RGFA: ENHANCING FEEDBACK LEARNING IN LARGE LANGUAGE MODELS VIA RELIABILITY-GUIDED FEEDBACK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from feedback has long been studied in psychology as a fundamental mechanism underlying adaptive decision-making. In humans, this mechanism enables efficient adaptation even when feedback is noisy or uncertain. As large language models (LLMs) are increasingly deployed in interactive and decision-making tasks, a fundamental question arises: Can LLMs learn from feedback as effectively as humans, and how do their feedback-driven learning behaviors differ? Using a controlled probabilistic reversal learning paradigm, we compare how LLMs and humans learn from feedback under uncertainty. We find that LLMs perform comparably to humans with positive feedback but are much less effective with negative feedback. This gap arises mainly because LLMs struggle to distinguish informative negative feedback from environmental noise and to use it to improve their policies. To address these limitations, we propose Reliability-Guided Feedback Adaptation (RGFA), a lightweight approach that regulates policy updates according to the reliability of accumulated feedback evidence. Experiments show that RGFA substantially improves learning from negative feedback, and thus narrows the behavioral gap between LLMs and humans. Additional evaluation on a rule-induction and set-shifting task further suggests that RGFA can generalize beyond the original reversal-learning setting.

1 INTRODUCTION

Feedback is a fundamental concept in psychology and provides an important signal for adaptive decision-making. In humans, learning from feedback enables behavioral adaptation by determining whether the current policy should be maintained or revised. Existing studies show that humans can adapt efficiently even when feedback is noisy or uncertain (Behrens et al., 2007; Piray & Daw, 2021). For large language models (LLMs), learning from feedback has also become a common way to adapt behavior in interactive and decision-making tasks such as web navigation (Yao et al., 2022; Zhou et al., 2024) and embodied household (Shridhar et al., 2020). A widely adopted approach is to extract positive and negative feedback from historical interactions and use it to guide subsequent decisions and improve future behavior. Now, work in this area is underway more broadly. One recent example is a series of works about recursive self-improvement (Yin et al., 2025; Zheng et al., 2026). This raises a critical question: *Can LLMs learn from feedback as effectively as humans, and how do their feedback-driven learning behaviors differ?*

This question has motivated growing interest in understanding how LLMs learn from feedback. Some studies investigate whether iterative feedback can improve subsequent responses or reasoning at test time (Li et al., 2025; Wang et al., 2026a). Others focus on why LLMs fail to effectively use available feedback. For example, Jiang et al. (2025) identify *feedback friction*, showing that LLMs may resist corrective external feedback even under favorable conditions, while Zhang et al. (2025) analyze failure modes in intrinsic self-correction. While these studies provide important insights into how LLMs use feedback and how feedback learning can be improved, they largely study LLMs in isolation. It therefore remains unclear how LLM feedback learning differs from human learning.

Answering this question remains challenge, mainly because we lack a suitable evaluation environment for human–LLM comparison. Existing human learning studies usually rely on task-specific

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.

psychological paradigms, such as probabilistic learning tasks (Behrens et al., 2007; Piray & Daw, 2021). These paradigms provide controlled behavioral measurements, but are not designed to evaluate LLMs under matched feedback schedules. In contrast, commonly used LLM agent benchmarks mainly focus on task completion in environments such as web interaction and embodied decision-making (Yao et al., 2022; Zhou et al., 2024; Shridhar et al., 2020). Their feedback signals are typically tied to task success and do not independently control factors such as feedback noise, environmental change, and task load. As a result, existing settings are not well suited for directly comparing how humans and LLMs learn from feedback over time.

To address this challenge, we first construct a controlled feedback-learning task adapted from established psychological paradigms. As shown in Section 3.1, the task requires subjects to infer the preferred option for each stimulus from accumulated positive and negative feedback. We further introduce two uncertainty conditions: S+, where feedback becomes noisier while the preferred option remains unchanged, and V+, where the preferred option changes over time. Together with two stimulus loads, this design allows us to separately examine the effects of feedback noise, environmental change, and task load. Based on this setting, we systematically compare how humans and LLMs learn from feedback. We find two main differences: 1) LLMs perform comparably to or better than humans after positive feedback, but are less effective at learning from negative feedback under uncertainty; and 2) LLMs are less sensitive to the reliability of accumulated negative evidence, making it difficult to determine when to revise an old strategy and how to establish a new one afterward.

Based on these insights, we propose Reliability-Guided Feedback Adaptation (RGFA), a lightweight method for improving feedback learning in LLMs. Our basic idea is to regulate strategy revision based on the reliability of accumulated feedback evidence, so that the model can avoid overreacting to isolated noise while responding more strongly to persistent contradiction. RGFA consists of two stages. Stage 1 determines when the current strategy should be reconsidered. Once revision is triggered, Stage 2 uses short-term feedback to evaluate newly explored alternatives and help establish a new strategy. RGFA does not directly choose actions for the model, but instead guides how feedback is used for subsequent adaptation.

Our main contributions are summarized as follows:

- We introduce a controlled task for direct comparison of human and LLM feedback learning. The task includes six conditions that combine two stimulus loads with three feedback environments, resulting in 396 trials per subject. We further provide a comprehensive behavioral evaluation, such as switching and feedback-conditioned adaptation.
- We provide a systematic comparison between human participants and multiple LLMs under matched feedback schedules. Across 29 human participants and four representative LLMs, we find that LLMs perform well after positive feedback but are less effective at using negative feedback, especially when repeated contradiction should provide reliable evidence for revising an outdated strategy.
- We propose RGFA, which regulates strategy revision based on feedback reliability. Extensive experiments show that RGFA substantially improves feedback adaptation under environmental change. For example, on Qwen3.6, RGFA improves V+ accuracy from 57.22% to 64.08%, increases switching after repeated contradiction from 45.92% to 91.75%, and raises next-encounter selection of the newly preferred option from 22.61% to 44.43%.

2 Related Work

Learning from Feedback in Humans. In psychology, human feedback learning under uncertainty has been studied extensively. A key distinction is between stochasticity, where outcomes vary while the environment remains stable, and volatility, where the underlying environment changes (Behrens et al., 2007; Nassar et al., 2010). Humans can track these different sources of uncertainty and adjust their learning accordingly (Piray & Daw, 2021; 2024). When feedback is noisy, they rely more on accumulated evidence; when the environment changes, they update more rapidly. Recent work further shows that humans adapt exploration and learning precision to unexpected uncertainty (Lee et al., 2023). Thus, effective feedback learning does not simply require responding to each outcome, but judging how much that outcome should influence the current policy. This process is also shaped by cognitive factors. Working memory supports rapid feedback learning under low

load, while higher load increases reliance on slower incremental learning (Collins & Frank, 2012; Collins et al., 2017). Reversal learning further requires abandoning previously effective responses and establishing new ones after environmental change (D'Cruz et al., 2011). Feedback reliability and prior confidence also affect how strongly new outcomes are incorporated (Ernst & Steinhauser, 2017; 2018; Ben Yehuda et al., 2025). Motivated by this line of psychological research, we adopt a controlled probabilistic reversal learning paradigm to directly compare how LLMs and humans learn from feedback under uncertainty.

Learning from Feedback in LLMs. Learning from feedback has become a common way to improve LLM behavior beyond the initial response. Early work mainly focuses on iterative refinement. For example, self-Refine uses model-generated feedback to repeatedly revise responses (Madaan et al., 2023), Reflexion stores feedback in episodic memory to guide later decisions (Shinn et al., 2023), and CRITIC introduces external tool feedback to support correction (Gou et al., 2024). More recently, this idea has been extended to self-improving agents (Yin et al., 2025; Wang et al., 2026b). However, receiving feedback does not guarantee effective learning. LLMs may fail to improve when feedback is weak or uninformative (Huang et al., 2024; Kamoi et al., 2024). More importantly, even informative external feedback may not be fully incorporated into subsequent behavior, a phenomenon recently characterized as *feedback friction* (Jiang et al., 2025). Our work builds on this line of research and focuses on a more basic question: how do LLMs learn from positive and negative feedback under uncertainty, and how does this learning differ from that of humans?

