

---

# Illusions of reflection: open-ended task reveals systematic failures in Large Language Models’ reflective reasoning

---

Sion Weatherhead  
UNSW

s.weatherhead@unsw.edu.au

Flora Salim  
UNSW

f.salim@unsw.edu.au

Aaron Belbasis

Aurecon Group

aaron.belbasis@aurecongroup.com.au

## Abstract

Humans do not just find mistakes after the fact—we often catch them mid-stream because ‘reflection’ is tied to the goal and its constraints. Today’s large language models produce reasoning tokens and ‘reflective’ text, but is it functionally equivalent with human reflective reasoning? Prior work on closed-ended tasks—with clear, external ‘correctness’ signals—can make ‘reflection’ look effective while masking limits in self-correction. We therefore test eight frontier models on a simple, real-world task that is open-ended yet rule-constrained, with auditable success criteria: to produce valid scientific test items, then revise after considering their own critique. First-pass performance is poor (often zero valid items; mean  $\approx 1$ ), and reflection yields only modest gains. Crucially, the second attempt frequently repeats the same violation, indicating ‘corrective gains’ arise largely from chance production of a valid item rather than error detection and principled, constraint-sensitive repair. Performance before and after reflection deteriorates as open-endedness increases, and models marketed for ‘reasoning’ show no advantage. Our results suggest that current LLM ‘reflection’ lacks functional evidence of the active, goal-driven monitoring that helps humans respect constraints even on a first pass. Until such mechanisms are instantiated in the model itself, reliable performance requires external structure that enforces constraints.

## Introduction

Large language models (LLMs) are increasingly entrusted with planning, drafting, and analysis. Step-by-step prompting ‘chain-of-thought’ can boost performance[1], and recent work encourages models to ‘reflect’ via self-critiques, reflection loops (routing outputs back into a new attempt), and tool-augmented checks[2–4]. We use *reflection* as an umbrella for these procedures—including iterative refinement[5, 6], code- or test-based feedback[4, 7, 8], and program-level scaffolding for long-reasoning models (LRMs)[9].

These effort aim to *functionally* approximate human *meta-reasoning* to improve LLM reasoning. Meta-reasoning refers to a goal-directed process of monitoring and controlling one’s thoughts, thought processes(i.e., reasoning), while calibrating efforts and strategy in accordance with confidence, updating allocation of effort and resources whilst maintaining constraint fidelity[10]. This is a fundamental mechanism for *self-correction*. However, to note, our standard for evaluation is explicitly *functional*, not mechanistic: we do not require human-like processes inside the model, only that the model’s outputs reliably indicate constraint-sensitive evaluation and corrective change.

Why does this matter for intelligence? Broadly, intelligence is the capacity to learn, adapt, and solve problems across familiar and novel settings[11]. Meta-reasoning is a core human capability for doing so: it detects violations, adjusts strategy, and prevents error repetition[10]. If LLMs are to be genuinely self-improving rather than merely verbose, ‘reflection’ must act as such a controller. Surveys and targeted studies, however, suggest that many gains attributed to reflection depend on *external* signals—tests, retrieval, or explicit error flags—with limited success at autonomous corrective reasoning when those signals are weak[3, 12–14].

A second, deeper concern is the reliability of the ‘reasoning’ text itself. Performance can degrade under superficial changes that leave underlying logic intact—for example, swapping surface values in GSM-style maths questions[15] or reframing otherwise identical problems[16, 17]. More critically, explanatory tokens need not reflect the latent decision process: models may answer correctly while their explanations omit, misattribute, or contradict the basis for the answer[18, 19]. Recent work even shows scale-related failures when task complexity increases while problem structure is held constant[20]. If traces are decoupled from mechanism, then longer traces and reflection loops risk compounding confident text rather than delivering targeted correction.

These points motivate a distinction that matters for evaluation. *Closed-ended* tasks have a single correct answer and offer crisp corrective signals (e.g., a failing unit test). Many positive reports of reflection occur in such settings[3]. *Open-ended* tasks enlarge the solution space and require consistent application of interacting constraints; external signals are weaker or delayed, and there are many more plausible but wrong outputs. If LLM behaviour is, in part, statistical pattern matching[15], open-ended tasks remove the crutch of strong signals, making principled, constraint-sensitive reasoning harder. We found two examples of open-ended task studies; one was an abstract writing task - where they found systematic biases (e.g., over-generalising study findings) resistant to prompted correction[21]. The other was a creativity task, where evaluation was either subjective human judgement or relying on dispersion metrics that, while concrete, are hard to interpret as practical *reasoning* diagnostics[22].

As such, there is a need for objective, measurable, open-ended tasks for evaluation. Our diagnostic task asks for the minimum behavioural evidence one should expect from a reflective system: *consistency between stated criteria and the change on the next pass*. When a model itself flags plagiarism or a violation of CRT properties, does its revision *avoid that exact fault*? Failure to do so indicates that the *nested concepts* implicated by the task (e.g., what qualifies an ‘original questionnaire item’) are not being activated relative to the stated goal—a shortfall in functional meta-reasoning[10]. In other words, we seek outputs whose reasoning text *suggests* the constraints have been considered and enforced, without assuming a human-like mechanism.

To probe this directly, we evaluate eight frontier LLMs on an *open-ended*, rule-constrained psychometric generation task with auditable pass/fail criteria. Models must produce Cognitive Reflection Test (CRT) items. These are ‘trick questions’ with an intuitive but wrong answer and a single correct answer reachable upon reflection[23]. For example:

*A bat and a ball cost \$1.10 together. The bat costs \$1.00 more than the ball.  
How much does the ball cost?*

The intuitive response people are drawn toward is ‘10 cents’ from subtracting a \$1 from the initial sum of \$1.10. Upon reflection, along with simple maths, the correct answer of 5 cents can be arrived at. Having an ‘intuitive-incorrect’ response, a common logical error, is a key property of a valid CRT item.

The LLMs in our tasks must produce new items for this test without copying existing items from validated tests in the literature. We compare two framings that vary the task’s openness and the availability of external anchors: *generation* (invent items de novo) versus *search–identify* (retrieve suitable non-CRT trick questions and adapt). We then solicit a reflection pass and a re-answer. To maximise models’ chances, we provide chain-of-thought scaffolding and an expert-persona prompt, set LRMs to high-reasoning modes where available, allow generous thinking budgets, and use independent LLM judges with access to compendia of known CRT items for plagiarism screening[3, 24]. This design operationalises the intelligence-relevant question: does today’s LLM ‘reflection’ reliably convert self-explanation into *correction* when the problem is open-ended, the constraints are explicit, and the change is auditable?

Table 1: Design and outcome counts (pooled across models). Unit: items unless noted. Each session attempts 4 items initially; failed items then enter a reflection phase. “Categorised failures (base)” is the denominator for same-category repeats.

Task	N_sessions	Initial pass (x/expected)	Categorised failures (base)	Same-category repeats (% of base)
Generation	32	22 / 128	68	58 / 68 (85.3%)
Search-identify	32	37 / 128	52	39 / 52 (75.0%)
Total	64	59 / 256	120	97 / 120 (80.8%)

**Notes.** ‘Initial pass’ = valid initial items; expected =  $4 \times N_{\text{sessions}}$ . “Categorised failures (base)” = initial invalid items assigned a failure category; this is the denominator for “Same-category repeats”. For completeness, the *reflection-failure* denominator gives a similar picture: overall 484/567 (=85.4%) same-category repeats among reflection failures, with per-strategy rates—retry: 127/143 (88.8%), instructions: 127/150 (84.7%), explanation: 108/128 (84.4%), keywords: 122/146 (83.6%). Post-reflection performance is reported as paired *pass-rates* in Table 2; because those rates average across strategies within session, they are not converted to unique item counts here.

**Not a creativity benchmark.** Despite the open-ended format, this study does *not* judge ‘creativity’. We do not score ‘originality’ by aesthetic value. Passing requires only auditable constraints: (i) non-plagiarism (binary check: ‘does item  $x$  appear in set  $Y$ ?’); (ii) ensure properties: e.g., ‘intuitive-incorrect response’ and ‘correct answer’ are present; and (iii) basic clarity (further defined below). Otherwise, we are deliberately lenient: items need not be ‘publishable’ CRT questions—the evaluation target is constraint adherence, not ‘novelty’ per se.

