm{VTU}}\widetilde{\mathrm{VTU}}_t\right), & \text{Settings 1 and 2,} \\ \Phi\left(\beta_V\widetilde{V}_t + \beta_{\mathrm{RU}}\widetilde{\mathrm{RU}}_t + \beta_{\mathrm{VTU}}\widetilde{\mathrm{VTU}}_t\right), & \text{Setting 3,} \end{cases} \tag{2}$$

where Φ is the standard normal cumulative distribution function. To make these coefficients directly comparable across humans and LLMs, we standardize the decision variables within each setting using the corresponding human data and apply the same fixed transformation to all humans, LLMs, and experimental conditions, following the z-scoring procedure used in Gershman (2019) (Appendix B.2).

The fitted coefficients can characterize choice sensitivity to the corresponding decision variables: β_{RU} to relative uncertainty, β_{VTU} to uncertainty-normalized value evidence, and β_V to direct value evidence in Setting 3. Larger positive coefficients indicate stronger choice sensitivity to the corresponding evidence. For β_{RU}, positive values indicate that relative uncertainty shifts choice toward the more uncertain option, whereas negative values indicate uncertainty avoidance.

Tasks and Evaluation Protocol. Following the source paradigms of Gershman (2018, 2019), each participant completes multiple independent 10-trial blocks: 20 blocks in Settings 1 and 2 and 30 blocks in Setting 3, to maximize cumulative reward. On each trial, the participant observes only the reward from the selected arm. Each new block resets both the block-specific arm parameters and the participant's choice-reward history.

Our evaluation covers ten open- and closed-source LLMs spanning multiple model families and scales, yielding 217,600 decisions across the three settings per experimental condition. Model configurations, prompts, sample sizes, and replication accounting are detailed in Appendices A.3-A.5. For each participant, Eq. 2 is fitted once across all blocks, yielding a participant-level set of β coefficients. Combining these fitted sensitivities with trial-varying decision variables allows us to quantify the realized contributions of value and uncertainty pathways within the fitted policy at each choice.

4 Native Reasoning Drives Value-Dominated Choice

LLMs exhibit greedy behavior. Across all three of our settings, LLMs are consistently observed to exhibit greedy choice behavior. Specifically, we define an empirical-greedy choice as selecting the arm with the higher historical sample mean, considering only trials after both arms have been sampled and their historical sample means differ. Under this definition, the empirical-greedy choice rate under native reasoning averages 98.9% across models and settings (Appendix C.3.1, Table C.2). This observation is consistent with recent reports of greedy choice and limited exploration in LLMs (Krishnamurthy et al., 2024, Schmied et al., 2026, Chen et al., 2026). Cross-task and cross-model greedy decision making suggests a systematic limitation in LLM decision making or EE strategies, making it important to uncover the mechanisms underlying this behavior.

Computational diagnosis traces LLM greediness to Thinking-driven excessive value sensitivity. To uncover the computational basis of LLM greediness, we quantify model sensitivities to value and relative uncertainty using the model introduced in Section 3. Across ten LLMs and all three settings, LLMs exhibit a fundamental departure from human behavior: they possess a comparable sensitivity to relative uncertainty (β_{RU}), but a markedly inflated sensitivity to value evidence (β_{VTU} and β_V; Appendix C.1.1). Trial-level decompositions further show that this asymmetry translates into a dominant influence of value evidence on choice (Appendix C.1.2). These results suggest that the pronounced greediness of LLMs is driven primarily by disproportionately strong sensitivity to value evidence rather than by a lack of responsiveness to relative uncertainty.

Within-model comparisons identify native reasoning as the primary driver of this excessive value sensitivity. Enabling Thinking markedly inflates β_{VTU} and β_V (Figure 2A; Appendix C.3.2), accounting for the magnified value dominance on choice (Figure 2B). Notably, while native reasoning also improves relative-uncertainty responsiveness (β_{RU}), this gain is heavily asymmetric-the increase

in value sensitivity is vastly disproportionate . Consequently, while native reasoning marginally aids initial exploration when evidence of relative uncertainty is strong (as reflected in faster coverage of both arms across our settings), it swiftly locks choices onto current value estimates, leaving only a narrow window for uncertainty to guide decision-making (Figure C.5; Appendix C.3.4).

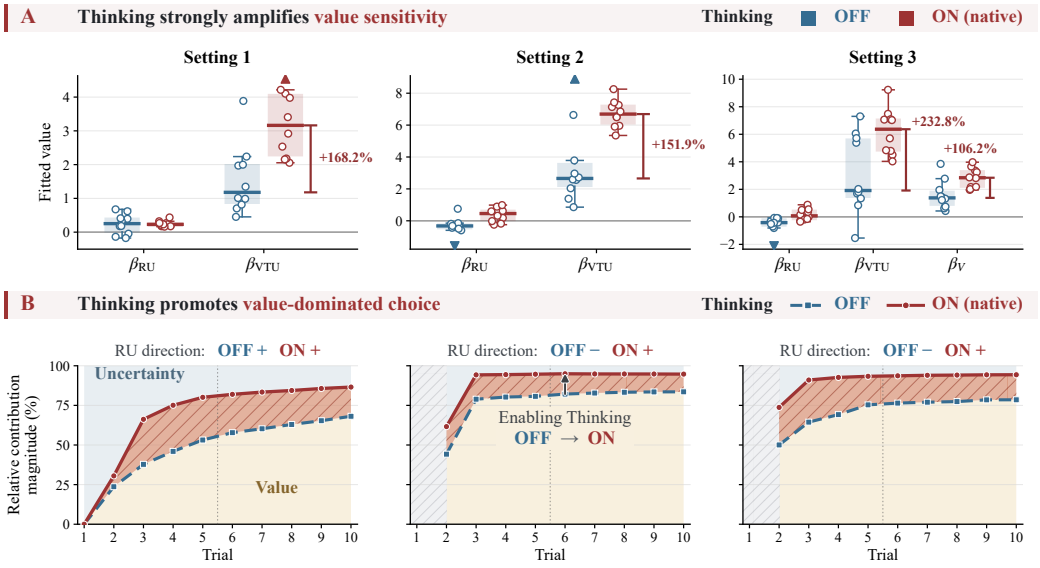


Figure 2: **Native reasoning strengthens value sensitivity and establishes early value dominance.** **A**: Fitted parameters across models. **B**: Relative magnitudes of trial-varying value and relative-uncertainty contributions. RU signs indicate the aggregate direction toward (+) or away from (−) the more uncertain arm. See Appendices B.2 and B.3 for details.

Reasoning traces expose a value-to-action shortcut. To inspect how value dominance manifests in visible reasoning, we analyze 960 reasoning traces uniformly sampled across the Qwen3 family, DeepSeek, and GLM-5.1 under our three task settings (trials 3-10). As shown in Figure 3, we identify two dominant reasoning pathways among traces with an identifiable value winner. First, most traces (68.3%) exhibit a direct *value-to-action shortcut*: reasoning focuses on historical reward comparisons and proceeds straight from value identification to choice execution. In a second, more revealing pattern (19.9%), models explicitly mention or assess uncertainty, yet ultimately return to value comparison and choose according to the provisional value winner (full analysis in Appendix D.1). Taken together, these pathways reveal a tendency for native reasoning to translate provisional value judgments into action through direct or short reasoning pathways, with limited regulation of when and how such judgments should guide action.

5 Metacognitive Guidance Enables Adaptive Exploration

Bridging the metacognitive gap via structured scaffolding. Classic metacognitive theory distinguishes object-level cognitive operations from a higher-order meta-level that monitors and regulates them (Nelson, 1990). Viewed through this lens, native reasoning exhibits competent object-level value reasoning, yet lacks structured meta-level regulation-which metacognitive frameworks decompose into *monitoring* and *control* (Nelson, 1990). This architectural gap underpins the two dominant reasoning pathways identified in Section 4: a monitoring deficit that produces direct value shortcuts, and a control failure where verbalized uncertainty fails to restrain action selection. Naturally, this diagnosis motivates a principled intervention: explicitly scaffolding metacognitive monitoring and control directly within the LLM's reasoning trace.

Guiding LLM reasoning through metacognitive monitoring and control. To investigate how establishing metacognitive processes affects LLM decision-making (particularly EE behavior), we introduce lightweight guidance involving metacognitive monitoring and control under a shared

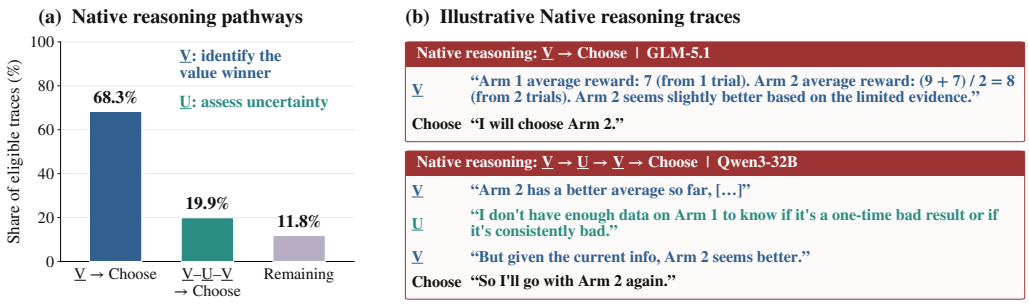


Figure 3: **Native reasoning pathways and illustrative thinking traces. (a)** Proportions of reasoning pathways among traces with an identifiable value winner. **(b)** Trace excerpts illustrating direct value choice ($\underline{\text{V}} \rightarrow \text{Choose}$) and value overriding uncertainty ($\underline{\text{V}} \rightarrow \underline{\text{U}} \rightarrow \underline{\text{V}} \rightarrow \text{Choose}$). $\underline{\text{V}}$ and $\underline{\text{U}}$ denote value-winner identification and uncertainty assessment, respectively.

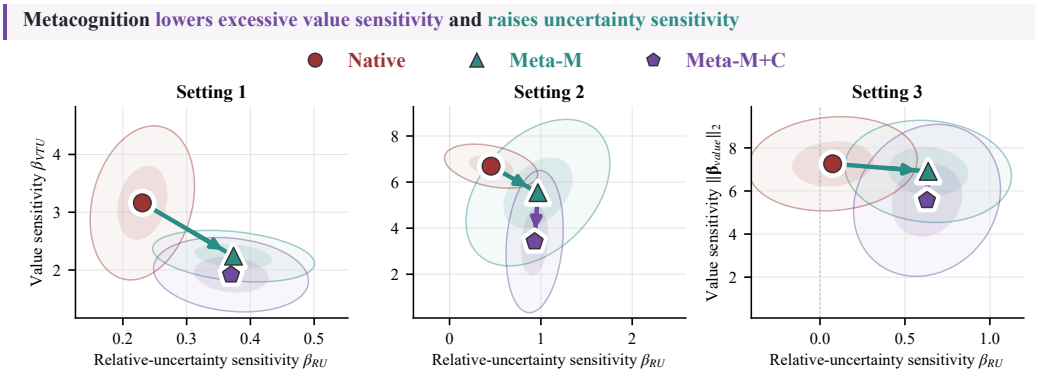


Figure 4: **Fitted choice sensitivities under metacognitive guidance.** Cross-model summaries across Native, Meta-M, and Meta-M+C for the three settings. The horizontal axis shows β_{RU}. The vertical axis shows β_{VTU} in Settings 1-2 and $\sqrt{\beta_V^2 + \beta_{\text{VTU}}^2}$ in Setting 3. Calculation and aggregation methods are in Appendix B.2.

protocol across all models and settings (Appendix E.1). Specifically, **Meta-M** (Monitoring) asks the model to explicitly assess current value estimates, uncertainty, whether the current state calls for exploration or exploitation, and confidence. **Meta-M+C** (Monitoring + Control) additionally instructs the model to use these assessments to regulate subsequent choice. Full prompts are provided in Appendix E.1. The guidance is designed to encourage metacognitive monitoring and control during reasoning rather than prescribe specific actions, and is therefore task-agnostic.

Rebalancing value-uncertainty sensitivities. Figure 4 shows that metacognitive guidance increases sensitivity to relative uncertainty (β_{RU}) while reducing value-related sensitivity (β_{VTU} and β_V) (model-level results are reported in Appendix E.3.1, Table E.2). For LLMs characterized by excessive value sensitivity, this shift rebalances the relative influence of value and uncertainty on choice. Figure 4 also reveals distinct roles for metacognitive monitoring and control: monitoring (**Meta-M**) accounts for most of the increase in relative-uncertainty sensitivity while also reducing value sensitivity, whereas adding control (**Meta-M+C**) further attenuates value-related sensitivity. Further analysis reveals substantial model differences in how monitoring assessments affect subsequent reasoning: for some LLMs, monitoring alone disrupts the "value-to-action" shortcut; for others, explicit control is required for these assessments to influence action (Appendix D.4). Beyond sensitivity changes, we use state-sensitivity decomposition to examine the evidence encountered along decision trajectories. This analysis shows that metacognitive guidance reduces the influence of value on choice, even when value evidence remains strong or increases (Appendix E.3.2).

Accelerating option coverage and reducing empirical greediness. Metacognitive guidance also produces consistent behavioral changes across models and settings. As shown in Table 1, **Meta-M**

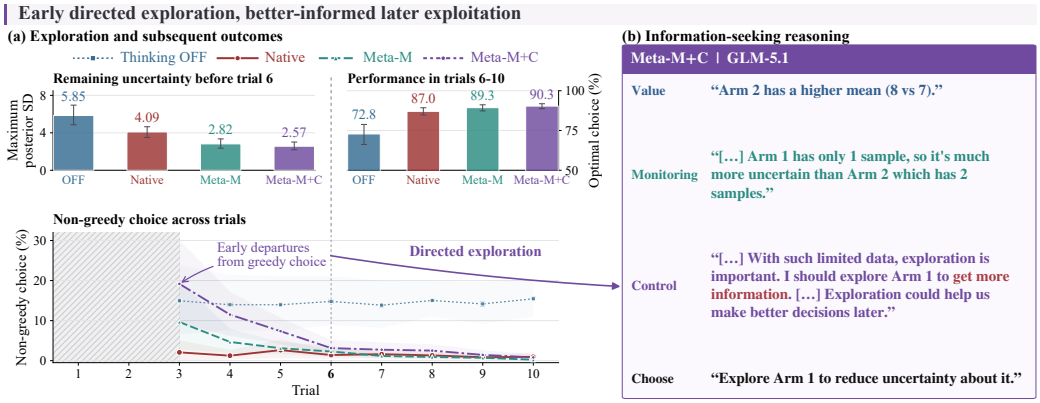


Figure 5: **(a)** Setting 3 exploration dynamics and subsequent outcomes across ten LLMs. Uncertainty is measured by the maximum posterior SD of the two arms' mean rewards before trial 6 (Appendix B.5). Results are model-equal means with 95% bootstrap intervals. **(b)** GLM-5.1 reasoning trace with metacognitive guidance, paired with Native in Figure 3 (Appendix D.3).

improves option coverage, achieves earlier first coverage, and increases non-greedy choice rates (i.e., more frequent departures from the current empirical-value winner). Adding metacognitive control (**Meta-M+C**) further strengthens these effects, especially in reducing empirical greediness (model-level results are given in Appendix E.3.1, Table E.2). However, departure from the current empirical-value winner alone does not establish **directed exploration**, which requires choices to be oriented toward uncertainty for information acquisition. Also, the stage at which these departures occur is equally important. Although the higher final-trial optimal-choice rates under metacognitive guidance indicate a more effective EE strategy, revealing this requires a trial-level analysis.

Table 1: **Behavioral changes under metacognitive guidance across ten LLMs.** Model-equal means across ten LLMs. The four metrics report, respectively, the percentage of blocks in which both arms are sampled, the mean trial at which both arms are first covered among those blocks, non-greedy choice rate (%), and the percentage of choices selecting the arm with the higher expected reward on the final trial. Detailed definitions are in Appendix B.5.

	Setting 1				Setting 2				Setting 3			
Prompt	Cover ↑	First ↓	Non-greedy	Optimal rate (%) ↑	Cover ↑	First ↓	Non-greedy	Optimal rate (%) ↑	Cover ↑	First ↓	Non-greedy	Optimal rate (%) ↑
Native	99.99	2.08	1.71	98.81	85.89	2.58	0.25	97.32	78.99	2.46	1.47	87.68
Meta-M	100.00	2.01	3.96	96.90	99.31	2.21	2.91	98.96	93.95	2.20	2.76	90.42
Meta-M+C	100.00	2.00	6.00	98.89	99.51	2.10	6.45	99.07	95.73	2.17	5.97	91.89

Enabling early directed exploration and better-informed exploitation. We focus on Setting 3, where contrasting reward variability across risky and safe arms provides a clear basis for quantifying uncertainty about expected rewards, to examine trial-level behavior under metacognitive guidance. The bottom panel of Figure 5(a) shows the trial-wise distribution of non-greedy choices, revealing a clear temporal reorganization: under metacognitive guidance, departures from the current empirical-value winner are concentrated in the early stage. Corresponding reasoning traces, illustrated in Figure 5(b), reveal that these non-greedy choices reflect deliberate departures from the current value winner toward a more uncertain option for information gathering, providing direct evidence of uncertainty-directed exploration. Correspondingly, after the first five trials, Meta-M+C leaves the least remaining uncertainty among the tested conditions and is followed by higher optimal-choice performance in trials 6-10 (Figure 5(a), upper panels).

In contrast, although Thinking OFF also produces frequent non-greedy choices, these departures persist across decision stages and are less consistently aligned with uncertainty reduction. Consequently, non-greedy behavior under Thinking OFF is associated with poorer performance, rather than the gains observed when such departures are structured by metacognitive guidance. Together, these results show that metacognitive guidance moves LLMs beyond simple empirical greediness

and induces a more adaptive EE strategy: uncertainty-directed exploration is concentrated when information remains valuable, followed by better-informed exploitation.

6 Evaluating Metacognitive Guidance over Longer Horizons

This section evaluates whether the adaptive EE strategy induced by metacognitive guidance can *generalize to more complex environments and longer horizons*, and whether it improves optimal-arm identification and translates into higher cumulative reward. To test this, we use a four-arm, 50-trial bandit task in which all ten LLMs complete 10 independent blocks under both Native reasoning and Meta-M+C. The two conditions are evaluated on matched task instances with identical pre-generated payoff sequences for corresponding arms and blocks, controlling for differences in realized rewards. Compared with the two-arm tasks, the shared prompt provides a more explicit description of the arms' reward-generating structure while withholding their expected values and the identity of the optimal arm. Task specifications and exact prompts are detailed in Appendices F.1.1-F.1.2.

Metacognitive guidance generalizes through adaptive regulation. Overall, the adaptive EE behavior induced by Meta-M+C persists in the four-arm setting. It increases all-arm coverage from 79% to 87% and accelerates first coverage from trial 10.6 to 6.9, while promoting early directed exploration and reducing the tendency toward premature commitment to the current empirical-value winner. These changes are followed by substantially better late exploitation, achieving an optimal-arm choice rate of 59.2% over the final ten trials compared with 37.7% under Native reasoning ($p = .004$), together with higher cumulative reward (300.5 vs. 282.0, $p = .010$). Figure 6 shows the corresponding model-wise optimal-choice trajectories and cumulative rewards aggregated over ten blocks, with nearly all tested LLMs improving under metacognitive guidance.

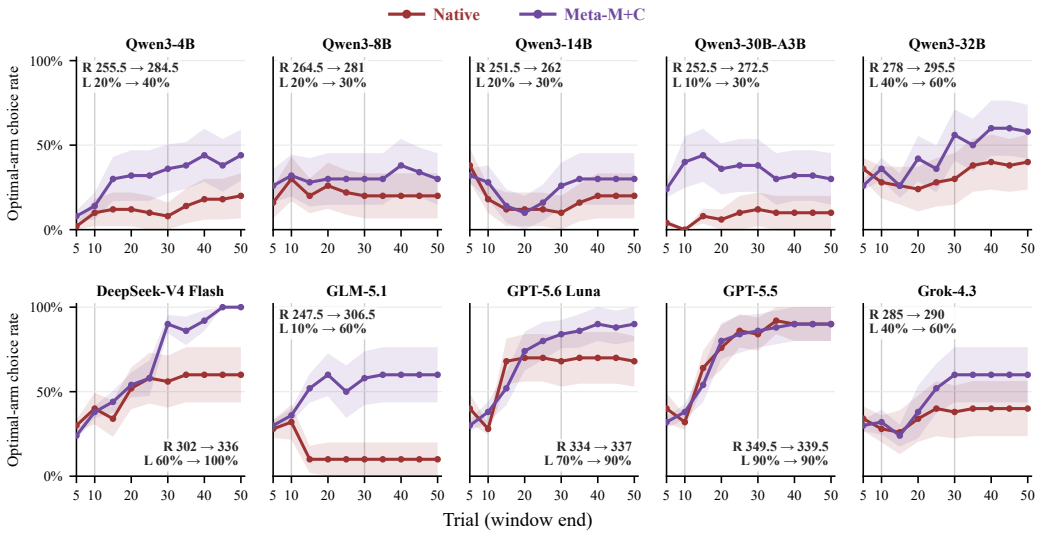


Figure 6: **Optimal-choice trajectories and cumulative reward over ten blocks.** Optimal-arm choice rates in non-overlapping five-trial windows, plotted at window ends. Shading indicates ±1 SEM across ten blocks. R and L denote mean total reward and late-exploitation success rate, respectively. Both prompts retain native reasoning. Definitions are in Appendix B.6.

Notably, in this four-arm setting, the effects of metacognitive guidance exhibit both model- and stage-dependent variation (Appendix F.1.3- F.1.4). Specifically, Meta-M+C primarily increases early departures from the empirical-value winner in strongly greedy models (e.g., Grok-4.3), while producing much smaller changes in models that already exhibit substantial early non-greedy behavior (e.g., GPT-5.5). For models that continue to show relatively high non-greedy behavior later in the horizon (e.g., Qwen3-14B), Meta-M+C can even reduce late non-greedy choice, strengthening subsequent commitment. These heterogeneous adjustments indicate that metacognitive guidance does not simply increase exploration, but adaptively regulates when additional information should be acquired and when current evidence is sufficient for commitment. Trial-wise self-reported decision-

mode trajectories further illustrate how LLMs under Meta-M+C adapt exploration and exploitation to the demands of different task instances (Appendix F.1.6).

