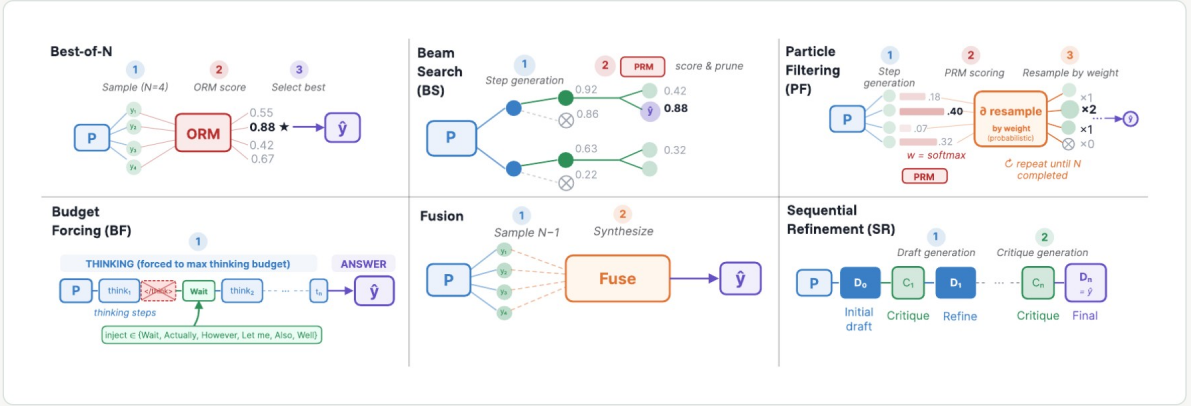
NEURIPS 2026

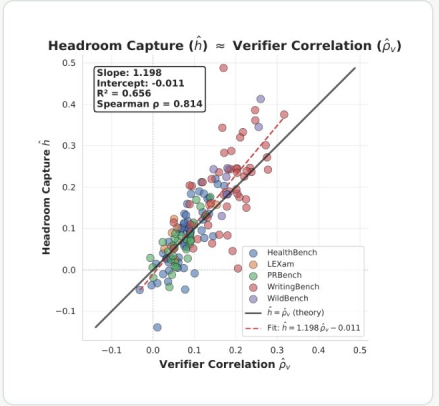
Test-Time Scaling in the Wild

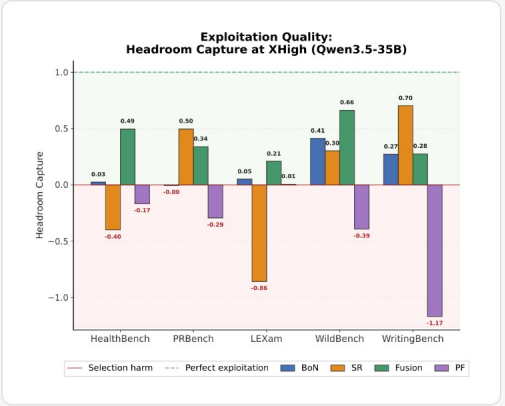
Why Exploitation, Not Exploration, Is the Bottleneck