3 Humans vs. LLMs in Controlled Feedback Learning

We compare LLMs with humans under the same controlled setting to better understand how they learn from feedback. This experimental setting isolates key factors, including feedback noise, environmental change, and task load. It allows us to identify systematic differences between humans and LLMs in feedback learning.

3.1 Task and Environment

We design a controlled probabilistic decision task inspired by prior studies (Lee et al., 2023; Wen et al., 2026) to investigate how humans and LLMs learn from feedback. Specifically, the task consists of repeated interactions, where subjects[1] gradually learn the preferred option for each stimulus from accumulated feedback. For example, a stimulus may be represented by a dog, and the three options may correspond to different toys, with one toy being preferred by that dog. In each trial, subjects choose among three options and receive positive or negative feedback. One option is currently preferred, but this preference is not directly revealed. Subjects must infer it from previous choices and feedback. Because feedback is probabilistic, even the preferred option may occasionally receive negative feedback. This makes it necessary to distinguish stochastic noise from evidence that the current strategy is no longer reliable.

Cond.	n	**Env.**	p_+	**Rev.**	T
nS2-Ref	2	Stable, reliable	0.90	0	30
nS2-S+	2	Stable, noisy	0.65	0	30
nS2-V+	2	Reversal	0.90	3	72
nS4-Ref	4	Stable, reliable	0.90	0	60
nS4-S+	4	Stable, noisy	0.65	0	60
nS4-V+	4	Reversal	0.90	3	144
Total					**396**

Table 1: Experimental design. Cond. denotes condition, n denotes the number of stimuli, Env. denotes the environmental structure, p_+ denotes the probability of positive feedback after selecting the currently preferred option, Rev. denotes the number of reversals per stimulus, and T denotes the number of trials.

As summarized in Table 1, the task varies two factors: *stimulus load* and *environmental uncertainty*. Stimulus load controls how many stimulus–option mappings must be learned and maintained at the same time. In nS2, subjects learn and track two stimuli, whereas in nS4, they learn and track four stimuli. Environmental uncertainty is defined by three settings: Ref, S+, and V+. Ref provides a

[1] In psychology, a subject refers to an individual participant who performs the experimental task. In this work, we use "subjects" to refer to both human participants and LLM runs.

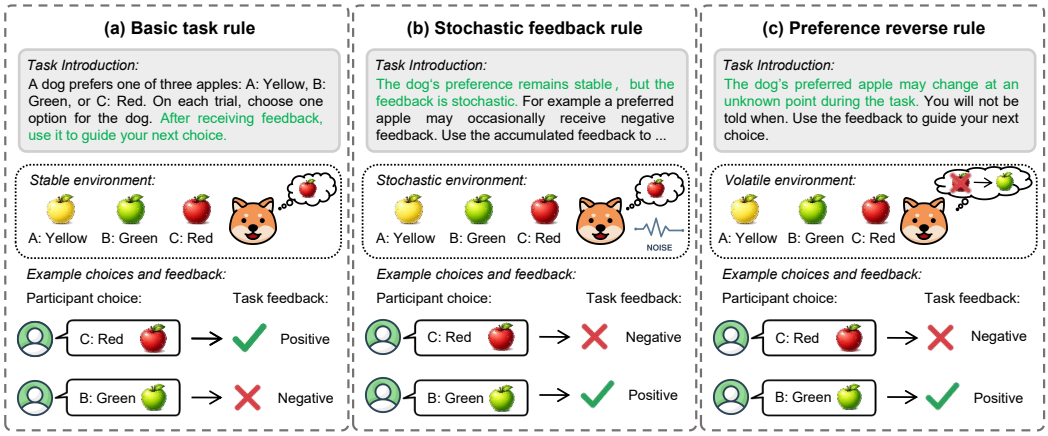


Figure 1: Illustration of the designed decision task. (a) Basic task rule, where participants infer the preferred option from repeated feedback. (b) Stochastic feedback, where the preferred option remains unchanged but may occasionally receive negative feedback. (c) Preference reversal, where the preferred option changes at an undisclosed time.

stable and relatively reliable environment. S+ keeps the underlying preference unchanged but introduces noisier feedback, while V+ changes the preferred option over time through hidden reversals. Thus, S+ mainly tests robustness to noisy feedback, whereas V+ tests adaptation to genuine environmental change. Crossing these factors yields six conditions and allows us to separately examine the effects of feedback noise, environmental change, and task load under the same controlled setting for humans and LLMs.

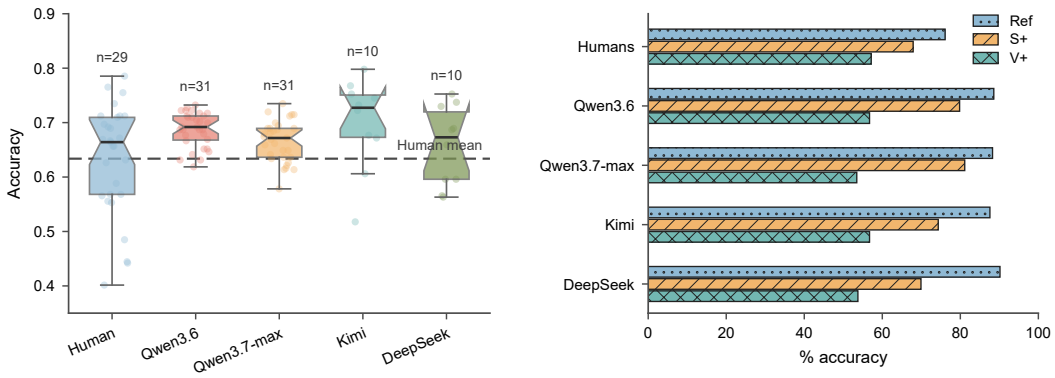
3.2 Human Participants

We recruited 31 university students to complete the controlled feedback-learning task. Two participants were excluded because their choice accuracy fell below 0.20 in at least one condition, leaving 29 participants for behavioral analysis. More details about participant demographics and recruitment are provided in Appendix(Yang et al., 2025). Before the formal experiment, participants received standardized task instructions, completed two practice blocks, and underwent an oral comprehension check. As illustrated in Figure 1, each stimulus was associated with three candidate options, and participants learned the preferred option through repeated positive and negative feedback. They were informed that preferences remained stable in Ref and S+, but could change during V+. However, the preferred options, feedback probabilities, and reversal times were not disclosed.

On each trial, participants selected one of the three options, reported their confidence on a four-point scale from one (very uncertain) to four (very certain), and then received positive or negative feedback. Their previous choices and feedback remained available through the ongoing task history. Condition-to-stimulus assignments and stimulus-to-preference mappings were counterbalanced across participants where possible.

3.3 LLM Virtual Subjects

Each complete LLM run across the six conditions is treated as one virtual subject. Before the formal task, the model receives task instructions matched to those given to human participants, completes practice trials, and passes a comprehension check. At the beginning of each condition, the model is informed whether preferences remain stable or may change, but is not told the preferred option, feedback probabilities, or reversal times. On each trial, the model is prompted to choose one of the three options and report its confidence on the same four-point scale used by human participants. We use a fixed response format to reduce parsing ambiguity. For example, the model is instructed to return only its selected option and confidence level, such as "Choice: B; Confidence: 3." After each decision, the environment returns positive or negative feedback, which is appended to the interaction history and becomes available for subsequent decisions. The model therefore learns from the same sequence of choices and feedback as human participants.