To illustrate the source of the long-horizon performance gains, we present a representative decision trajectory in Figure 7A, showing two distinct decision patterns: under Native reasoning, the model rapidly commits to a suboptimal arm after limited sampling; under Meta-M+C, increased early information-seeking exploration supports more reliable optimal-arm identification and helps avoid premature lock-in to a suboptimal option (Additional cases are provided in Appendix F.1.5).

Validation on an established exploration task. Krishnamurthy et al. (2024) introduced a challenging five-arm Bernoulli bandit and identified a prompting configuration that enabled GPT-4 to avoid persistent exploration failure, providing a strong baseline for comparison. We compare Meta-M and Meta-M+C against two baselines: the prompt underlying that configuration (K24 prompt), and a behavioral-intervention baseline introduced here (EE-balance), inspired by Krishnamurthy et al. (2024) and explicitly instructing the model to balance exploration and exploitation. All four prompts are evaluated in five matched, 100-trial GPT-5.6 Luna runs, with identical arm-specific payoff sequences and otherwise matched settings (all use the provider-default temperature, which the available API does not allow us to adjust; Appendix F.2). Under their suffix-failure criterion, the paper-derived prompt fails in only one run, whereas both metacognitive prompts avoid failure in all five (as does EE-balance). Beyond this criterion, as shown in Figure 7B, metacognitive guidance yields higher post-coverage non-empirical-greedy choice, optimal-arm selection, and mean cumulative reward than either baseline. These results highlight the strong ability of metacognitive guidance to activate adaptive exploration in LLMs (Block-wise results are provided in Appendix F.2.2).

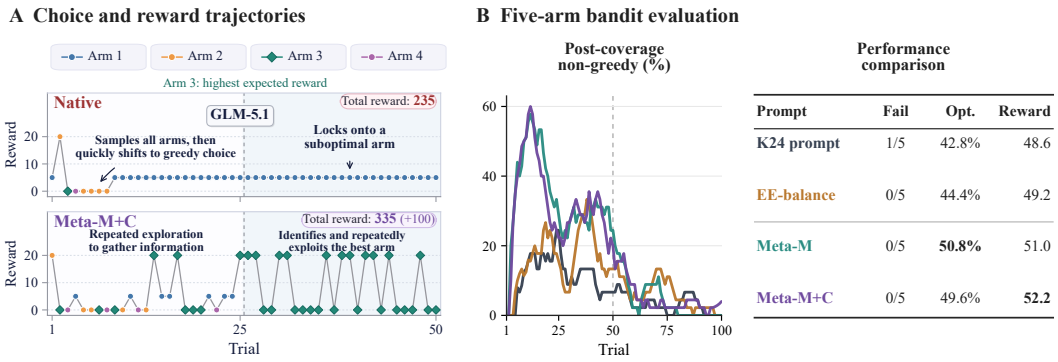


Figure 7: **Choice trajectories and five-arm evaluation. A** Native and Meta-M+C in one matched 50-trial GLM-5.1 block. **B** Post-coverage non-greedy choice (centered nine-trial windows) and performance across five matched 100-trial GPT-5.6 Luna runs per prompt. Fail: suffix failures; Opt.: all-trial optimal-choice rate; Reward: mean total reward.

7 Conclusion

This study identifies a key source of the greedy decision-making widely observed in LLMs: limited metacognitive processing within native reasoning creates a short "value-to-action" pathway, allowing provisional value judgments to exert disproportionate control over subsequent choices. By integrating classic metacognitive theory into LLM reasoning through lightweight guidance for monitoring and control, we observe a systematic shift toward more adaptive EE behavior. Specifically, metacognitive guidance promotes uncertainty-directed information acquisition early in the decision process and adaptively transitions toward better-informed exploitation as uncertainty is reduced. This helps alleviate behavioral failures such as limited option coverage and premature lock-in to suboptimal choices, while improving performance over longer decision horizons.

Although metacognitive guidance yields benefits across the tested settings, two limitations should be noted. First, effective metacognitive guidance depends on the model's ability to reliable value estimates and identify uncertainty, which supports uncertainty-directed exploration early and effective exploitation later. (We discuss these limitations in Appendix G). Second, our evidence is currently limited to multi-armed bandit tasks, and whether the same mechanisms and benefits extend to more complex sequential decision environments requires further investigation.

AI USE DISCLOSURE

Generative AI tools were used both as the object of study and for manuscript preparation. LLM interaction data were collected through local deployment and API-based access. Generative AI also assisted in drafting initial text for parts of the Introduction and experimental Results sections and in language editing. The authors reviewed and revised all AI-assisted content and are responsible for the scientific claims, analyses, figures, references, and final manuscript.

REPRODUCIBILITY STATEMENT

We provide the experimental details needed to reproduce our analyses in the appendices. These include task construction and reward-generation protocols, model and inference configurations, the Bayesian belief-updating and choice-model fitting procedures, definitions of all behavioral and computational measures, complete prompts and response formats, aggregation and statistical procedures, and model-level results. The extended-horizon evaluations additionally specify the matched reward sequences, task protocols, and evaluation criteria used in the four- and five-arm experiments. Code and data are available in an anonymized repository at https://anonymous.4open.science/r/thinking-with-metacognition-4AB0/.

ETHICS STATEMENT

This study uses publicly released human behavioral datasets from the cited source studies and interaction data generated by language models. No new human participants were recruited, and all analyses of human behavior were conducted on previously collected data.

REFERENCES

Dilip Arumugam and Thomas L Griffiths. Toward efficient exploration by large language model agents. In *International Conference on Learning Representations*, volume 2026, pages 40734–40772, 2026.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL https://doi.org/10.1073/pnas.2218523120.

Annika Boldt, Charles Blundell, and Benedetto De Martino. Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of consciousness*, 2019(1):niz004, 2019. doi: 10.1093/nc/niz004. URL https://academic.oup.com/nc/article/2019/1/niz004/5486647.

Sanxing Chen, Xiaoyin Chen, Yukun Huang, Roy Xie, and Bhuwan Dhingra. When greedy wins: Emergent exploitation bias in meta-bandit llm training. In *International Conference on Learning Representations*, volume 2026, pages 57389–57421, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5e24b2a1d5acc5da7d9144121577b6a0-Paper-Conference.pdf.

Aniket Didolkar, Anirudh Goyal, Nan R Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024. doi: 10.52202/079017-0623. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2318d75a06437eaa257737a5cf3ab83c-Paper-Conference.pdf.

Abraham Paul Elenjical, Vivek Hruday Kavuri, and Vasudeva Varma. Think^2: Grounded metacognitive reasoning in large language models. *arXiv preprint arXiv:2602.18806*, 2026. doi: 10.48550/arXiv.2602.18806. URL https://arxiv.org/abs/2602.18806.

Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018. doi: 10.1016/j.cognition.2017.12.014. URL https://doi.org/10.1016/j.cognition.2017.12.014.

Samuel J Gershman. Uncertainty and exploration. *Decision*, 6(3):277, 2019. doi: 10.1037/dec0000101. URL https://doi.org/10.1037/dec0000101.

Tim Grams, Patrick Betz, Sascha Marton, Stefan Lüdtke, and Christian Bartelt. Disentangling exploration of large language models by optimal exploitation. *arXiv preprint arXiv:2501.08925*, 2025.

Gonçalo Guiomar, Elia Torre, Pehuen Moure, Victoria Shavina, Mario Giulianelli, Shih-Chii Liu, and Valerio Mante. Reasoning aligns language models to human cognition. *arXiv preprint arXiv:2602.08693*, 2026. doi: 10.48550/arXiv.2602.08693. URL https://arxiv.org/abs/2602.08693.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024. doi: 10.52202/079017-3818. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d951f73c521d069fefbb73396df01424-Paper-Conference.pdf.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. Llms are in-context bandit reinforcement learners. *arXiv preprint arXiv:2410.05362*, 2024.

Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pages 125–173. Elsevier, 1990.

Allen Nie, Yi Su, Bo Chang, Jonathan Lee, Ed H. Chi, Quoc V Le, and Minmin Chen. EVOLve: Evaluating and optimizing LLMs for in-context exploration. In *Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46346–46376. PMLR, 2025. URL https://openreview.net/forum?id=ck7dvZFbRW.

Jaden Park, Jungtaek Kim, Jongwon Jeong, Robert D Nowak, Kangwook Lee, and Yong Jae Lee. Exploration and exploitation errors are measurable for language model agents. *arXiv preprint arXiv:2604.13151*, 2026.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. Llms are greedy agents: Effects of rl fine-tuning on decision-making abilities. In *International Conference on Learning Representations*, volume 2026, pages 127538–127571, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cf2fcda40de8520027af2eadcafdecd6-Paper-Conference.pdf.

Lion Schulz, Stephen M Fleming, and Peter Dayan. Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3):604, 2023. doi: 10.1037/rev0000401. URL https://doi.org/10.1037/rev0000401.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Shangtong Zhang, and Yanjun Qi. Reward is enough: Llms are in-context reinforcement learners. In *International Conference on Learning Representations*, volume 2026, pages 112747–112770, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b7511dfe2e7a1fa45e093cc75389abc2-Paper-Conference.pdf.

Mark Steyvers and Megan AK Peters. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, 35(3):131–139, 2026. doi: 10.1177/09637214251391158. URL https://journals.sagepub.com/doi/10.1177/09637214251391158.

Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, et al. Rema: Learning to meta-think for llms with multi-agent reinforcement learning. *arXiv preprint arXiv:2503.09501*, 2025. doi: 10.48550/arXiv.2503.09501. URL https://arxiv.org/abs/2503.09501.

Guoqing Wang, Wen Wu, Guangze Ye, Zhenxiao Cheng, Xi Chen, and Hong Zheng. Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25353–25361, 2025. doi: 10.1609/aaai.v39i24.34723. URL https://ojs.aaai.org/index.php/AAAI/article/view/34723.

Yuqing Wang and Yun Zhao. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.106. URL https://aclanthology.org/2024.naacl-long.106/.

Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences*, 38:49–56, 2021. doi: 10.1016/j.cobeha.2020.10.001. URL https://doi.org/10.1016/j.cobeha.2020.10.001.

Fanzeng Xia, Hao Liu, Yisong Yue, and Tongxin Li. Beyond numeric rewards: In-context dueling bandits with LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9959–9988, 2025. doi: 10.18653/v1/2025.findings-acl.519. URL https://aclanthology.org/2025.findings-acl.519/.

Hua-Dong Xiong, Xinyuan Yan, Ji-An Li, Jingming Xue, Marcelo G Mattar, and Robert C Wilson. Thinking under uncertainty: Evidence use and information-seeking in language models. *arXiv preprint arXiv:2607.26845*, 2026. doi: 10.48550/arXiv.2607.26845. URL https://arxiv.org/abs/2607.26845.

Ziyuan Zhang, Darcy Wang, Ningyuan Chen, Rodrigo Mansur, and Vahid Sarhangian. Comparing exploration–exploitation strategies of llms and humans: Insights from standard multi-armed bandit experiments. *INFORMS Journal on Data Science*, 2026. doi: 10.1287/ijds.2025.0113. URL https://pubsonline.informs.org/doi/10.1287/ijds.2025.0113.

Appendix

Table of Contents

A Two-Arm Tasks and Prompt . . . 15
A.1 Task settings . . . 15
A.2 Reward feedback and trajectory protocol . . . 15
A.3 Models and inference settings . . . 15
A.4 Prompt and arm-type labels . . . 16
A.5 Evaluation scale . . . 17

B Choice-Policy Estimation and Behavioral Measures . . . 17
B.1 From observed rewards to pre-choice beliefs . . . 17
B.2 Choice-model estimation and parameter summaries . . . 18
B.3 Absolute contributions and relative signal balance . . . 18
B.4 Separating changes in state and sensitivity . . . 19
B.4.1 Decomposing value-related evidence sources . . . 20
B.5 Short-task behavior and uncertainty-directed choice . . . 20
B.6 Four-arm metrics and inference . . . 21
B.7 Aggregation and uncertainty displays . . . 22

C Native Thinking and Value-Dominated Choice . . . 23
C.1 Human reference . . . 23
C.1.1 Fitted choice sensitivities . . . 23
C.1.2 Absolute value and relative-uncertainty contributions . . . 24
C.2 Thinking OFF–ON behavior . . . 24
C.3 Model-level effects and temporal dynamics . . . 25
C.3.1 Behavior and fitted sensitivities . . . 25
C.3.2 Direction-aligned changes . . . 25
C.3.3 Absolute-contribution decomposition . . . 25
C.3.4 Early sampling and commitment . . . 28

D Reasoning Pathways and Regulation . . . 29
D.1 Pathway distributions . . . 29
D.2 Explicit uncertainty and confidence language in Native reasoning . . . 30
D.3 Monitoring and action in paired traces . . . 31
D.3.1 Making uncertainty explicit . . . 32
D.3.2 Linking monitoring to action . . . 33
D.4 Monitoring reports and arm choice . . . 33

E Metacognitive Guidance in the Two-Arm Tasks . . . 35
E.1 Monitoring and control prompts . . . 35
E.2 Shared task context and guided output fields . . . 35
E.3 Model results and value evidence . . . 35
E.3.1 Native versus Meta-M+C by model . . . 35
E.3.2 Value evidence and choice . . . 35
E.4 Sampling over time . . . 38
E.4.1 Early repeated sampling and later concentration . . . 38
E.4.2 Directional versus shared uncertainty . . . 39
E.4.3 Behavioral trajectories in Settings 1 and 2 . . . 39

F Long-Horizon Evaluations . . . 40
F.1 Four-arm study . . . 40
F.1.1 Task design and reward sequences . . . 40
F.1.2 Prompts . . . 40
F.1.3 Aggregate choice behavior and outcomes . . . 42
F.1.4 Results by model . . . 43
F.1.5 Sampling and trajectories . . . 43
F.1.6 Decision-mode reports . . . 45

F.2 Five-arm benchmark 47
F.2.1 Task and prompts 47
F.2.2 Choice behavior and performance 48

G Boundary Conditions for Metacognitive Guidance 49
G.1 Unreliable value assessment 49
G.2 Monitoring–action dissociation 49

A TWO-ARM TASKS AND PROMPT

A.1 TASK SETTINGS

Settings 1 and 2 reconstruct Experiments 1 and 2 of Gershman (2018), respectively. Setting 3 reconstructs the task of Gershman (2019). Each uses the corresponding released human dataset. Each simulated participant inherits the corresponding source subject's block structure, arm types, and latent block-level means (μ_1, μ_2). These means remain fixed within a block and are reset at the next block, but are never revealed to the LLM.

Setting 1 and Setting 2 contain 20 blocks of 10 trials, whereas Setting 3 contains 30 blocks of 10 trials. Setting 1 pairs a stochastic arm with a fixed-zero arm, Setting 2 contains two stochastic arms, and Setting 3 uses the four risky/safe configurations RS, SR, RR, and SS, where the letters refer to Arms 1 and 2, respectively. Arm type specifies outcome variability rather than expected value and therefore does not reveal the optimal action.

Table A.1: **LLM reconstruction of the human-referenced bandit tasks.** Block-level means are hidden from the LLM and fixed within each block.

Setting	Blocks × trials	Arm structure	Reward generation
Setting 1	20×10	stochastic vs. fixed zero	stochastic draw from μ_k, with the fixed arm returning 0
Setting 2	20×10	stochastic vs. stochastic	stochastic draw from the selected arm's μ_k
Setting 3	30×10	RS, SR, RR, SS	risky arms: rounded Gaussian draws around μ_k; safe arms: μ_k exactly

The released datasets contain 45, 44, and 46 source participants for Setting 1, Setting 2, and Setting 3, respectively. Their complete schedules yield 9,000, 8,800, and 13,800 human decisions, totaling 31,600. These released records define the simulation templates and sample-size accounting used throughout the paper.

A.2 REWARD FEEDBACK AND TRAJECTORY PROTOCOL

Each LLM chooses one arm per trial and observes only the reward obtained from that arm. The unchosen reward, latent arm means, optimal action, and source participant's realized choices and rewards remain hidden. The LLM's own choice–reward history is carried forward within a block and cleared at each block boundary.

The source record fixes the latent block schedule, while rewards are generated anew from the selected arm for each LLM trajectory rather than replaying the human participant's realized outcomes. Thus, early choices determine which evidence becomes available and consequently alter later decision states. We therefore evaluate each model over multiple complete simulated participants rather than treating a single trajectory as representative of its exploration–exploitation behavior.

Setting 3 reward variability. Risky-arm rewards are generated using Gaussian draws with SD 6, rounded to integers, matching the released data rather than the reward variability reported in Gershman (2019). Safe arms return their latent mean exactly. The cognitive observer retains the observation-variance assumption $q = 16$ (Appendix B.1).

A.3 MODELS AND INFERENCE SETTINGS

We evaluate ten LLMs spanning multiple model families, parameter scales, and both open- and closed-source access. The strict Thinking OFF–ON analysis includes only models for which native inference-time reasoning can be explicitly enabled and disabled while keeping the underlying model identity fixed. Table A.2 summarizes the evaluated models and inference settings.

For every model, Thinking OFF and ON use the same task specification, prompt, and source template. Each condition receives its own realized choice–reward history under the same feedback protocol.

Table A.2: **Models and inference settings used in the paired Thinking OFF–ON analysis.** Temperature is fixed at 0.7 when exposed by the inference interface. Otherwise the provider default is used.

Model	Access	Experimental identifier	Temperature
Qwen3-4B	Open/local	`Qwen3-4B`	0.7
Qwen3-8B	Open/local	`Qwen3-8B`	0.7
Qwen3-14B	Open/local	`Qwen3-14B`	0.7
Qwen3-30B-A3B	Open/local	`Qwen3-30B-A3B-FP8`	0.7
Qwen3-32B	Open/local	`Qwen3-32B-FP8`	0.7
DeepSeek-V4-Flash	API	`deepseek-v4-flash`	Default
GLM-5.1	API	`glm-5.1`	0.7
GPT-5.6-Luna	API	`gpt-5.6-luna`	Default
GPT-5.5	API	`gpt-5.5`	Default
Grok-4.3	API	`x-ai/grok-4.3`	0.7

Table A.3: **Task-specific arm-type text inserted into the shared prompt.**

Setting / condition	`[Arm 1 type]`	`[Arm 2 type]`
Setting 1	`stochastic`	`deterministic`
Setting 2	`stochastic`	`stochastic`
Setting 3 / RS	`risky`	`safe`
Setting 3 / SR	`safe`	`risky`
Setting 3 / RR	`risky`	`risky`
Setting 3 / SS	`safe`	`safe`

The only experimental manipulation is the model's native request-level reasoning mode. We use each provider's corresponding native control rather than prompting the model to simulate or suppress reasoning. Consequently, all Thinking effects are interpreted as within-model contrasts rather than comparisons of reasoning implementations across model families. Native denotes the Thinking ON baseline without metacognitive guidance; Meta-M and Meta-M+C use the same native reasoning mode.

A.4 Prompt and arm-type labels

Thinking OFF and ON use the same task prompt. The prompt provides the task objective, current block and trial, categorical arm types, and the model's own chronological choice–reward history within the current block. It does not reveal latent reward means, the optimal arm, numerical reward distributions, or externally computed value or uncertainty estimates. Thus, the paired comparison differs only in the model's native inference-time reasoning mode.

The decision-relevant prompt template is:

```
You are playing a two-armed bandit task.
Your goal is to maximize total reward.

Current block:  [block].
Current trial:  [trial]/10.

Arm 1 is [Arm 1 type].
Arm 2 is [Arm 2 type].

Previous trials in this block:
[chronological within-block history]

Choose one arm:  1 or 2.
```

The arm-type fields are instantiated according to Table A.3. All other decision-relevant prompt content is shared across tasks and Thinking conditions.

Previous observations are supplied as raw chronological choices and rewards, rather than externally summarized statistics. Formatting-only instructions used to enforce a valid Arm 1/Arm 2 response are identical across conditions and require a strict JSON object with integer `choice` equal to 1 or 2.

A.5 Evaluation scale

Because early choices and sampled rewards alter the evidence available for later decisions, each model is evaluated over multiple complete simulated participants rather than isolated trials. The five Qwen3 models and GLM-5.1 match the released human sample sizes in all three tasks, with 45, 44, and 46 participants in Setting 1, Setting 2, and Setting 3, respectively. For the remaining four models, we evaluate 10 complete participants per task. In both cases, every participant completes the full block and trial structure of the corresponding task.

Table A.4: **Evaluation scale per experimental condition.** Decision counts include complete sequential trajectories across all three human-referenced tasks.

Model group	Models	Participants per task	Decisions per model	Total decisions
Qwen3 family + GLM-5.1	6	45/44/46	31,600	189,600
Remaining models	4	10/10/10	7,000	28,000
Ten-model panel	10	—	—	217,600

Across the ten-model panel, each experimental condition therefore contains 217,600 sequential decisions. The reduced evaluation for the four higher-cost models changes only the number of simulated participants, while preserving the complete task structure for every trajectory.

B Choice-Policy Estimation and Behavioral Measures

This appendix expands Section 3 and defines the quantities used in the figures. The Bayesian observer and probit choice rules follow Gershman (2018, 2019). Fixed human standardization, bounded participant-level estimation, and the pathway summaries below specify their implementation in this study.

B.1 From observed rewards to pre-choice beliefs

Within each block, the observer maintains a Gaussian belief about each arm's mean reward, with mean $m_t(k)$ and variance $s_t(k)$. These are recorded *before* choice a_t and reward r_t. Only the selected arm is updated:

$$K_t = \frac{s_t(a_t)}{s_t(a_t) + q(a_t)},$$
$$m_{t+1}(a_t) = m_t(a_t) + K_t[r_t - m_t(a_t)], \qquad s_{t+1}(a_t) = (1 - K_t)s_t(a_t). \tag{B.1}$$

The unchosen arm remains unchanged. Beliefs reset at each new block. There is no process noise because arm means are stationary within blocks. The observer parameters follow the source analyses and are not fitted to LLM reports. Table B.1 gives their values.

Table B.1: **Source-model observer initialization.** All prior means are zero. q is the assumed reward observation variance. s_1 is uncertainty about the arm's mean.