Davide Romano*, Kanak Raj*, Jerrod Parker, Daniele Giofrè

Thomson Reuters · *equal contribution







TTS works when the verifier works

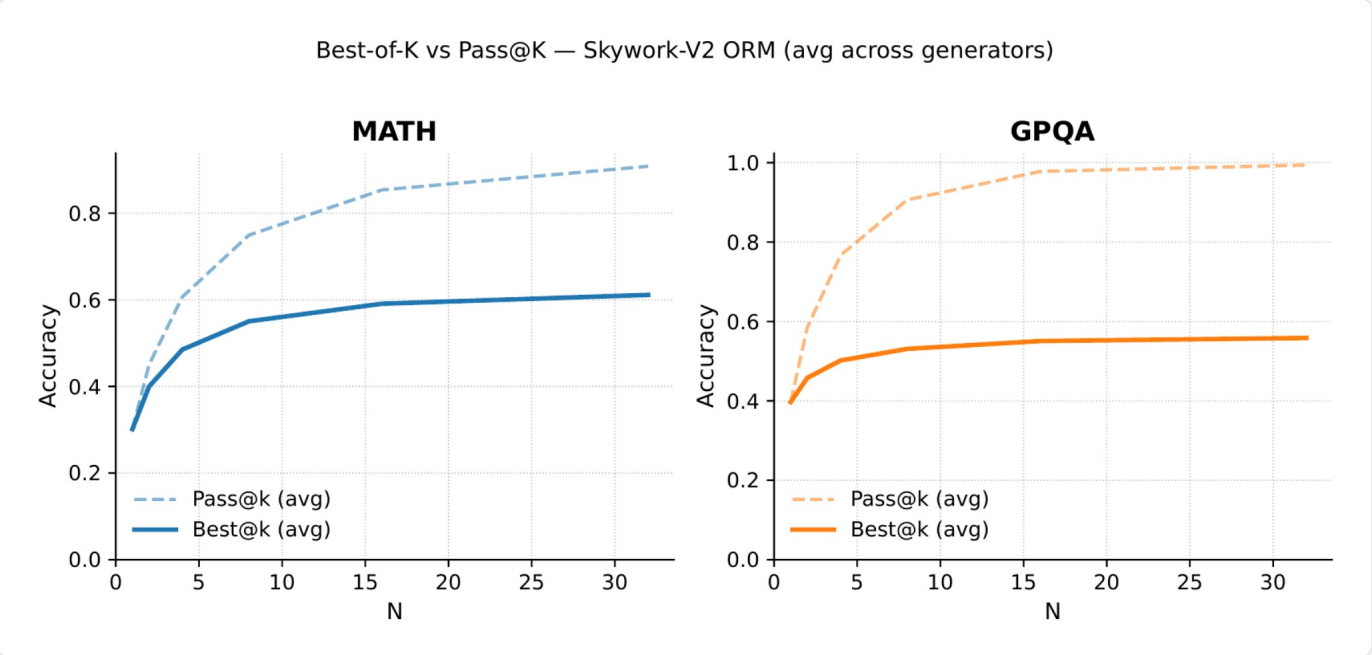


Fig. 10 · Best-of-N with the Skywork-V2 reward model on MATH and GPQA (weaker generator): selected accuracy (solid) rises with N.

- Math, code, multiple choice: binary outcomes and abundant labels.
- Reward models trained on them rank candidates reliably.
- Medicine, law, finance, writing: graded against expert rubrics, no exact match.

At matched compute, how do TTS methods compare on open-ended tasks, and where does their budget pay off: exploration or exploitation?

Framework: exploration vs. exploitation

$$Q(T) - \mu = \underbrace{(Q^*(T) - \mu)}_{\text{exploration headroom}} + \underbrace{(Q(T) - Q^*(T))}_{\text{exploitation effect} \le 0} \tag{1}$$

$$h = (Q(T) - \mu) \,/\, (Q^*(T) - \mu) \tag{2}$$

μ single-sample expected quality

Q* oracle: best candidate in the pool, bias-corrected for judge noise

Q realised: score of the method's actual output

$h = 1$: oracle returned · $h = 0$: no gain over one sample · $h < 0$: compute hurts