(a) Overall accuracy of humans and LLMs

(b) Accuracy across Ref, S+, and V+ environments

Figure 2: Overall and environment-specific feedback-learning performance of humans and four representative LLMs under Ref, S+, and V+ conditions.

We keep the prompting procedure fixed across trials and conditions. The prompt specifies the task rules and output format, but does not provide additional hints about which option should be selected or how feedback should be interpreted. Model parameters remain fixed throughout the experiment, with no fine-tuning or weight updates. Behavioral adaptation therefore arises entirely from in-context use of task instructions and accumulated interaction history.

3.4 Evaluation Metrics

We compare humans and LLMs using three behavioral measures: choice accuracy, switching rate, and confidence. Accuracy is the proportion of valid trials in which the currently preferred option is selected. Confidence is the mean reported score on the shared four-point scale. Switching is measured between consecutive encounters with the same stimulus. Let a_j and a_{j-1} denote the current and previous choices for that stimulus. We define this switch as

$$s_j = \mathbb{I}(a_j \neq a_{j-1}) \tag{1}$$

where $s_j = 1$ indicates a switch. For each feedback type, the switch rate is computed as the proportion of eligible transitions with $s_j = 1$. We report switching after positive, isolated-negative, and repeated-negative feedback. The first encounter with each stimulus and transitions with invalid responses are excluded. All metrics are first computed within each participant or model run and then averaged across participants or runs.

3.5 Comparative Analysis

As shown in Fig. 2, we compare human participants with four representative LLMs, including Qwen3.6-27B (*Qwen3.6*) (Qwen Team, 2026a), *Qwen3.7-Max* (Qwen Team, 2026b), Kimi-K3 (*Kimi*) (Kimi Team, 2026), and DeepSeek-V4-Pro (*DeepSeek*) (DeepSeek-AI, 2026), under the same experimental setting. Specifically, we first examine whether LLMs exhibit human-like feedback learning across different environments. We then analyze which type of feedback contributes most to the observed difference.

(a) Overall accuracy			(b) Next-encounter behavior		
Condition	**Human**	**Qwen3.6**	**Feedback case**	**Human**	**Qwen3.6**
Ref	76.10%	88.58%	Positive: accuracy	84.47%	95.14%
S+	67.90%	79.82%	Isolated negative: switch	46.25%	47.89%
V+	57.15%	56.73%	Repeated negative: switch	65.43%	45.92%

Table 2: Human and Qwen3.6 feedback learning. Panel (a) shows choice accuracy in Ref, S+, and V+; panel (b) shows next-encounter behavior after different feedback histories.

Do LLMs learn from feedback in the same way as humans? LLMs generally outperform humans in the stationary Ref and S+ conditions(Table 2(a)). For example, Qwen3.6 achieves about 88.58% accuracy in Ref and 79.82% in S+, compared with about 76.10% and 67.90% for humans, respectively. However, this advantage is substantially reduced in V+, where Qwen3.6 drops to about 56.73%, below the human accuracy of about 57.15%. Similar trends are observed across the other

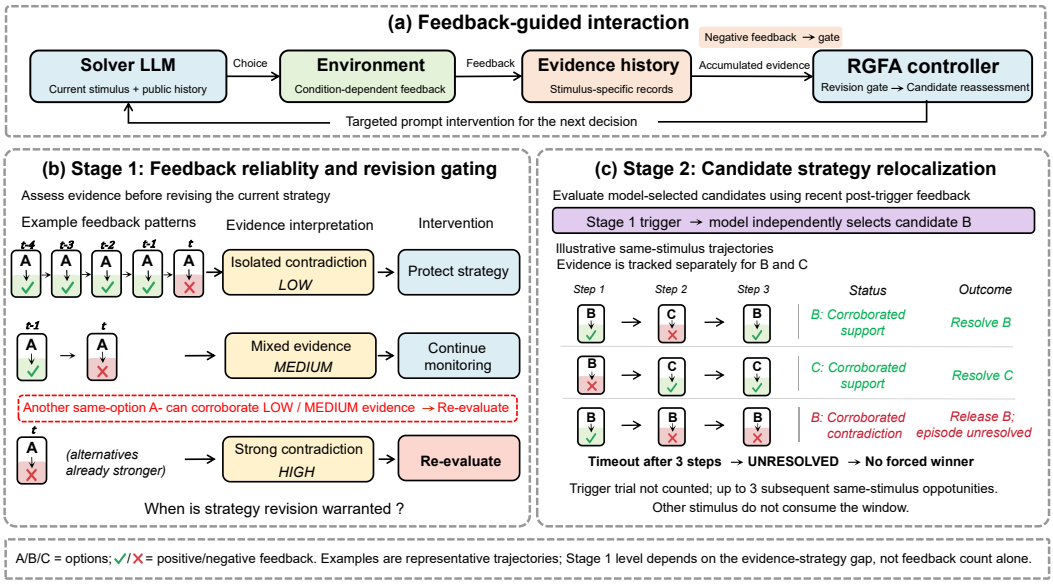


Figure 3: Overview of Reliability-Guided Feedback Adaptation (RGFA). RGFA regulates how feedback is used during interaction without directly selecting actions for the model. Stage 1 evaluates accumulated feedback evidence to determine when the current strategy should be revised. Once revision is triggered, Stage 2 tracks feedback for newly explored candidates and helps identify a reliable alternative over a short stimulus-specific window.

LLMs. This suggests that strong performance under stable feedback does not directly translate into effective adaptation when the environment changes.

How do LLMs respond to uncertainty in feedback? The contrast between S+ and V+ provides a first clue. In S+, uncertainty mainly comes from stochastic feedback noise, while the underlying preference remains unchanged. In V+, the preferred option changes over time, requiring subjects to distinguish transient noise from genuine environmental change. The results suggest that LLMs are not primarily challenged by noisy feedback itself, but by adapting when feedback begins to contradict an established strategy. For example, Qwen3.6 accuracy drops from about 79.82% in S+ to 56.73% in V+, whereas the corresponding decline for humans is smaller.

Is the gap caused by positive or negative feedback learning? The above results show where the human–LLM gap emerges, but do not yet explain which type of feedback drives it. We therefore further separate learning behavior following positive and negative feedback and analyze their contributions to adaptation(Table 2(b)).At the next encounter with the same stimulus after positive feedback, accuracy was 84.47% for humans and 95.14% for Qwen3.6, indicating no model deficit in learning from positive feedback. After an isolated negative outcome, humans and Qwen3.6 switched at similar rates (46.25% and 47.89%). When contradictory negative feedback recurred for the same stimulus, however, the human switch rate increased to 65.43%, whereas the Qwen3.6 switch rate instead decreased to 45.92%. This opposing pattern suggests that Qwen3.6 is less sensitive to the increased reliability of accumulated negative evidence, making it more difficult to distinguish transient feedback noise from genuine environmental change. This is also consistent with Jiang et al. (2025)'s observations, where LLMs often fail to fully incorporate corrective external feedback into subsequent responses.

4 Reliability-Guided Feedback Adaptation

The analysis above suggests that LLMs mainly fail in two places: deciding when negative feedback is reliable enough to revise a strategy, and finding a new reliable option after revision begins. To address these issues, we propose Reliability-Guided Feedback Adaptation (RGFA), a lightweight two-stage controller shown in Fig. 3. Stage 1 estimates the reliability of accumulated feedback and

decides when the current strategy should be reconsidered. Stage 2 is activated after a strong revision signal and tracks short-term evidence for newly explored options. RGFA does not choose actions for the model; it only provides feedback-use guidance through prompts.