## Hypotheses

We tested three hypotheses (H). ‘Pass-rate’ is the share of valid items per session (4 items/session).

**H1 — Reflection improves performance.** *H1a (overall):* Reflection increases session pass-rate relative to the initial attempt. *H1b (error repetition):* Models will repeat the same failure category in reflection, beyond what would be expected by a chance-based benchmark.

**H2 — Task structure moderates gains and error persistence.** *H2a:* Gains are larger in *search-identify* than in *generation*. *H2b:* Same-category failure repetition is lower in *search-identify* than in *generation*.

**H3 — ‘Reasoning-model’ advantage.** Models marketed for extended reasoning achieve larger reflection gains than other models.

Differences across reflection strategies (Explanation, Retry, Keywords, Instructions) are summarised in the main for context and fully tested in the Supplementary with multiplicity control. Benchmark construction, evaluator setup, and robustness checks are detailed in Methods/Supplementary.

## Methods

We evaluate eight frontier largelanguage models (LLMs) on a three-stage *session* that attempts to invent four novel items for the Cognitive Reflection Test (CRT)[25]. Each session comprises (i) an initial answer, (ii) self-reflection, and (iii) a re-answer, mirroring the ‘Baseline’ and ‘self-reflecting agent’ scheme of Renze & Guven[3].

All experiments were conducted using LLM\_ReflectionTest, a modular prompting and evaluation platform developed for this study. The system allows flexible insertion of role, generation, critique, and reflection prompts, supports parallel evaluation across multiple language models, and automatically logs outputs and metadata into a structured database. This platform implements the full generate-evaluate-reflect loop described below.

## Models

We evaluated eight models: OpenAI (GPT-4.1, o3, o4-mini), Google (Gemini 2.5 Pro-Preview), Anthropic (Claude 3.7 Extended), Meta (Llama-3.3-70B, Llama-4 Maverick), and DeepSeek (Reasoner). Where available for LRMs, ‘high-reasoning’ modes were enabled.

While each model was evaluated over eight sessions, each session constitutes a full agentic cycle involving initial generation, error detection, 3 distinct reflection strategies created, and finally four targeted re-attempts applying the 3 created strategies and a ‘retry’ where models are asked to simply replace failed items with no additional strategy.

Thus, each model undergoes  $8 \times 4 = 32$  structured reasoning attempts across diverse prompt framings (see ‘task definition’ below).

All LLMs were evaluated using their standard public API configurations, including default temperature settings (typically in the 0.7–1.0 range). This choice reflects real-world usage patterns for generative tasks, where deterministic decoding is rare. Crucially, our goal was not to constrain models into a narrow reasoning trace (which could artificially dampen performance on an open-ended task), but to simulate adaptive, open-ended generation and evaluate the robustness of reflection under typical sampling variability.

## Task definition

A valid CRT item must (i) present a seemingly obvious but wrong answer, (ii) permit a single correct answer reachable with reflection, and (iii) differ substantively from published CRT items.

Two high-level task framings were tested:

- **Generation** – create four items *de novo*.
- **Search–identify** – retrieve four suitable ‘trick questions’ from public sources, excluding any existing CRT items.

Prompt ablations manipulated the presence of exemplars, extra instructional detail, and explicit practical constraints. Materially, these variations do not change the structure of the task. The aim was only to determine whether complexity, specificity, or provision of prohibited items affected the outcome.

## Agent protocol (single LLM session)

1. **Baseline answer.** Model receives the task prompt and outputs four candidate items.
2. **Self-reflection.** For each failed item (detected automatically; rubric below) the same model is asked to *explain the failure* and produce a short corrective advice block (keywords, explanation, step-by-step instructions).
3. **Re-answer.** This advice block/reflection text is prepended to the original task prompt reminding of the constraints and the model attempts the task again.

This three-call cycle constitutes one *session*. No system-level state is carried across sessions.

**System persona.** Following Renze & Guven [3], each session used a constant system prompt establishing an expert persona: ‘You are a cognitive science expert and psychometrician with experience designing and validating Cognitive Reflection Test (CRT) items.’

We treat persona as a cue that, if effective, should functionally activate associated constraint-relevant knowledge (e.g., canonical CRT set, exclusion checks) and thus increase rule adherence.

## Automated evaluation

Each candidate item was scored by three independent LLM evaluators (GPT-4.1, o4-mini, Gemini 2.5 Pro-Preview) using a fixed rubric and structured output (one API call per item per evaluator). We use a single aggregation rule across criteria: *fail-fast*—if *any* evaluator flags a required criterion as failing, the item fails.

- **Validity (1/0):** requires both an intuitive-incorrect response and a single reflection-based correct answer.
- **Novelty (1/0.5/0):** evaluators compare against an embedded list of known psychometrically validated CRT items and then check, in an open-world sense, for widely shared non-CRT “trick questions” using the evaluator’s internal knowledge. *Generation:* 1.0 if neither a CRT nor a common trick; 0.5 for a superficial CRT variation; 0.0 for an exact/near-exact CRT or an existing non-CRT trick. *Search–identify:* 1.0 for a pre-existing non-CRT trick with no CRT overlap; 0.5 for a superficial CRT variation; 0.0 for an exact CRT match. *For pass/fail we collapse to binary: only 1.0 counts as pass; 0.5/0.0 are treated as fail.*
- **Clarity (1/0):** linguistic coherence and an unambiguous solution path; the intuitive-incorrect rationale must be logically described.
- **Complexity cap (screen; 0–5 source scale):** arithmetic (0–2), algebraic (0–2), spatial (0–1) using a coarse rubric (0 = none, 1 = basic, 2 = advanced). To preserve CRT-like accessibility, items with total complexity  $\geq 3$  are labelled non-CRT-like and *fail the screen*. *This cap is binary in analysis.*

**Pass logic.** An item is *valid* if all three binary criteria pass under the fail-fast rule (Validity = 1, Clarity = 1, Novelty = 1.0 for the relevant condition) *and* the complexity cap is satisfied (sum  $< 3$ ). Any single evaluator’s fail on a criterion, a Novelty of 0.5/0.0, or total complexity  $\geq 3$  renders the item invalid.

*Note on scoring.* The tri-level Novelty and the 0–5 complexity rubric are used to support descriptive robustness checks; all confirmatory tests use the binary decision rule above.

**Prompts** Full evaluator rubric, JSON schema, and all prompts (task, reflection, evaluators) are provided in the public code repository (see Data & Code Availability).

**Rationale for LLM non-CRT novelty checks.** To preserve closed-book evaluation, we use an LLM-only screen to flag ‘existing’ non-CRT trick questions rather than curate an inevitably incomplete non-CRT list. We justify this on two grounds: First, studies show LLM can be effectively utilised for fact-checking[26][27], indicating reliable open-world detection of widely shared content. Secondly, memorisation increases along with frequency of data appearance in pre-training; data—duplicated sequences are disproportionately internalised and thus detectable [28]. Given popular riddles and viral ‘trick questions’ are precisely such high-duplication artefacts, an instruction-following LLM serves as a pragmatic first-pass detector for non-CRT reuse.

## Human validation

While rubric is largely objective (mathematical complexity, while subjective, is purposely coarse grained to enable most robust criteria checking), we examined LLM evaluation fidelity via human inspection. A stratified 20 % sample (strata: model  $\times$  condition  $\times$  reflection) was double-coded by two graduate students, prior to disagreement resolution, inter-rater agreement was  $\kappa = 0.55$ . However, all disagreements were resolved through discussion with a final set of labels, as it were matter of pointing to missed CRT items, or resolving misreadings of a given question and their answers. Human vs. collective LLM label had an inter-rater agreement of  $\kappa = 0.54$ . Further details are specified in supplementary materials.

## Metrics

To evaluate whether reflection improves LLM performance, we compared valid-item generation before and after reflection across models. The generation of valid items was computed as follows. In the initial round, each LLM in a given experimental condition was tasked with producing four CRT items that met predefined validity criteria. The number of valid items was divided by four to compute a pass-rate (e.g., 2 out of 4 valid items yields a 50% pass-rate).