Setup: five TTS families, matched compute

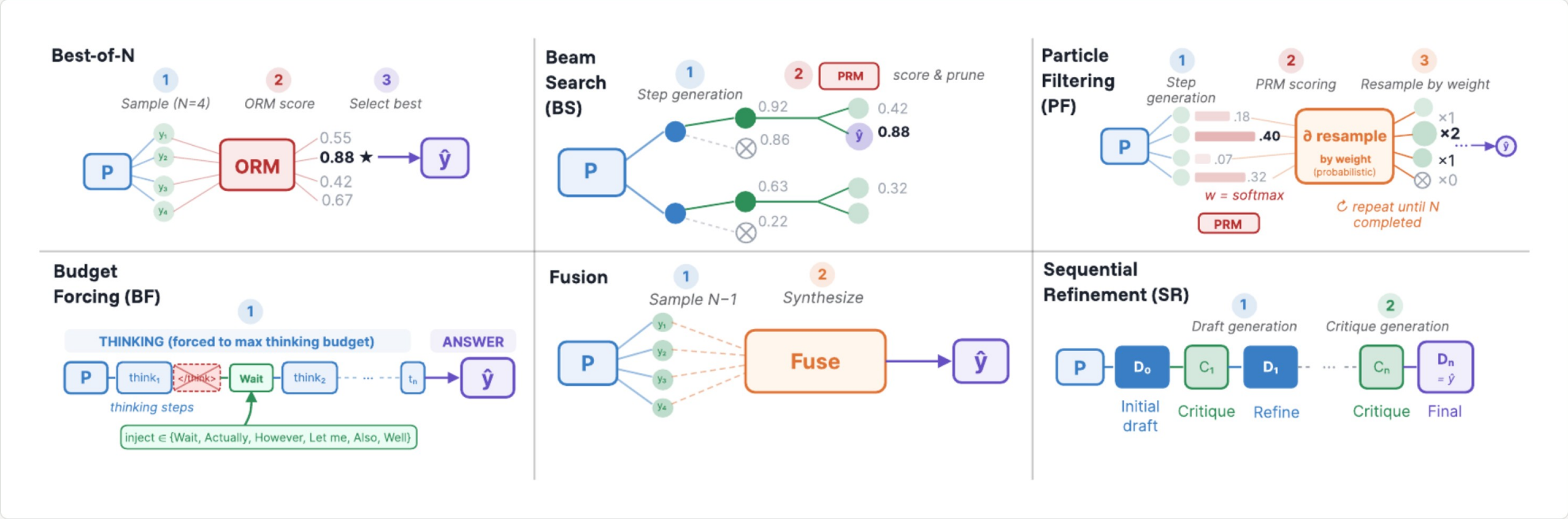


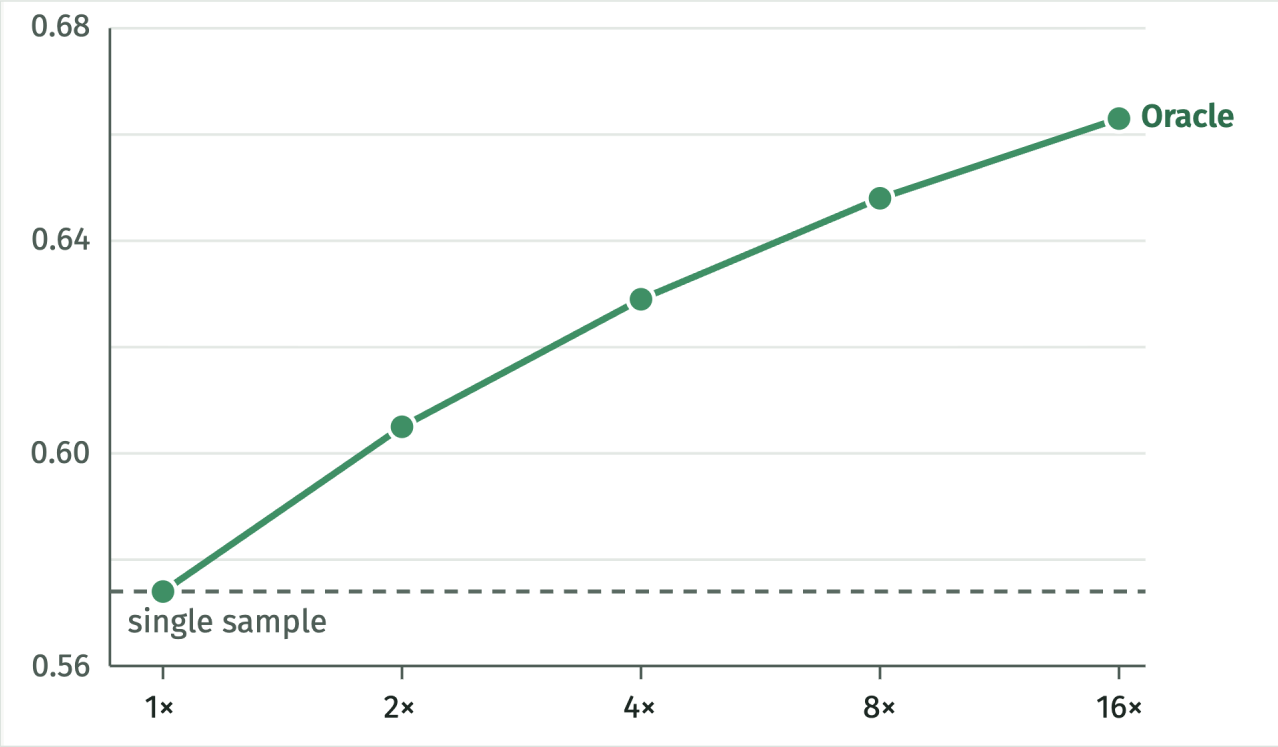
Fig. 1 · Methods compared. Budget = generator output tokens at four levels, matching Best-of-N with N = 2, 4, 8, 16.