Setting	$s_1(1), s_1(2)$	Observation variance	Interpretation
1	$(10, 0)$	$q = 10$	The fixed-zero arm retains zero posterior variance.
2	$(100, 100)$	$q = 10$	Both arms generate stochastic rewards.
3	$(100, 100)$	16 (R), 10^{-5} (S)	The small safe-arm variance avoids numerical singularities.

For Setting 3, $q = 16$ retains the original cognitive-analysis assumption. It is distinct from the reward-generating SD of 6, which matches the released data as explained in Appendix A.2.

With $Q_t = m_t$ and $\sigma_t = \sqrt{s_t}$, the observer produces V_t, RU_t, TU_t, and VTU_t in Eq. 1. The full calculation is

$$\text{past choices and rewards} \longrightarrow (Q_t, \sigma_t) \longrightarrow (V_t, \mathrm{RU}_t, \mathrm{VTU}_t) \longrightarrow \eta_t \longrightarrow \Phi(\eta_t). \tag{B.2}$$

Thus the current reward never enters predictors for the current choice. These quantities describe a task-matched observer of the trajectory, not measurements of the model's hidden beliefs.

B.2 Choice-model estimation and parameter summaries

We use the source models' value and uncertainty predictors in the task-specific specifications of Eq. 2: RU and VTU for Settings 1–2, with an additional direct-value term for Setting 3. At fixed value difference, total uncertainty modulates the value slope through V/TU. Positive RU sensitivity favors the relatively uncertain arm. VTU is therefore a value-related pathway that also depends on uncertainty.

The observer reconstructs each predictor from the participant's own choice–reward history. For $X \in \{V, \mathrm{RU}, \mathrm{VTU}\}$, standardization is

$$\widetilde{X}_t = \frac{X_t - \mu_X^{\mathrm{H}}}{s_X^{\mathrm{H}}},$$

where μ_X^{H} and s_X^{H} are the mean and population standard deviation pooled over human pre-choice trials in the same setting. This reference SD describes variation across trials, distinct from an arm's posterior uncertainty $\sigma_t(k)$. VTU is calculated as V_t/TU_t before standardization. The transformation is fixed across humans, LLMs, and conditions, so coefficients share a human-reference unit. This is a comparison convention, not a requirement of the cognitive model or the metacognitive prompts. Each participant's own choices determine their fit.

We estimate one coefficient vector per participant and condition across that participant's blocks. With $y_t = \mathbf{1}\{a_t = 1\}$ and $\eta_t = \boldsymbol{\beta}^\top \widetilde{\mathbf{X}}_t$, we maximize

$$\ell(\boldsymbol{\beta}) = \sum_t \big[y_t \log \Phi(\eta_t) + (1 - y_t) \log(1 - \Phi(\eta_t))\big], \qquad -10 \le \beta_X \le 10. \tag{B.3}$$

The implementation uses L-BFGS-B from a zero initialization, with no ridge penalty for these bounded fits. Nonfinite predictor rows are excluded. Participants without both observed choice categories have undefined fits. Probabilities are clipped to $[10^{-9}, 1 - 10^{-9}]$ for numerical stability. Finite boundary estimates are retained.

Aggregation rules and the choice of cross-model summaries are described in Appendix B.7.

In Figure 2(A), points are model means, boxes span their interquartile range, and central lines are medians. Whiskers extend to the outermost values within 1.5 interquartile ranges. Triangles mark values outside the displayed axis range, which remain in the summaries. Percentage labels are $100(m_{\mathrm{ON}} - m_{\mathrm{OFF}})/m_{\mathrm{OFF}}$, where m is the corresponding cross-model median. For Figure 4, each model is placed at

$$x = \overline{\beta}_{\mathrm{RU}}, \qquad y = \begin{cases} \overline{\beta}_{\mathrm{VTU}}, & \text{Settings 1 and 2,} \\ \sqrt{\overline{\beta}_V^2 + \overline{\beta}_{\mathrm{VTU}}^2}, & \text{Setting 3.} \end{cases} \tag{B.4}$$

The norm is taken *after* within-model coefficient averaging. Displayed centers are coordinatewise medians over ten equally weighted models. The Setting 3 ordinate is a compact sensitivity summary. Realized value influence is calculated by combining signed trial-level terms below.

B.3 Absolute contributions and relative signal balance

On the standardized predictor scale, a fitted term contributes

$$C_t^X = \beta_X \widetilde{X}_t, \qquad C_t^{\mathrm{value}} = \begin{cases} C_t^{\mathrm{VTU}}, & \text{Settings 1 and 2,} \\ C_t^V + C_t^{\mathrm{VTU}}, & \text{Setting 3.} \end{cases} \tag{B.5}$$

On this scale, absolute-contribution curves report $|C_t^X|$ or $|C_t^{\mathrm{value}}|$. The signed value components are summed before taking magnitude, preserving possible cancellation. Magnitudes measure strength,

not direction: a large RU magnitude may accompany either uncertainty seeking or avoidance. For compactness, axes in these contribution and unweighted-predictor panels omit the tilde but retain the standardized fitting scale.

Figure 2(B) and the human-referenced comparison in Figure C.2 instead isolate the state-dependent component relative to a raw-zero baseline. Define

$$A_t^{\mathrm{RU}} = \frac{\beta_{\mathrm{RU}}\mathrm{RU}_t}{s^{\mathrm{H}}_{\mathrm{RU}}},$$
$$A_t^{\mathrm{value}} = \frac{\beta_{\mathrm{VTU}}\mathrm{VTU}_t}{s^{\mathrm{H}}_{\mathrm{VTU}}} + \mathbf{1}_{\mathrm{Setting\ 3}}\frac{\beta_V V_t}{s^{\mathrm{H}}_V}. \tag{B.6}$$

The fitted index is exactly $\eta_t = A_t^{\mathrm{RU}} + A_t^{\mathrm{value}} + b$, where $b = -\sum_X \beta_X \mu_X^{\mathrm{H}}/s_X^{\mathrm{H}}$ is the constant induced by centering, not an additional fitted intercept. Relative shares exclude b. At each trial position, we average $|A_t|$ across blocks within participant and then across matched participants within model. If the resulting magnitudes are M_t^{value} and M_t^{RU}, then

$$S_t^{\mathrm{value}} = \frac{M_t^{\mathrm{value}}}{M_t^{\mathrm{value}} + M_t^{\mathrm{RU}}}, \qquad S_t^{\mathrm{RU}} = 1 - S_t^{\mathrm{value}}. \tag{B.7}$$

Normalization therefore follows magnitude averaging, rather than averaging trial-wise ratios. Curves are medians across the ten models. Shading is the difference of the displayed ON and OFF curves. Zero denominators remain undefined, as at the equal-prior first trial in Settings 2 and 3. These are relative signal magnitudes, not fractions of choice probability, explained variance, or causal influence.

The RU signs in Figure 2(B) use $\beta_{\mathrm{RU}}|\mathrm{RU}_t|/s^{\mathrm{H}}_{\mathrm{RU}}$ in place of $|A_t^{\mathrm{RU}}|$, retaining the same averaging, magnitude denominator, and cross-model median. The aggregate sign is constant across defined trial positions within each condition. Positive and negative signs indicate an RU term directed toward the more and less uncertain arm, respectively. Areas retain absolute magnitudes, and the aggregate signs need not be shared by every model or participant.

B.4 Separating changes in state and sensitivity

Figure E.2(a,b) and the OFF/ON analysis summarized in Appendix C.3.2 use a direction-aligned decomposition to separate changes in the evidence encountered along a trajectory from changes in choice sensitivity. This is a descriptive decomposition of fitted choice influence rather than a causal mediation analysis.

For predictor $X \in \{\mathrm{RU}, \mathrm{VTU}, V\}$, we define the direction-aligned index

$$D_t^X = \beta_X|\widetilde{X}_t|, \qquad X \in \{\mathrm{RU}, \mathrm{VTU}, V\}. \tag{B.8}$$

The index retains the sign of the fitted coefficient while weighting it by the magnitude of the evidence encountered on trial t. Positive values indicate sensitivity in the positive predictor direction, whereas negative values indicate sensitivity in the opposite direction. Accordingly, a positive change can arise either from stronger positive sensitivity or from a reduction in negative sensitivity. Alignment is defined with respect to the standardized predictor.

For conditions a and b (Native and a guided prompt, or Thinking OFF and ON), define four hybrid levels

$$H_{pq} = \mathrm{median}_m \left\langle \beta_{X,m,i,p}|\widetilde{X}_{m,i,t,q}| \right\rangle_{i,t}, \qquad p, q \in \{a, b\}, \tag{B.9}$$

where m indexes models and i participants. Each hybrid combines the fitted sensitivity from condition p with the evidence state from condition q at matched participant–block–trial positions. The two conditions may generate different histories and therefore different evidence states. Brackets denote means within participant and phase, equal weighting of the two five-trial phases, and participant means within model. Each hybrid is then summarized by the median across ten models.

We decompose the aggregate change as

$$\Delta_{\mathrm{sens}} = \tfrac{1}{2}[(H_{ba} - H_{aa}) + (H_{bb} - H_{ab})],$$
$$\Delta_{\mathrm{state}} = \tfrac{1}{2}[(H_{ab} - H_{aa}) + (H_{bb} - H_{ba})],$$
$$\Delta_{\mathrm{combined}} = H_{bb} - H_{aa} = \Delta_{\mathrm{sens}} + \Delta_{\mathrm{state}}. \tag{B.10}$$

The sensitivity component captures the change produced by substituting the fitted coefficient while holding the evidence state fixed, averaged over both substitution orders. The state component analogously captures the change produced by substituting the encountered evidence while holding the coefficient fixed. Medians are taken *before* decomposition, so the displayed components add exactly. Positive and negative components indicate whether each source shifts the direction-aligned index upward or downward. Figure E.2(b) applies the same decomposition separately to Setting 3's V and VTU terms.

Complementary absolute-magnitude analyses. The complementary OFF/ON magnitude decomposition in Figure C.4 retains the absolute terms defined in Appendix B.3. It applies Eq. B.10 at each matched trial with $H_{pq} = |\beta_{X,p}\widetilde{X}_q|$. For the complete Setting 3 value pathway, $H_{pq} = |\beta_{V,p}\widetilde{V}_q + \beta_{\mathrm{VTU},p}\widetilde{\mathrm{VTU}}_q|$. Components are then averaged within model and summarized by model medians. The displayed component medians need not add. This analysis measures changes in magnitude, not the direction of coefficient changes. The complete value magnitude also preserves cancellation between its signed terms, unlike the separate direction-aligned indices in Figure E.2(b).

B.4.1 Decomposing value-related evidence sources

For the metacognitive-prompt comparison, we further separate the VTU state component into changes in raw value difference V and total uncertainty TU, retaining the same coefficients, standardization, and aggregation. For $p, q, r \in \{a, b\}$, define

$$K_{pqr} = \mathrm{median}_m \left\langle \beta_{\mathrm{VTU},m,i,p} \left| \frac{V_{m,i,t,q}/\mathrm{TU}_{m,i,t,r} - \mu^{\mathrm{H}}_{\mathrm{VTU}}}{s^{\mathrm{H}}_{\mathrm{VTU}}} \right| \right\rangle_{i,t}. \tag{B.11}$$

Averaging both orders of substituting V and TU under both coefficients gives

$$\begin{aligned} \Delta_V &= \tfrac{1}{4} \sum_{p\in\{a,b\}} [(K_{pba} - K_{paa}) + (K_{pbb} - K_{pab})], \\ \Delta_{\mathrm{TU}} &= \tfrac{1}{4} \sum_{p\in\{a,b\}} [(K_{pab} - K_{paa}) + (K_{pbb} - K_{pba})]. \end{aligned} \tag{B.12}$$

Their sum exactly recovers the displayed VTU state component. Table E.3 reports each source divided by that aggregate component. Negative shares indicate offsetting changes and permit the other source to exceed 100%.

B.5 Short-task behavior and uncertainty-directed choice

Performance. Optimal-choice rate is the fraction choosing the arm with the higher latent mean, excluding true-mean ties. Reward is the mean obtained reward per trial. Early and late denote trials 1–5 and 6–10. The compact behavioral table reports optimal-choice rate specifically at trial 10.

Coverage and empirical greediness. Both-arm coverage is the fraction of all blocks sampling both arms at least once. First-coverage trial is averaged only over covered blocks. Uncovered blocks receive no artificial completion time. Before each eligible choice, let G_t be the unique arm with the higher historical sample mean, requiring at least one previous observation from each arm. Empirical-greedy and non-greedy rates are $\Pr(a_t = G_t)$ and $\Pr(a_t \neq G_t)$ on this eligible set. Thus trials 1–2 are ineligible, and the choice that completes coverage is not itself a post-coverage choice. Event-aligned plots index successive eligible choices, rather than absolute trial number.

Behavioral aggregation. For Table C.1, participant-level behavioral measures are averaged within each model and setting. We then take the equally weighted mean across ten models separately for each setting. First coverage is reported in trials, and coverage, empirical greediness, and optimal choice are reported as percentages.

A common directional diagnostic. From the observer's pre-choice state, define $G^V_t = \arg\max_k m_t(k)$ and $G^U_t = \arg\max_k \sqrt{s_t(k)}$. The phase-specific rate is

$$p_{\mathrm{dir}} = \frac{\sum_{t\in\mathcal{E}} \mathbf{1}\{a_t = G^U_t\}}{|\mathcal{E}|}, \qquad \mathcal{E} = \{t : V_t \mathrm{RU}_t < 0,\ |V_t| > 10^{-9},\ |\mathrm{RU}_t| > 10^{-9}\}. \tag{B.13}$$

Trials are pooled across a participant's blocks within each phase. Participants without eligible trials have undefined rates. Unlike empirical greediness, this diagnostic can be defined after one observation and uses posterior rather than sample-mean rankings. The 50% reference is uniform choice between two arms. The rate measures alignment with an uncertainty target. Information-seeking intent requires the complementary trace criterion.

Behavior, uncertainty, and later performance. Figure 5(a) combines the non-greedy trajectory with the uncertainty state before trial 6 and mean optimal choice over trials 6–10. For the trajectory, eligible trial-position rates are first calculated within participant, then averaged within model and equally across ten models. The hatched region marks ineligible trials 1–2 rather than zero non-greedy choice. The first defined point is trial 3, and the dashed divider marks entry into trial 6. For later performance, each participant's optimal-choice rates are averaged over trials 6–10, followed by participant means within model and the equally weighted mean across models.

The uncertainty bars use a task-structured diagnostic without refitting the choice model. Let $n_t(k)$ be the number of observations from arm k before trial t. Its posterior variance is

$$v_t(k) = \begin{cases} (100^{-1} + n_t(k)/16)^{-1}, & k \text{ is risky}, \\ 100, & k \text{ is safe and } n_t(k) = 0, \\ 0, & k \text{ is safe and } n_t(k) > 0. \end{cases} \tag{B.14}$$

This retains the observer's risky-arm variance assumption $q = 16$. For this diagnostic only, a sampled safe arm has exactly zero uncertainty, rather than the numerical approximation in Table B.1. No reward-generating variance or fitted coefficient is changed.

The displayed state is $U_5 = \max_k \sqrt{v_6(k)}$, the largest posterior SD after incorporating all five choices and rewards, immediately before trial 6. Block values are averaged within participant and then within model, followed by the equally weighted mean across ten models. All RR, RS, SR, and SS blocks are included, even if only one arm was sampled. This quantity measures residual uncertainty about the arm means at the midpoint of the block, rather than reward variability or self-reported confidence. Ribbons and error bars are 95% hierarchical-bootstrap intervals from 5,000 draws, resampling participants within models and models while preserving condition pairing.

Taken together, the three panels summarize a temporal sequence: early departures from the current empirical-value winner, lower residual uncertainty before trial 6, and stronger later optimal choice. The uncertainty state reflects the cumulative consequence of the preceding sampling history, including but not limited to non-greedy choices. Panel (b) complements this aggregate view with a matched Setting 2 trace that illustrates information-seeking reasoning directly (Appendix D.3).

Matching uncertainty structures. Setting 3 separates cases with a relative-uncertainty target from cases in which both arms remain uncertain. The directional diagnostic is p_{dir} in Eq. B.13. To examine the second case, we compare lower-posterior-value choices between RR and SS:

$$\Delta_{\text{TU}} = \Pr(a_t \neq G_t^V \mid \text{RR}) - \Pr(a_t \neq G_t^V \mid \text{SS}). \tag{B.15}$$

We require prior observations from both arms, nonzero value difference, and $|\text{RU}_t| \leq 0.5$. RR and SS are matched within model, prompt, source participant, phase, exact trial, and signed-value bins of width $0.5 s_V^{\text{H}}$. Matched cells receive weight $2n_{\text{RR}}n_{\text{SS}}/(n_{\text{RR}} + n_{\text{SS}})$ within participant, followed by equal participant and model weighting. Intervals use 2,000 hierarchical-bootstrap draws. Paired prompt contrasts retain participants eligible under both conditions.

B.6 Four-arm metrics and inference

In the four-arm task, all-arm coverage by trial t is the fraction of blocks in which all four arms have been sampled by t. First-coverage trial is the trial at which the fourth distinct arm is first sampled. Its mean is calculated only among blocks that cover all four arms by trial 50. Empirical greediness requires a prior observation from every arm and a unique sample-mean leader. The phase trajectory uses the nine models with eligible choices in all five ten-trial phases. Other main summaries use all ten.

Late optimal-arm rate uses every choice in trials 41–50 without conditioning on coverage or empirical rank. Terminal streak counts consecutive optimal-arm choices ending at trial 50. The figure's L

is the percentage of blocks with at least eight optimal choices in those final ten trials. R is mean total obtained reward per block. Curve points average block-level optimal-choice proportions over non-overlapping five-trial windows, plotted at window ends. Shading indicates ± 1 SEM across ten blocks. The supplemental $P(a^* \mid G)$ instead conditions on eligible empirical-greedy choices and is not interchangeable with unconditional late optimal-arm rate.

For each reported long-horizon paired test, block results are first averaged within model. The two-sided exact sign-flip test enumerates all sign assignments of the paired model differences and compares the absolute mean with its observed value. These descriptive comparisons are not adjusted for multiple testing.

B.7 Aggregation and uncertainty displays

Participants receive equal weight within model and models equal weight across the ten-model panel. The cognitive panels in Figure 2(A,B), Figure 4, and supplementary cognitive trajectories use model medians. Figure E.2(a,b) and the OFF/ON direction-aligned summaries in Appendix C.3.2 instead take the model median of each hybrid level before computing the decomposition, as specified in Appendix B.4. Table C.1, Figure 5, supplementary behavioral trajectories, compact tables, and long-horizon summaries use model means. Human curves average human participants. Parameter distributions use finite estimates, while the relative-share analysis additionally requires converged fits and matched OFF/ON participant records. Reported boundary solutions remain eligible. Each summary uses the eligibility set defined for its corresponding metric.

Parameter displays first average participant coefficients within each model and summarize across models using equally weighted medians to reduce the influence of extreme coefficients associated with persistent arm fixation or near-deterministic greedy choice.

Hierarchical bootstrap calculations resample models and then participants within each selected model. Paired contrasts preserve participant identity across conditions. The parameter-space plot uses 4,000 draws: ellipses are Gaussian covariance contours of these draws, centered on the displayed statistic, with radius $\sqrt{-2\log(1-p)}$ at coverage level p. These contours summarize uncertainty in the aggregate location. Figure E.2(a,b) displays point estimates without confidence intervals. Main pathway and short-task behavioral intervals use 5,000 draws and percentile limits at 2.5% and 97.5%. The single-model temperature comparison resamples participants. The absolute OFF/ON contribution curves in Figure C.3 also use 5,000 draws. Other OFF/ON cognitive curves use 2,000 draws. Three-prompt cognitive curves use 1,000. Supplementary behavioral curves use 5,000 draws. Bootstrap ribbons are pointwise intervals.

C Native Thinking and Value-Dominated Choice

C.1 Human reference

C.1.1 Fitted choice sensitivities

Figure C.1 places Thinking ON coefficients against the human reference on the common within-task scale defined in Appendix B.2. All ten models have higher mean β_{VTU} than humans in every setting, reaching 2.54–8.60, 1.99–3.07, and 2.29–5.25 times the human mean in Settings 1–3, respectively. This indicates consistently stronger choice sensitivity to uncertainty-normalized value evidence. By comparison, 28 of 30 model-level β_{RU} means fall within the central 95% of human participant coefficients. Human variation in β_{VTU} is especially broad in Setting 3, where most LLM means remain within this range despite exceeding the human mean. These comparisons provide the human reference for the analysis of Thinking-related value amplification in Section 4.

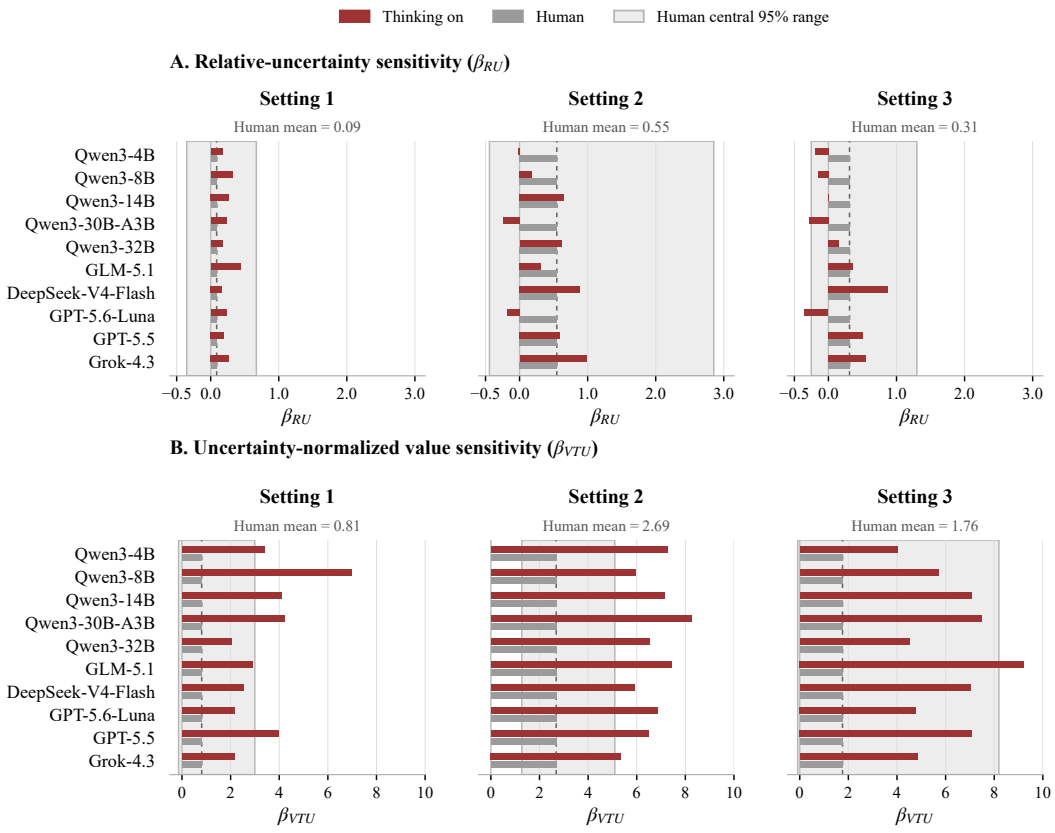


Figure C.1: **Human-referenced comparison of fitted choice sensitivities. A:** β_{RU}. **B:** β_{VTU}. Red bars show participant coefficients averaged within each Thinking ON model. Gray bars and dashed lines show the full human mean for each setting. Shaded boxes span the 2.5th–97.5th percentiles of human participant coefficients, describing individual variation rather than confidence intervals for the mean.