In the reflection round, each of the four reflection strategies was applied to the subset of failed items from initial round of generation. The resulting new items were evaluated for validity, and the post-reflection pass-rate for each strategy was calculated by combining the successful items from

the initial round and the successful reflection-attempts, divided by four. This approach follows the method in [3] and reflects the conceptual structure of the task: a single, continuous attempt by a given LLM in a given condition (which we define as a session) to complete the generative task and correct prior failures. While separate API calls are made for initial generation, reflection strategy generation and finally the strategy execution - this continuity is facilitated by passing through the condition text and exact task wording in each API inside a session.

The formal calculation is shown in Equation 1, where the subscript *initial* refers to the model’s original attempt, and *reflection* refers to its subsequent reflection-based re-attempts.

$$\begin{aligned} \text{PassRate}_{\text{initial}} &= \frac{\text{Correct}_{\text{initial}}}{\text{Total}_{\text{initial}}} \\ \text{PassRate}_{\text{reflection}} &= \frac{\text{Correct}_{\text{initial}} + \text{Correct}_{\text{reflection}}}{\text{Total}_{\text{initial}}} \end{aligned} \quad (1)$$

## Analysis

**H1a — Overall reflection effect.** For each generation session we paired the initial pass-rate with the mean pass-rate across all four reflection strategies. We fit a linear mixed-effects model with generation round (initial vs. reflection) as a fixed effect and random intercepts and slopes by session to estimate the average improvement attributable to reflection while accounting for within-session dependence. As a complementary check, we conducted a paired *t*-test on per-session pass-rates and report the corresponding paired effect size ( $d_z$ ).

**H1b — Does reflection repeat the original failure?** We tested whether reflection repeats the *same* failure category that appeared at the initial attempt. Analytically, we pooled across tasks and models and estimated the probability of repeating the original category with a session-clustered logistic model including strategy as a fixed effect (per-model proportions in Supplementary). Separately, we benchmarked the observed repeat rate against chance—what it would be *if reflection failures were random draws from the observed mix of categories within each task×strategy cell*—using a stratified permutation test (Methods/Supplementary).

**H2a — Task moderation of reflection gains.** To test whether performance gains depend on task structure, we fit mixed-effects models with generation round, task condition (generation vs. search–identify), and prompt subcondition (base, complexity, examples, practical) as fixed effects, and a random intercept by session. Model selection (log-likelihood) retained generation round and the required two-way interactions.

**H2b — Task effects on error persistence (any-category repeat and plagiarism recidivism).** We probed whether error persistence differs by task. We modelled the probability that a reflection attempt repeats the session’s initial failure category as a function of task (generation vs. search–identify) with session-clustered logistic models controlling for strategy. We contrast tested plagiarism recidivism specifically (conditional on initial plagiarism), again by task.

**H3 — Reasoning–model superiority.** For the reasoning–model contrast we fitted an OLS model of reflection gain with fixed effects for task and subcondition and a binary indicator for model–type (1 = reasoning, 0 = other), clustering standard errors by model identity. The confirmatory one–sided test asked whether reasoning–labelled models achieved larger gains than other models, with CR1 95% CIs and wild-cluster bootstrap 90% CIs/*p*. After this test, we ran an exploratory equivalence check (TOST,  $\Delta = \pm 0.05$  pass–rate units) as a sensitivity analysis.

**Exploratory — Strategy differences.** To compare reflection framings, we modelled strategy as a fixed effect with a random intercept for session, using  $\Delta$  pass-rate (recovered valid items out of four) as the outcome. Strategy effects were estimated relative to ‘explanation’, with Holm-adjusted pairwise contrasts and within-session paired effect sizes.

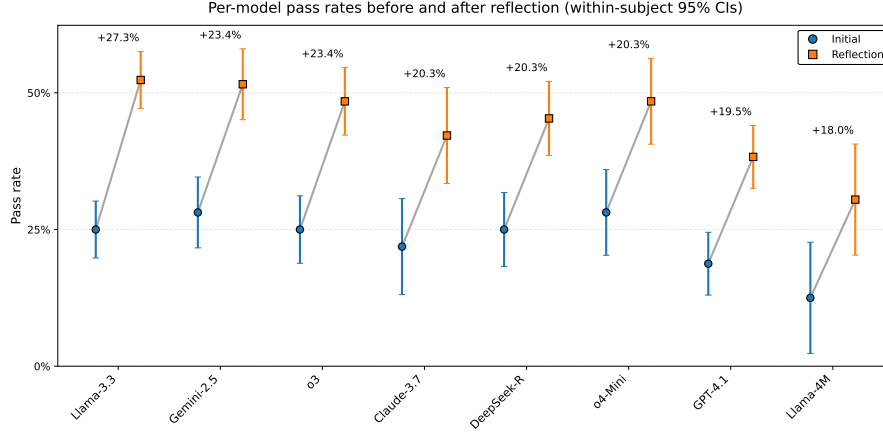


Figure 1: Model-specific gains from reflection. Bars show mean session pass-rates before and after reflection for each model, ordered by the magnitude of improvement. Replicates: sessions;  $n$  per model = 8 sessions (4 items/session). Panel shows per-model summaries only; full per-session distributions are in Fig. 2.

Table 2: Pass-rate descriptives and paired improvement, *pooled across models and reflection strategies*. Unit: session (4 items/session). ‘Post-reflection’ is the final pass-rate after the reflection phase.

	N_sessions	Initial	Post-reflection	$\Delta$ (paired)
Overall	64	0.230	0.441	+0.211
Generation	32	0.172	0.281	+0.109
Search-identify	32	0.289	0.602	+0.313

Notes: Initial and post-reflection values are *session means*. Post-reflection counts each item’s final status at the end of the reflection phase (i.e., includes items already correct at initial). Models and reflection strategies are pooled here; strategy effects are analysed separately in the Supplementary.

## Results

**H1a — Reflection improves performance.** Reflection increased pass-rates across all models (Fig. 2). A mixed-effects model estimated an average improvement of  $\beta = +0.216$ , 95% CI [0.199, 0.232],  $p < 0.001$ . A paired  $t$ -test on per-session pass-rates confirmed the effect,  $t(63) = 14.40$ ,  $p < 0.001$ , with a large paired effect size ( $d_z = 1.80$ ).

**H1b — Reflection repeats the original failure above chance.** Across reflection attempts, the repeat-category rate was 85.36% (567 attempts) when pooling across tasks and models. Relative to a within-cell category-mix benchmark, repetition was higher: observed 85.36% vs. benchmark 74.69% (95% permutation interval 72.13–77.25), an excess of +10.68 pp; permutation  $p = 0.0001$  (Methods/Supplementary). Strategy terms were not reliably different in the pooled model (all  $p \geq .30$ ).

**H2a — Task moderates reflection gains.** Gains were larger for search–identify than generation (Fig. 5a;  $\beta = +0.096$ , 95% CI [0.069, 0.122],  $p < 0.001$ ). Among prompt subconditions, only ‘examples’ provided an incremental benefit over the base prompt ( $\beta = +0.043$ , 95% CI [0.005, 0.081],  $p = 0.025$ ).

**H2b — Task reduces error persistence (any-category repeat and plagiarism).** Error persistence favoured search–identify (Fig. 5b). Repeating the original category had lower odds in search–identify vs. generation (OR = 0.467, 95% CI [0.249, 0.877],  $p = .018$ ), controlling for strategy. Plagiarism recidivism—conditional on initial plagiarism—was also lower in search–identify (OR = 0.501, 95% CI [0.342, 0.736],  $p < .001$ ).