Benchmarks HealthBench, LEXam, PRBench, WildBench, WritingBench

Generators Qwen3.5 9B and 35B-A3B, OLMo3 7B and 32B, Gemma-4 check

Judge Qwen3.5-397B, validated against native judges, physicians and GLM5.2

Exploration scales: pools keep improving

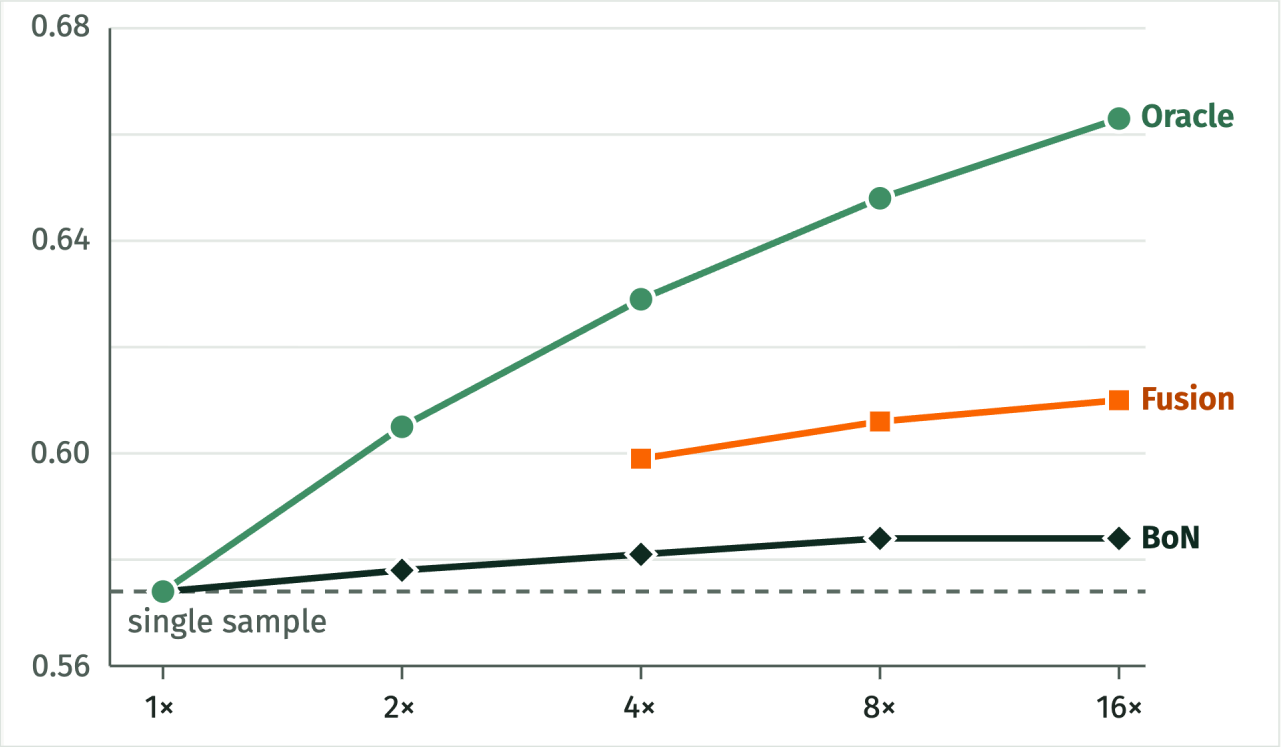


Overall score vs. compute (× single-sample tokens, log scale), Qwen3.5-35B-A3B, bias-corrected oracle. Dashed: single sample.

Overall, 16×	Score	Δ
Single sample	0.574	—
Oracle	0.663	+0.089

The best candidate in the pool improves steadily with compute, as on deterministic tasks.

Exploitation does not: outputs barely move



Overall score vs. compute (× single-sample tokens, log scale), Qwen3.5-35B-A3B, bias-corrected oracle. Dashed: single sample.

Overall, 16×	Score	Δ
Single sample	0.574	—
Oracle	0.663	+0.089
Fusion	0.610	+0.036
Best-of-N (ORM)	0.584	+0.010

Realised quality stagnates for selection; even Fusion leaves most of the headroom in the pool.