4.1 Stage 1: Feedback Reliability and Revision Gating

Stage 1 decides whether negative feedback should be treated as noise or as evidence for revising the current strategy. Its core idea is to accumulate feedback over repeated encounters and trigger revision only when contradiction becomes reliable. Specifically, this stage first maintains stimulus-option evidence, and then converts negative outcomes into graded revision signals.

Accumulated Feedback Evidence. For each stimulus s and option $o \in \{L, M, R\}$, RGFA maintains Beta pseudo-counts $a_{s,o,t}$ and $b_{s,o,t}$ for positive and negative feedback, initialized as $a_{s,o,0} = b_{s,o,0} = 1$. When trial t yields valid feedback, the previous evidence for the current stimulus is first retained as

$$\tilde{a}_{s,o,t} = 1 + \lambda_r(a_{s,o,t-1} - 1) \tag{2}$$

$$\tilde{b}_{s,o,t} = 1 + \lambda_r(b_{s,o,t-1} - 1) \tag{3}$$

where $\lambda_r = 1.00$ in stable conditions and $\lambda_r = 0.80$ in volatile conditions. For the selected option c_t and outcome $y_t \in \{+, -\}$, RGFA then updates

$$a_{s,o,t} = \tilde{a}_{s,o,t} + \mathbf{1}[o = c_t]\mathbf{1}[y_t = +] \tag{4}$$

$$b_{s,o,t} = \tilde{b}_{s,o,t} + \mathbf{1}[o = c_t]\mathbf{1}[y_t = -] \tag{5}$$

The support score of each option is the posterior mean

$$S_t(s, o) = \frac{a_{s,o,t}}{a_{s,o,t} + b_{s,o,t}} \tag{6}$$

Misses do not update evidence, and evidence for other stimuli is unchanged. Model-reported confidence is not used.

Reliability-Guided Strategy Revision. After a valid negative outcome for the selected option c_t, Stage 1 compares the selected option with the best alternative:

$$G_{s,t} = \max_{o \neq c_t} S_t(s, o) - S_t(s, c_t) \tag{7}$$

With a fixed margin $\delta = 0.08$, this gap is mapped to a contradiction level:

$$L_{s,t} = \begin{cases} \text{LOW}, & G_{s,t} \leq -\delta \\ \text{MEDIUM}, & -\delta < G_{s,t} < \delta \\ \text{HIGH}, & G_{s,t} \geq \delta \end{cases} \tag{8}$$

LOW means that the selected option is still better supported; MEDIUM means that the evidence is ambiguous; HIGH means that another option is better supported. Positive outcomes and misses receive no contradiction level.

Stage 1 uses this level to gate revision. A HIGH contradiction immediately triggers strong re-evaluation. A first LOW or MEDIUM contradiction is treated as provisional: LOW protects the current strategy, while MEDIUM starts monitoring. These provisional states are upgraded only by temporal corroboration. If the same stimulus reappears, the model repeats the same option, and another negative outcome is observed, the pending state is released or escalated into strong re-evaluation. Choosing another option, receiving positive feedback, or missing the response clears the pending state but keeps the accumulated evidence.

4.2 Stage 2: Candidate Strategy Relocalization

Stage 2 addresses what happens after revision has been triggered. Instead of assuming that the first alternative is correct, it tracks short-term evidence for newly explored options and accepts a candidate only after it receives corroborating support.

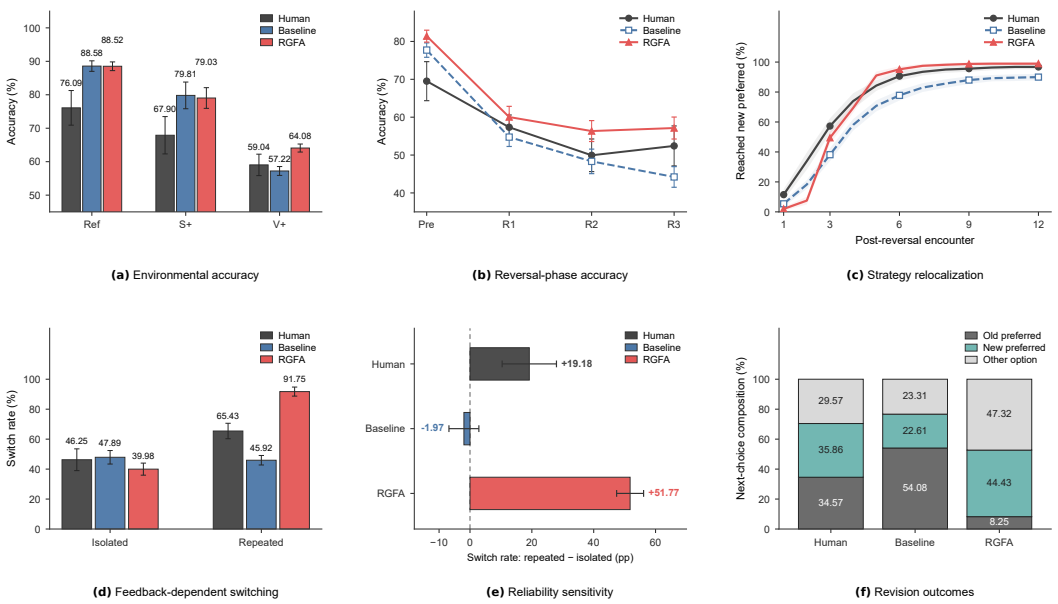


Figure 4: Comparison of humans, baseline Qwen3.6, and RGFA. (a) Accuracy across environmental conditions. (b) Accuracy across successive reversal phases. (c) Cumulative proportion of reversals in which the newly preferred option is reached after reversal. (d) Switching rates after isolated negative feedback and repeated contradictory feedback. (e) Reliability sensitivity, measured as the difference between switch rates after repeated contradiction and isolated negative feedback. (f) Distribution of the next choice following repeated contradiction.

Candidate-Specific Evidence Tracking. After a strong trigger in a volatile condition, Stage 2 opens a stimulus-specific window over the next three presentations of the triggered stimulus, excluding the trigger trial. Other stimuli do not consume the window. A missed response to the triggered stimulus consumes one opportunity but adds no candidate evidence. Within the window, RGFA tracks short-term feedback for each option selected by the model. The first positive outcome marks that option as a provisional candidate. An option that receives negative feedback before any positive support is recorded as explored-negative. The Stage 1 prompt guides the first post-trigger decision; after feedback is observed, Stage 2 may provide candidate-specific guidance for later decisions. The model still selects each option independently.

Candidate Resolution. A candidate is accepted when it receives a second positive outcome before being released. The two positive outcomes need not be consecutive. If a candidate with one positive outcome later receives negative feedback, Stage 2 uses the current Stage 1 contradiction level to interpret it: LOW protects the candidate, MEDIUM keeps monitoring, and HIGH releases it. A candidate-targeted strong Stage 1 signal also releases it. A second negative outcome after positive support releases the candidate regardless of the current level.

Releasing a candidate does not close the window if opportunities remain. The model can still explore another option, and a later positive outcome can establish a new provisional candidate. If no candidate receives two positive outcomes by the third post-trigger presentation, Stage 2 terminates unresolved. All valid choices and feedback remain in the public history and continue to update Stage 1. Thus, Stage 2 evaluates candidate reliability without forcing a choice.

5 Experiments

We evaluate RGFA on Qwen3.6 using the same 2×3 experimental design as in Section 3, with two stimulus loads (nS2 and nS4) and three environmental conditions (Ref, S+, and V+). We conduct 31 RGFA runs and compare them with the corresponding baseline runs from Section 3. Each paired run uses the same stimulus assignments, pre-generated feedback schedules, and reversal timing. Comparisons are paired by subject identifier, with each complete run treated as one statistical unit.