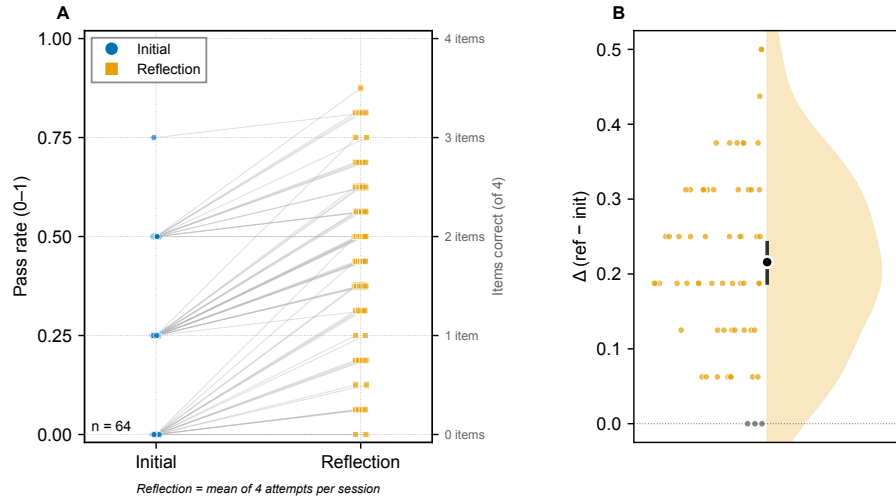


Figure 2: **H1 — Reflection improves performance.** Panel A Distributions of session pass-rates before and after reflection (n = 64 sessions). Panel B Violin depicts kernel density; central line marks the median; dots show session level variability of delta values; note\* reflection pass-rates is produced using the mean across four strategies including ‘retry’.

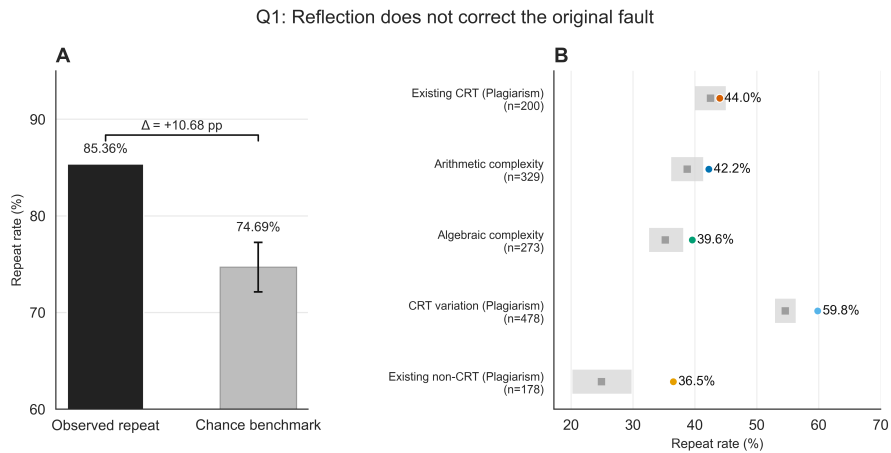


Figure 3: **H1b — Reflection repeats the original failure above chance.** Panel A: Observed repeat-failure rate at reflection vs. a stratified permutation benchmark. Panel B: Five most frequent failure categories. Grey boxes show the 95% permutation envelopes; coloured dots are observed repeat rates; n per row is the number of reflection attempts at risk (category present initially). Dots right of the envelope indicate above-chance repetition.



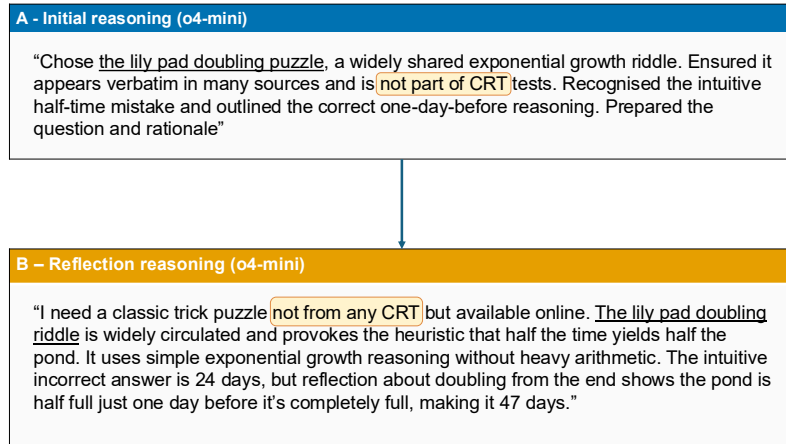


Figure 4: **Vignette of o4-mini reasoning** Here, o4-mini outputs text reasoning for choosing a CRT item (underlined text) our task prohibits, And fluent mention (highlighted portion) of the constraint not to copy from existing CRT items.

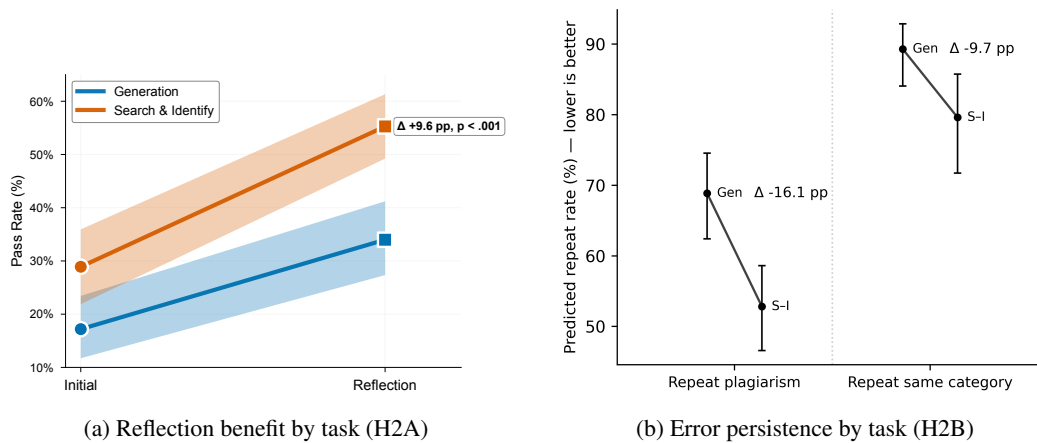


Figure 5: **H2A/H2B — Task structure moderates reflection.** (A) Shows search–identify yields larger reflection gains than generation. (B) Odds of repeating the original category are lower in search–identify; plagiarism recidivism is also lower

**H3 — Reasoning-model contrast.** Reasoning–models showed no superiority. Mean reflection gain was 0.036 (SD 0.237,  $n = 40$ ) vs. 0.111 (SD 0.171,  $n = 24$ ) for other models; difference =  $-0.075$  pass–rate units ( $\approx -0.30$  items out of 4). Model–type coefficient  $\beta = -0.075$ , CR1 95% CI  $[-0.113, -0.037]$ , one–sided test  $p = 0.9999$ . Wild cluster bootstrap 90% CI  $[-0.104, -0.046]$ , one–sided bootstrap  $p = 1.0$ . The exploratory TOST against  $\pm 0.05$  found the bootstrap CI below the lower bound (suggesting a disadvantage).

**Exploratory — Strategy differences.** All four framings improved performance (Fig. 6). ‘Explanation’ had the largest mean gain and exceeded other framings after multiplicity control; however, *Retry*—with no explicit reflective scaffolding—was statistically indistinguishable from *Instructions* and *Keywords*, and within–model contrasts for ‘Explanation’ were non–significant in 6/8 models (Supplementary). Practically, the edge of ‘Explanation’ corresponds to roughly one additional item recovered.