The gap holds on every benchmark

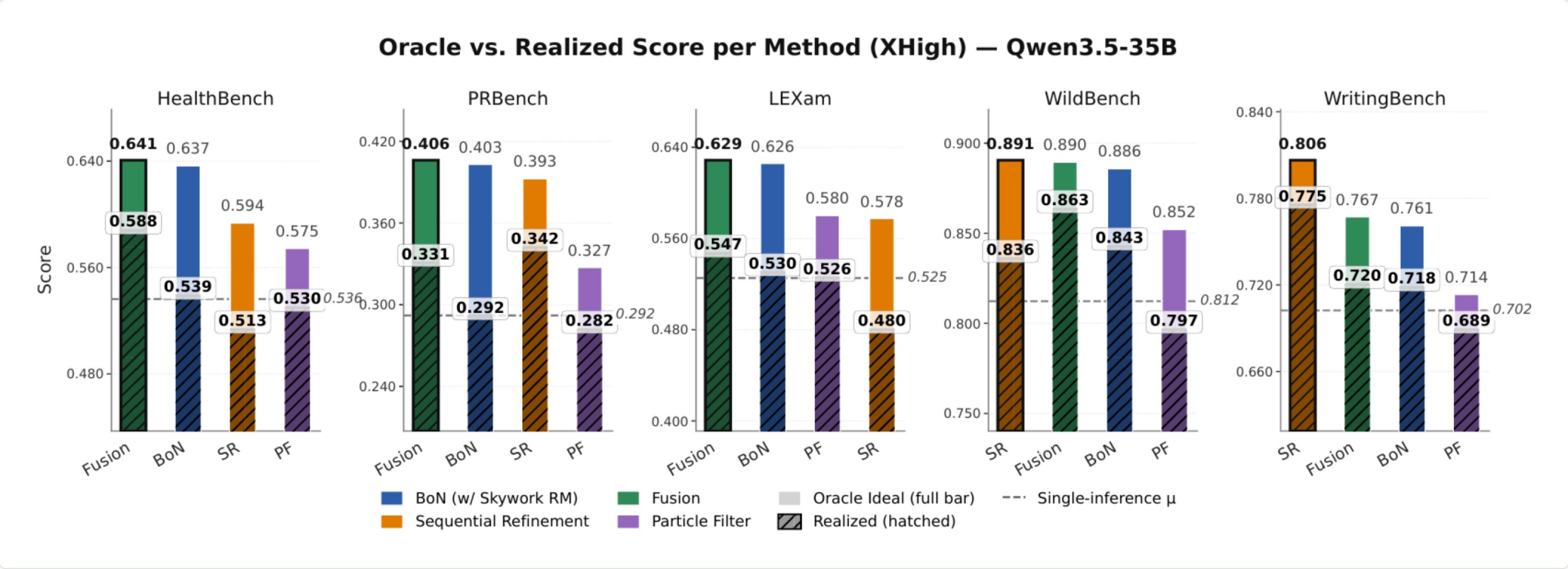


Fig. 2 · Oracle (full bar) vs. realised (hatched) per method at XHigh, Qwen3.5-35B-A3B; dashed line = single-sample μ. Fusion and BoN pools are nearly identical, their outputs are not.

Selection is bounded by verifier correlation

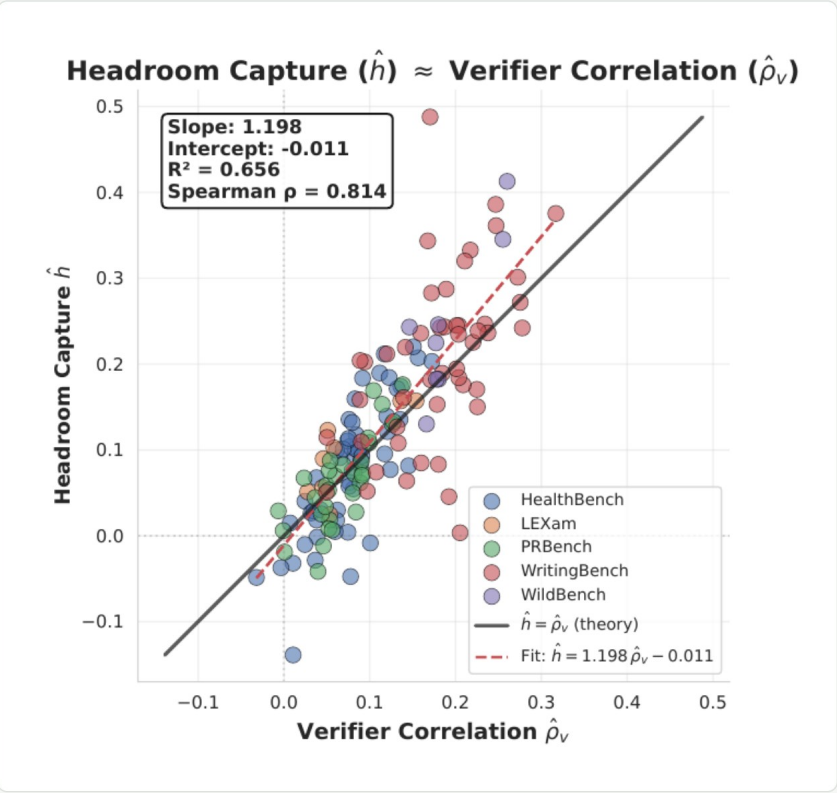


Fig. 3 (left) · 152 (generator, RM, benchmark) points; solid line is the identity h = ρ.

Result (Eq. 4)

$$h \approx \rho = \mathrm{Corr}[\, V(y), V^*(y) \,]$$

Best-of-N headroom capture equals the verifier's correlation with true quality under i.i.d. sampling; judge noise attenuates both sides equally (App. F.2).