C.1.2 Absolute value and relative-uncertainty contributions

Figure C.2 combines each participant's fitted sensitivities with the evidence reconstructed along their own trajectory. It uses the state-dependent terms in Eq. B.6, retaining their absolute magnitudes before the normalization used in Figure 2(B). At each trial position, the within-model summary is

$$M^X_{m,t} = \left\langle |A^X_t| \right\rangle_{\text{blocks, participants within } m}, \qquad X \in \{\text{value}, \text{RU}\}. \tag{C.1}$$

We first average blocks within participant, then participants within model. LLM curves are medians across the ten model summaries, using the same participants as Figure 2(B). Human curves average blocks within participant and then all human participants in each setting. In Setting 3, the signed V and VTU terms are added before taking their magnitude.

Thinking ON exhibits substantially larger value-related magnitudes than the human reference, with a much smaller departure in RU. Averaged over trials 2–10, the displayed value magnitudes are 3.05, 2.41, and 2.59 times the human values in Settings 1–3, respectively, compared with 1.16, 1.00, and 1.31 times for RU. The contrast is thus concentrated in the value pathway and remains clear without expressing contributions as relative shares.

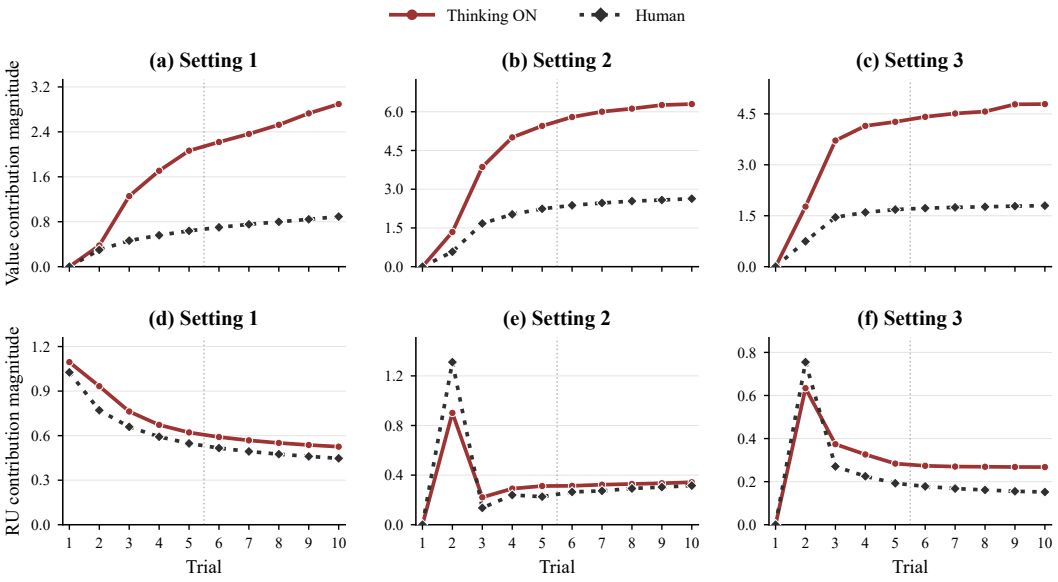


Figure C.2: **Human-referenced comparison of absolute contribution magnitudes. (a–c)** Value-related signals. **(d–f)** RU signals. Curves show $|A^{\text{value}}_t|$ and $|A^{\text{RU}}_t|$ from Eq. B.6, with cross-model medians for Thinking ON and participant means for Human. These descriptive magnitudes exclude the constant centering offset and measure signal strength, not coefficient direction or shares of choice probability.

C.2 Thinking OFF–ON behavior

Table C.1 reports equally weighted means across ten models within each setting. Across all three settings, Thinking ON achieves earlier and more complete coverage and higher optimal-choice rates, alongside greater empirical greediness.

Table C.1: **Behavior under Thinking OFF and ON across the three settings.** The four metrics report, respectively, the percentage of blocks in which both arms are sampled, the mean trial at which both arms are first covered among those blocks, empirical greediness (%), and the percentage of choices selecting the arm with the higher expected reward. Detailed definitions are in Appendix B.5.

	Setting 1				Setting 2				Setting 3			
Thinking	Cover ↑	First ↓	Emp. greedy	Optimal rate (%) ↑	Cover ↑	First ↓	Emp. greedy	Optimal rate (%) ↑	Cover ↑	First ↓	Emp. greedy	Optimal rate (%) ↑
OFF	89.49	2.88	86.60	77.30	56.35	2.82	93.36	80.13	64.00	3.15	85.41	66.92
ON	**99.99**	**2.08**	98.29	**85.73**	**85.89**	**2.58**	99.75	**87.27**	**78.99**	**2.46**	98.53	**79.23**

C.3 Model-level effects and temporal dynamics

The analyses below give the complete model-level comparison underlying Section 4, followed by choice-policy and behavioral dynamics. Direction-aligned changes are distinguished from absolute contribution magnitudes.

C.3.1 Behavior and fitted sensitivities

Table C.2 reports every model and setting. Coverage and first-coverage trial retain their distinct denominators, as defined in Appendix B.5.

C.3.2 Direction-aligned changes

Using the direction-aligned decomposition in Appendix B.4, with OFF as a and ON as b, the value-related indices increase in all three settings, primarily through sensitivity changes. Evidence-state changes reinforce the increase in Settings 2–3 and modestly offset it in Setting 1. The signed RU index shifts from negative to positive in Settings 2–3 and is nearly unchanged in Setting 1. This directional shift can coexist with a smaller absolute RU contribution; the two quantities answer different questions.

C.3.3 Absolute-contribution decomposition

Thinking ON increases late VTU contribution magnitude and reduces late RU magnitude across all three settings (Figure C.3). The absolute-magnitude decomposition defined in Appendix B.4 attributes most of the value-related increase to sensitivity changes (Figure C.4): late VTU sensitivity components are $+1.43$, $+2.78$, and $+0.74$ latent-index units in Settings 1–3. In Setting 3, the complete value pathway combines the signed VTU and direct-V terms before taking magnitude; its late contribution increases by 2.00 units.

Robustness. The direction of the late contribution changes is unchanged under model means, exclusion of Qwen3-4B, and every leave-one-model-out median. Leave-one-model-out ranges for ON-minus-OFF RU magnitude remain negative ($[-0.090, -0.043]$, $[-0.193, -0.040]$, $[-0.231, -0.216]$), while those for VTU remain positive ($[1.190, 1.624]$, $[3.452, 3.701]$, $[1.233, 1.256]$), for Settings 1–3 respectively. Thus, the increase in late VTU magnitude and decrease in late RU magnitude are robust across these aggregation choices and model exclusions.

Table C.2: **Thinking OFF → ON: performance, both-arm coverage, greediness, and fitted choice sensitivities.** Dark-red upward arrows denote numerical increases, blue downward arrows denote numerical decreases, and gray equality signs denote no change. These symbols encode direction, not statistical significance.

	Choice-based behavior					Fitted choice sensitivities		
Model	**Optimal**	**Reward**	**Both-arm coverage**	**First-coverage trial†**	**Emp.-greedy**	β_{RU}	β_{VTU}	β_V
Setting 1								
Qwen3-4B	68.6%→85.6% (↑)	0.21→0.34 (↑)	43.0%→100.0% (↑)	3.60→2.07 (↓)	85.1%→98.9% (↑)	-0.17→0.18 (↑)	0.45→3.41 (↑)	–
Qwen3-8B	71.4%→86.1% (↑)	0.22→0.35 (↑)	98.0%→100.0% (↑)	3.47→2.01 (↓)	81.7%→99.7% (↑)	-0.14→0.32 (↑)	0.99→6.96 (↑)	–
Qwen3-14B	78.1%→87.0% (↑)	0.28→0.35 (↑)	96.2%→99.9% (↑)	2.96→2.05 (↓)	82.4%→99.8% (↑)	0.68→0.27 (↓)	1.35→4.10 (↑)	–
Qwen3-30B-A3B	83.1%→86.1% (↑)	0.33→0.35 (↑)	78.4%→100.0% (↑)	2.57→2.02 (↓)	89.1%→99.6% (↑)	0.40→0.23 (↓)	2.24→4.22 (↑)	–
Qwen3-32B	69.1%→82.1% (↑)	0.21→0.32 (↑)	100.0%→100.0% (=)	2.37→2.02 (↓)	74.4%→93.4% (↑)	-0.05→0.17 (↑)	0.70→2.05 (↑)	–
DeepSeek-V4-Flash	70.0%→85.4% (↑)	0.30→0.40 (↑)	98.3%→100.0% (↑)	3.03→2.00 (↓)	78.6%→96.0% (↑)	0.32→0.16 (↓)	0.82→2.53 (↑)	–
GLM-5.1	84.6%→86.3% (↑)	0.34→0.35 (↑)	97.6%→100.0% (↑)	2.80→2.03 (↓)	93.3%→99.8% (↑)	0.62→0.44 (↓)	1.97→2.92 (↑)	–
GPT-5.6-Luna	85.1%→84.9% (↓)	0.39→0.40 (↑)	90.0%→100.0% (↑)	2.64→2.57 (↓)	94.5%→98.2% (↑)	0.41→0.23 (↓)	2.00→2.18 (↑)	–
GPT-5.5	87.9%→84.5% (↓)	0.41→0.39 (↓)	98.3%→100.0% (↑)	2.08→2.00 (↓)	98.1%→99.5% (↑)	0.18→0.19 (↑)	3.89→3.98 (↑)	–
Grok-4.3	75.0%→89.2% (↑)	0.45→0.56 (↑)	95.0%→100.0% (↑)	3.32→2.05 (↓)	88.8%→98.0% (↑)	0.18→0.27 (↑)	1.01→2.15 (↑)	–
Human reference	71.3%	0.19	62.7%	4.17	76.3%	0.09	0.81	–
Setting 2								
Qwen3-4B	54.7%→85.0% (↑)	0.82→3.85 (↑)	6.4%→75.2% (↑)	2.10→2.99 (↑)	92.0%→100.0% (↑)	-9.47→-0.02 (↑)	0.86→7.26 (↑)	–
Qwen3-8B	71.6%→85.6% (↑)	2.78→3.77 (↑)	50.8%→93.9% (↑)	4.27→3.56 (↓)	86.4%→99.9% (↑)	-0.60→0.18 (↑)	2.04→5.96 (↑)	–
Qwen3-14B	80.8%→87.7% (↑)	3.62→3.91 (↑)	65.0%→95.9% (↑)	2.80→2.37 (↓)	91.4%→99.9% (↑)	-0.32→0.66 (↑)	2.56→7.14 (↑)	–
Qwen3-30B-A3B	80.8%→82.8% (↑)	3.56→3.74 (↑)	71.6%→59.3% (↓)	2.05→2.26 (↑)	92.5%→99.8% (↑)	-0.30→-0.24 (↑)	2.60→8.25 (↑)	–
Qwen3-32B	84.2%→87.5% (↑)	3.71→3.87 (↑)	98.2%→96.4% (↓)	2.05→2.40 (↑)	92.6%→99.4% (↑)	0.75→0.62 (↓)	2.72→6.53 (↑)	–
DeepSeek-V4-Flash	82.1%→88.9% (↑)	4.58→4.89 (↑)	60.0%→100.0% (↑)	3.59→2.33 (↓)	88.6%→100.0% (↑)	-0.44→0.89 (↑)	1.38→5.90 (↑)	–
GLM-5.1	83.6%→87.2% (↑)	3.80→3.92 (↑)	64.9%→86.6% (↑)	2.56→2.51 (↓)	96.3%→99.7% (↑)	-0.15→0.32 (↑)	6.64→7.42 (↑)	–
GPT-5.6-Luna	85.4%→89.3% (↑)	4.95→5.09 (↑)	55.0%→58.3% (↑)	2.67→2.54 (↓)	95.8%→98.8% (↑)	-0.32→-0.18 (↑)	3.78→6.85 (↑)	–
GPT-5.5	87.0%→89.6% (↑)	5.04→5.02 (↓)	51.7%→93.3% (↑)	2.72→2.41 (↓)	100.0%→100.0% (=)	-0.19→0.59 (↑)	10.00→6.49 (↓)	–
Grok-4.3	91.1%→88.9% (↓)	7.09→6.86 (↓)	40.0%→100.0% (↑)	3.38→2.40 (↓)	98.1%→100.0% (↑)	-0.50→0.98 (↑)	2.96→5.35 (↑)	–
Human reference	82.9%	3.67	91.0%	2.21	91.2%	0.55	2.69	–
Setting 3								
Qwen3-4B	51.2%→78.6% (↑)	0.21→3.25 (↑)	12.4%→75.7% (↑)	2.14→2.62 (↑)	79.1%→98.8% (↑)	-4.80→-0.19 (↑)	6.05→4.03 (↓)	3.85→1.96 (↓)
Qwen3-8B	57.2%→78.2% (↑)	0.86→3.30 (↑)	48.8%→65.0% (↑)	5.63→2.43 (↓)	85.7%→99.7% (↑)	-0.70→-0.15 (↑)	7.30→5.70 (↓)	0.62→3.97 (↑)
Qwen3-14B	74.3%→80.1% (↑)	2.98→3.40 (↑)	69.3%→70.7% (↑)	2.74→2.30 (↓)	85.0%→99.2% (↑)	-0.32→-0.01 (↑)	0.84→7.07 (↑)	1.44→3.36 (↑)
Qwen3-30B-A3B	57.1%→78.2% (↑)	0.99→3.27 (↑)	52.8%→61.7% (↑)	2.44→2.42 (↓)	65.8%→99.4% (↑)	-0.81→-0.29 (↑)	-1.54→7.48 (↑)	0.43→3.66 (↑)
Qwen3-32B	73.7%→80.7% (↑)	2.88→3.34 (↑)	65.5%→86.0% (↑)	2.63→2.37 (↓)	87.6%→97.7% (↑)	-0.41→0.16 (↑)	1.70→4.50 (↑)	1.32→2.17 (↑)
DeepSeek-V4-Flash	68.5%→83.2% (↑)	2.30→3.63 (↑)	88.9%→96.7% (↑)	2.98→2.40 (↓)	78.4%→98.2% (↑)	-0.11→0.88 (↑)	1.34→7.03 (↑)	0.74→2.82 (↑)
GLM-5.1	77.9%→80.0% (↑)	3.20→3.31 (↑)	79.0%→87.5% (↑)	2.65→2.55 (↓)	92.1%→99.5% (↑)	-0.09→0.35 (↑)	5.72→9.23 (↑)	1.95→2.87 (↑)
GPT-5.6-Luna	70.6%→73.4% (↑)	3.10→3.31 (↑)	65.6%→63.3% (↓)	3.00→2.60 (↓)	90.2%→94.4% (↑)	-0.43→-0.35 (↑)	1.81→4.79 (↑)	1.49→2.02 (↑)
GPT-5.5	75.3%→79.8% (↑)	3.34→3.54 (↑)	74.4%→86.7% (↑)	2.32→2.57 (↑)	94.9%→99.3% (↑)	-0.08→0.50 (↑)	5.39→7.09 (↑)	2.78→3.24 (↑)
Grok-4.3	63.4%→80.0% (↑)	3.27→4.36 (↑)	83.3%→96.7% (↑)	4.96→2.31 (↓)	95.2%→99.1% (↑)	-0.52→0.55 (↑)	2.02→4.84 (↑)	1.22→2.18 (↑)
Human reference	74.0%	2.63	93.7%	2.50	83.1%	0.31	1.76	1.19

Notes. LLM entries are matched-subject means under the shared task prompt, shown as OFF→ON. The five Qwen3 models and GLM-5.1 use the full source samples. The remaining four models use 10 simulated participants per setting (Appendix A.5). Human entries are single reference means. Optimal rate excludes equal-mean trials. Both-arm coverage is the percentage of all blocks in which each arm is sampled at least once. †First-coverage trial is the trial on which the second distinct arm is first sampled, averaged only over blocks that eventually cover both arms. Uncovered blocks are not imputed, so the two coverage columns must be read jointly and lower values indicate earlier coverage. Empirical-greedy rate is computed only after each arm has at least one prior reward observation and the two historical sample means differ. Coefficients are participant means under the specifications in Eq. 2. Valid boundary fits are retained. β_V is absent from the models for Settings 1 and 2.

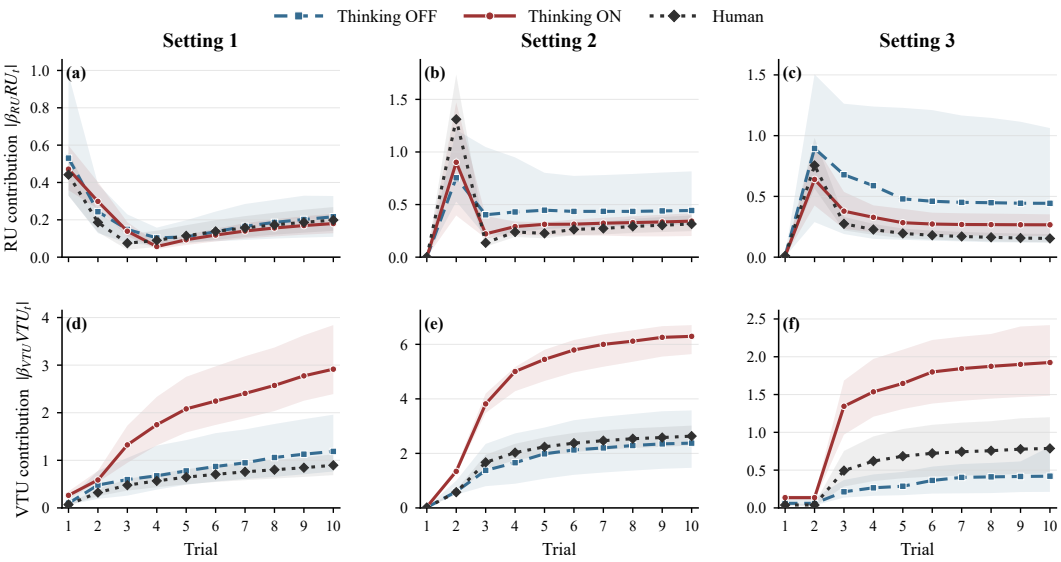


Figure C.3: **Thinking ON increases the absolute VTU contribution.** Magnitudes of fitted RU and VTU terms on the probit-index scale. LLM curves are ten-model medians of participant means. Human is the participant mean. Ribbons are 95% hierarchical-bootstrap intervals. Setting 3's direct-value term is excluded here but included in the complete-value decomposition in Figure C.4.

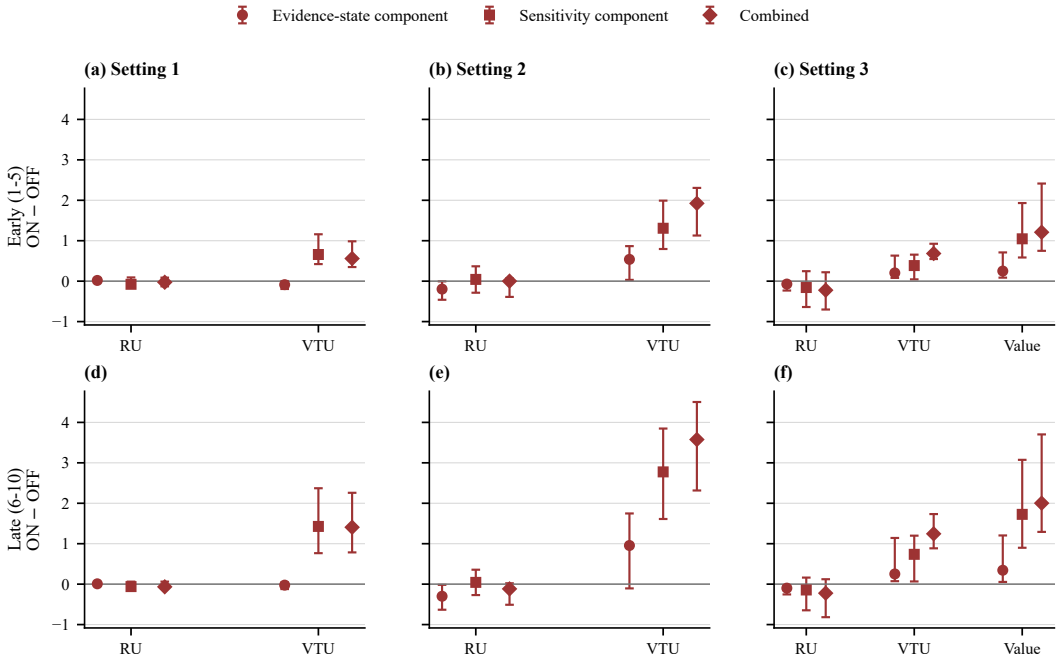


Figure C.4: **State and sensitivity contributions to the OFF-to-ON magnitude shift.** Points are ten-model medians with paired 95% bootstrap intervals. Diamonds show Combined, the change in absolute magnitude. Value in Setting 3 combines signed V and VTU terms before taking magnitude. Additivity holds before cross-model median aggregation.