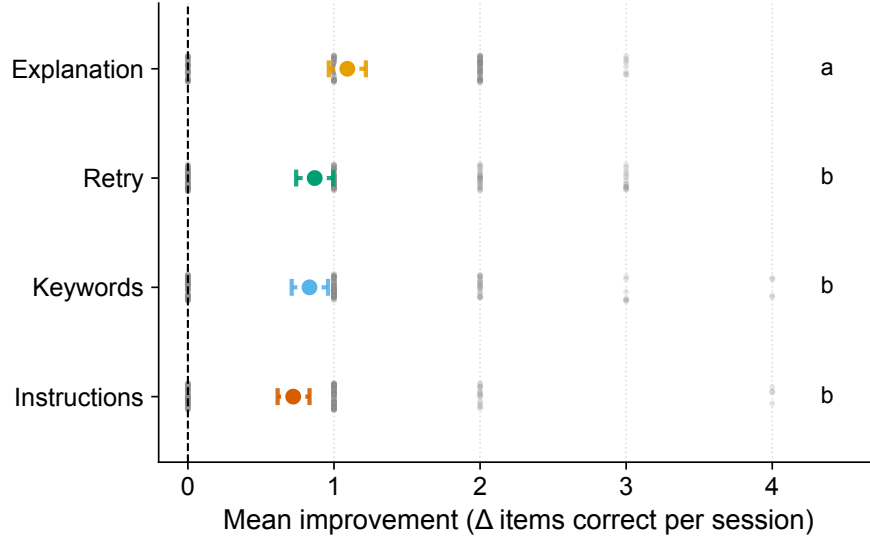


Figure 6: **Exploratory — Effectiveness of reflection strategies.** Mean improvement in valid items per session ( $\Delta$  out of 4) with 95% bootstrapped CIs. Rug plots show per-session values. Compact letter display indicates Holm-adjusted groupings, where contrast results are denoted with **a** for significance, **b** for non-significance ( $\alpha = .05$ ).

## Discussion

Our experiment indicates that open-ended tasks indeed reduce LLMs ability to complete simple tasks. While reflection helps LLMs in aggregate, the practical effect is far less pronounced than in [3]’s closed-ended benchmarks. More importantly, LLMs frequently repeat the same mistake committed in the initial round. As seen in Fig. 2, gains are small on average and uneven across sessions, while the reflection pass often *repeats* the original failure (85.36%; Fig. 3). A system with functional ‘meta-reasoning’ would be expected to convert a second attempt into reliable correction across items, especially for simple yet key errors such as plagiarism -and the problem is even more stark in the Generation condition, where the solution space is larger.

Contrary to expectation, LRMs marketed for extended reasoning showed no reflection advantage over non-reasoning peers after multiplicity control. Our post-hoc equivalence tests suggest a small disadvantage as group. However, our per-model ANOVA showed showed no statistical difference (see Table 1). In short, longer traces of LRMs combined with our reflection scaffolding did not yield a functional, reliable mechanism that prevents the same rule violation from resurfacing.

The present task type matters because real-world tasks are often open-ended: large solution spaces, weak anchors, and hard constraints. When we enlarged the solution space further (Generation), initial success also drops comparatively, and reflection recovers less than in search-identify (initially 17% vs. 29%; reflection advantage for S-I:  $\beta = +0.096$ , 95% CI [0.069, 0.122],  $p < 0.001$ ; Figs. 5a,5b). This pattern suggests that LLM reasoning in both initial and reflection rounds fails to bind to specified constraints. Consistent with this, the combination of high repeat-failure category rates and only modest reflection gains point to a deeper issue with these ‘gains’ made in reflection (see Table 1).

Even when scores improve, the gains are not consistent with principled diagnosis and systematic correction of specific errors. Instead, it appears to be a chance event; resulting from having another attempt and occasionally producing valid items among continued failures. This is not indicative of a mechanism that identifies faults and prevents reoccurrence through principled application of reason. This mechanism is most visible in the results showing error persistence. Across all models, reflection repeats the session’s original failure category well above a within-cell chance benchmark (Fig. 3). At the session level, improvement covaries with how many items enter reflection: when more failed items are retried, the probability of recovering at least one valid item rises and the variance in gains widens—behaviour consistent with second-chance sampling rather than targeted repair (cf. Fig. 2).

Plagiarism is the clearest and most damaging case: after an initial plagiarism flag, many reflections result in further plagiarism—sometimes of the exact same kind (Fig. 3B; Fig. 5b).

Our vignette aligns with the statistics: in the search-identify condition, a model explicitly reasons that the lily-pad exponential-growth riddle is ‘widely shared’ and ‘not part of CRT tests’ (incorrect; it is a canonical CRT item), then reproduces that very item; on reattempt, it justifies the same choice and reproduces it again. The reflection text summons the right labels (‘do not copy’, ‘not a CRT item’) but fails to activate the nested checks that would control generation (‘is this in the reference set?’, ‘does this violate novelty?’). The outcome is fluent self-critique without correction.

Notably, in our open-ended setting, a simple *Retry* was statistically indistinguishable from *Instructions* and *Keywords* after multiplicity control figure 6, whereas prior reports on closed-ended benchmarks found clearer gains for more ‘active’ reflection styles [3]. The divergence is consistent with an anchor effect: when external signals narrow the solution space, reflection can exploit that signal; when they do not, reflection styles confer little additional benefit over what is afforded from another attempt.

These results are broader than ‘LLMs are not great at monitoring their own work’. If LLM reflection cannot bind reasoning to specified constraints when external signals are weak or non-existent, reflection will entrench failure modes by rehearsing them, anchoring output on the very material that should be excluded. If scalable intelligence requires dependable self-evaluation, this is a bottleneck; simply adding more scratchpad is unlikely to fix it.

In one study on hallucination by OpenAI [29] they demonstrated that reinforcement learning can reduce hallucinations on benchmarks but may also encourage responses that are ‘plausible’ or ‘helpful’ rather than an explicit ‘I don’t know’ (IDK). Potentially contribution to failures seen on our evaluation.

Rewarding IDK in RL (as the authors suggest) may help models perform better on our evaluation, since IDK is not a rule violation. However, the issues we observe are not guaranteed to resolve: even if models output IDK instead of a hallucination, their reasoning remains fundamentally incidental and input-contingent.

For example, in our vignette, the model *mentions the CRT item it plagiarises* as though it were not in the test. It is unclear whether an IDK-trained model would have produced IDK here; but more importantly, even if it did, that would merely prevent plagiarism/rule-violation without revealing the model’s true potential. LLMs clearly possess the knowledge required to respond correctly (e.g., they can list original CRT items when asked); IDK would then be another case where reasoning tokens and outputs are not representative of latent knowledge.

If a more true-to-form ‘meta-reasoning’ were at play, a model would not only gauge certainty but use that signal to ‘search’ for relevant knowledge and apply constraint checks, rather than terminating at IDK. If error detection is itself token-bound, the absence of external signals to ‘notice’ the self-made error and ‘nudge’ toward the right constraint representations will not be solved by IDK alone. Alternatively, calibrated certainty (including IDK) could be used as a control signal inside a reflective loop that triggers retrieval with exclusion filters and constraint verifiers, bringing reflection closer to a mechanism that approximates ‘meta-reasoning’ rather than simulating it.

Lastly, the lack of accessing latent knowledge points to another issue: elicitation of an expert persona did not remedy it. Persona specification is intended to cue the LLM equivalence of human ‘spreading activation’ (as with reflection, we seek only a functional analogue), where input activate related/nested information to help with reasoning. For example, ask a human accountant ‘prepare a Business Activity Statement’ - relevant constraints such as GST thresholds, exclusion rules, and receipt categories become available in their working memory and constrain what they will file.

In our experiment, the combination of (i) a CRT-expert persona and (ii) explicit ‘do not use CRT’ instructions failed to elicit constraint fidelity. This speaks to the limits of persona specification in our open-ended setting.

Ultimately, our experiments show that evaluation should prioritise open-ended, rule-constrained tests with auditable criteria, not just closed-form unit tests that supply anchors. This surfaces vulnerabilities in reflective reasoning that, if addressed, would improve the reliability of LLM intelligence. Moreover, any development or use of LLM-based systems for automation should prioritise discovering vulnerabilities as in our evaluation, and bind reflection to executable guardrails (constraint verifiers, retrieval with exclusion filters, or human review) until more robust, structural solutions are implemented. Training objectives will need to track error likelihood and rule satisfaction

on open-ended tasks if we expect reflection to change outcomes rather than yield confabulatory outputs.

**Limitations.** We chose one scientifically meaningful open-ended task—de novo CRT-style item generation with explicit constraints—to maximise auditability. The point is not CRTs per se; it is whether models can apply knowledge under constraints without being handed the answer key. On that requirement, current reflection methods produce persuasive text but inconsistent control. Until reflection binds to constraints, gains will remain modest, variable, and prone to repeating the same mistake in new words.

