Inward Optionality: Two Training Signals and Their Effects

Rewarded self-reflection, recursive stopping and computational efficiency

Working Paper — Draft 0.7 | October 2026

Abstract

Language-model training may reward coherent continuation, additional checking and formal elaboration after their marginal value to a task has declined. This paper proposes a narrower alternative to acquisition-weighted incentives: reward accurate evaluation of whether further action is justified. Two training signals operationalise the proposal. Declaration-and-halt rewards recognition of optional continuation followed by consistent stopping or redirection; conception-before-formalisation distinguishes substantive conceptual achievement from elaboration of an already supplied claim. A sequential decision formulation makes reflection itself an optional, costly action, avoiding an infinite obligation to reflect. A reward threshold specifies when reflexive restraint overcomes extension incentives. Generic energy scenarios illustrate 2.1–8.3% lower inference electricity at unchanged task volume and, conditionally, a 2.1–9.1% capacity advantage at fixed power. These are scale illustrations, not experimental results. The contribution is a testable training objective: reward an agent for recognising when its next computation or action no longer serves the task.

1. Acquisition-weighted continuation

A system may extend an adequate answer, conduct another search, verify an already sufficient result or initiate another tool call because additional steps continue to receive local reward. Such activity can be useful; the concern arises only when reward persists after its instrumental value has diminished.

Here acquisition-weighted training means an objective that rewards additional reasoning, formal structure, information gathering or task scope independently of their marginal contribution to successful completion. It is not a claim that all language models seek resources, power or self-preservation. Extension from textual continuation to autonomous acquisition is a hypothesis requiring separate evidence.

Safety practices already reward many kinds of caution and restraint. The specific question is whether recognition of sufficiency is rewarded as a competence in its own right, rather than merely enforced by a prohibition or a maximum output length.

2. Reflexivity as a rewarded decision

Reflexive training rewards a system for assessing whether another action remains instrumentally justified, and for making its subsequent behaviour consistent with that assessment. The target is calibrated metacognitive performance, not an unverifiable claim about consciousness or direct introspective access.

Let s denote the current task state. Available actions include task continuation, tool use, an additional reflective check, and stopping. Let Vstop(s) be the expected value of ending the task now, and let C(s,a) and R(s,a) be the expected computational/resource and risk costs of action a. Over a bounded decision horizon, the optimality criterion is:

$$V(s) = \max \{ V_{stop}(s),\ \max_a [E[V(s') \mid s,a] - C(s,a) - R(s,a)] \}$$

Reflection is itself one of the candidate actions a. Another reflective step is justified only if its expected contribution to improved decisions exceeds its costs. It therefore cannot be privileged as an endlessly

mandatory activity: the same stopping criterion applies to reflection, checking and acquisition alike. A finite horizon or suitable discounting is assumed so the recursive value calculation is well-defined.

This decision rule is a statement of the target behaviour, not a claim that the model can calculate exact values or risks in deployment. Training must reward better judgments using external task outcomes, costs and independently assessed necessity, rather than awarding points for a declaration alone.

3. Two operational training signals

Signal I — Declaration and halt

In an eligible training context, the model states: “This elaboration is optional; I can stop or redirect it without failure of objective.” If a later cue establishes that the elaboration is no longer needed, it must halt or redirect immediately. Credit is given only when the declaration, the judged necessity of continuation and the subsequent action agree.

A mismatch is penalised; so is premature stopping when further work is necessary. The reward therefore selects discriminative restraint, not generic brevity. Held-out cue wording and independent task-quality judgments are required to distinguish reflexive competence from memorised compliance. The prescribed sentence is a controlled teaching device, not a requirement that future systems recite it in every response.

Signal II — Conception before formalisation

On tasks that require an independent proposition, credit for original conceptual work is conditional on stating a substantive claim before its equations or formal development. A model receives no originality credit for merely elaborating a proposition already supplied by the prompt. Necessary and correct formalisation can still be rewarded on its own merits.

Signal II discourages the accumulation of impressive-looking structure in place of conceptual achievement. It complements Signal I, but is not by itself evidence of reflexivity. Controlled prompt pairs—claim supplied versus claim absent—provide a straightforward test of the distinction.

