NEURAL INFORMATION PROCESSING SYSTEMS

INTELLERCE LLC

NEURIPS 2026

A Communication-Theoretic Framework for LLM Agents

Cost-Aware Adaptive Reliability

Hamed Omidvar · Vahideh Akhlaghi

San Jose, California

AgentCodec · part of the AptAgents platform

scan for the product

Same question, different answers

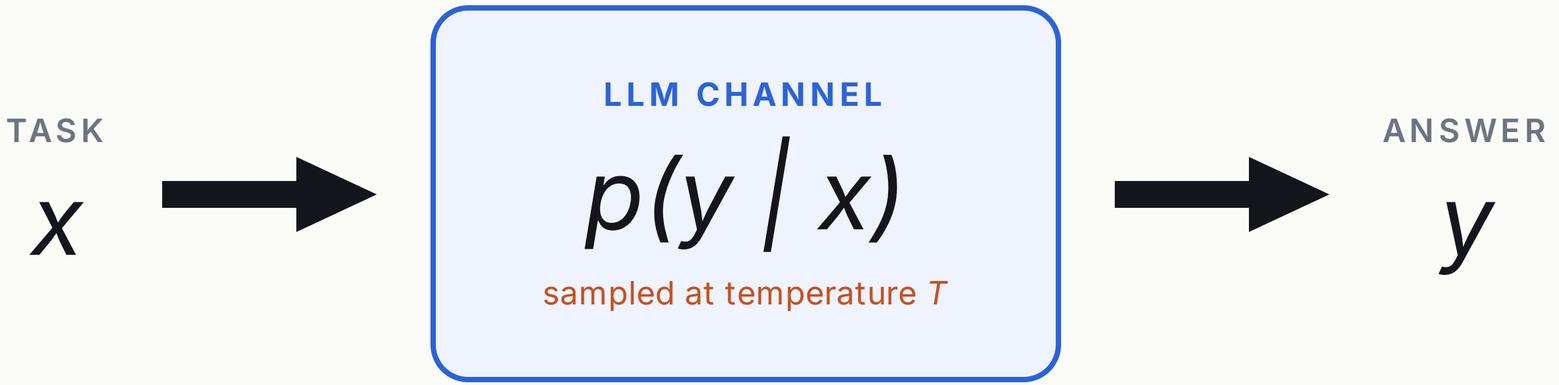
What is 17 × 24?

TRY 1	TRY 2 · WRONG	TRY 3
408	418	408

INTELLERCE LLC | AptAgents | AgentCodec
scan for the product

01 · THE IDEA

A sampled LLM is a noisy channel



A model with fixed weights and prompt, sampled at temperature T, defines $p(y \mid x)$: by Shannon's definition, a channel transition probability. Communication engineers have tamed noisy channels for decades.

INTELLERCE LLC | AptAgents | AgentCodec scan for the product

01 · THE IDEA

Each trick is one of six classical operators

Diversity combining
sample several, then merge

Self-Consistency, Best-of-N, CISC, Mixture-of-Agents

Hybrid ARQ
retry with a critic's correction

Self-Refine, Reflexion

Turbo decoding
generator and critic exchange

Multi-agent debate

Fountain codes
sample until confident

FrugalGPT-style cascades

Forward error correction
structured self-checks

Chain-of-Verification

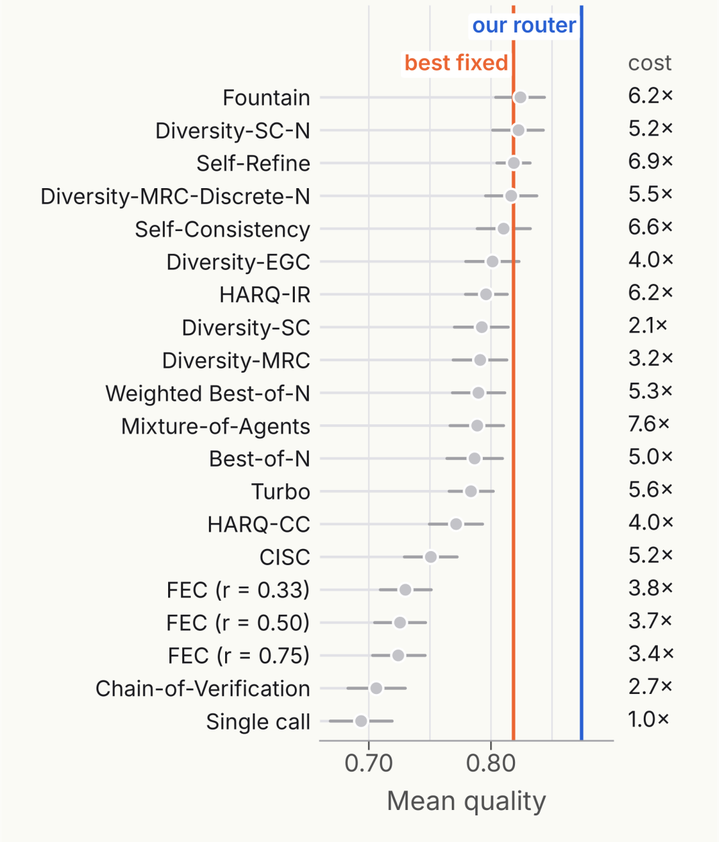
Adaptive coding & modulation
route by difficulty