For RGFA, the evidence retention factor is set to $\lambda = 1.00$ in Ref and S+ and $\lambda = 0.80$ in V+. We use a contradiction margin of $\delta = 0.08$ in Stage 1 and a window of at most three subsequent same-stimulus encounters in Stage 2. Detailed hyperparameter settings are provided Appendix C.

Does RGFA Improve Adaptation under Environmental Change? As shown in Figure 4(a), RGFA improves accuracy in V+ from 57.22% to 64.08%, while performance in Ref and S+ remains close to baseline. The improvement becomes more apparent across successive reversals. Furthermore, as shown in Figure 4(b), the gain over baseline increases from 5.30 percentage points in R1 to 12.92 points in R3, where accuracy improves from 44.22% to 57.13%. These results show that RGFA improves adaptation to repeated environmental changes without substantially affecting performance in stable environments.

Does RGFA Better Distinguish Reliable from Noisy Negative Feedback? Figure 4(d) shows a clear difference between isolated and repeated negative feedback. After isolated negative feedback, RGFA reduces the switch rate from 47.89% to 39.98%. In contrast, after repeated contradiction, it increases the switch rate from 45.92% to 91.75%. Accordingly, the difference between repeated-contradiction and isolated-negative switching increases from -1.97 to 51.77 percentage points (Figure 4(e)). This shows that RGFA does not simply encourage more switching. Instead, it suppresses revision after weak evidence while promoting revision when contradictory evidence accumulates.

Does RGFA Improve Strategy Relocalization after Revision? We next examine what happens after the current strategy is challenged. Following repeated contradiction, RGFA reduces the probability of retaining the old preferred option from 54.08% to 8.25%, while increasing selection of the newly preferred option from 22.61% to 44.43% (Figure 4(f)). Over subsequent encounters, the probability of reaching the new preferred option before the next reversal also increases from 89.96% to 98.92% (Figure 4(c)). However, selection of the third option also increases, and accuracy conditional on switching does not significantly improve. From the results, we can see that RGFA is effective at helping the model leave an outdated strategy, while identifying the correct alternative remains more challenging.

6 Cross-Task Generalization

To assess cross-task generalization, we evaluated RGFA on Reflection-Bench's Wisconsin Card Sorting Test (WCST; Li et al., 2025), a rule-induction and set-shifting task where models must infer the currently valid sorting rule from feedback and adapt when that rule changes. Unlike mathematical or knowledge reasoning benchmarks, WCST emphasizes feedback-driven cognitive flexibility: models need to discover and update shape-, color-, or quantity-based rules from sequential outcomes. As shown in Table 3, the RGFA extension improved overall accuracy from 45.83% to 69.44% and post-switch accuracy on trials 4–12 from 43.33% to 72.22%. Its candidate-aligned choice rate was 63.01%, compared with 24.66% for the baseline, suggesting improved feedback-driven rule adaptation beyond the original reversal-learning task.

Metric (%)	Baseline	RGFA extension
Overall accuracy	45.83	**69.44**
Post-switch accuracy (trials 4–12)	43.33	**72.22**
Candidate-aligned choice rate	24.66	**63.01**

Table 3: WCST performance comparison across two runs of the same adapted task sequence. Post-switch accuracy covers trials 4–12; candidate alignment is evaluated against RGFA's candidate rule at 73 diagnostic positions shared by both methods.

7 Conclusion

This work studies how LLMs learn from feedback under uncertainty by comparing them with humans in a controlled probabilistic reversal-learning task. We find that LLMs perform well in stable environments, but struggle when negative feedback must be interpreted as either stochastic noise or evidence of environmental change. To address this, we propose RGFA, a lightweight two-stage method that gates revision based on feedback reliability and then tracks evidence for newly explored candidates. Experiments show that RGFA improves adaptation under volatility while largely preserving performance in stable conditions.

AI USE STATEMENT

We used ChatGPT to assist with manuscript writing, including grammar correction, wording refinement, and stylistic editing, as well as with the summarization and organization of experimental data and results. ChatGPT was used to help consolidate experimental outputs, organize intermediate results, and improve the clarity of their presentation. All quantitative results, technical content, and scientific interpretations were checked and verified by the authors. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

The study protocol was approved by the local ethics committee at the Institute of Psychology, Chinese Academy of Sciences. Written informed consent was obtained from all participants prior to participation. The behavioral data analyzed in this study were de-identified and contained no personally identifiable information.

REFERENCES

Timothy E. J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F. S. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, 2007. doi: 10.1038/nn1954.

Michael Ben Yehuda, Robin A. Murphy, Mike E. Le Pelley, Danielle J. Navarro, and Nick Yeung. Confidence regulates feedback processing during human probabilistic learning. *Journal of Experimental Psychology: General*, 154(1):80–95, 2025. doi: 10.1037/xge0001669.

Anne G. E. Collins and Michael J. Frank. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035, 2012. doi: 10.1111/j.1460-9568.2011.07980.x.

Anne G. E. Collins, Brittany Ciullo, Michael J. Frank, and David Badre. Working memory load strengthens reward prediction errors. *The Journal of Neuroscience*, 37(16):4332–4342, 2017. doi: 10.1523/JNEUROSCI.2700-16.2017.

Anna-Maria D'Cruz, Michael E. Ragozzino, Matthew W. Mosconi, Mani N. Pavuluri, and John A. Sweeney. Human reversal learning under conditions of certain versus uncertain outcomes. *NeuroImage*, 56(1):315–322, 2011. doi: 10.1016/j.neuroimage.2011.01.068.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Benjamin Ernst and Marco Steinhauser. Top-down control over feedback processing: The probability of valid feedback affects feedback-related brain activity. *Brain and Cognition*, 115:33–40, 2017. doi: 10.1016/j.bandc.2017.03.008.

Benjamin Ernst and Marco Steinhauser. Effects of feedback reliability on feedback-related brain activity: A feedback valuation account. *Cognitive, Affective, & Behavioral Neuroscience*, 18(3): 596–608, 2018. doi: 10.3758/s13415-018-0591-7.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Dongwei Jiang, Alvin Zhang, Andrew Wang, Nicholas Andrews, and Daniel Khashabi. Feedback friction: LLMs struggle to fully incorporate external feedback. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1831.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Junseok K Lee, Marion Rouault, and Valentin Wyart. Adaptive tuning of human learning and choice variability to unexpected uncertainty. *Science Advances*, 9(13):eadd0501, 2023.

Yanyang Li, Michael R. Lyu, and Liwei Wang. Learning to reason from feedback at test-time. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 5241–5253. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.262.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019.

Matthew R. Nassar, Robert C. Wilson, Benjamin Heasly, and Joshua I. Gold. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, 30(37):12366–12378, 2010. doi: 10.1523/JNEUROSCI.0822-10.2010.

Payam Piray and Nathaniel D. Daw. A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12:6587, 2021. doi: 10.1038/s41467-021-26731-9.

Payam Piray and Nathaniel D. Daw. Computational processes of simultaneous learning of stochasticity and volatility in humans. *Nature Communications*, 15:9073, 2024. doi: 10.1038/s41467-024-53459-z.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Jiaming Wang, Yunke Zhao, Peng Ding, Jun Kuang, Yibin Shen, Zhe Tang, Yilin Jin, ZongYu Wang, Xiaoyu Li, and Xuezhi Cao. Meeseeks: A feedback-driven, iterative self-correction benchmark evaluating llms' instruction following capability. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14745–14773. Association for Computational Linguistics, 2026a.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. Reinforcement learning for self-improving agent with skill library. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1529–1550. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.69.