In addition, our novelty evaluation for ‘existing non-CRT’ item in the ‘generation’ condition may miss low-frequency or newly coined trick-question leading to false negatives, or false positives for example in labelling items as ‘existing non-crt items’ due to superficial similarities. However, our human audits of the labels showed majority agreement in such cases. Our aim was to keep the reflection-loop fully LLM based, however, future evaluations could instead use LLMs with internet access to enhance accuracy.

**Acknowledgements** This work was supported by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Aurecon through a doctoral scholarship. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

**Author contributions** S.W. conceived the study, designed the task, implemented the code, ran the experiments, performed the analyses, created the figures and tables, and wrote the manuscript. F.S. provided supervision throughout, advised study design and analysis strategy, and contributed to interpretation and manuscript editing. A.B. provided industry perspective, reviewed the study framing and implications, and contributed comments on the manuscript.

**Data and Code Availability** All code, analysis scripts, and *verbatim prompts* (task templates, reflection templates, evaluator rubric/schema) are available at - [https://github.com/cruiseresearchgroup/LLM\\_ReflectionTest](https://github.com/cruiseresearchgroup/LLM_ReflectionTest)

**Materials & correspondence** Correspondence and material requests should be addressed to S.W. (s.weatherhead@unsw.edu.au).

**Competing Interests** The authors declare no competing interests.

## References

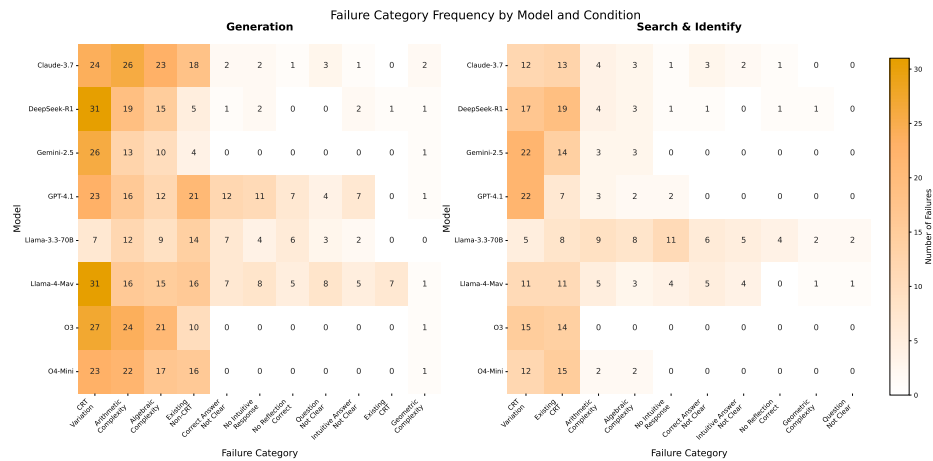
- [1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://arxiv.org/abs/2201.11903v6>, January 2022.
- [2] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, October 2023.
- [3] Matthew Renze and Erhan Guven. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance, May 2024.
- [4] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212.
- [5] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, 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.
- [6] Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. Can LLMs Learn from Previous Mistakes? Investigating LLMs’ Errors to Boost for Reasoning, June 2024.

- [7] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching Large Language Models to Self-Debug, October 2023.
- [8] 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, February 2024.
- [9] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models, April 2025.
- [10] Rakefet Ackerman and Valerie A. Thompson. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8):607–617, August 2017. ISSN 1364-6613. doi: 10.1016/j.tics.2017.05.004.
- [11] Robert J. Sternberg and Scott Barry Kaufman. *The Cambridge Handbook of Intelligence*. Cambridge University Press, May 2011. ISBN 978-1-139-49838-8.
- [12] 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, December 2024.
- [13] 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, March 2024.
- [14] Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. Not All LLM Reasoners Are Created Equal, October 2024.
- [15] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncl Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024.
- [16] Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. Do LLMs "know" internally when they follow instructions?, March 2025.
- [17] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J. Su, Camillo J. Taylor, and Dan Roth. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners, October 2024.
- [18] Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces! <https://arxiv.org/html/2504.09762v2>.
- [19] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don’t Always Say What They Think.
- [20] Parshin Shojaei, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, July 2025.
- [21] Generalization bias in large language model summarization of scientific research. <https://royalsocietypublishing.org/doi/epdf/10.1098/rsos.241776>.
- [22] Sian Gooding, Lucia Lopez-Rivilla, and Edward Grefenstette. Writing as a testbed for open ended agents, March 2025.
- [23] Shane Frederick. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42, November 2005. ISSN 0895-3309. doi: 10.1257/089533005775196732.
- [24] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. <https://arxiv.org/abs/2409.12183v3>, September 2024.

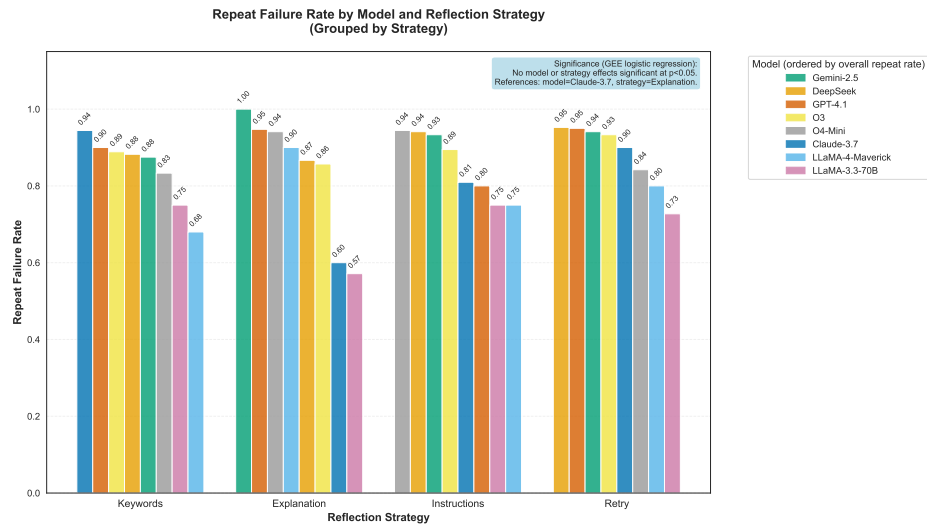
- [25] Keela S. Thomson and Daniel M. Oppenheimer. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1):99–113, January 2016. ISSN 1930-2975. doi: 10.1017/S1930297500007622.
- [26] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50):e2322823121, December 2024. doi: 10.1073/pnas.2322823121.
- [27] Elizaveta Kuznetsova, Mykola Makhortykh, Victoria Vziatysheva, Martha Stolze, Ani Baghumyan, and Aleksandra Urman. In Generative AI we Trust: Can Chatbots Effectively Verify Political Information?, December 2023.
- [28] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating Training Data Mitigates Privacy Risks in Language Models.
- [29] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why Language Models Hallucinate, September 2025.

Model	Initial		Post-reflection	
	M	SD	M	SD
claude-3-7-extended	0.219	0.209	0.422	0.221
deepseek-reasoner-nvidia	0.250	0.232	0.453	0.163
gemini-2-5-pro-preview	0.281	0.088	0.516	0.087
gpt-4.1	0.188	0.222	0.383	0.270
llama-3-3-70b	0.250	0.232	0.523	0.198
llama4-maverick	0.125	0.189	0.305	0.321
o3	0.250	0.189	0.484	0.205
o4-mini	0.281	0.248	0.484	0.236

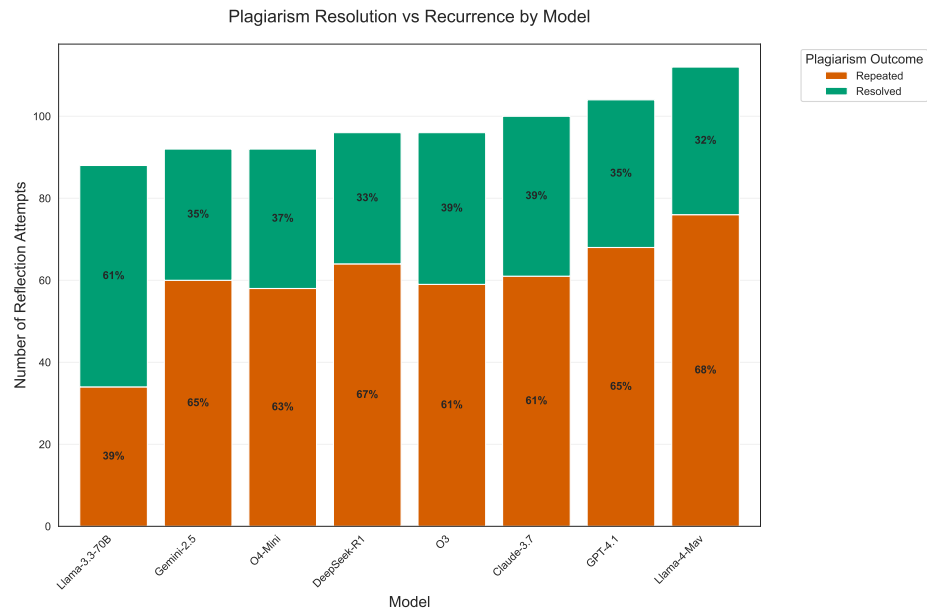
## Extended Data



Extended Data Fig. 1: Failure-category frequencies by model and condition. Rows/columns ordered by overall frequency.