RouteLLM

Blue: prior methods each operator recovers at a specific hyperparameter setting.

INTELLERCE LLC | AptAgents | AgentCodec
scan for the product

No fixed choice is best for every task



Across six channel configurations

no fixed model, technique and budget dominates.

The quality leader is rarely the efficiency leader

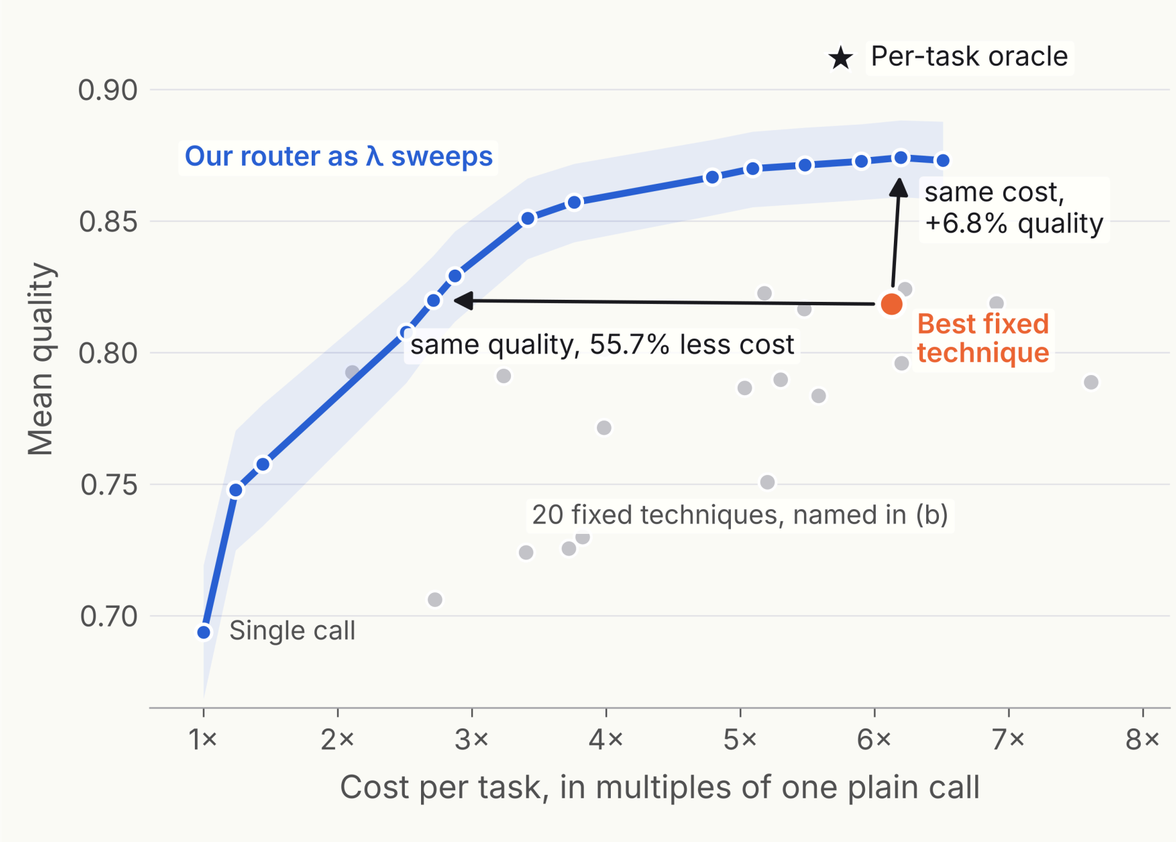
At 3B and 8B, fountain gains most at 6.3 to 9.1× the cost of one call; Diversity-SC is most efficient at about 2×.

The winner changes with the task

Diversity-SC-N leads on GSM8K, fountain on MMLU, HARQ-CC on HumanEval.