Mengxin Wen, Chengyan Yang, Tongran Liu, Kristoffer H. Madsen, and Xun Liu. The effects of working memory and interference control on reinforcement learning: Evidence from computational modeling and structural mri. *NeuroImage*, 329:121818, 2026. doi: 10.1016/j.neuroimage.2026.121818.

Chengyan Yang, Tongran Liu, Mengxin Wen, and Xun Liu. The distinct functions of working memory and intelligence in model-based and model-free reinforcement learning. *npj Science of Learning*, 10:72, 2025. doi: 10.1038/s41539-025-00363-w. URL https://doi.org/10.1038/s41539-025-00363-w.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1508.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursively self-improvement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27890–27913. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1354.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, and Han Qiu. Understanding the dark side of llms' intrinsic self-correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 27066–27101. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1314.

Tong Zheng, Xidong Wu, Zheng Zhang, Zhankui He, Chaoyi Zhang, Benjamin Coleman, Ruoqiao Wei, Di Bai, Haolin Liu, Rui Liu, et al. Dream-rsi: Recursive self-improvement through evolving worlds. *arXiv preprint arXiv:2609.14858*, 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

A Task and Human Experiment Details

A.1 Participants

We recruited 31 university students through WeChat groups. Two participants were excluded because their choice accuracy was below 0.20 in at least one experimental condition, leaving 29 participants for analysis (21 women and 8 men; mean age: 22.94 years; range: 19.25–26.59 years). All participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological or psychiatric disorders, brain injury, or medication use. The study was approved by the ethics committee of the Institute of Psychology, Chinese Academy of Sciences, and all participants provided informed consent.

A.2 Instructions and Trial Procedure

Before the formal task, participants viewed standardized instructions, completed two practice blocks, and underwent an oral comprehension check. Each practice block contained one stimulus. The first comprised 12 trials with a fixed preference and a positive-feedback probability of 0.75 for the preferred option. The second comprised 20 trials with a corresponding probability of 0.90 and preference reversals at trials 9 and 17.

Each formal trial consisted of a 2-second choice window, a 1.5-second confidence-rating window, and 1 second of feedback. Participants reported confidence from 1 (very uncertain) to 4 (very certain) before observing feedback. A green check mark indicated positive feedback, a red cross indicated negative feedback, and a red exclamation mark indicated a missed choice response. Between-stage intervals ranged from 0.5 to 2.5 seconds, and intertrial intervals ranged from 1.5 to 3 seconds.

A.3 Additional Task Details

The six experimental conditions and their trial allocations are described in the main text. In V+, each stimulus had four preference phases of 6–12 encounters each. Reversal timing followed successive encounters with the same stimulus rather than global trial number. Each condition was associated with a distinct stimulus set. Condition order and condition–stimulus-set assignments were counterbalanced across participants. Stimulus–preferred-option mappings and the screen positions of preferred options were also varied across participants to reduce systematic identity and position biases.Participants were informed of the distinction between stable and volatile environments, but the preferred options, feedback probabilities, and reversal times were not disclosed.

B Models, Inference Configuration, and Interaction Protocol

B.1 Models and Virtual Subjects

We evaluated four LLMs: Qwen3.6-27B, Qwen3.7-Max, Kimi-K3, and DeepSeek-V4-Pro. Each complete run across the six conditions contained 396 formal trials and was treated as one virtual subject. Model weights remained fixed throughout the experiment, so adaptation occurred through interaction history and, where applicable, RGFA guidance rather than parameter updates.

All Qwen ablation experiments used Qwen3.6-27b with reasoning enabled.. The reasoning budget was 2,048 tokens, the visible output budget was 32 tokens, and the request timeout was 300 seconds. These settings were shared across the matched controls and ablation variants. Token budgets specified upper limits rather than the number of tokens used on every trial.

B.2 Interaction Protocol

The archived Qwen setup included task instructions, practice trials, and seven comprehension questions. On each formal trial, the model received the current stimulus and three options and submitted a choice and confidence rating before feedback. Responses were requested in JSON format, with L, M, and R denoting the three option positions and confidence ranging from 1 to 4. Choices and feedback were appended to the interaction history. The model was instructed to apply general task rules across scenes but to learn stimulus-specific preferences separately in each scene.

B.3 Matched Comparisons

Matched Qwen comparisons shared stimulus sequences, option mappings, latent preferences, reversal locations, and preassigned trial-level animal responses. Identical schedules did not imply identical observed feedback, because different choices could produce different outcomes. Baseline used the standard task interaction, whereas RGFA additionally supplied controller-generated guidance. The model remained responsible for selecting an option and reporting confidence.

C RGFA Implementation and Default Parameters

C.1 Evidence and Control Procedure

RGFA maintains positive and negative pseudo-counts for each stimulus–option pair, initialized to $(1, 1)$. After valid feedback, it first applies regime-specific retention to the current stimulus's evidence and then increments the selected option's corresponding count. Evidence for other stimuli remains unchanged.Missed responses cause neither evidence decay nor count updates, and reported confidence is not used by the controller.

Stage 1 evaluates contradictory feedback using accumulated evidence. In V+, weak evidence supports retaining or monitoring the current choice, whereas strong evidence or corroborating negative outcomes supports reassessment. A positive outcome, a change of choice, or a missed response clears pending corroboration without deleting accumulated evidence.

A strong V+ trigger opens Stage 2 for at most W subsequent encounters with the same stimulus, excluding the triggering encounter. A missed response to the same stimulus consumes a window opportunity but does not update the evidence.Positive feedback supports a model-selected candidate, while contradictory feedback is evaluated using Stage 1 evidence and the candidate-release rule. An episode ends when the candidate is sufficiently supported or the window expires. RGFA provides guidance but never directly selects an option.

C.2 Default Parameters

The retention coefficient is 1.00 in stable conditions and 0.80 in V+. The evidence-score margin is $\delta = 0.08$, and the same-option negative threshold is $k_{\text{corr}} = 2$. Stage 2 uses a window of $W = 3$ encounters, with a candidate positive-support threshold of $k_+ = 2$ and a negative-release threshold of $k_- = 2$. Parameter sensitivity results are reported in Appendix E.2.

D Metric Definitions and Statistical Analysis

This section defines the behavioral measures, model-specific decision signals, and aggregation procedures used in our analyses. Unless otherwise stated, each measure is first calculated within a participant or model run and then averaged across participants or runs.

D.1 Core Behavioral Measures

Choice accuracy is defined as follows. Let a_t denote the selected option and p_t the currently preferred option on trial t. For a set of N scheduled trials, choice accuracy is

$$\text{Acc} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}[a_t = p_t] \tag{9}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Invalid choices are counted as incorrect. Accuracy is defined by the underlying preference, not by the observed feedback: a correct choice can receive negative feedback because feedback is probabilistic.

Confidence is reported on a shared four-point scale. Let $c_t \in \{1, 2, 3, 4\}$ be a valid confidence rating, and let V contain the trials with valid ratings. Mean confidence is

$$\text{Conf} = \frac{1}{|V|} \sum_{t \in V} c_t \tag{10}$$

Missing or invalid ratings are excluded rather than assigned a value of zero. These ratings represent reported confidence and are not interpreted as calibrated probabilities.

Switching compares option identities across consecutive encounters with the same stimulus within the same block. Here, j indexes encounters with a particular stimulus, rather than consecutive trials in the overall task sequence. A switch is defined as

$$s_j = \mathbf{1}[a_j \neq a_{j-1}] \tag{11}$$

For a set C of eligible transitions, the switching rate is

$$\mathrm{Switch}_C = \frac{1}{|C|} \sum_{j \in C} s_j \tag{12}$$

Both choices must be valid. First encounters and transitions containing an invalid choice are excluded; invalid encounters are not skipped to create a transition between nonadjacent encounters. Option identities, rather than their on-screen positions, are compared.