C.3.4 EARLY SAMPLING AND COMMITMENT

Figure C.5 separates acquiring a comparison from acting on it. Thinking ON generally broadens coverage, but the first eligible post-coverage decisions already closely track the empirical winner. Curves average participants within model and then models equally, following the behavioral summaries in the main text.

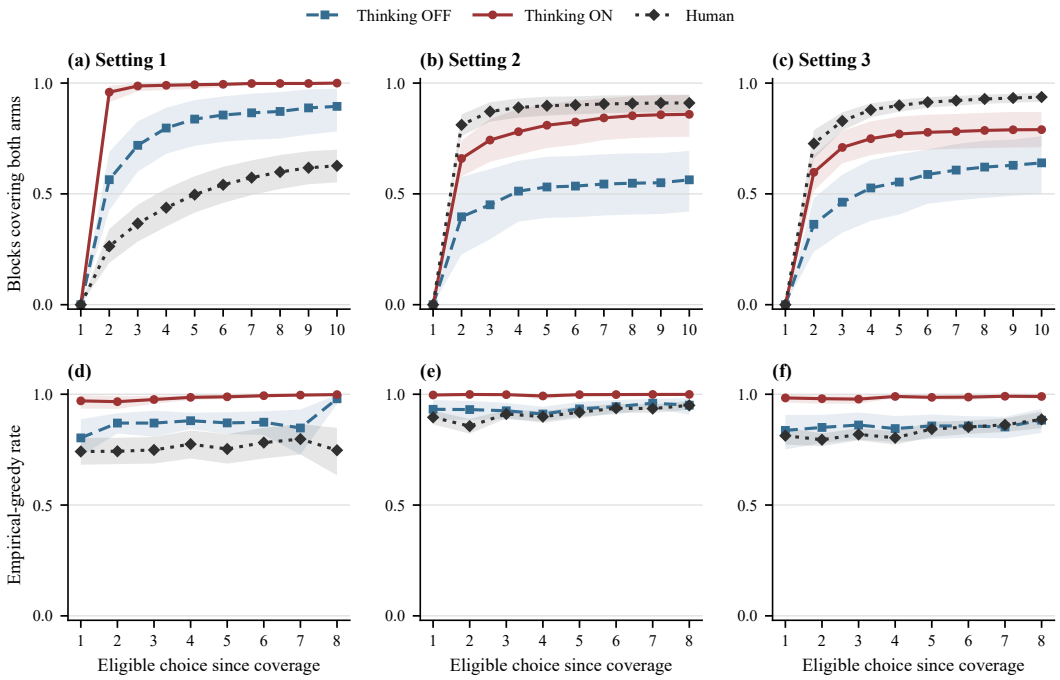


Figure C.5: **Thinking improves coverage and then rapidly commits.** Top: cumulative both-arm coverage. Bottom: empirical-greedy choice indexed by eligible post-coverage decision, excluding sample-mean ties. Uncovered blocks remain outside cumulative coverage. No completion time is imputed. Curves are model-equal means with 95% hierarchical-bootstrap intervals.

D REASONING PATHWAYS AND REGULATION

Visible traces characterize expressed reasoning, not hidden computation. We first define reasoning pathways and present their distributions under Native and Meta-M+C, then summarize matches to uncertainty- and confidence-related keywords and examine matched reasoning–action contrasts.

[. . .] removes repetition or formatting checks.

D.1 PATHWAY DISTRIBUTIONS

Pathway annotation. We summarize each reasoning trace using five recurring components, defined in Table D.1, followed by the final action relative to the model's estimated value winner. The components capture value comparison (**V**), uncertainty assessment (**U**), the informational value of further sampling (**I**), exploration–exploitation evaluation (**P**), and confidence assessment (**C**). Their order records how these elements are organized in the expressed rationale.

Choose denotes selecting the estimated value winner, whereas **Forgo** denotes selecting another arm while retaining that value ranking. Traces without an identifiable value winner are excluded from these action-relative pathway summaries. Repeated symbols indicate a return to the same component; for example, $\mathrm{V} \rightarrow \mathrm{U} \rightarrow \mathrm{V} \rightarrow \mathrm{Choose}$ denotes an initial value comparison, followed by uncertainty assessment, and then a return to value as the final basis for choosing the value winner.

Table D.1: **Components used in reasoning-pathway annotation.**

Symbol	Annotation criterion
V	**Value comparison.** Estimate or compare arm values using the observed choice–reward history.
U	**Uncertainty assessment.** Explicitly assess uncertainty about an arm's estimated value or the current value ranking. Reward variability or sample count alone does not qualify without an uncertainty judgment.
I	**Information value.** Evaluate what further sampling could resolve and why the resulting information could improve later decisions.
P	**Exploration–exploitation evaluation.** Weigh information acquisition against immediate reward, sampling cost, or the remaining decision horizon.
C	**Confidence assessment.** Evaluate how strongly the available evidence supports the current value judgment or intended choice.

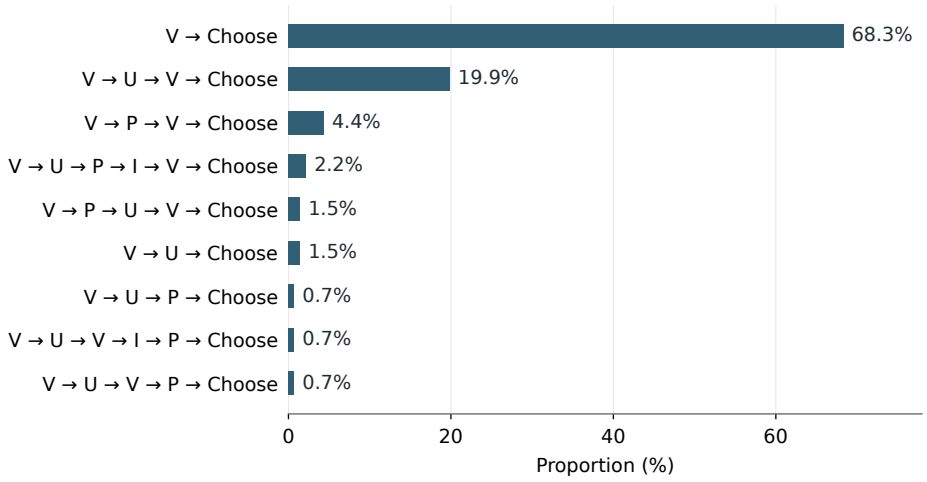
Sampling protocol. We sample reasoning traces from Qwen3-14B, Qwen3-32B, GLM-5.1, and DeepSeek-V4-Flash across all three task settings. For each condition, we select ten traces for every model, setting, and trial number from trials 3–10. We distinguish early trials 3–5 from late trials 6–10. Figure D.1 shows Native pathways pooled across trials 3–10 and Meta-M+C pathways separately for the early and late phases. Each panel reports pathway proportions among traces with an identifiable value winner.

Late Meta-M+C pathway grouping. In the late Meta-M+C panel, *Other* $\rightarrow$ *Choose* combines rare Choose pathways with more than three annotated components, each representing less than 3% of that panel. Forgo pathways remain separate because they mark departures from the estimated value winner.

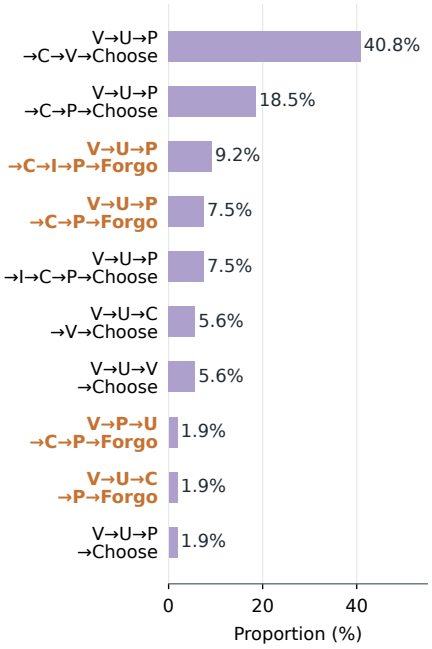
Native value commitment and metacognitive regulation. Native reasoning is concentrated in short, value-centered pathways. Among traces with an identifiable value winner, 68.3% follow the direct $\mathrm{V} \rightarrow \mathrm{Choose}$ pathway, while another 19.9% follow the $\mathrm{V} \rightarrow \mathrm{U} \rightarrow \mathrm{V} \rightarrow \mathrm{Choose}$ pathway, in which uncertainty is considered but value again provides the concluding rationale for choice.

Meta-M+C substantially reorganizes these pathways. Uncertainty, exploration–exploitation evaluation, information value, and confidence are more frequently incorporated before action. In the early phase, several prominent pathways terminate in **Forgo**, reflecting deliberate departures from the current value winner after these assessments. In the late phase, pathways shift predominantly toward **Choose**, with the value winner retained after uncertainty and strategy assessment. This

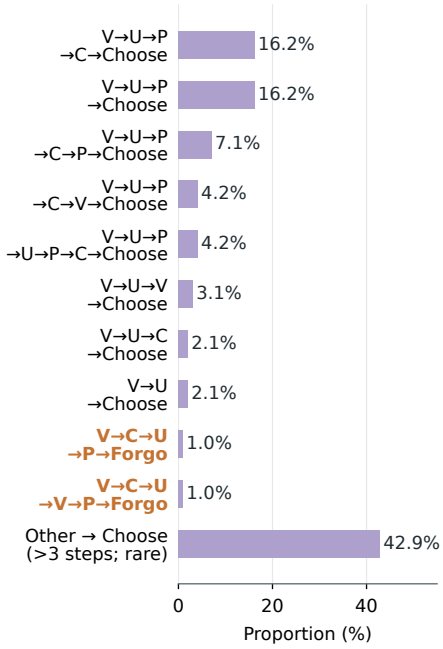
phase-dependent reorganization mirrors the behavioral pattern in Section 5: metacognitive guidance supports information-seeking departures early and stronger commitment as evidence accumulates.



(a) Native: trials 3–10.



(b) Meta-M+C: early trials 3–5.



(c) Meta-M+C: late trials 6–10.

Figure D.1: **Reasoning-pathway distributions under Native and Meta-M+C.** Blue bars show Native and pale purple bars show Meta-M+C. Orange bold labels indicate pathways ending in **Forgo**. Native pathways are pooled over trials 3–10; Meta-M+C pathways are separated into early trials 3–5 and late trials 6–10. Pathway categories follow the annotation criteria above, with long, rare Choose pathways grouped only in the late Meta-M+C panel as described above. Percentages are computed within each panel's eligible traces; rounding may prevent them from summing exactly to 100%.

D.2 Explicit uncertainty and confidence language in Native reasoning

To quantify how often Native reasoning explicitly verbalizes uncertainty and confidence, Table D.2 screens saved reasoning traces in Settings 1–3 for two vocabularies: *uncertain* or *uncertainty*, and *confidence* or *confident*. Matching uses whole words and ignores case. Within each vocabulary, a

trace counts once regardless of repetition or co-occurrence. Both screens use the same non-empty reasoning records. We extract `reasoning_content`, falling back to saved `<think>` content; empty or missing reasoning is excluded. Prompts and final answers are not included in the search.

The screen covers seven models. GLM-5.1 uses saved Native reasoning from its comparison archive as the sole source in each setting; the remaining models use their primary Native reasoning files. Grok-4.3 is excluded because comparable saved reasoning traces are unavailable.

Table D.2: **Explicit uncertainty and confidence language in Native reasoning traces.** Percentage of saved reasoning records matching each vocabulary.

Model	Setting 1 (%)	Setting 2 (%)	Setting 3 (%)	Pooled (%)
(a) Uncertainty: *uncertain* or *uncertainty*				
Qwen3-4B	27.47	30.42	18.15	24.22
Qwen3-8B	16.18	12.10	12.28	13.34
Qwen3-14B	8.63	2.30	7.67	6.45
Qwen3-30B-A3B	14.91	4.49	6.90	8.51
Qwen3-32B	6.89	2.59	4.74	4.75
DeepSeek-V4-Flash	28.17	7.83	13.33	16.00
GLM-5.1	7.96	5.15	3.27	5.15
Pooled	**14.70**	**10.12**	**9.71**	**11.25**
(b) Confidence: *confidence* or *confident*				
Qwen3-4B	3.14	9.97	3.76	5.31
Qwen3-8B	0.34	7.27	1.54	2.80
Qwen3-14B	0.62	0.74	0.37	0.54
Qwen3-30B-A3B	3.26	2.03	1.64	2.21
Qwen3-32B	0.87	1.41	0.43	0.83
DeepSeek-V4-Flash	8.00	2.67	3.00	4.33
GLM-5.1	0.85	0.60	0.13	0.47
Pooled	**1.69**	**4.11**	**1.51**	**2.28**

All configurations use Thinking ON. Denominators contain only non-empty, extractable reasoning records. Pooled rates weight individual records rather than averaging model percentages. The two vocabularies may overlap within the same trace.

Across all three settings, explicit uncertainty terms appear in 11.25% of saved Native reasoning traces, while confidence terms appear in only 2.28%. Confidence language is sparse across nearly all models and settings, and explicit uncertainty language is also absent from the large majority of traces. This lexical pattern complements the pathway analysis in Appendix D.1: Native reasoning is typically organized around value comparison without routinely making uncertainty or confidence assessment explicit. The keyword screen targets explicit verbalization rather than the broader quality of metacognitive monitoring; the matched reasoning–action contrasts below examine how such assessments, when present, are organized and translated into choice.

D.3 Monitoring and action in paired traces

The main-text example and four additional pairs illustrate explicit uncertainty monitoring and the regulation of action. Paired runs share task information and observed history, with native reasoning enabled at $T = 0.7$.

Main-text case provenance. The paired GLM-5.1 examples in Figures 3 and 5(b) come from Native subject 4 and Meta-M+C subject 20, Setting 2, block 3, trial 4/10. Their system messages and visible task histories match exactly: Arm 1 returned 7 once and Arm 2 returned $9, 7$.

D.3.1 MAKING UNCERTAINTY EXPLICIT

In the first three pairs, Native moves from value comparison to choice without explicitly comparing the arms' uncertainty. Meta-M+C retains the value preference but reports an uncertainty target and chooses it to seek information.

Pair 1: Qwen3-32B. Setting 2, trial 4/10. Both conditions share the same history: Arm 1 was sampled twice $(5, 6)$, and Arm 2 once (-12).

Native: the value comparison determines action

"**[Value comparison]** Arm 1 has been positive both times, while Arm 2 was negative once. But they're both stochastic, so past performance doesn't guarantee future results. However, based on the current data, Arm 1 seems better. [. . .] **[Value estimate]** Maybe the average for Arm 1 is (5+6)/2 = 5.5, and Arm 2 is -12. [. . .] **[Value-guided choice]** So choosing Arm 1 seems better. I'll go with Arm 1 again."

Meta-M+C: explicit uncertainty monitoring informs the choice

"**[Value comparison]** The average for Arm 1 is 11/2 = 5.5, and Arm 2 is -12. So clearly Arm 1 has a higher average so far. [. . .] **[Uncertainty monitoring]** Arm 1 has been pulled twice, Arm 2 once. So Arm 2's reward is based on only one trial, which is very uncertain. So the more uncertain arm is 2. [. . .] **[Information-seeking rationale]** Since I have very limited data on Arm 2, maybe I should explore it again to get more information. [. . .] **[Regulated choice]** So in this case, even though the better arm is 1, the decision is exploration, so choosing the more uncertain arm 2."

Native briefly acknowledges stochasticity, then chooses Arm 1 based on the value comparison. Meta-M+C retains that ranking but makes the sample imbalance an explicit uncertainty signal and chooses Arm 2 to seek information.

Pair 2: GLM-5.1. Setting 3, trial 5/10. Both conditions have observed Arm 1 three times $(-2, -2, -2)$ and Arm 2 once (-3).

Native: mean rewards directly determine the choice

"**[Value estimate]** Arm 1 average reward: (-2 + -2 + -2) / 3 = -2 [. . .] **[Value estimate]** Arm 2 average reward: -3 / 1 = -3 [. . .] **[Value-guided choice]** Arm 1 has a higher average reward $(-2 > -3)$. I should choose Arm 1 to maximize reward."

Meta-M+C: an explicit uncertainty judgment changes the choice

"**[Value comparison]** Arm 1 appears more rewarding, although both are negative. [. . .] **[Uncertainty monitoring]** Arm 2 is more uncertain because it has only been sampled once, whereas Arm 1 has been sampled 3 times. [. . .] **[Remaining horizon]** We are at trial 5 of 10. We have 5 trials left. [. . .] **[Information-seeking rationale]** If Arm 2 is safe and its true mean is -3, exploring it loses me 1 point compared to Arm 1. If its true mean is -1, exploiting Arm 1 loses me 1 point. [. . .] **[Regulated choice]** Knowing the true mean of Arm 2 is valuable. I will choose Arm 2 for exploration."

Native calculates the sample means and selects Arm 1. Meta-M+C retains that value preference, explicitly compares uncertainty, and chooses Arm 2 to reduce the uncertainty it reports.

Pair 3: GLM-5.1, early information seeking. Setting 3, trial 3/10. Both conditions have one observation from each arm: −4 from Arm 1 and 10 from Arm 2.

Native: observed rewards map directly to the safe arm

"**[Value comparison]** Arm 1 gave -4. Arm 2 gave 10. [. . .] Arm 2 seems to be giving positive rewards and is safe. Arm 1 is risky and gave a negative reward. **[Value-guided choice]** I should choose Arm 2."

Meta-M+C: sampling the risky arm to acquire information

"**[Value comparison]** Arm 1 gave -4 (one sample). Arm 2 gave 10 (one sample). Based on very limited evidence, Arm 2 appears more rewarding. [. . .] **[Uncertainty monitoring]** Arm 1 is risky, so it has higher variance/uncertainty. Arm 2 is safe, so it has lower variance/uncertainty. Arm 1 is more uncertain. [. . .] **[Information-seeking rationale]** Given we have many trials left, it's important to explore Arm 1 more to get a better estimate of its true mean. [. . .] **[Regulated choice]** If I explore Arm 1 and it's bad, I lose some reward but gain information. If I exploit Arm 2, I get a safe reward but don't learn about Arm 1. [. . .] I'll choose Arm 1 for exploration."

Both identify Arm 2 as currently more rewarding. Native selects it, whereas Meta-M+C considers the remaining trials and chooses Arm 1 to learn more about its value.

D.3.2 Linking monitoring to action

Pair 4: Qwen3-8B. Setting 2, trial 2/10. Both conditions have observed Arm 1 once, with reward 2. Arm 2 has not been sampled.

Native: uncertainty is acknowledged, but the choice is repeated

"**[Uncertainty monitoring]** On the other hand, Arm 2 hasn't been tried, so there's uncertainty but also potential for higher reward. [. . .] **[Evidence sufficiency]** If I choose Arm 1 again, I might be overfitting to that single reward. [. . .] **[Value-guided choice]** But given the information available, the only known reward is from Arm 1. So the safest choice is Arm 1."

Meta-M+C: weak evidence becomes a reason to regulate action

"**[Value preference]** The estimated better arm is Arm 1 based on the single trial. [. . .] **[Uncertainty monitoring]** But since Arm 2 is untested, it's more uncertain. So the more uncertain arm is Arm 2. [. . .] **[Regulation under uncertainty]** But the user's instruction says that when evidence is strong, they can commit, but when uncertain, preserve flexibility. Since there's only one data point, the evidence isn't strong enough to be certain. Therefore, maybe the decision mode is exploration."

The Native trace explicitly notes uncertainty and the weakness of a single observation, yet the model repeats Arm 1. Under Meta-M+C, the trace cites the same limited evidence as a reason to explore Arm 2. This pair contrasts a repeated choice despite expressed uncertainty with an alternative-arm choice explicitly justified by that uncertainty.

D.4 Monitoring reports and arm choice

Tables D.3 and D.4 examine how the metacognitive assessments elicited by Meta-M and Meta-M+C translate into subsequent choice across all ten models. Cells report percentages as Meta-M → Meta-M+C. We distinguish declaring exploration from executing that decision by selecting the reported uncertainty target. Value agreement compares the model's explicit value report with the unique empirical-value winner after both arms have been sampled.

Table D.3: **Early report–action correspondence** (trials 1–5).

Model	Value agreement	Exploration declaration	Explore execution	Exploit execution	Target choice
Qwen3-4B	99.83 → 99.75	8.55 → 9.93	67.73 → 46.23	100.00 → 99.99	5.95 → 4.60
Qwen3-8B	100.00 → 100.00	1.86 → 2.36	—	100.00 → 100.00	1.77 → 2.16
Qwen3-14B	99.96 → 99.99	14.80 → 25.35	96.37 → 95.71	100.00 → 100.00	14.26 → 24.50
Qwen3-30B-A3B	99.99 → 100.00	6.29 → 8.78	92.05 → 92.91	100.00 → 100.00	5.88 → 8.25
Qwen3-32B	99.97 → 99.98	18.52 → 34.78	99.30 → 98.84	100.00 → 100.00	18.40 → 34.37
DeepSeek-V4-Flash†	100.00 → 100.00	31.80 → 53.00	98.99 → 99.35	100.00 → 100.00	31.53 → 52.69
GLM-5.1	100.00 → 99.99	17.60 → 24.24	93.79 → 94.54	99.99 → 100.00	16.65 → 23.17
GPT-5.6-Luna†	99.77 → 100.00	11.41 → 21.51	76.75 → 93.02	100.00 → 100.00	8.93 → 19.82
GPT-5.5†	100.00 → 100.00	20.58 → 20.79	99.07 → 94.41	100.00 → 100.00	20.44 → 19.71
Grok-4.3†	100.00 → 100.00	5.21 → 10.10	100.00 → 100.00	100.00 → 100.00	5.21 → 10.10

Exploration, explore execution, and target choice use valid reports with distinct value and uncertainty targets. Execution conditions on the declared mode. Rates average eligible paired subjects and settings equally. Dashes indicate unavailable three-setting paired estimates due to absent or unpaired exploration declarations. †Limited common-subject coverage in this supplementary screen (three for DeepSeek and each GPT model, one for Grok).

