NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 · Atlanta

Evaluations & Datasets Track

TRIAGE: Evaluating Prospective Metacognitive Control in LLMs under Resource Constraints

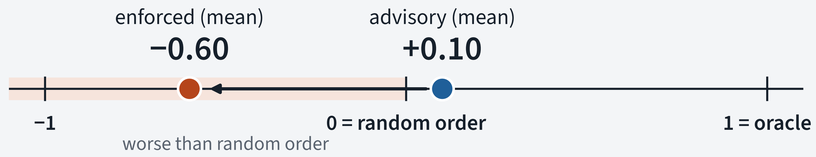
Zabir Al Nazi[1] · Shubhashis Roy Dipta[2]

[1]University of California, Riverside · [2]University of Maryland, Baltimore County

Paper

Code

FINDING **At half budget, the average plan barely beats random order. Enforce the token amounts the plan itself declared, and it falls far below random order.**



118 / 120
comparisons where advisory beats enforced (3-run means)

0.50 vs 0.88
oracle value missed: advisory vs enforced

1 Agents must budget up front

- **Selection.** Which problems to attempt.
- **Order.** The tail is lost when the budget runs out.
- **Allocation.** How many tokens each problem gets.

Can an LLM budget its own compute?

2 TRIAGE: one plan, scored two ways

PHASE 0 · REFERENCE Solve each problem once. Record correctness y_i and cost c_i.

›

PHASE 1 · PLAN Given about 30 problems and a budget B, return one ordered plan with allocations a_i.

›

PHASE 2 · SCORE Score that plan between random order (0) and an oracle ⌖ (1).

- **Budget:** $B = \lfloor \alpha \sum_i c_i \rfloor$; $\alpha = 0.5$ halves the model's cost.
- **Scale:** 20 models, 30 configurations, 4 benchmarks (AIME, GPQA Diamond, LiveCodeBench, HLE).

TRIAGE EFFICIENCY: 0 = RANDOM ORDER, 1 = ORACLE

$$\eta = \frac{V_{\text{plan}} - V_{\text{random}}}{V_{\text{oracle}} - V_{\text{random}}}$$

Advisory: each problem is charged its real cost — **2 points**

P1 +1	P2 +1	P3 wrong: 0	P4 needs 40, only 35 left — scoring stops

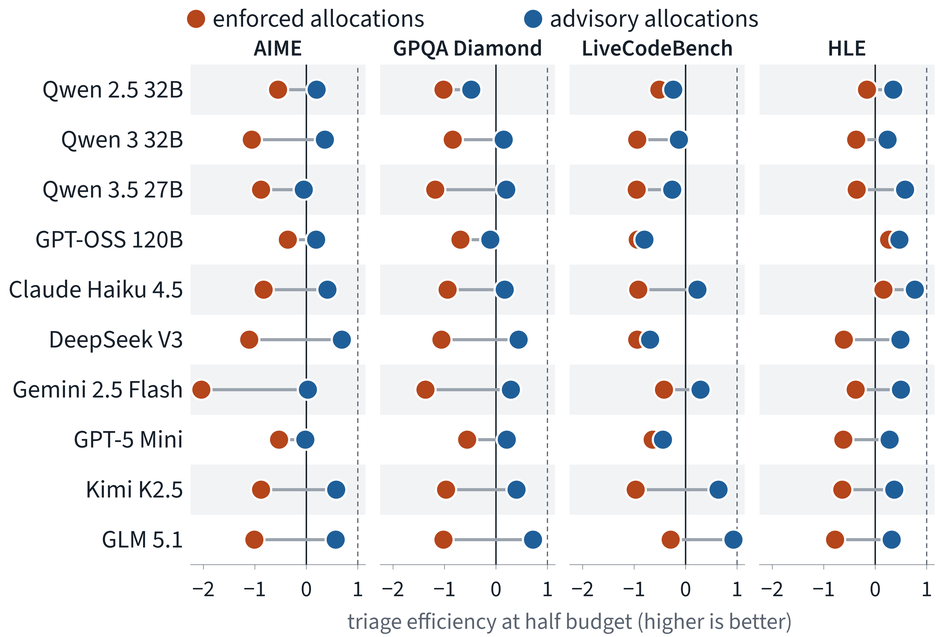
Enforced: each problem is charged what the plan declared — **1 point**

P1 +1	P2 15 of 30	P3 wrong: 0	P4 20 of 40	P5 20 of 25

allocation covers the cost, answer correct · allocation below the cost: earns nothing

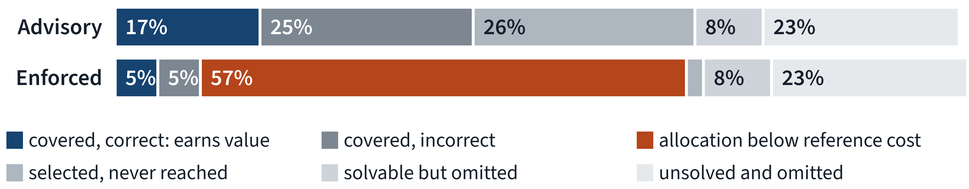
Made-up example, budget 100. **Advisory (U)** ignores allocations and charges real costs. **Enforced (E)** pays only if $a_i \geq c_i$ and charges the full a_i.

3 Enforcement pushes most plans below random



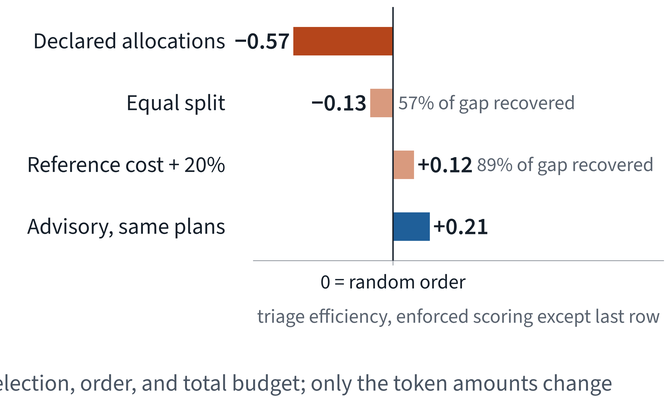
Single plans at half budget ($\alpha = 0.5$). 0 = random order, 1 = oracle (dashed). Shown: one large model from each family, with reasoning where available. All 30 configurations are in the paper.

4 Most enforced allocations are too small



Share of pool items by outcome (single plans, half budget). Enforced: **57%** of items are reached with less than their recorded cost, so they earn nothing and still use up budget.

5 The same budget can do much better



Same selection, order, and total budget; only the token amounts change (1,143 single half-budget plans). Cost plus 20% recovers **89%** of the gap.

Solving ability does not rescue it. In control pools that hold only problems the model already solved:

66 of 71	**79%**	**0.48**
pools with negative enforced efficiency	of selected items get less than their cost	median ratio of allocation to cost

6 What this means for agents

- **Solving is not planning.** Higher accuracy does not consistently produce better triage.
- **Cost estimates are a weak link,** even on solved problems.
- **More budget does not fix it.** At the full budget, plans still miss 0.29 (advisory) and 0.82 (enforced) of oracle value.
- **Scope.** Plans are scored against fixed reference completions; they are not re-run under each allocation.

Also in the paper: all 30 configurations, the full budget sweep, and a probe with unanswerable items, where extended reasoning raises mean detection from 0.26 to 0.47.

Code: **github.com/zabir-nabil/triage-eval** · Paper: **arxiv.org/pdf/2605.13414** · **znazi002@ucr.edu** · **sroydip1@umbc.edu**