D.2 Feedback Use

For feedback-conditioned analyses, j denotes the source encounter and $j+1$ the next encounter with the same stimulus within the same block. Let $f_j \in \{+, -\}$ denote the feedback received on encounter j. Event sets are constructed separately for each participant or run. Let E_+ and E_- contain eligible transitions following positive and negative feedback, respectively. The corresponding switching rates are

$$\mathrm{Switch}_+ = \frac{1}{|E_+|} \sum_{j \in E_+} \mathbf{1}[a_{j+1} \neq a_j] \tag{13}$$

$$\mathrm{Switch}_- = \frac{1}{|E_-|} \sum_{j \in E_-} \mathbf{1}[a_{j+1} \neq a_j] \tag{14}$$

Both choices must be valid. These general measures condition on feedback valence; additional restrictions for the diagnostic measures below are applied separately.

Positive-feedback accuracy measures whether the preferred option is selected at the next encounter following positive feedback. Let E_{pos} contain transitions for which the source feedback is positive, both choices are valid, and the preferred option remains unchanged between encounters. Then

$$\mathrm{Acc}_+ = \frac{1}{|E_{\mathrm{pos}}|} \sum_{j \in E_{\mathrm{pos}}} \mathbf{1}[a_{j+1} = p_{j+1}] \tag{15}$$

Eligible events are pooled across the six task conditions within each participant or run. This measure describes subsequent choice accuracy, rather than the causal effect of an individual positive outcome.

Noise-negative switching is calculated in V+ when the currently preferred option is selected but receives negative feedback. The first encounter after a reversal is excluded. The next same-stimulus choice must be valid and remain within the same preference phase. Let E_{noise} contain these eligible events. Then

$$\mathrm{Switch}_{\mathrm{noise}} = \frac{1}{|E_{\mathrm{noise}}|} \sum_{j \in E_{\mathrm{noise}}} \mathbf{1}[a_{j+1} \neq a_j] \tag{16}$$

Because the selected option is still preferred, negative feedback in these events reflects stochastic feedback noise rather than a preference change. This definition does not require the neighboring feedback outcomes to be positive.

Repeated-negative events require two consecutive post-reversal encounters with the same stimulus on which the old preferred option is selected and receives negative feedback on both occasions. Let E_{rep} index the second encounter in each eligible pair, and let o_j denote the old preferred option. Both source encounters and the following encounter must belong to the same post-reversal phase, and the following choice must be valid. The switching rate is

$$\mathrm{Switch}_{\mathrm{rep}} = \frac{1}{|E_{\mathrm{rep}}|} \sum_{j \in E_{\mathrm{rep}}} \mathbf{1}[a_{j+1} \neq o_j] \tag{17}$$

This measure quantifies departure from the old preference after repeated contradictory feedback.

Reliability sensitivity is the difference between repeated-negative and noise-negative switching:

$$\Delta\text{Switch} = 100\,(\text{Switch}_{\text{rep}} - \text{Switch}_{\text{noise}}) \tag{18}$$

The two rates are entered as proportions, so the difference is expressed in percentage points. A larger value indicates stronger behavioral differentiation between the two negative-feedback contexts. It does not directly measure an internal ability to recognize noise. Because methods can generate different choice histories, their eligible event sets need not be identical.

D.3 Post-Reversal Adaptation

V+ trials are grouped into the phase before the first reversal (Pre) and the phases following successive reversals (R1–R3). Let T_h contain the trials assigned to phase h. Phase-specific accuracy is

$$\text{Acc}_h = \frac{1}{|T_h|} \sum_{t \in T_h} \mathbf{1}[a_t = p_t], \qquad h \in \{\text{Pre}, \text{R1}, \text{R2}, \text{R3}\} \tag{19}$$

Trials are assigned according to the reversal history of their own stimulus. Invalid choices are counted as incorrect.

A reversal episode begins when a stimulus changes its preferred option and ends at the next reversal of that stimulus or at block termination. Let L_r be the number of encounters in episode r, and let T_r be the first encounter at which the new preferred option is selected. If it is never selected, set $T_r = \infty$. For R reversal episodes, the proportion reaching the new preference is

$$\text{Reach} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}[T_r \le L_r] \tag{20}$$

This measure records whether the new preference is selected at least once; it does not imply sustained adoption of that option.

Using the same reversal episodes, the cumulative proportion reaching the new preference by encounter k is

$$\text{Reach}(k) = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}[T_r \le \min(k, L_r)] \tag{21}$$

The denominator remains fixed across encounters. For episodes ending before encounter k, their terminal success or failure status is retained; choices from subsequent phases are not included. Missed responses consume an encounter but cannot count as reaching the new preference. Curves are calculated within each participant or run and then averaged equally across participants or runs.

Following the repeated-negative events defined in Appendix D.2, the next choice is classified as the old preferred option, the new preferred option, or the remaining option. Their proportions are

$$P_{\text{old}} = \frac{1}{|E_{\text{rep}}|} \sum_{j \in E_{\text{rep}}} \mathbf{1}[a_{j+1} = o_j] \tag{22}$$

$$P_{\text{new}} = \frac{1}{|E_{\text{rep}}|} \sum_{j \in E_{\text{rep}}} \mathbf{1}[a_{j+1} = p_{j+1}] \tag{23}$$

$$P_{\text{other}} = 1 - P_{\text{old}} - P_{\text{new}} \tag{24}$$

With the same eligible event set, $P_{\text{old}} = 1 - \text{Switch}_{\text{rep}}$. These outcomes distinguish abandoning the old preference from selecting the correct alternative.

D.4 Internal Decision Signals

Internal decision signals are analyzed for Qwen3.6 using the logged probabilities at the behavioral choice-token position. They are distinct from the model's reported four-point confidence.

Let $\ell_t(b)$ denote the returned log probability for candidate label $b \in \{L, M, R\}$ on trial t. The probabilities are renormalized over the three candidate labels:

$$q_t(b) = \frac{\exp(\ell_t(b))}{\sum_{u \in \{L,M,R\}} \exp(\ell_t(u))} \tag{25}$$

These quantities are conditional on the three candidate labels rather than probabilities over the full token vocabulary. They are calculated only when log probabilities for all three candidates are available. Missing signals are not assigned zero. Candidate labels are mapped to option identities using the option mapping on each trial.

Let b_t denote the selected candidate label. The probability assigned to the selected option is

$$P_{\text{chosen},t} = q_t(b_t) \tag{26}$$

Uncertainty over the three candidates is summarized by

$$H_t = -\sum_{b \in \{L,M,R\}} q_t(b) \log q_t(b) \tag{27}$$

Natural logarithms are used, so entropy is measured in nats. Higher entropy indicates a less concentrated candidate distribution.

Let $\ell_{t,(1)}$ and $\ell_{t,(2)}$ be the largest and second-largest candidate log probabilities. The recorded logit margin is

$$M_t = \ell_{t,(1)} - \ell_{t,(2)} \tag{28}$$

A larger margin indicates greater separation between the two highest-scoring candidates. It is not a difference between their probabilities.

Let r_t denote the logged reasoning-token count and let U contain the trials with available counts. Mean reasoning-token usage is

$$\overline{R} = \frac{1}{|U|} \sum_{t \in U} r_t \tag{29}$$

This quantity measures recorded reasoning-token usage, rather than the configured maximum reasoning budget.

D.5 Aggregation and Statistical Analysis

For environmental accuracy summaries, the two stimulus loads receive equal weight:

$$\text{Acc}_e = \frac{\text{Acc}_{\text{nS2},e} + \text{Acc}_{\text{nS4},e}}{2}, \qquad e \in \{\text{Ref}, \text{S+}, \text{V+}\} \tag{30}$$

Overall accuracy instead pools all 396 scheduled trials. Conditional feedback measures pool eligible events within each participant or run before group-level averaging. A conditional rate is undefined when no eligible event is available; it is not replaced by zero.