Measured

- Mean Spearman ρ between RM and judge scores: 0.12 (Skywork), 0.11 (Llama).
- Fit $h = 1.20\,\rho - 0.01$, $R^2 = 0.66$: the data track the identity.
- So Best-of-N captures ~15% of headroom: near-random selection.

Stronger verifiers and search don’t help

Best-of-16 selectors (Table 5)

Headroom capture, mean of HealthBench, PRBench, LEXam · Qwen3.5-35B-A3B, XHigh

Skywork-V2 8B (ORM)	2%
Llama-3.1 70B (ORM)	7%
Qwen3.5-35B (itself)	1%
DeepSeek-V4-Flash	0%
Fusion (synthesis)	**35%**

No selector exceeds 9% on any of the three; Fusion reaches 21–49%.

Tree search collapses diversity

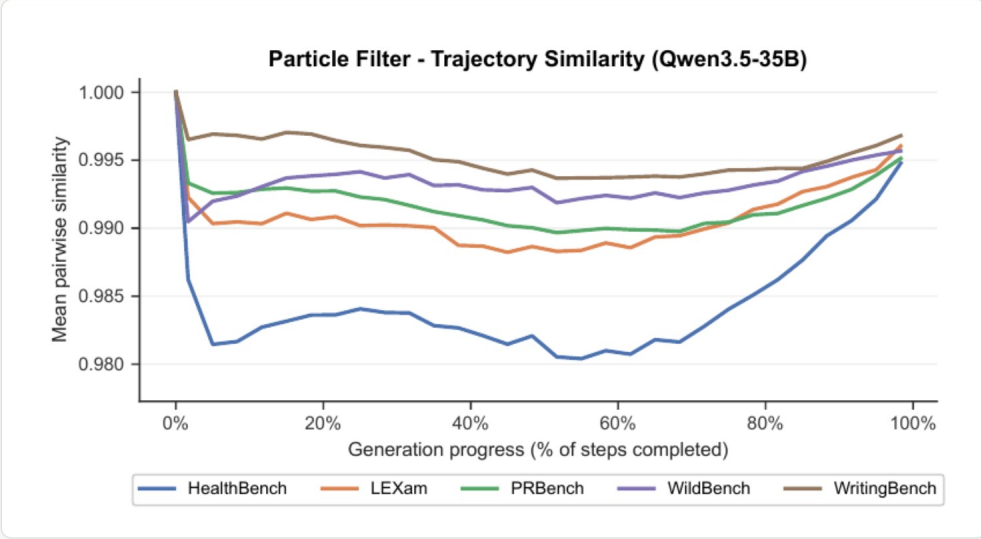


Fig. 4 (right) · Pairwise similarity of 16 particle-filter trajectories stays above 0.98: effectively one candidate.