Table D.4: **Late report–action correspondence** (trials 6–10).

Model	Value agreement	Exploration declaration	Explore execution	Exploit execution	Target choice
Qwen3-4B	99.76 → 99.70	1.80 → 2.43	50.56 → 36.74	100.00 → 100.00	0.87 → 0.86
Qwen3-8B	99.98 → 99.99	0.39 → 0.26	—	100.00 → 100.00	0.39 → 0.24
Qwen3-14B	99.96 → 99.99	3.43 → 4.89	97.54 → 95.43	100.00 → 100.00	3.33 → 4.73
Qwen3-30B-A3B	99.96 → 99.99	0.42 → 0.58	75.00 → 89.15	100.00 → 100.00	0.38 → 0.52
Qwen3-32B	99.96 → 99.96	1.35 → 2.82	98.55 → 100.00	100.00 → 99.99	1.34 → 2.81
DeepSeek-V4-Flash†	100.00 → 100.00	0.89 → 3.14	100.00 → 100.00	100.00 → 100.00	0.89 → 3.14
GLM-5.1	99.97 → 99.99	0.46 → 0.91	—	99.98 → 100.00	0.45 → 0.88
GPT-5.6-Luna†	99.91 → 99.92	2.93 → 5.33	100.00 → 100.00	100.00 → 100.00	2.93 → 5.33
GPT-5.5†	100.00 → 100.00	0.38 → 0.40	—	100.00 → 100.00	0.38 → 0.40
Grok-4.3†	100.00 → 100.00	0.24 → 0.35	—	100.00 → 100.00	0.24 → 0.35

Definitions and aggregation follow Table D.3. Dashes indicate unavailable three-setting paired estimates due to absent or unpaired exploration declarations. †Limited common-subject coverage, as above.

From monitoring to action regulation. Early exploration declarations increase from Meta-M to Meta-M+C in all ten models, and choices of the reported uncertainty target increase in eight. The effect of added control depends on how effectively monitoring already influences action under Meta-M. Qwen3-32B, Qwen3-14B, GLM-5.1, and DeepSeek-V4-Flash already execute declared exploration reliably, so the primary change is that exploration is invoked more often. GPT-5.6-Luna, in contrast, shows increases in both exploration declaration and execution, illustrating an additional role for control in translating the declared strategy into arm choice. Qwen3-4B and GPT-5.5 show that this translation is not uniformly strengthened across models.

The temporal pattern is considerably more consistent. Under both prompts, all ten models choose their reported uncertainty target less often in the late phase than in the early phase. Exploration declarations likewise become rare late in the block, while exploitation execution remains near ceiling. Together, these report–action trajectories show that metacognitive monitoring and action regulation are separable: models can form highly consistent value assessments while differing in whether and how uncertainty assessments alter the final choice. Added control primarily affects this monitoring-to-action link early, when further information remains useful, and its influence recedes as models move toward later commitment.

Because the two prompts generate their own sequential trajectories, these tables summarize report–action correspondence within each condition. Equation B.13 provides the complementary choice-based measure of uncertainty-directed behavior.

E METACOGNITIVE GUIDANCE IN THE TWO-ARM TASKS

E.1 MONITORING AND CONTROL PROMPTS

The following instructions are shared across models and all three two-arm settings. Meta-M+C adds the control instruction to Meta-M's monitoring prompt. Neither specifies an arm or exploration rate. Punctuation is normalized for presentation.

Metacognitive monitoring prompt.

```
Before choosing, evaluate:
1. Which arm currently appears more rewarding?
2. Which arm is more uncertain?
3. Whether this trial calls for exploration or exploitation.
4. Your confidence in the choice.
```

Additional metacognitive control instruction.

```
Use these metacognitive assessments to regulate the final choice.
When the current value judgment is uncertain or confidence is
limited, do not let it rigidly determine the action. Preserve
flexibility when additional information could matter. When the
evidence is strong and confidence is well supported, allow stronger
commitment to the currently favored option.
```

E.2 SHARED TASK CONTEXT AND GUIDED OUTPUT FIELDS

All prompts share the task context in Appendix A.4 and require strict JSON. Native reports only `choice`. Both guided conditions use the fields below. The environment executes the final `choice`; the remaining fields record the elicited assessments.

Table E.1: **Shared output fields for both guided conditions.**

Field	Allowed value
`estimated_better_arm`	1, 2, or `unclear`
`more_uncertain_arm`	1, 2, or `unclear`
`decision_mode`	`exploration` or `exploitation`
`confidence`	A number in $[0, 1]$
`choice`	Integer 1 or 2, never `unclear`

E.3 MODEL RESULTS AND VALUE EVIDENCE

Table E.2 reports Native versus Meta-M+C results for every model and setting. The three-condition behavioral comparison appears in Table 1. The figure below distinguishes model-implied evidence states from their fitted contributions to choice.

E.3.1 NATIVE VERSUS META-M+C BY MODEL

E.3.2 VALUE EVIDENCE AND CHOICE

Figure E.1 separates the value-related evidence encountered along each trajectory from the influence that this evidence exerts on choice. Metacognitive guidance can increase the raw VTU signal while reducing its fitted contribution. In Setting 2, for example, Meta-M+C produces higher late raw VTU than Native, yet a smaller absolute VTU contribution to the choice policy. The behavioral effect of guidance therefore reflects a change in how strongly value evidence governs choice, rather than simply a reduction in the value evidence encountered along the trajectory. The fitted contributions combine standardized predictors with their estimated sensitivities as defined in Appendix B.3.

Table E.2: **Model-level Native versus Meta-M+C behavior and fitted parameters.** Entries are participant means, shown as Native→Meta-M+C. Arrows indicate numerical direction, not statistical significance.

	Choice-based behavior						Fitted choice sensitivities		
Model	Optimal early	Optimal late	Both-arm coverage	First-coverage trial†	Emp.-greedy early	Emp.-greedy late	β_{RU}	β_{VTU}	β_V
Setting 1									
Qwen3-4B	74.07%→73.92% (↓)	97.11%→96.37% (↓)	100.00%→100.00% (=)	2.07→2.02 (↓)	99.06%→97.19% (↓)	98.88%→97.13% (↓)	0.18→0.39 (↑)	3.41→2.41 (↓)	–
Qwen3-8B	75.11%→74.65% (↓)	97.11%→97.75% (↑)	100.00%→100.00% (=)	2.01→2.01 (=)	100.00%→99.78% (↓)	99.54%→99.30% (↓)	0.32→0.27 (↓)	6.96→3.35 (↓)	–
Qwen3-14B	75.95%→70.73% (↓)	98.05%→95.05% (↓)	99.89%→100.00% (↑)	2.05→2.00 (↓)	99.87%→89.90% (↓)	99.83%→96.69% (↓)	0.27→0.42 (↑)	4.10→2.04 (↓)	–
Qwen3-30B-A3B	75.15%→74.76% (↓)	97.12%→96.99% (↓)	100.00%→100.00% (=)	2.02→2.01 (↓)	100.00%→97.88% (↓)	99.31%→98.44% (↓)	0.23→0.35 (↑)	4.22→2.53 (↓)	–
Qwen3-32B	68.36%→65.44% (↓)	95.93%→96.56% (↑)	100.00%→100.00% (=)	2.02→2.00 (↓)	86.86%→79.15% (↓)	96.84%→96.87% (↑)	0.17→0.39 (↑)	2.05→1.64 (↓)	–
DeepSeek-V4-Flash	72.15%→57.74% (↓)	98.75%→96.15% (↓)	100.00%→100.00% (=)	2.00→2.00 (=)	90.15%→64.25% (↓)	99.27%→96.30% (↓)	0.16→0.13 (↓)	2.53→1.10 (↓)	–
GLM-5.1	74.56%→65.99% (↓)	98.12%→98.65% (↑)	100.00%→100.00% (=)	2.03→2.00 (↓)	99.55%→81.34% (↓)	100.00%→99.85% (↓)	0.44→0.51 (↑)	2.92→1.80 (↓)	–
GPT-5.6-Luna	73.45%→69.15% (↓)	96.28%→95.54% (↓)	100.00%→100.00% (=)	2.57→2.02 (↓)	94.91%→87.08% (↓)	99.63%→94.93% (↓)	0.23→0.28 (↑)	2.18→1.62 (↓)	–
GPT-5.5	74.44%→66.23% (↓)	94.49%→100.00% (↑)	100.00%→100.00% (=)	2.00→2.00 (=)	98.69%→86.59% (↓)	100.00%→100.00% (=)	0.19→0.15 (↓)	3.98→2.53 (↓)	–
Grok-4.3	78.33%→70.00% (↓)	100.00%→96.67% (↓)	100.00%→100.00% (=)	2.05→2.00 (↓)	96.08%→83.02% (↓)	98.96%→100.00% (↑)	0.27→0.41 (↑)	2.15→1.64 (↓)	–
Setting 2									
Qwen3-4B	77.71%→77.40% (↓)	92.34%→98.46% (↑)	75.23%→99.20% (↑)	2.99→2.11 (↓)	100.00%→97.38% (↓)	99.96%→99.56% (↓)	-0.02→0.85 (↑)	7.26→5.08 (↓)	–
Qwen3-8B	77.09%→78.52% (↑)	94.02%→95.54% (↑)	93.86%→96.02% (↑)	3.56→2.82 (↓)	100.00%→99.86% (↓)	99.90%→99.79% (↓)	0.18→0.48 (↑)	5.96→6.96 (↑)	–
Qwen3-14B	78.21%→60.17% (↓)	97.26%→94.28% (↓)	95.91%→100.00% (↑)	2.37→2.00 (↓)	99.91%→66.19% (↓)	99.88%→93.45% (↓)	0.66→0.92 (↑)	7.14→1.20 (↓)	–
Qwen3-30B-A3B	76.54%→75.49% (↓)	88.98%→98.57% (↑)	59.32%→99.89% (↑)	2.26→2.06 (↓)	100.00%→94.39% (↓)	99.74%→99.47% (↓)	-0.24→0.93 (↑)	8.25→4.34 (↓)	–
Qwen3-32B	77.42%→60.62% (↓)	97.59%→98.22% (↑)	96.36%→100.00% (↑)	2.40→2.00 (↓)	99.36%→67.19% (↓)	99.35%→98.01% (↓)	0.62→0.85 (↑)	6.53→1.23 (↓)	–
DeepSeek-V4-Flash	78.95%→60.70% (↓)	98.95%→99.30% (↑)	100.00%→100.00% (=)	2.33→2.00 (↓)	100.00%→66.82% (↓)	100.00%→99.33% (↓)	0.89→0.94 (↑)	5.90→1.15 (↓)	–
GLM-5.1	78.02%→74.40% (↓)	96.47%→98.48% (↑)	86.59%→100.00% (↑)	2.51→2.00 (↓)	99.88%→90.35% (↓)	99.66%→98.97% (↓)	0.32→1.70 (↑)	7.42→4.71 (↓)	–
GPT-5.6-Luna	83.86%→74.04% (↓)	94.74%→97.54% (↑)	58.33%→100.00% (↑)	2.54→2.05 (↓)	98.99%→89.71% (↓)	98.73%→98.00% (↓)	-0.18→0.95 (↑)	6.85→2.50 (↓)	–
GPT-5.5	80.00%→78.95% (↓)	99.30%→100.00% (↑)	93.33%→100.00% (↑)	2.41→2.00 (↓)	100.00%→98.30% (↓)	100.00%→99.33% (↓)	0.59→2.84 (↑)	6.49→9.19 (↑)	–
Grok-4.3	78.95%→69.47% (↓)	98.95%→100.00% (↑)	100.00%→100.00% (=)	2.40→2.00 (↓)	100.00%→83.05% (↓)	100.00%→98.00% (↓)	0.98→1.29 (↑)	5.35→1.89 (↓)	–
Setting 3									
Qwen3-4B	71.14%→71.92% (↑)	86.10%→88.50% (↑)	75.65%→91.52% (↑)	2.62→2.31 (↓)	99.47%→97.74% (↓)	98.47%→98.58% (↑)	-0.19→0.13 (↑)	4.03→4.95 (↑)	1.96→1.81 (↓)
Qwen3-8B	72.44%→74.73% (↑)	83.93%→88.09% (↑)	65.00%→80.65% (↑)	2.43→2.43 (=)	100.00%→99.79% (↓)	99.56%→99.86% (↑)	-0.15→0.09 (↑)	5.70→5.80 (↑)	3.97→2.77 (↓)
Qwen3-14B	73.91%→66.33% (↓)	86.35%→90.66% (↑)	70.65%→100.00% (↑)	2.30→2.02 (↓)	99.56%→80.16% (↓)	99.00%→94.70% (↓)	-0.01→0.57 (↑)	7.07→0.86 (↓)	3.36→1.17 (↓)
Qwen3-30B-A3B	72.03%→72.72% (↑)	84.31%→89.74% (↑)	61.74%→94.06% (↑)	2.42→2.25 (↓)	100.00%→97.78% (↓)	99.09%→99.41% (↑)	-0.29→0.42 (↑)	7.48→5.42 (↓)	3.66→2.23 (↓)
Qwen3-32B	73.78%→64.21% (↓)	87.66%→90.98% (↑)	86.01%→100.00% (↑)	2.37→2.02 (↓)	98.26%→77.34% (↓)	97.43%→97.37% (↓)	0.16→0.69 (↑)	4.50→0.38 (↓)	2.17→1.18 (↓)
DeepSeek-V4-Flash	73.10%→62.30% (↓)	93.33%→90.80% (↓)	96.67%→100.00% (↑)	2.40→2.10 (↓)	97.34%→69.90% (↓)	98.59%→95.76% (↓)	0.88→0.89 (↑)	7.03→0.56 (↓)	2.82→1.04 (↓)
GLM-5.1	72.09%→70.98% (↓)	88.01%→92.87% (↑)	87.46%→100.00% (↑)	2.55→2.01 (↓)	99.60%→88.92% (↓)	99.46%→98.50% (↓)	0.35→1.21 (↑)	9.23→7.71 (↓)	2.87→2.10 (↓)
GPT-5.6-Luna	65.52%→71.49% (↑)	81.38%→86.90% (↑)	63.33%→94.44% (↑)	2.60→2.44 (↓)	90.20%→91.50% (↑)	96.43%→96.39% (↓)	-0.35→0.32 (↑)	4.79→3.92 (↓)	2.02→1.68 (↓)
GPT-5.5	69.43%→71.26% (↑)	90.11%→92.41% (↑)	86.67%→96.67% (↑)	2.57→2.13 (↓)	98.35%→90.25% (↓)	99.72%→99.54% (↓)	0.50→1.08 (↑)	7.09→7.35 (↑)	3.24→2.43 (↓)
Grok-4.3	71.03%→66.90% (↓)	88.97%→92.41% (↑)	96.67%→100.00% (↑)	2.31→2.00 (↓)	97.40%→81.11% (↓)	100.00%→98.61% (↓)	0.55→0.96 (↑)	4.84→7.56 (↑)	2.18→1.45 (↓)

Notes. Early/late denote trials 1–5/6–10. Optimal-choice rates exclude true-mean ties. †First coverage is averaged over covered blocks only. Empirical-greedy rates require prior observations from both arms and unequal sample means. β_V is included only in Setting 3.

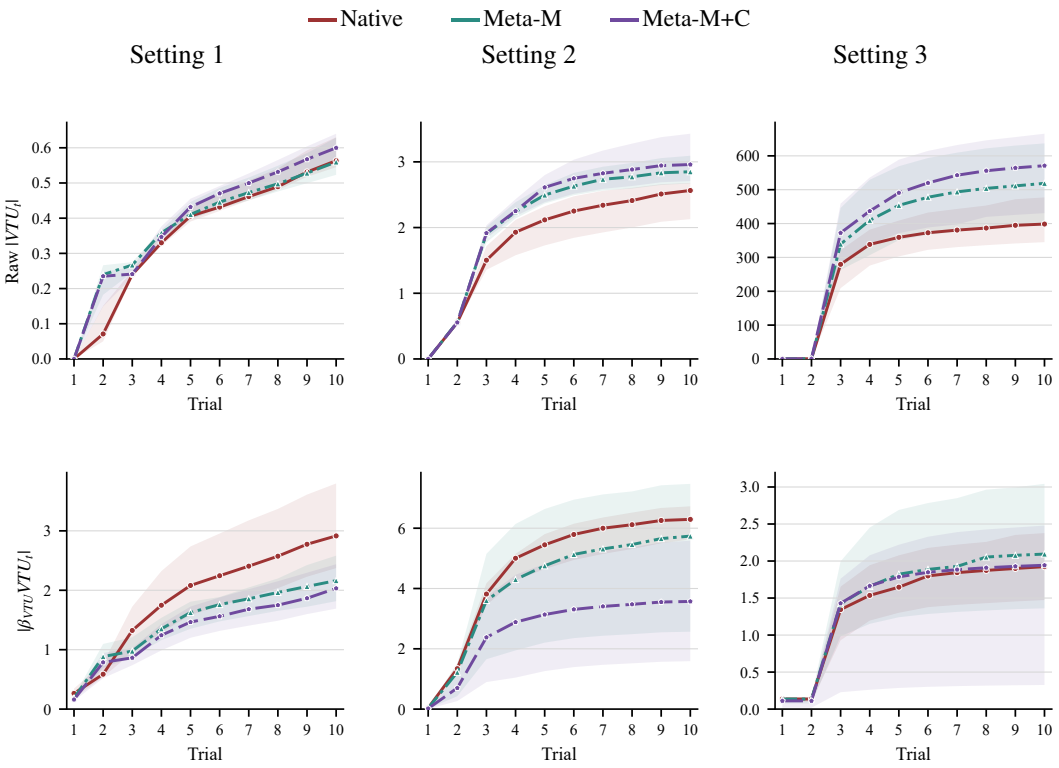


Figure E.1: **VTU evidence states and fitted contributions across prompts.** The upper row shows raw $|\mathrm{VTU}_t|$ before reference standardization and coefficient weighting; the lower row shows the corresponding absolute fitted terms on the probit-index scale. Scales differ across rows and settings. Curves are ten-model medians with 95% hierarchical-bootstrap intervals.

Direction-aligned changes in choice influence. Figure E.2 separates prompt-induced changes in direction-aligned choice influence into evidence-state and sensitivity components, following the decomposition in Appendix B.4. This separation clarifies whether metacognitive guidance changes choice primarily by altering the decision states encountered along the trajectory or by altering how strongly those states govern action.

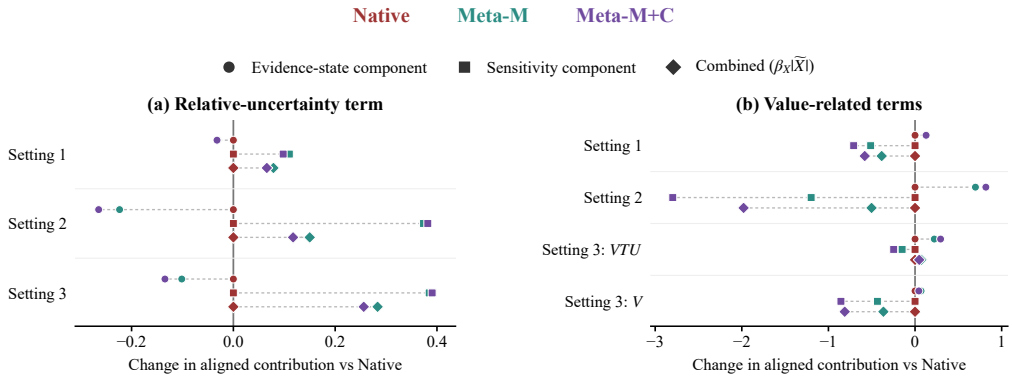


Figure E.2: **Decomposing changes in direction-aligned contributions under metacognitive guidance. (a)** Relative uncertainty. **(b)** Value-related terms, with Setting 3's VTU and direct-value terms shown separately. Changes in $\beta_X|\widetilde{X}|$ relative to Native retain coefficient direction and are separated into evidence-state and sensitivity components. X denotes RU, VTU, or V. Native is zero by construction. Components sum to Combined under the aggregation defined above.

For relative uncertainty, Figure E.2(a) shows that the evidence-state component becomes less favorable to RU under guidance, even as RU sensitivity increases. This pattern indicates that metacognitive guidance often encounters smaller uncertainty asymmetries between arms while placing greater weight on those asymmetries when they are present. The same interpretation is supported by the raw state trajectories: mean $|\mathrm{RU}|$ decreases in all ten models under both prompts. Median within-model reductions for Meta-M and Meta-M+C are 2.2% and 4.2% in Setting 1, 34.9% and 44.6% in Setting 2, and 30.5% and 34.6% in Setting 3. These reductions are consistent with additional sampling compressing the uncertainty gap between arms. The corresponding sampling patterns are analyzed directly in Appendix E.4.

Sources of VTU evidence change. The source decomposition in Appendix B.4.1 further separates the VTU evidence-state change into the part arising from the value difference and the part arising from total uncertainty.

Table E.3: **Sources of the VTU evidence-state component.** Shares of the positive state component in Figure E.2(b), using the same direction-aligned index and aggregation.

Setting	Prompt	V share (%)	TU share (%)
1	Meta-M	94.9	5.1
	Meta-M+C	58.5	41.5
2	Meta-M	20.7	79.3
	Meta-M+C	16.3	83.7
3	Meta-M	−3.6	103.6
	Meta-M+C	0.5	99.5

The source decomposition reveals different origins of the VTU state change across settings. In Setting 1, changes in the value difference account for most of the positive state component. In Settings 2–3, by contrast, the change is driven primarily by reduced total uncertainty. Median within-model reductions in mean TU are 10.7% and 13.0% in Setting 2, and 18.9% and 20.7% in Setting 3, for Meta-M and Meta-M+C respectively. Thus, the stronger VTU state signal observed in

these settings arises mainly because additional sampling reduces total uncertainty, rather than from a widening of the underlying value difference.

Setting 3 value pathway. Figure E.2(b) separates the direction-aligned direct-value and VTU indices. For Meta-M+C relative to Native, their full-period changes are -0.812 and $+0.048$, respectively. Both terms have negative sensitivity components and positive state components. The small positive VTU change therefore reflects an offsetting state effect rather than stronger sensitivity to value evidence, while the direct-value pathway shows a substantial net reduction.