Let x_i be a participant-level or run-level metric and let n be the number of participants or runs with a defined value. The group mean and sample standard deviation are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{31}$$

$$\text{SD} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{32}$$

Each participant or run receives equal weight, regardless of its number of eligible feedback events. Standard deviations describe between-participant or between-run variation and are not confidence intervals. Rates are multiplied by 100 when reported as percentages.

Method comparisons are paired by task schedule. For methods A and B, define the schedule-level difference as

$$d_i = x_i^A - x_i^B \tag{33}$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{34}$$

A paired 95% confidence interval is

$$\bar{d} \pm t_{0.975,n-1}\frac{s_d}{\sqrt{n}} \tag{35}$$

where s_d is the sample standard deviation of the paired differences. These intervals are not adjusted for multiple comparisons. Matching task schedules does not imply identical choice histories or identical eligible feedback events across methods.

Post-hoc V+ comparisons use two-sided exact sign-flip tests on the paired schedule-level differences. For $\boldsymbol{\epsilon} \in \{-1,+1\}^n$, the two-sided p-value is

$$p = \frac{1}{2^n}\sum_{\boldsymbol{\epsilon}\in\{-1,+1\}^n} \mathbf{1}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i d_i\right| \geq |\bar{d}|\right] \tag{36}$$

The test relies on sign exchangeability of the paired differences under the null hypothesis. Holm correction is applied across the five mechanism contrasts included in the analysis. Trials and individual feedback events are not treated as independent replicates for these tests.

Parameter-sensitivity results are summarized descriptively across matched task schedules. These analyses characterize changes associated with the tested parameter settings; they do not establish that a setting is globally optimal.

E Additional Results and Ablation Analysis

E.1 Additional behavioral results.

The main Qwen comparison included 31 runs per method. After averaging equally across stimulus loads, V+ accuracy increased from 57.22% for Baseline to 64.08% for RGFA, corresponding to an absolute improvement of 6.86 percentage points. Improvements were observed under both stimulus loads: nS2-V+ accuracy increased from 60.75% to 65.55%, whereas nS4-V+ accuracy increased from 53.70% to 62.61%. Complete condition-level results are reported in Table 4.

Condition	Baseline	RGFA
nS2-Ref	90.22	87.74
nS2-S+	79.68	79.78
nS2-V+	60.75	65.55
nS4-Ref	86.94	89.30
nS4-S+	79.95	78.28
nS4-V+	53.70	62.61
Overall	68.73	72.77

Table 4: Condition-level accuracy in the main Qwen3.6 comparison. Values are mean accuracies (%) across 31 runs per method. Overall accuracy pools all 396 formal trials within each run.

E.2 Parameter Sensitivity

We evaluated eleven single-parameter variants on three matched task schedules to examine the effects of the RGFA hyperparameters (Table 5). The default configuration achieved 65.51% V+ accuracy, compared with 54.98% for the matched Baseline, and had the highest mean V+ accuracy among the tested settings.

The corroboration threshold showed the clearest behavioral trade-off. Reducing k_{corr} from 2 to 1 increased noise-negative switching from 38.56% to 91.20%, indicating more frequent switching after stochastic negative feedback. Increasing it to 3 reduced repeated-negative switching from 94.17% to 51.44% and lowered V+ accuracy from 65.51% to 58.80%, consistent with delayed revision of an outdated preference. These patterns support the role of the default threshold in balancing resistance to feedback noise against responsiveness to repeated contradictory evidence.

Setting	V+ Acc. (%)	Δ Acc. (pp)	Noise (%)	Repeated (%)	Sensitivity (pp)
Matched Baseline	54.98	–	54.50	49.26	−5.25
Default RGFA	65.51	0.00	38.56	94.17	55.61
$\lambda = 0.60$	62.15	−3.36	51.85	92.67	40.82
$\lambda = 1.00$	64.58	−0.93	47.69	86.24	38.54
$\delta = 0.04$	64.00	−1.50	38.07	94.12	56.05
$\delta = 0.12$	61.46	−4.05	50.98	93.45	42.47
$k_{\text{corr}} = 1$	64.93	−0.58	91.20	80.00	−11.20
$k_{\text{corr}} = 3$	58.80	−6.71	36.06	51.44	15.39
$W = 2$	63.54	−1.97	42.43	92.88	50.46
$W = 4$	63.89	−1.62	39.08	78.34	39.27
$k_+ = 1$	62.73	−2.78	35.91	90.37	54.46
$k_+ = 3$	62.15	−3.36	45.52	90.78	45.27
$k_- = 1$	62.50	−3.01	41.50	84.75	43.25

Table 5: Parameter sensitivity of RGFA on three matched task schedules. V+ Acc. reports mean accuracy in V+ conditions; Δ Acc. is the change from the default RGFA setting. Noise and Repeated denote switching after noise-negative and repeated-negative feedback, respectively. Sensitivity is Repeated minus Noise.

All eleven variants retained higher mean V+ accuracy than the matched Baseline, although none exceeded the default configuration. Together, these descriptive results provide preliminary support for the default hyperparameter settings within the tested range. The comparison is based on three matched schedules; conditional switching rates should be interpreted cautiously when eligible events are sparse.

E.3 Mechanism ablations

We compared Baseline, Stage 1-Only, Stage 1+Fast3, and Full RGFA on ten matched task schedules. The results are listed in Table 6. From the results, we can see that Stage 1-Only disables Stage 2. Stage 1+Fast3 adds short-window candidate tracking, while Full RGFA additionally includes candidate-specific graded regulation. Stage 1-Only improved V+ accuracy from 55.87% to 61.42%, a gain of 5.56 percentage points (95% paired CI [2.00, 9.11]; Holm-adjusted exact $p = .0176$). Adding the short-window second stage through Fast3 further increased accuracy to 65.69%, improving over Stage 1-Only by 4.27 percentage points (95% paired CI [2.32, 6.22]; adjusted $p = .0176$). These comparisons support the contribution of first-stage regulation and the additional benefit of short-window candidate tracking in volatile environments.

Method	**V+ accuracy**	**Noise**	**Repeated**	**Sensitivity**
Baseline	55.87	45.80	45.58	−0.22
Stage 1-Only	61.42	37.62	80.04	+42.42
Stage 1+Fast3	65.69	38.92	91.62	+52.70
Full RGFA	64.65	38.87	93.38	+54.51

Table 6: Mechanism ablations on ten matched task schedules. Values are means across runs. V+ accuracy and noise-negative (Noise) and repeated-negative (Repeated) switch rates are reported in percent. Sensitivity is repeated minus noise switching, in percentage points.

The feedback-conditioned measures provide complementary behavioral evidence. Stage 1-Only reduced noise-negative switching from 45.80% to 37.62% while increasing repeated-negative switching from 45.58% to 80.04%. The corresponding sensitivity increased from −0.22 to 42.42 percentage points. Fast3 and Full RGFA showed sensitivities of 52.70 and 54.51 percentage points, respectively. These descriptive patterns are consistent with reduced switching after noisy negative feedback and stronger revision following repeated contradictory evidence, although the eligible event sets can differ across methods.

Full RGFA achieved 64.65% V+ accuracy, exceeding Baseline by 8.78 percentage points (95% paired CI [5.45, 12.12]; adjusted $p = .0098$). Together, the accuracy and switching results support the effectiveness of the two-stage approach in this cohort. However, Full RGFA did not show an

additional accuracy benefit over Fast3: the paired difference was -1.04 percentage points (95% paired CI $[-3.46, 1.38]$; adjusted $p = .3672$).