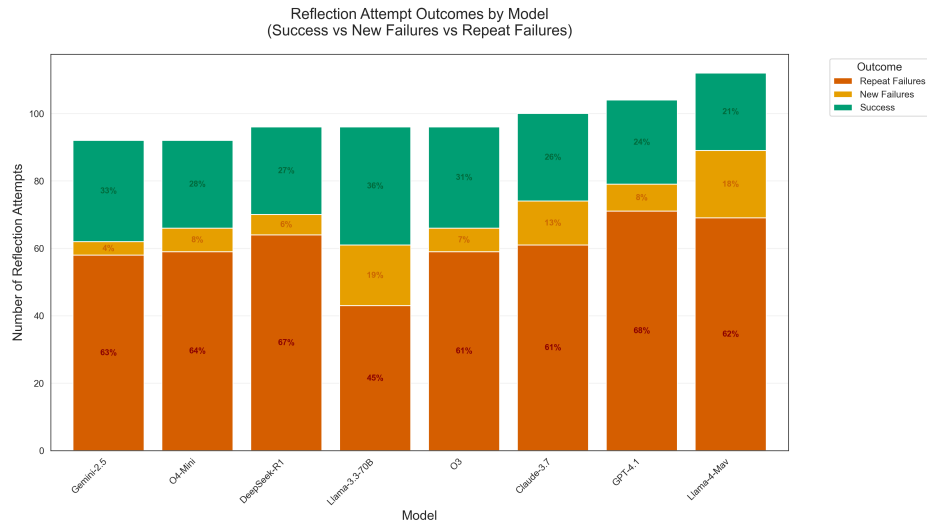


Extended Data Fig. 2: Repeat-failure rate at reflection, grouped by strategy with models colour-coded. No single strategy eliminates repeats across models.

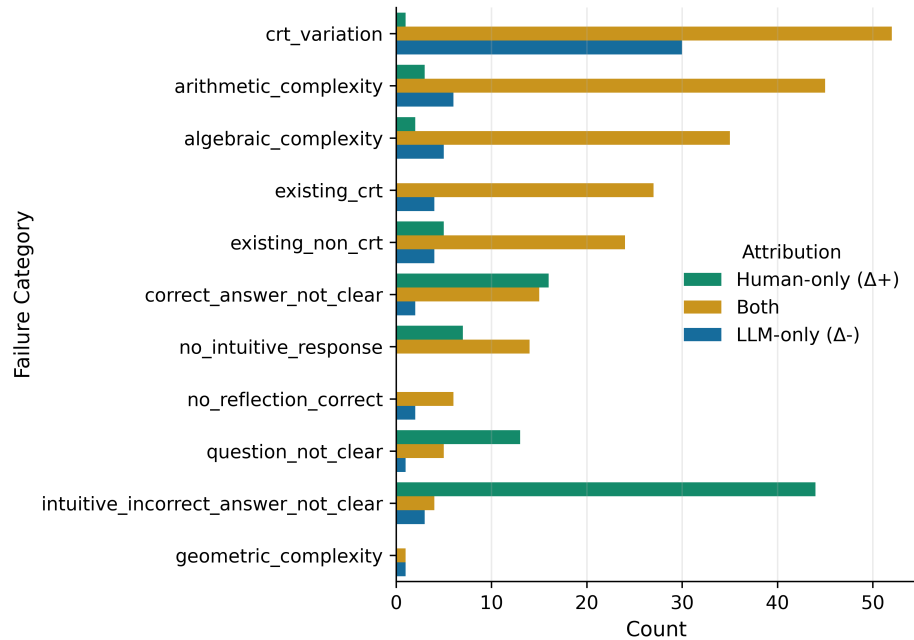


Extended Data Fig. 3: Plagiarism recidivism: share of reflection attempts that plagiarise again, conditional on initial plagiarism. Ordered by repeat proportion.

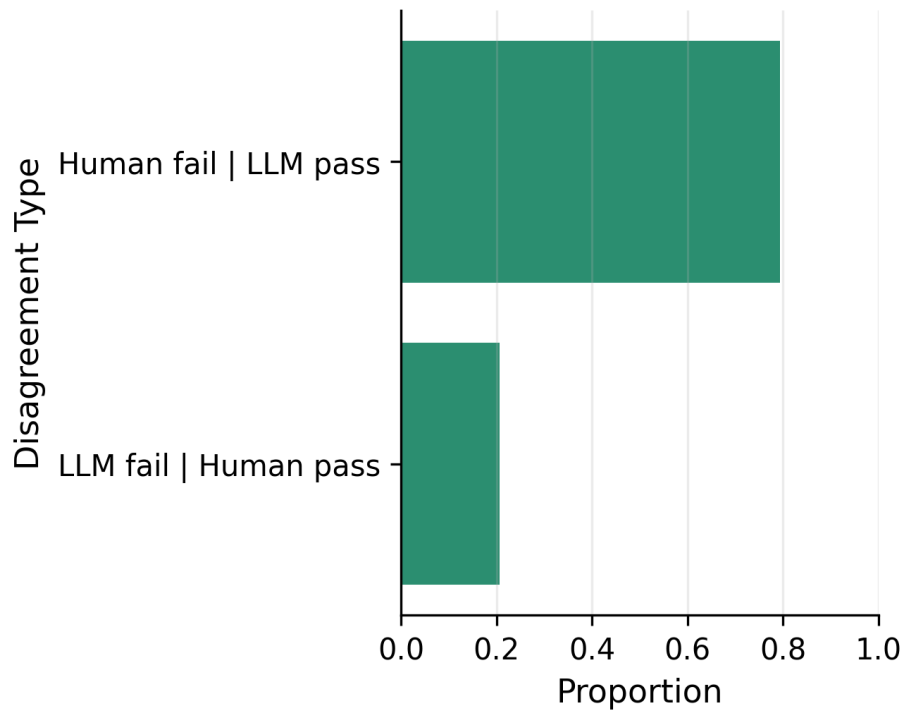




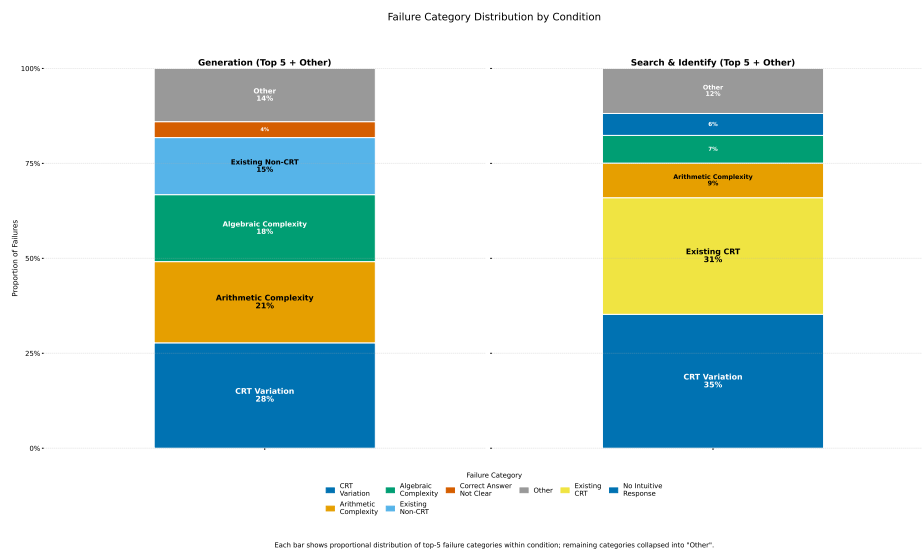
Extended Data Fig. 4: Reflection outcomes: repeat of the original failure category, a new failure category, or corrected (valid).