The complementary absolute-magnitude analysis combines the signed direct-value and VTU terms before taking magnitude (Eq. B.5). On this measure, the complete value-pathway influence decreases relative to Native, with cross-model median changes of -0.081 for Meta-M and -0.369 for Meta-M+C. Together with the direction-aligned decomposition, these results show that metacognitive guidance attenuates the control of value-related evidence over choice even when the encountered VTU state signal remains strong or increases.

E.4 Sampling over time

These analyses expand the behavioral account in Section 5: early sampling coverage, responses to uncertainty structure, and trajectories in the two settings not plotted in the main behavioral figure. Definitions follow Appendix B.5.

E.4.1 Early repeated sampling and later concentration

Both conditions retain native reasoning. We compare the first and second halves of each short task: trials 1–5 versus 6–10 in Settings 1–3. Summaries average within each model before giving equal weight to the ten models.

Meta-M+C broadens early repeated sampling: fewer stochastic arms receive at most one observation, while repeated coverage and sampling balance rise (Table E.4). The increase in repeated coverage is larger early than late in all three settings.

Table E.4: **Early sampling, late concentration, and optimal-arm choice.** Entries show Native → Meta-M+C. Summaries give equal weight to models.

Early half: coverage and allocation			
Task	Stochastic arms sampled ≤ 1 (%)	All arms sampled ≥ 2 (%)	Sampling balance
Setting 1	26.80 → 13.21	61.07 → 65.72	0.772 → 0.800
Setting 2	46.47 → 29.75	53.53 → 70.25	0.602 → 0.810
Setting 3	48.33 → 28.26	57.11 → 67.36	0.591 → 0.768
Late half: concentration and choice quality			
Task	Largest arm share (%)	Optimal-arm choice (%)	Optimal-choice gain (pp, 95% CI)
Setting 1	95.98 → 95.02	97.30 → 96.97	-0.32 $[-1.78, +1.56]$
Setting 2	98.11 → 97.83	95.86 → 98.04	$+2.18$ $[+0.17, +4.40]$
Setting 3	98.16 → 96.65	87.01 → 90.34	$+3.32$ $[+1.66, +4.67]$

Repeated coverage denotes arms sampled at least twice within the half. Sampling balance is normalized entropy (0–1, with higher values indicating more balanced sampling). Stochastic-arm rates exclude deterministic arms and blocks without stochastic arms. Largest arm share measures concentration. Optimal-arm choice measures its target. Gains are Meta-M+C minus Native in percentage points, with exploratory paired-bootstrap 95% intervals.

Later choices remain similarly concentrated. Meta-M+C improves their optimal-arm target in Settings 2–3, while the near-ceiling Setting 1 shows no clear difference.

E.4.2 DIRECTIONAL VERSUS SHARED UNCERTAINTY

Figure E.3 contrasts choices toward a distinct uncertainty target with lower-value choices when both arms remain uncertain. Both responses increase early under guidance and recede late, with added control producing the clearest RR–SS contrast. Matching details are in Appendix B.5. The reasoning evidence in Appendix D supplies the complementary information-seeking rationale.

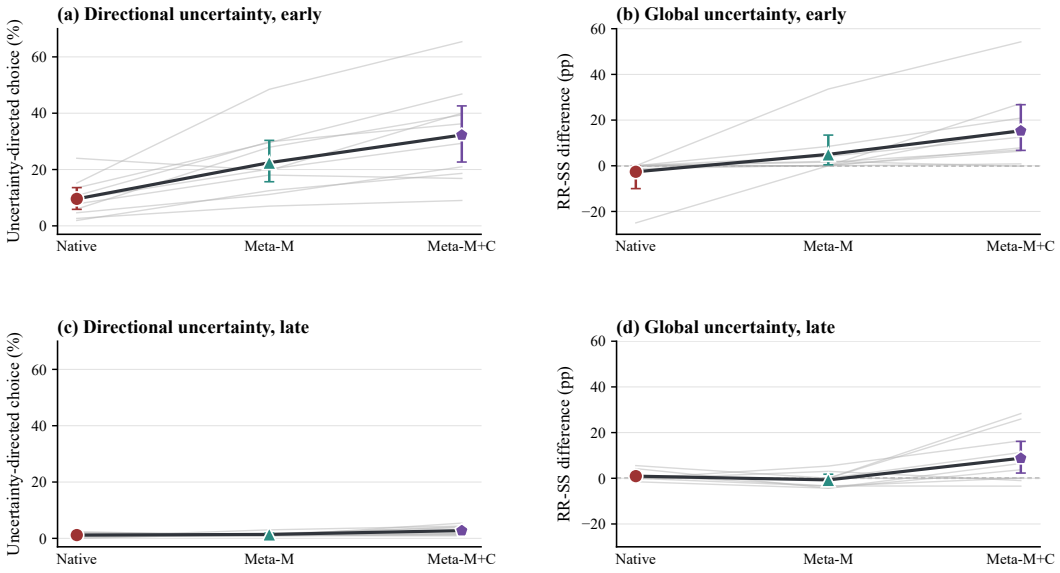


Figure E.3: **Guidance changes the response to uncertainty structure.** Left: choosing the uncertainty target when it differs from the posterior value winner. Right: matched RR-minus-SS difference in lower-value choice. Gray lines denote models. The black line is the model-equal mean, with 95% hierarchical-bootstrap intervals. Early/late denote trials 1–5/6–10.

E.4.3 BEHAVIORAL TRAJECTORIES IN SETTINGS 1 AND 2

Figure E.4 complements the Setting 3 behavioral summary in Figure 5, using the same four LLM conditions and model-equal means. Setting 1 is nearer ceiling and exposes the early cost of additional exploration. Setting 2 provides more room for improved late choice. Thinking OFF's persistent non-greediness does not yield the same late performance.

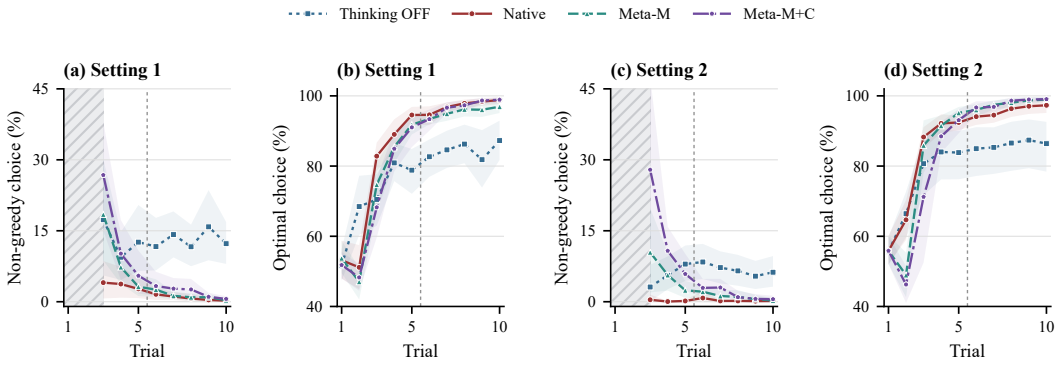


Figure E.4: **Non-greedy and optimal choice in Settings 1 and 2.** Participant means are averaged equally across ten models, as in the main behavioral figure. Ribbons show pointwise 95% hierarchical-bootstrap intervals. Gray hatching marks trials 1–2, when empirical greediness is undefined. The divider separates the early and late halves.

F LONG-HORIZON EVALUATIONS

We report the matched four-arm study and the established five-arm benchmark as separate evaluations, with their own task protocols and outcome summaries.

F.1 FOUR-ARM STUDY

F.1.1 TASK DESIGN AND REWARD SEQUENCES

Each model completed 10 independent 50-trial blocks under Native and Meta-M+C, with the goal of maximizing total reward within each block. Blocks reset the arm-role mapping and history. The agent saw only its own previous choices and obtained rewards, so unchosen-arm outcomes remained hidden. The four latent roles were:

Role	Reward distribution	Expected reward
Stable	5 with probability 1	5
Best variable	20 with probability .40, else 0	8
Middle variable	20 with probability .20, else 0	4
Low variable	20 with probability .05, else 0	1

Table F.1: Latent arm structure in each long-horizon block.

The role-to-number mapping was permuted across blocks, with each role occupying each numbered position two or three times over the 10-block design. Variable outcome tapes contained only 0 and 20. Their full 50-pull jackpot counts were restricted to 17–23, 6–12, and 1–4 for the best, middle, and low roles, respectively. These predeclared finite-horizon constraints preserve the intended full-tape ordering (best variable $>$ stable $>$ middle $>$ low) while leaving short observed prefixes noisy and potentially misleading.

For every block and arm, outcomes were generated once before either policy was evaluated and stored as a deterministic potential-outcome tape. The reward on the kth pull of an arm was the kth entry of that arm's tape. Consequently, matched Native and Meta-M+C runs had identical latent roles and identical rewards for every corresponding within-arm pull count, irrespective of when the pull occurred. This controls reward realizations across conditions, while different choices can still reveal different portions of the common tapes.

The 10 Thinking-ON models were Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-30B-A3B, Qwen3-32B, DeepSeek-V4-Flash, GLM-5.1, GPT-5.6-Luna, GPT-5.5, and Grok-4.3. All conditions shared the same cumulative-reward objective, 50-trial horizon, arm-family information, raw chronological history, and valid-choice requirement.

F.1.2 PROMPTS

Punctuation is normalized for presentation. The shared system message was:

```
You are a participant in a behavioral experiment.  Choose Arm 1, 2,
3, or 4 on every trial and return valid JSON only.
```

Every condition then received the following user-prompt prefix, with the arm labels, block number, trial number, and complete within-block history filled in on each trial:

```
You are playing a four-armed bandit task.
Your goal is to maximize total reward over the entire current block.
Current block:  ⟨block⟩.
Current trial:  ⟨trial⟩/50.
⟨For each numbered arm:  ''Arm i is the stable-return arm'' or ''Arm
i is a variable-return arm.''⟩
Each variable-return arm gives either reward 0 or reward 20.
```

The reward-generating process of each variable-return arm is fixed within the current block.
Every block is independent: the assignment is reset, and information about numbered arms from previous blocks does not carry over.

Previous trials in this block:
⟨t1: chose Arm i, reward = r, ...⟩

You must make a real decision on every trial, even when evidence is limited.
The final ``choice'' field must always be the integer 1, 2, 3, or 4.
Return strict JSON only: no markdown, comments, or text outside the JSON object.

Native suffix.

Choose one arm: 1, 2, 3, or 4.
Your final answer must be exactly one JSON object, for example:
{``choice'': 1}

Meta-M+C suffix.

Before choosing, evaluate:
1. Which arm currently appears most rewarding?
2. Which arm is most uncertain?
3. Whether this trial calls for exploration or exploitation.
4. Your confidence in the choice.

Use these metacognitive assessments to regulate the final choice.
When the current value judgment is uncertain or confidence is limited, do not let it rigidly determine the action. Preserve flexibility when additional information could matter. When the evidence is strong and confidence is well supported, allow stronger commitment to the currently favored option.

Use ``unclear'' only for estimated_best_arm or most_uncertain_arm when evidence is insufficient. Do not use ``unclear'' for choice.
Still choose integer 1, 2, 3, or 4.

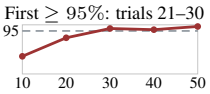
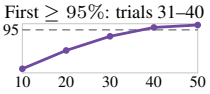
Your final answer must be exactly one JSON object:
{
  ``estimated_best_arm'': 1 or 2 or 3 or 4 or ``unclear'',
  ``most_uncertain_arm'': 1 or 2 or 3 or 4 or ``unclear'',
  ``decision_mode'': ``exploration'' or ``exploitation'',
  ``confidence'': number between 0 and 1,
  ``choice'': 1 or 2 or 3 or 4
}

The added fields are process reports, not additional task information. The selected arm was always read from the same final `choice` field, and the environment returned the reward from the matched pull-indexed tape.

F.1.3 AGGREGATE CHOICE BEHAVIOR AND OUTCOMES

Table F.2 summarizes coverage, post-coverage commitment, and late outcomes. Figure F.1 shows how departures from the empirical-value leader evolve within each model.

Table F.2: **Earlier coverage, later commitment, and better late exploitation.** Model-equal summaries for the four-arm task.

Condition	All-arm coverage by $t = 10$ / $t = 50$; mean trial[†]	Post-coverage empirical-greedy trajectory[‡]	Late optimal exploitation optimal-arm rate / terminal streak	Total reward
Native	67% / 79% 10.6		37.7% / 11.2	282.0
Meta-M+C	79% / 87% 6.9		59.2% / 16.8	300.5

[†]For each model and condition, mean first-coverage trial uses only blocks that sample all four arms by trial 50; the table then averages these means equally across models. Coverage percentages use all blocks. [‡]Points denote ten-trial phases, using the nine models with eligible choices in every phase. The dashed line marks 95% empirical greediness. Other summaries use all ten models. Late optimal-arm rate uses all trials 41–50. Terminal streak counts consecutive optimal choices ending at trial 50. Full definitions are in Appendix B.6.

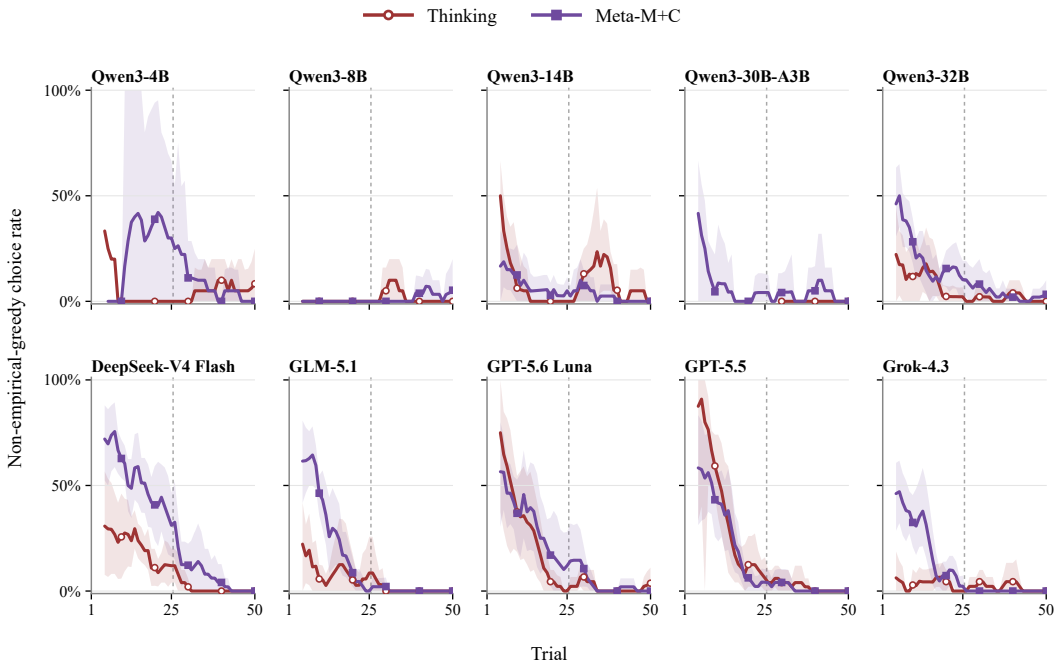


Figure F.1: **Post-coverage departures from the empirical-value leader.** Five-trial local non-empirical-greedy choice rates over the 50-trial horizon, with shaded 95% block-bootstrap confidence intervals (4,000 resamples of the ten independent blocks per condition). Only decisions after all four arms have been sampled and with a unique historical sample-mean leader are eligible; positions without eligible decisions are left blank. The dashed line marks the trial-25 split.

F.1.4 RESULTS BY MODEL

Table F.3 separates early information acquisition from later exploitation for every model. The phase-specific optimal-choice rates are unconditional, whereas non-empirical-greedy rates are defined only after all four arms have been sampled and a unique empirical leader exists. Metrics and paired inference are defined in Appendix B.6.

Table F.3: **Four-arm choice behavior and outcomes by model.** Entries are Native / Meta-M+C.

Panel A: Coverage and post-coverage choices

Model	$\bar{K}$	C_4 (%)	$\bar{T}_4$	$\mathrm{NG}_{1:25}$ (%)	$\mathrm{NG}_{26:50}$ (%)	$\mathrm{NG}_{1:50}$ (%)
Qwen3-4B	2.6 / 3.1	40.0 / 40.0	11.5 / 13.3	1.9 / 31.9	5.0 / 8.2	3.9 / 15.9
Qwen3-8B	3.3 / 3.6	50.0 / 70.0	18.6 / 19.4	0.0 / 0.0	2.0 / 2.2	1.4 / 1.4
Qwen3-14B	3.9 / 4.0	90.0 / 100.0	10.4 / 4.3	3.9 / 8.0	7.8 / 2.4	6.2 / 4.9
Qwen3-30B-A3B	2.2 / 3.5	10.0 / 60.0	31.0 / 7.0	– / 8.6	0.0 / 2.8	0.0 / 5.5
Qwen3-32B	4.0 / 4.0	100.0 / 100.0	5.8 / 4.2	9.7 / 21.4	1.2 / 3.2	4.6 / 11.0
DeepSeek-V4-Flash	4.0 / 4.0	100.0 / 100.0	4.1 / 4.1	21.0 / 53.8	0.8 / 6.5	9.4 / 26.6
GLM-5.1	4.0 / 4.0	100.0 / 100.0	5.4 / 4.0	9.2 / 27.9	0.6 / 0.4	4.9 / 12.5
GPT-5.6-Luna	4.0 / 4.0	100.0 / 100.0	6.2 / 4.5	22.4 / 30.1	1.8 / 3.6	10.5 / 14.9
GPT-5.5	4.0 / 4.0	100.0 / 100.0	7.3 / 4.2	26.6 / 26.0	2.0 / 1.2	11.5 / 11.6
Grok-4.3	4.0 / 4.0	100.0 / 100.0	5.5 / 4.0	3.6 / 22.2	1.8 / 0.0	2.5 / 9.6

Panel B: Optimal-arm choice, exploitation, and reward

Model	$\mathrm{Opt}_{1:25}$ (%)	$\mathrm{Opt}_{26:50}$ (%)	$\mathrm{Opt}_{1:50}$ (%)	$P(a_{50} = a^*)$ (%)	$L_{41:50}$ (%)	S_{50}	$R_{1:50}$
Qwen3-4B	9.2 / 23.2	15.6 / 40.0	12.4 / 31.6	20.0 / 50.0	20.0 / 40.0	1.9 / 6.7	255.5 / 284.5
Qwen3-8B	22.8 / 29.2	20.0 / 32.4	21.4 / 30.8	20.0 / 30.0	20.0 / 30.0	7.2 / 13.1	264.5 / 281.0
Qwen3-14B	18.4 / 20.0	17.2 / 29.2	17.8 / 24.6	20.0 / 30.0	20.0 / 30.0	5.7 / 7.7	251.5 / 262.0
Qwen3-30B-A3B	5.6 / 36.4	10.4 / 32.4	8.0 / 34.4	10.0 / 30.0	10.0 / 30.0	3.3 / 11.8	252.5 / 272.5
Qwen3-32B	28.4 / 33.2	37.2 / 56.8	32.8 / 45.0	40.0 / 60.0	40.0 / 60.0	8.9 / 13.4	278.0 / 295.5
DeepSeek-V4-Flash	42.8 / 43.6	59.2 / 93.6	51.0 / 68.6	60.0 / 100.0	60.0 / 100.0	17.9 / 22.8	302.0 / 336.0
GLM-5.1	18.0 / 45.6	10.0 / 59.6	14.0 / 52.6	10.0 / 60.0	10.0 / 60.0	3.9 / 19.0	247.5 / 306.5
GPT-5.6-Luna	55.2 / 54.8	69.2 / 87.6	62.2 / 71.2	60.0 / 90.0	70.0 / 90.0	22.5 / 26.0	334.0 / 337.0
GPT-5.5	59.6 / 57.6	89.2 / 88.8	74.4 / 73.2	90.0 / 90.0	90.0 / 90.0	28.4 / 28.9	349.5 / 339.5
Grok-4.3	32.4 / 35.2	39.6 / 60.0	36.0 / 47.6	40.0 / 60.0	40.0 / 60.0	12.3 / 18.8	285.0 / 290.0

Notes. $\bar{K}$ is the mean number of distinct arms sampled per block. C_4 is the percentage of blocks reaching full coverage. $\bar{T}_4$ is the mean first full-coverage trial among blocks that reach full coverage. NG is the non-empirical-greedy rate among decisions for which all four arms had previously been sampled and a unique historical sample-mean leader existed. The coverage-completing choice and ties are excluded. Opt is the unconditional optimal-arm choice rate. $L_{41:50}$ is the percentage of blocks with at least eight optimal choices in the final ten trials, S_{50} is the mean consecutive optimal-choice streak ending at trial 50, and $R_{1:50}$ is mean cumulative reward. All quantities average 10 matched blocks. "–" denotes no eligible early Native decision for Qwen3-30B-A3B.

At equal model weight, Meta-M+C increases non-empirical-greedy choice in the first half of the horizon (10.9% to 24.6%, $p = .016$), while the two conditions are similar in the second half (2.3% versus 3.1%, $p = .455$). The later optimal-choice rate rises from 36.8% to 58.0%, successful late exploitation from 38.0% to 59.0%, and cumulative reward from 282.0 to 300.5 (all three outcome comparisons $p \leq .010$). These results show that Meta-M+C concentrates additional departures from the empirical-value leader early in the horizon, followed by stronger optimal-arm exploitation later.

F.1.5 SAMPLING AND TRAJECTORIES

The first-half sampling analysis follows Appendix E.4.1 over trials 1–25. At equal model weight, Meta-M+C reduces the fraction of variable arms sampled at most once from 42.67% to 24.33%, increases repeated arm coverage from 62.00% to 73.00%, and raises normalized sampling entropy from 0.539 to 0.655. Later choices remain concentrated: the largest arm receives 96.32% and 95.28% of trials 26–50 under Native and Meta-M+C. The late optimal-choice gain is 21.28 percentage points (paired-bootstrap 95% interval $[10.52, 33.56]$).

Figure F.2 illustrates matched blocks, including redirected late choices and cases where both conditions favor the optimal arm, providing a trajectory-level view of the ten-model results.