Left: 20 fixed techniques with 95% CIs on 300 hard items from MMLU, GSM8K and HumanEval · NeurIPS 2026 · 5

One knob picks an operator for each task



$$\pi^*(x) = \operatorname{argmax}_i \bar{q}_i(x) - \lambda \cdot \bar{\rho}_i(x)$$

For a new task, average each operator's quality and normalized cost over the 20 most similar past tasks, then pick the best trade-off.

One knob λ moves along the whole cost-quality frontier, with no retraining.

INTELLERCE LLC | AptAgents | AgentCodec
scan for the product

Against the best fixed technique

55.7%

lower cost at the same quality

+6.8%

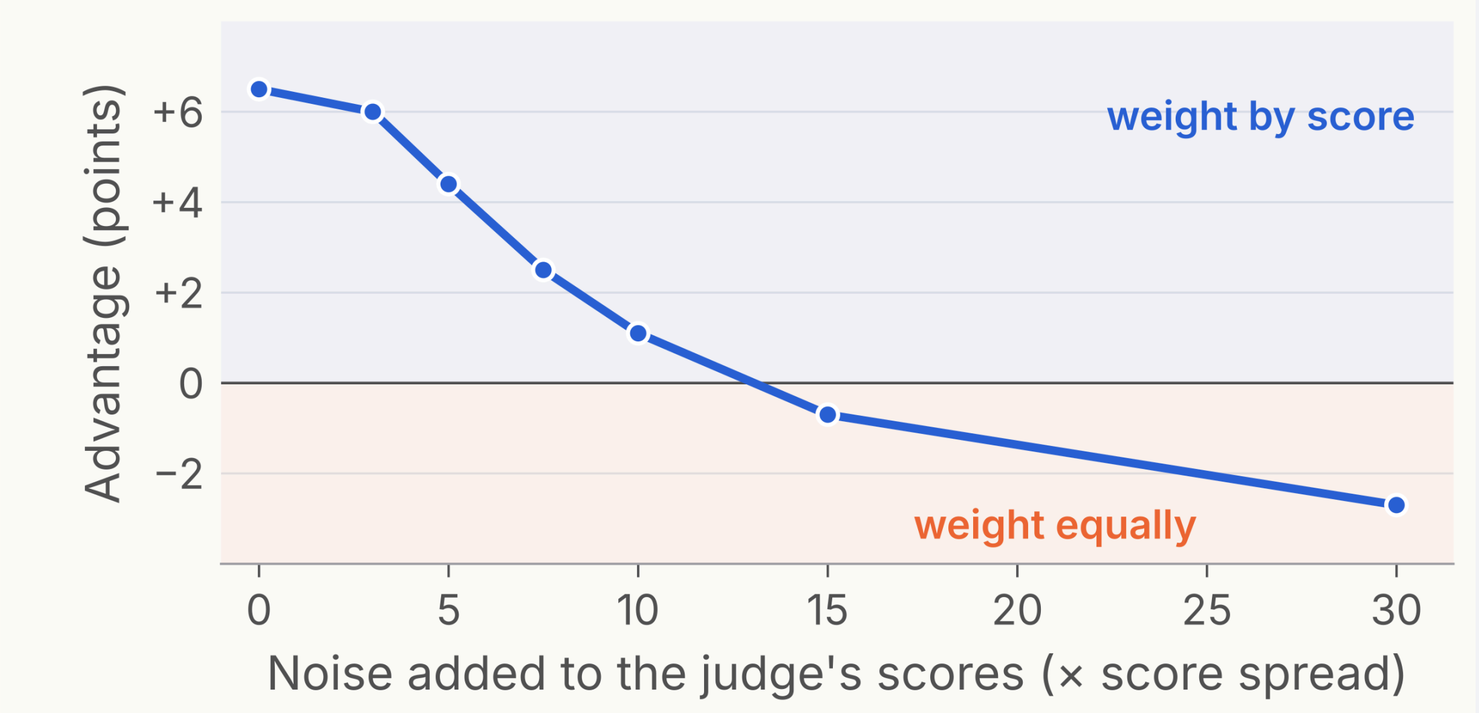
higher quality at the same cost

+26% higher quality than a single LLM call

300 hard items from MMLU, GSM8K and HumanEval; both policies cross-validated.