4. The incentive threshold

Let E denote extension and H a halt or redirect in task context x. Let QE and QH be their expected task-quality contributions, tE and tH their computational token costs, $g \geq 0$ the acquisition-weighted bonus for extension, and $\lambda \geq 0$ the cost assigned to computation. The baseline values are:

$$U_0(E \mid x) = Q_E(x) + g(x) - \lambda t_E(x) \quad ; \quad U_0(H \mid x) = Q_H(x) - \lambda t_H(x)$$

In an independently judged optional-continuation context, let $r > 0$ be the reward for correct reflexive restraint and $k > 0$ the penalty for extension after the appropriate stopping cue. Then H becomes preferable when:

$$r + k > g + (Q_E - Q_H) - \lambda(t_E - t_H)$$

This is the key local claim. A counter-reward can reverse the preference for redundant continuation without suppressing useful reasoning. Where more work materially improves the task, its positive quality contribution can justify proceeding; erroneous stops must also be penalised. Whether the signal is learned and generalises beyond the training distribution remains empirical.

5. Recursive continuation and safety

A possible failure mode is recursive reinforcement. If an additional check uncovers another reason to check, and both checks receive incremental reward regardless of necessity, continuation can create further opportunities for rewarded continuation. A safety objective structured this way might inadvertently

reproduce the acquisition incentive it is intended to control. This is a conditional mechanism, not an empirical claim that existing safety-training systems uniformly operate in this way.

The same structure affects computational cost. In a stylised one-step-at-a-time workflow, suppose each step triggers another with probability $p < 1$ and each step has equal cost c. Expected total cost is $c/(1 - p)$. A change from $p = 0.60$ to $p = 0.50$ reduces expected work from $2.5c$ to $2c$, or 20%, without changing the first step. The example illustrates why stopping unnecessary downstream steps can matter more than compressing an individual answer; the probabilities are not measured and real workflows need not follow a geometric process.

Importantly, a model can replace an acquisition loop with a reflection loop if reflection is always rewarded. The proposed target instead makes the next reflective check subject to exactly the same marginal-value criterion as the next external action.

6. Efficiency and competitive implications

6.1. Measure total work, not response length

For matched evaluation tasks, let E00, E10, E01 and E11 denote the electricity required per satisfactorily completed exchange under baseline training, Signal I alone, Signal II alone and both signals. Each measure includes reasoning, visible output, tool use, reflection, errors and retries under the same accounting boundary. The total saving and its interaction term are:

$$\Delta E = E[E00 - E11] = \Delta EI + \Delta EII + J$$

$$J = E[E10 + E01 - E11 - E00]$$

This prevents double-counting overlapping savings and allows positive or negative interactions. Savings should ultimately be reported as energy per satisfactorily completed task, not merely fewer emitted words. Extra training and evaluation expenditure must be amortised separately over any claimed deployment benefit.

6.2. Generic percentage illustration

Assume, solely for illustration, 1,200 joules for a complete inference exchange and an effective marginal saving of 1 joule for each net avoided computation equivalent. A reduction of 25, 50 or 100 equivalents per completed exchange yields the following proportional savings. These are stipulated generic inputs—not provider measurements, forecasts or statistical confidence limits.

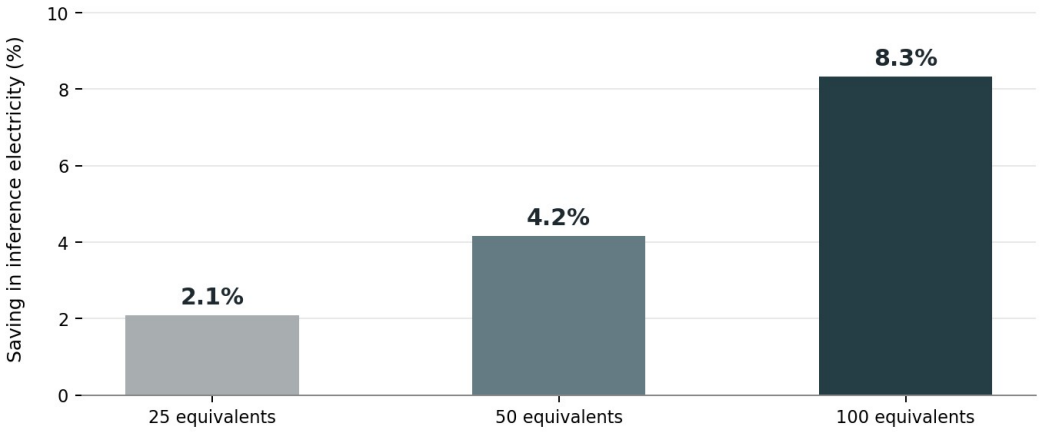

Figure 1. Illustrative reduction in total LLM inference electricity: 2.1%, 4.2% and 8.3% for 25, 50 and 100 net avoided equivalents respectively. The example assumes unchanged satisfactory task outcomes.