Fusion helps, but captures only ~40%

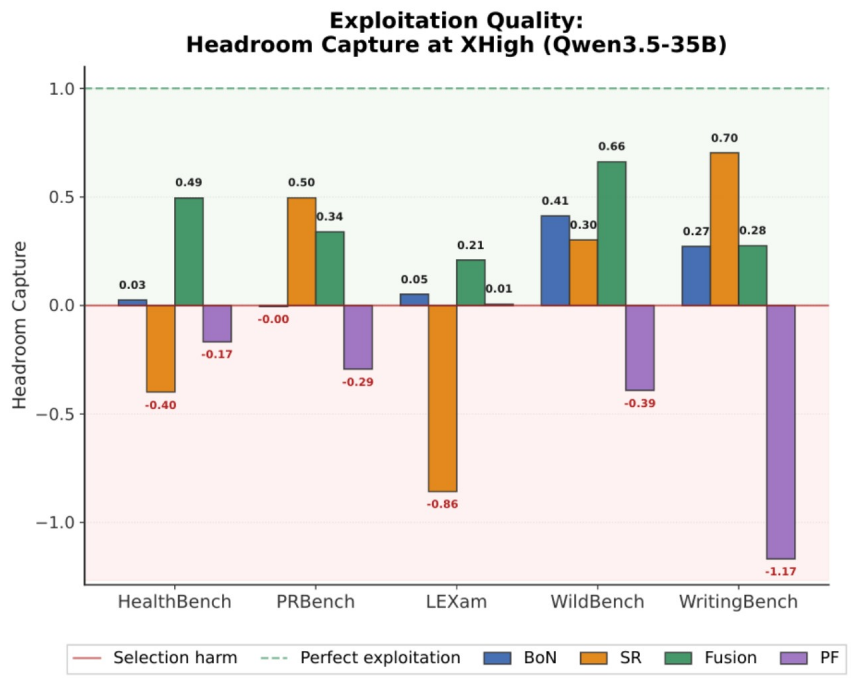


Fig. 3 (right) · Headroom capture at XHigh. Fusion (green) is positive everywhere; SR (orange) swings from −0.86 to +0.70.

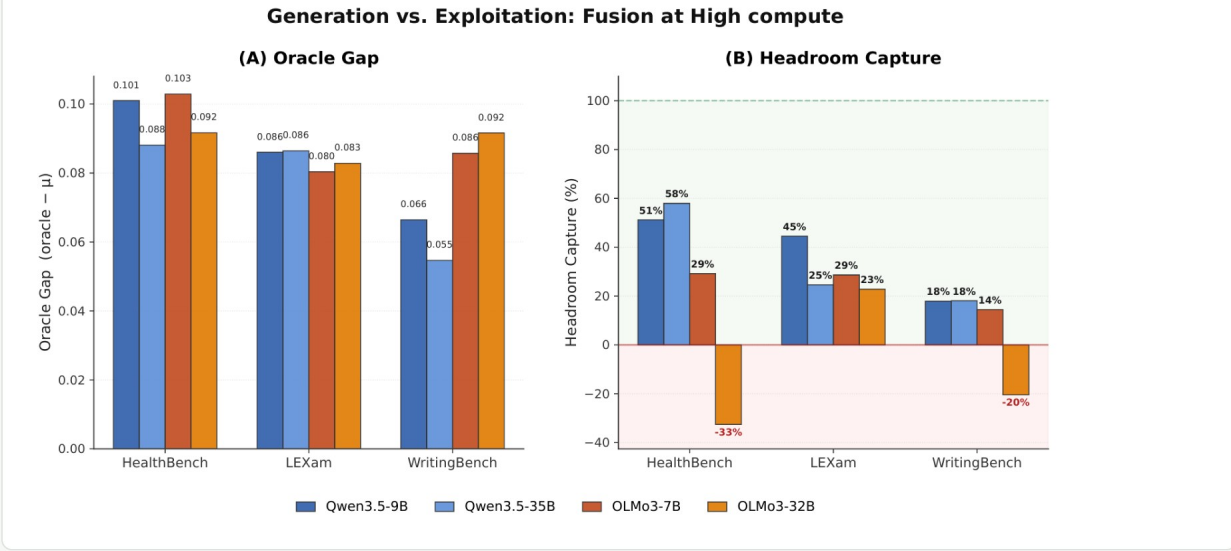


Fig. 19 · Similar oracle gaps across families, very different capture: OLMo3-32B fuses worse than a random candidate.

> **Same model selecting from the same pool (USC) stays within 0.01 of BoN: synthesis is the gain.**

TAKEAWAY

The pool is not the bottleneck. Choosing from it is.

Before spending compute: measure your verifier's correlation on a small sample.

Open problems: verifiers calibrated for rubrics, and exploitation beyond selection and synthesis.

Poster: Sydney, Hall 1–4 · Wed 10:00 AEDT · Session 3