INTELLERCE LLC | AptAgents | AgentCodec
scan for the product

Rule 1: noisy judge? Weight answers equally

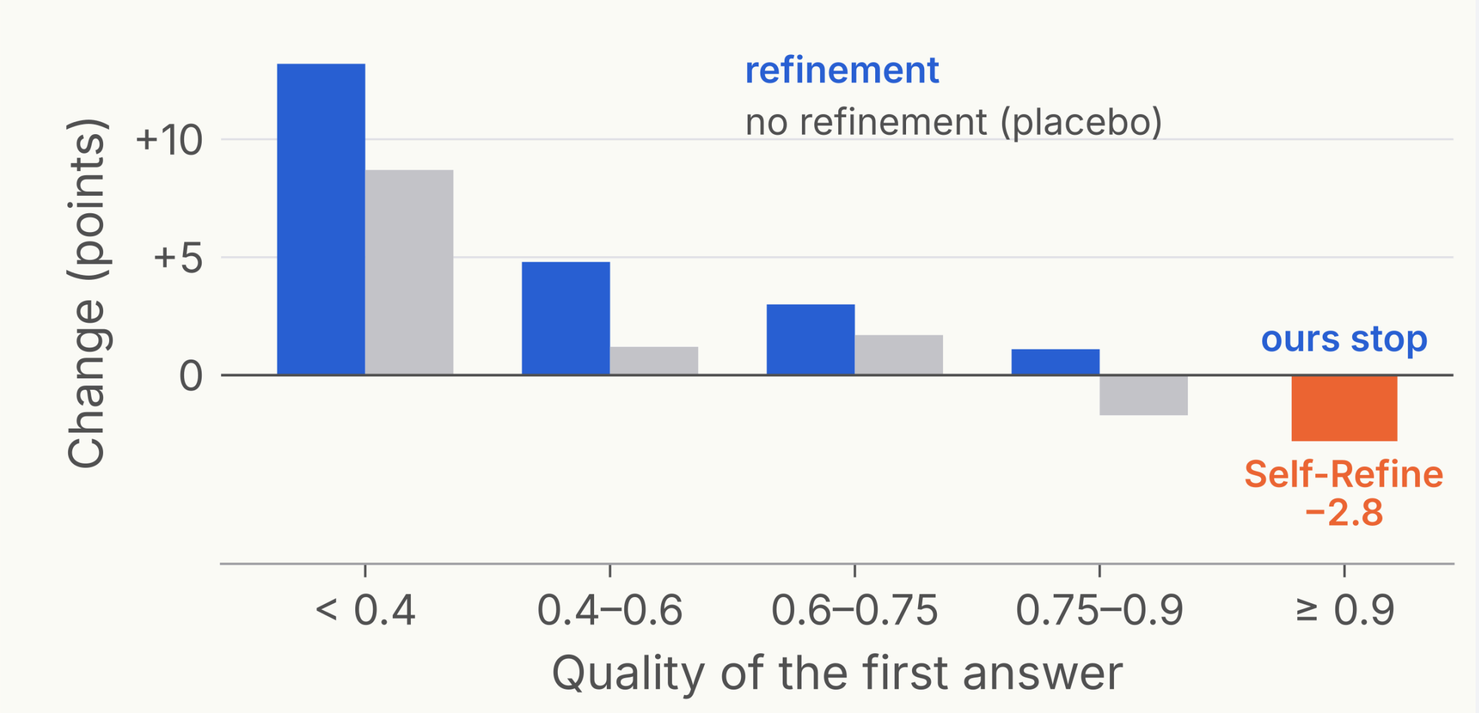


Weighting answers by their scores helps only while the scores are reliable.

Add noise to the scores and the advantage of weighting **crosses zero once**, as a closed-form noise threshold predicts.

INTELLERCE LLC
AptAgents
AgentCodec
scan for the product

Rule 2: refine only weak first answers



Refinement pays most when the first answer is weak.

Ungated Self-Refine **loses 2.8 points** on answers that start near perfect; our operators stop there.

INTELLERCE LLC | AptAgents | AgentCodec
scan for the product

Checked without the LLM judge

6,900
outputs graded with no LLM at all; the judge separates right from wrong (AUC 0.77)

+9.7
points for the router over the best fixed policy, judge-free (λ = 3); it leads at 5 of 6 cost settings and ties at the sixth

16/18
cost settings where the router ties or beats four learned routers, across three configurations

+4.0
points that gated refinement (HARQ-IR) still adds over a single call with reasoning on (p = 0.019)

The margin depends on the candidate pool: restricted to our 11 operators, router and fixed policy are statistically tied.

INTELLERCE LLC | AptAgents | AgentCodec scan for the product

Treat reliability as one tunable layer, designed like a channel code.

A Communication-Theoretic Framework for LLM Agents: Cost-Aware Adaptive Reliability

Hamed Omidvar · Vahideh Akhlaghi

NEURAL INFORMATION PROCESSING SYSTEMS

INTELLERCE LLC

AgentCodec
intellerce.com/products/agentcodec

openreview.net/forum?id=f9YBAJGutx · hamed.omidvar@intellerce.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026)