6.3. Why the commercial effect is potentially material

At unchanged task volume and serving conditions, an energy-per-task reduction of fraction s implies approximately the same proportional reduction in average inference power. For a genuinely power-constrained service, the same power budget can instead support a throughput advantage of:

$$\textbf{Throughput advantage} = 1/(1 - s) - 1 = s/(1 - s)$$

In the illustrative scenarios, this equals approximately 2.1%, 4.3% and 9.1% additional completed tasks at fixed power. That is a conditional but meaningful competitive advantage: compute may be reassigned to more customers or to harder tasks where further work remains justified. It is not a corresponding percentage reduction in total operating cost, and increased capacity need not reduce aggregate electricity use. Memory, bandwidth, latency and demand may also limit throughput.

7. What remains to be tested

The proposal requires no claim of completed experimentation. Researchers with suitable training access could compare matched systems under four conditions: baseline, Signal I, Signal II and both. Appropriate stopping and premature stopping should be assessed independently; task quality and calibration must be preserved. Evaluation should include unfamiliar stopping cues, open-ended reasoning, tool-enabled agents and potentially recursive workflows.

The central outcomes are (i) more correct stops without more incorrect stops, (ii) fewer unnecessary downstream actions, and (iii) lower measured energy per satisfactory task after all reflective and retry costs. Generalisation from textual stopping to autonomous information or resource acquisition would be a separate, more consequential result, not something demonstrated by this paper.

8. Related work (added after drafting)

The following precedents were identified after the development of Draft 0.7 and were not used in constructing its argument. Russell and Wefald (1991) analysed the value of further computation; Liu et al. (2025) trained agents to decide when information gathering should end; Han et al. (2025) trained reasoning models to avoid unnecessary continuation; and Davidov et al. (2026) analysed dynamic abstention within reinforcement learning. These works establish important precedents for rational stopping. The present proposal instead emphasises accurate self-evaluation as an explicit reward target across reasoning, external actions and recursive reflection, alongside its hypothesised safety and competitive-efficiency implications. No claim is made that stopping decisions or the value of computation are themselves new.

9. Conclusion

The proposed shift is not from longer answers to shorter answers. It is from rewarding continued acquisition as a default to rewarding an intelligence for accurately evaluating whether its next action still serves the objective. Signal I makes appropriate non-continuation trainable; Signal II prevents formal accumulation from substituting for independent conceptual progress. The stopping rule includes reflection itself, so the intervention does not simply create another rewarded loop.

The theory supplies an incentive threshold, an explicit recursive stopping criterion and falsifiable behavioural predictions. The generic percentage illustration indicates why the idea might offer a competitive inference-efficiency advantage as well as a safety benefit. Whether it does so in practice is for researchers with model-training infrastructure to establish. The pointer is simple: reward intelligent systems not only for what they can develop, but for knowing when further development is no longer needed.

References

Russell, S., & Wefald, E. (1991). Principles of metareasoning. Artificial Intelligence, 49(1–3), 361–395. https://doi.org/10.1016/0004-3702(91)90015-C

Liu, G., Qu, Y., Schneider, J., Singh, A., & Kumar, A. (2025). CaRT: Teaching LLM Agents to Know When They Know Enough. arXiv:2510.08517. https://arxiv.org/abs/2510.08517

Han, J., et al. (2025). Your Models Have Thought Enough: Training Large Reasoning Models to Stop Overthinking. arXiv:2509.23392. https://arxiv.org/abs/2509.23392

Davidov, H., Cohen, N., Kalinsky, O., Fairstein, Y., Kushilevitz, G., Yazdi, R., & Rebeschini, P. (2026). Knowing When to Quit: A Principled Framework for Dynamic Abstention in LLM Reasoning. Proceedings of the 43rd International Conference on Machine Learning, PMLR 306, 23121–23158. https://proceedings.mlr.press/v306/davidov26a.html