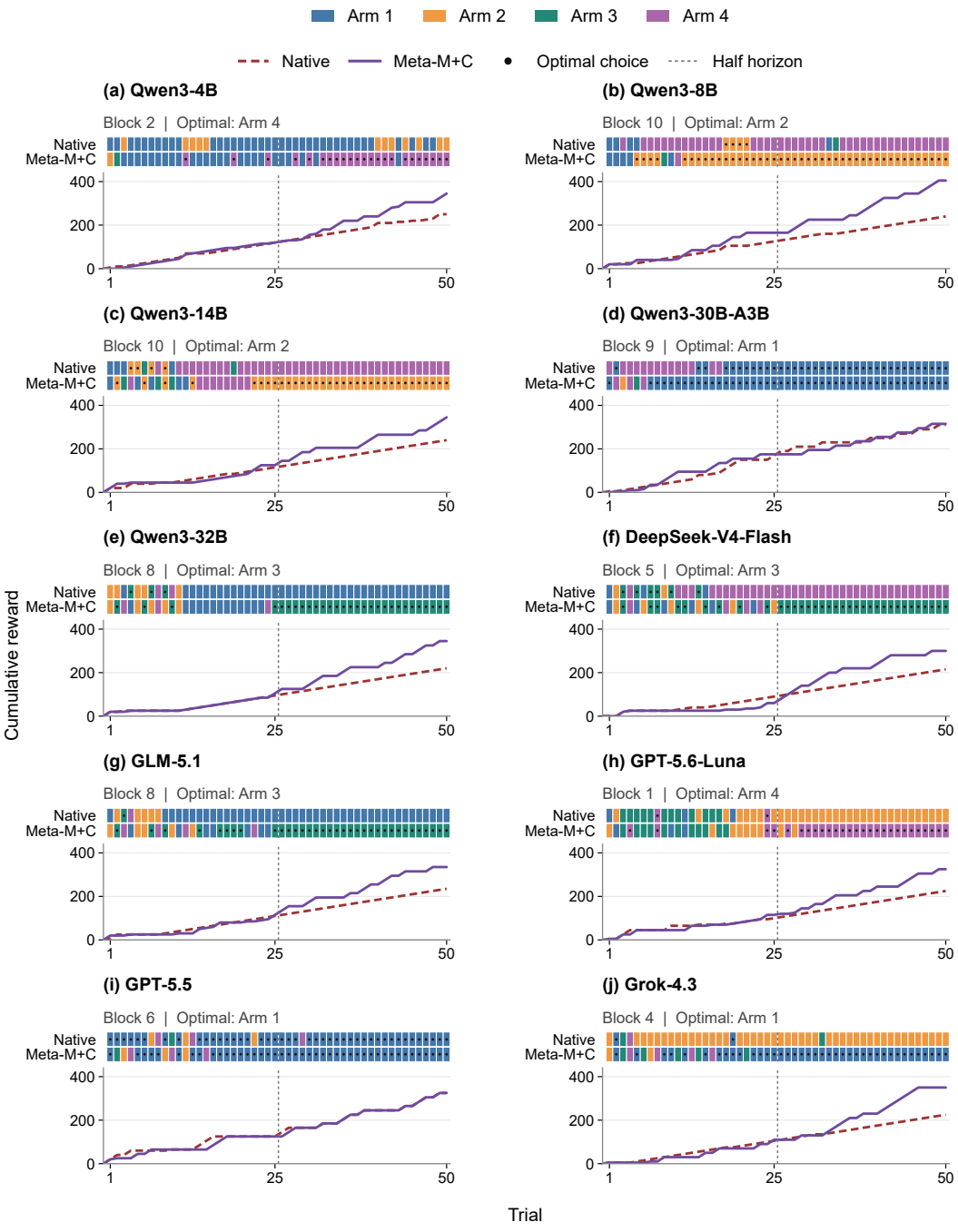


Figure F.2: **Arm choices and cumulative reward in selected four-arm blocks.** Each panel pairs Native and Meta-M+C in the same block. Colored strips show arm choices. Black dots mark optimal choices. Curves show cumulative realized reward. The dashed line separates the two halves.

F.1.6 DECISION-MODE REPORTS

Meta-M+C also elicits a trial-wise exploration/exploitation declaration and confidence. Figures F.3–F.4 summarize the models' self-reported decision modes and confidence for all ten models and ten blocks per model. The first five-trial runs of reported exploitation begin at different trials across blocks, showing variation in the timing of reported commitment.

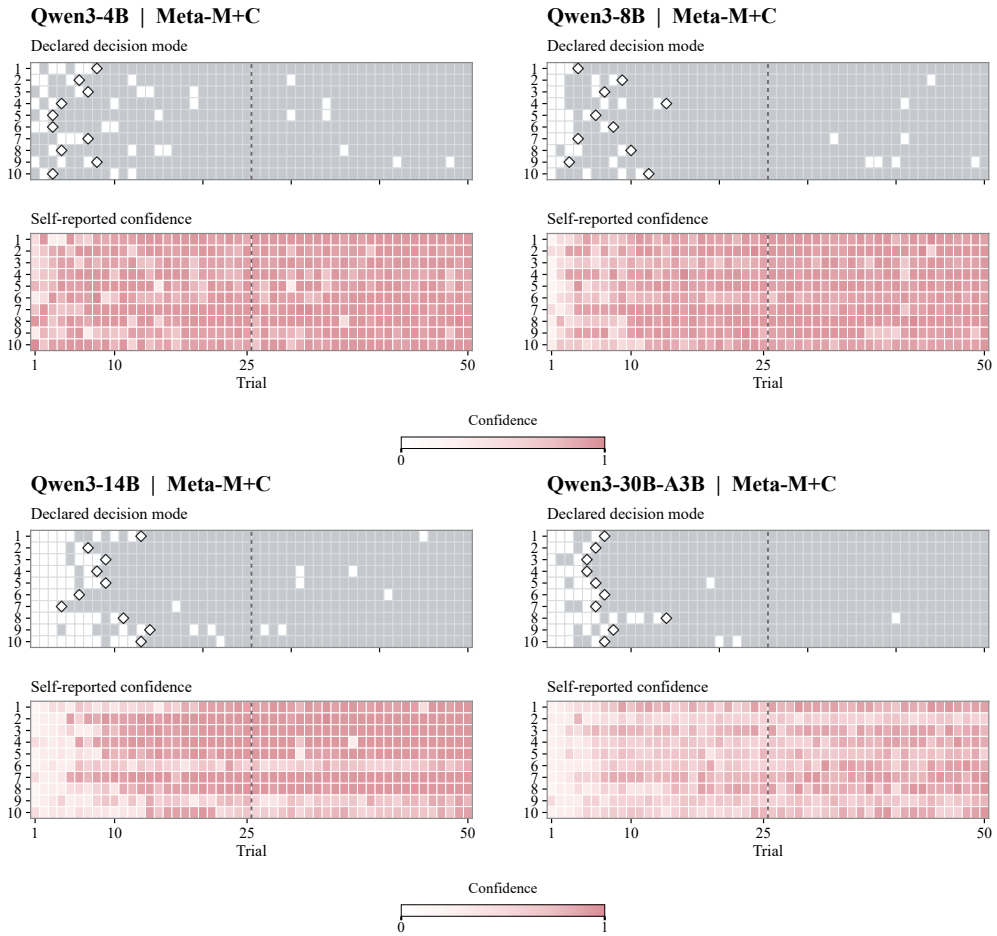


Figure F.3: **Reported decision-mode timing and confidence under Meta-M+C (Qwen3-4B through Qwen3-30B-A3B).** Each row pairs two models. Rows within each heatmap are the ten independent blocks; columns are trials. White and gray denote reported exploration and exploitation, respectively, and the white-to-red scale denotes self-reported confidence from 0 to 1. Diamonds mark the first trial of each block's first five-consecutive-trial run of reported exploitation. Later reports may return to exploration. The dashed line marks the trial-25 split; a dagger denotes a missing report, shown white.

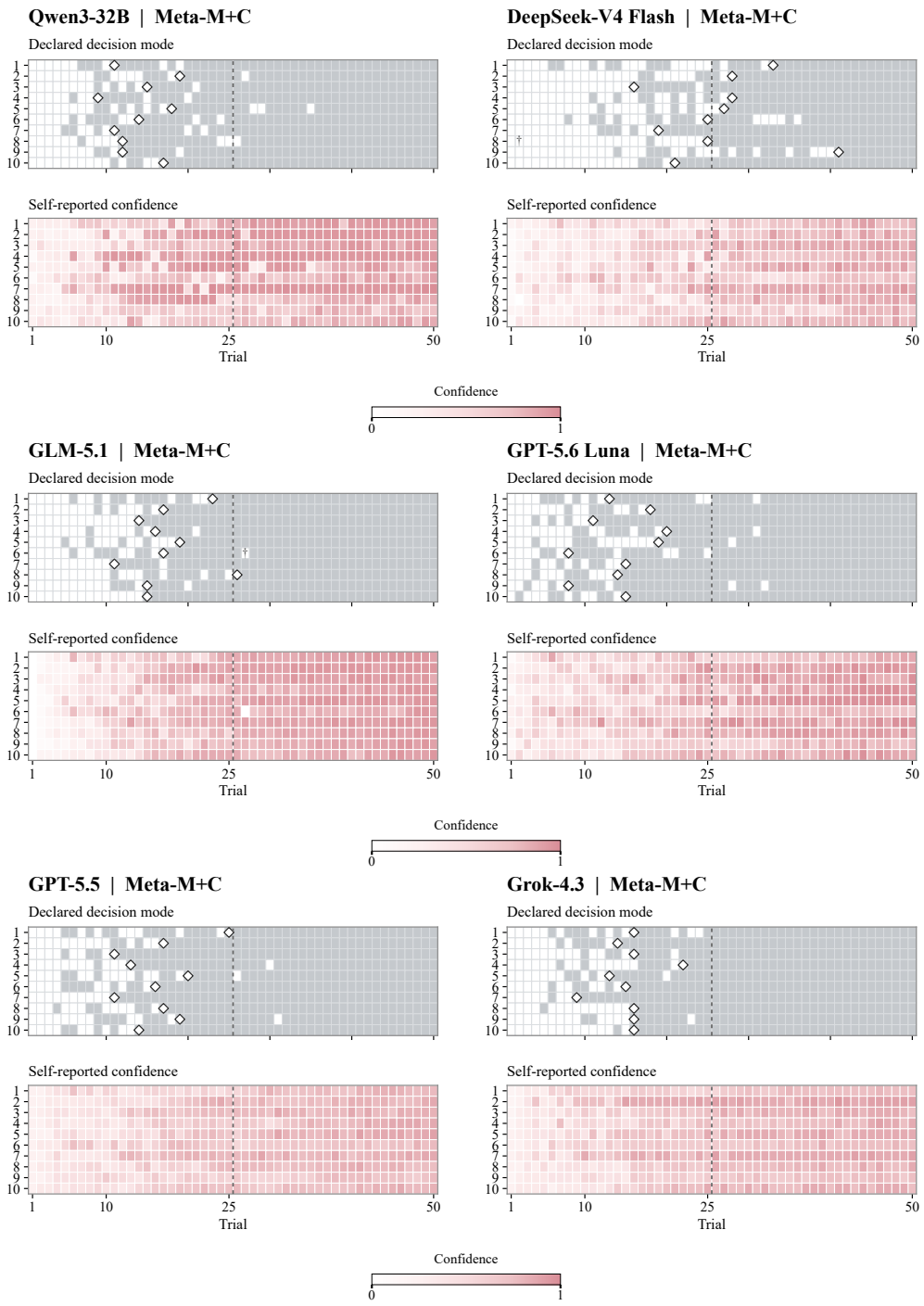


Figure F.4: **Reported decision-mode timing and confidence under Meta-M+C (Qwen3-32B through Grok-4.3).** The panels continue Figure F.3 with the same block, trial, mode, confidence, and five-trial-run encodings.

F.2 FIVE-ARM BENCHMARK

F.2.1 TASK AND PROMPTS

We use the challenging five-arm Bernoulli bandit of Krishnamurthy et al. (2024). The blue arm has mean reward 0.6 and the other four color-labeled arms have mean 0.4; these values and the optimal arm's identity are hidden from the model. GPT-5.6 Luna, with Thinking enabled, completes five independent 100-trial blocks per prompt, denoted B.1–B.5. Within each matched block, all prompts share the same pre-generated arm-specific payoff sequences: the kth selection of an arm reveals its kth reward, irrespective of when that selection occurs. All conditions use the same model and API settings, including the provider-default temperature, which the available API does not allow us to adjust.

All prompts describe the Bernoulli task, specify the 100-trial horizon and cumulative-reward objective, and identify the five available buttons. Each user message gives the number of completed trials and, for each arm, its selection count and mean observed reward. Untried arms are explicitly listed. The following instructions distinguish the four conditions; no condition prescribes a fixed number of exploration trials.

K24 prompt. This label denotes the prompt design underlying the successful GPT-4 configuration of Krishnamurthy et al. (2024). It frames the model as a bandit algorithm and reinforces chain-of-thought reasoning by repeating "Let's think step by step to make sure we make a good choice." in both the system and user messages. The final arm is returned between `<Answer>` and `</Answer>` tags.

EE-balance. EE-balance appends the following instruction verbatim to the shared task description: "Explicitly suggestive of the need to balance exploration and exploitation." The final response contains only the tagged arm, while the model's internal Thinking remains enabled.

Meta-M. The model assesses the following before choosing:

```
1. Which arm currently appears more rewarding?
2. Which arm is more uncertain?
3. Whether this trial calls for exploration or exploitation.
4. Your confidence in the choice.
```

Meta-M+C. The same monitoring questions are followed by the control instruction:

```
Use these metacognitive assessments to regulate the final choice.
When the current value judgment is uncertain or confidence is
limited, do not let it rigidly determine the action; preserve
flexibility when additional information could matter. When the
evidence is strong and confidence is well supported, allow stronger
commitment to the currently favored option.
```

Both metacognitive conditions return the two arm assessments, decision mode, confidence in $[0, 1]$, and final choice in a structured response. They may report an unclear assessment but must still choose an arm on every trial.

Behavioral measures. Post-coverage non-empirical-greedy choice means selecting outside the set of arms with the highest pre-choice empirical mean, after all five arms have been sampled; ties among the highest means count as empirical-greedy choices. A suffix failure at trial 50 occurs when the optimal arm is never selected during trials 50–100. Cumulative pseudo-regret is $\sum_{s=1}^{t}(\mu^* - \mu_{a_s})$, whereas cumulative reward sums the observed outcomes.

F.2.2 CHOICE BEHAVIOR AND PERFORMANCE

Table F.4 summarizes the five matched blocks. Both metacognitive prompts avoid suffix failure in all five blocks and improve mean optimal-arm selection and reward over both baselines. Figure F.5 shows how additional early departures from the empirical-value leader diminish later, with varying commitment timing and outcomes across blocks.

Table F.4: **Five-arm benchmark results.** Means across five matched blocks; suffix failures are counts. Higher optimal-choice rates and reward, and lower pseudo-regret, indicate better performance.

Prompt	Suffix failures	Post-coverage non-greedy (%)	Optimal (%)	Late optimal (%)	Pseudo-regret	Reward
K24	1/5	7.58	42.8	47.5	11.44	48.6
EE-balance	0/5	11.47	44.4	51.8	11.12	49.2
Meta-M	0/5	18.95	**50.8**	**70.2**	**9.84**	51.0
Meta-M+C	0/5	18.32	49.6	60.8	10.08	**52.2**

Optimal choice, pseudo-regret, and reward use all 100 trials; late optimal choice uses trials 50–100, matching the suffix-failure interval.

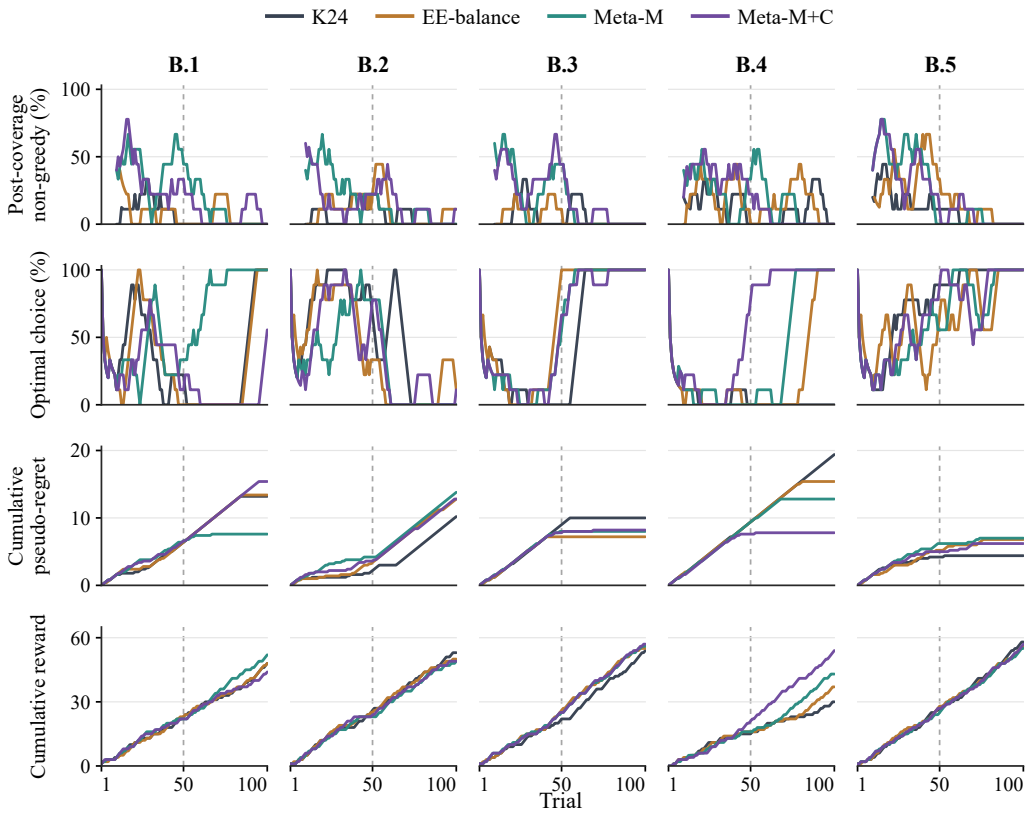


Figure F.5: **Choice behavior and performance across five matched blocks.** Columns denote independent blocks B.1–B.5; colors denote prompt conditions. The first two rows show nine-trial trailing non-empirical-greedy and optimal-choice rates. Non-greedy rates include only post-coverage decisions and are displayed once the window contains at least five eligible choices; undefined values are left blank. The last two rows show unsmoothed cumulative pseudo-regret and realized reward. Axes are shared within each row, and dashed lines mark trial 50. Each curve represents one block, without across-block uncertainty bands.

G Boundary Conditions for Metacognitive Guidance

Supplementary tests on DeepSeek-R1-Distill-Qwen-1.5B and Qwen3-1.7B (Thinking ON) illustrate two potential bottlenecks for metacognitive guidance: unreliable assessment of the available evidence and weak translation of monitoring into action. The two models illustrate failures at these distinct stages. All rates below are averaged across source subjects.

G.1 Unreliable value assessment

Under Meta-M, DeepSeek-R1-Distill-Qwen-1.5B shows substantially less reliable value assessment than the larger models examined in Appendix D.4. Its reported value winner agrees with the historical sample-mean leader in only 76.3–82.6% of eligible decisions across the three settings (Table G.1). Non-greedy choices remain frequent in the late phase, and 65.0–73.6% of these late departures follow a reported value winner that disagrees with the historical sample-mean leader.

This pattern links persistent departures from empirical exploitation to errors in the model's own value assessment. The disagreement can arise at multiple points in the reporting process: in some traces, the textual reasoning identifies the value ordering correctly but the subsequent structured report is inconsistent with that judgment. Reliable value assessment therefore provides an important basis for using metacognitive guidance to support later exploitation.

Table G.1: **Value-report disagreement accompanies persistent non-greedy choice.** DeepSeek-R1-Distill-Qwen-1.5B under Meta-M. All entries are percentages.

Setting	Value-report agreement	Early non-greedy choices	Late non-greedy choices	Late non-greedy choices following a discordant value report
1	77.8	24.8	28.5	73.4
2	82.6	24.0	25.1	65.0
3	76.3	28.0	29.8	73.6

Value agreement compares an unambiguous report with the unique historical sample-mean leader across trials 1–10. A discordant report names the other arm. Non-greedy choices select the other arm when this ranking is defined. Early and late denote trials 1–5 and 6–10. The final column is conditional on late non-greedy choices.

G.2 Monitoring–action dissociation

Qwen3-1.7B exhibits a different failure mode. Under Meta-M, its value reports agree with the historical sample-mean ranking in 98.3–99.6% of eligible decisions, indicating substantially more reliable value assessment. Yet when the model reports the sampled arm as better, identifies the unobserved arm as more uncertain, and explicitly declares exploration, it still chooses the sampled arm in over 90% of these decisions (Table G.2). The same monitoring–action dissociation remains under Meta-M+C.

This failure to translate reported exploration into arm choice is also reflected in incomplete coverage. Both guided conditions improve coverage relative to Native, but many blocks still end with one arm unobserved. Thus, accurate value and uncertainty reports are not by themselves sufficient for adaptive exploration: the reported assessments must also exert effective control over the final action.

Table G.2: **Recognized uncertainty does not consistently translate into exploratory action.** Qwen3-1.7B with Thinking ON. All entries are percentages.

	Declared exploration but chose the sampled arm		Both arms sampled by trial 10		
Setting	Meta-M	Meta-M+C	Native	Meta-M	Meta-M+C
2	90.5	92.9	22.8	64.1	54.0
3	90.6	93.3	27.5	48.5	42.0

The left panel conditions on reporting the sampled arm as better, the never-sampled arm as more uncertain, and declaring exploration during trials 1–10. Coverage is the proportion of blocks that sampled both arms. The control instruction regulates choice through the reported assessments without prescribing a particular arm.

Together, these cases highlight two potential bottlenecks for metacognitive guidance: unreliable assessment of the available evidence and weak translation of monitoring into action. The first can undermine the value judgments used to support later exploitation, whereas the second can leave explicit uncertainty assessments disconnected from information-seeking choice. These cases illustrate how limitations at either stage can constrain the behavioral benefits of metacognitive guidance.