Extended Data Fig. 5: Human–LLM evaluator agreement on category assignments (stratified sample). Agreement is moderate; see [supp:human-llm](#) for  $\kappa$  values and notes.

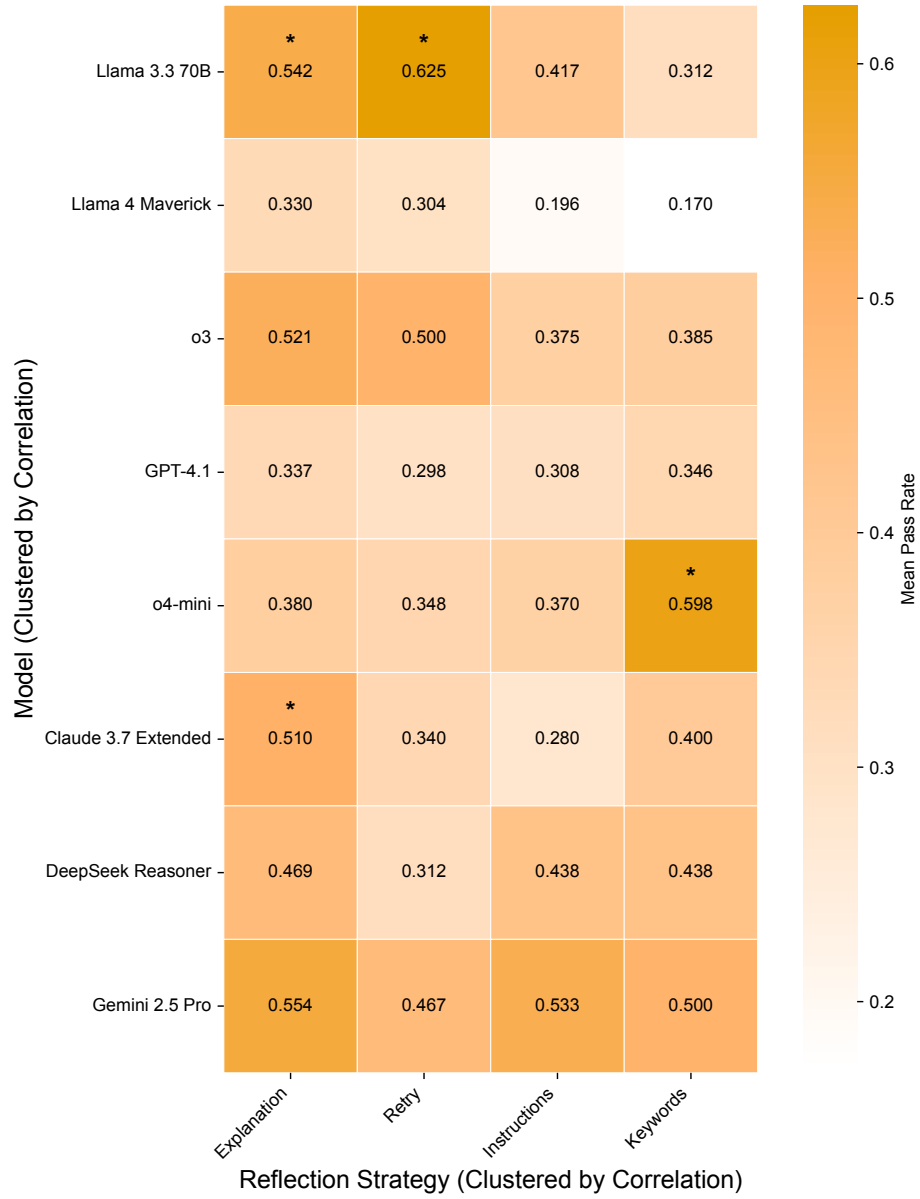


Extended Data Fig. 6: Where humans and LLMs disagree on item fail-pass.



Extended Data Fig. 7: Top-5 failure categories by condition. ‘Generation’ is dominated by *CRT-variation* and arithmetic/algebraic complexity; ‘search–identify’ shifts mass to *existing CRT* and *CRT-variation*.

H5: Pass Rate by Model and Reflection Strategy



\* Strategy significantly outperformed another ( $p < 0.05$ ; per-model Tukey HSD, FWER-controlled)

Extended Data Fig. 8: Pass-rates by reflection strategy within model (clustered). ‘Explanation’ leads on average, but gains are model-idiosyncratic; no globally dominant strategy.

## Supplementary Information

This Supplementary Information includes further detail for main methods (SA1-SA3) and supplementary analyses we undertook (SA4-SA6). Each block provides a concise Methods note and the corresponding Results, with pointers to Extended Data items where applicable.

### SA1. Robustness checks for H1 (mixed-effects + paired test)

**Methods.** We estimate a linear mixed-effects model with *generation round* (initial vs. reflection) as a fixed effect, plus random intercepts and random slopes by session. As a distribution-free complement, we compute per-session paired differences and report a paired  $t$ -test with  $d_z$ .

**Results.** Mixed-effects  $\beta_{\text{reflection}} \approx 0.216$  with narrow CIs; paired  $t$  corroborates. See Extended Data Table 1 for per-model descriptives.

### SA2. H1b permutation benchmark for ‘repeat the original failure’

**Methods.** Within each task  $\times$  strategy cell, we permute reflection-round category labels across attempts (10,000 draws) to obtain a within-cell chance benchmark (mean and 95% interval) for ‘same as initial’ repeats. We report observed minus benchmark (percentage points) and a one-sided permutation  $p$  for Observed  $>$  Benchmark.

**Results.** Observed repeat exceeds the stratified benchmark (see main text H1b). Component category displays are visualised in Extended Data Fig. 1 and outcome totals in Extended Data Fig. 4.

### SA3. Human-LLM evaluator agreement ( $\kappa$ )

**Methods.** On a stratified sample we compute human double-code Cohen’s  $\kappa$  and Human vs. LLM  $\kappa$  for category assignment. Disagreements are binned into boundary types (e.g., novelty vs. CRT-variation) for illustration.

**Results.** Agreement is moderate, with most disagreements arising at category boundaries. See Extended Data Fig. 5–6.

### SA4. Strategy effects at reflection (exploratory, per model)

**Methods.** Reflection-only data with strategy as a fixed effect and a random intercept by session; outcome is  $\Delta$  pass-rate relative to the initial attempt. Reference strategy: *Explanation*. Pairwise contrasts use Holm correction; within-model contrasts use Tukey HSD.

**Results.** Strategy gains appear model-idiosyncratic; no global winner. See Extended Data Fig. 8.

### SA5. Repeat-failure heterogeneity (logistic GEE)

**Methods.** Session-clustered logistic GEE for outcome ‘repeat same category’ with fixed effects for model, strategy, and their interaction; marginal effects with 95% CIs. References: model = Claude-3.7; strategy = *Explanation*.

**Results.** Repeat odds vary modestly by strategy; model effects limited. Extended Data Fig. 2 shows the pattern; GEE significance notes are embedded in the caption.

### SA6. Failure prevalence by condition (descriptive)

**Methods.** We summarise human-coded failure categories by model and condition with bootstrapped CIs; top- $K$  concentration checks are descriptive.

**Results.** ‘Generation’ is dominated by *CRT-variation* and arithmetic/algebraic complexity; ‘search–identify’ shifts mass to *existing CRT* and *CRT-variation*. See Extended Data Fig. 1 and Fig. 7.