EAGer: Entropy-Aware Generation for Adaptive Inference-Time Scaling

Daniel Scalena[1,2], Leonidas Zotos[2], Elisabetta Fersini[1], Malvina Nissim[2], Ahmet Üstün[3]

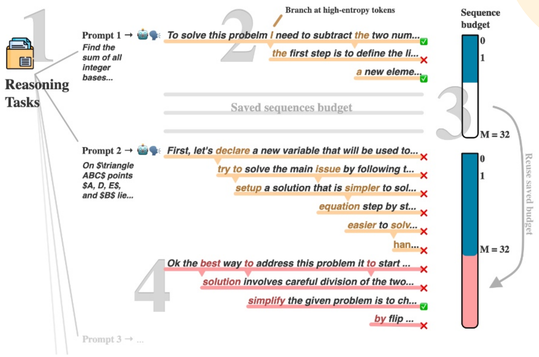
1 UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA 2 university of groningen 3

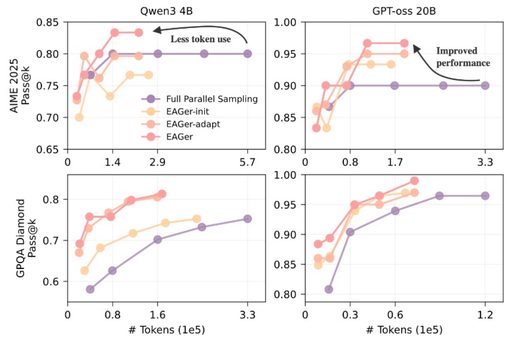
ICML International Conference On Machine Learning

Departure: Redundancy

Test-time scaling samples M parallel reasoning chains per prompt regardless of difficulty, wasting compute on easy problems while starving hard ones of exploration.

- CoT reasoning is dominated by predictable, low-information tokens;
- Parallel sampling re-generates near-identical prefixes across chains;
- Fixed budget M ignores that prompt difficulty varies widely.

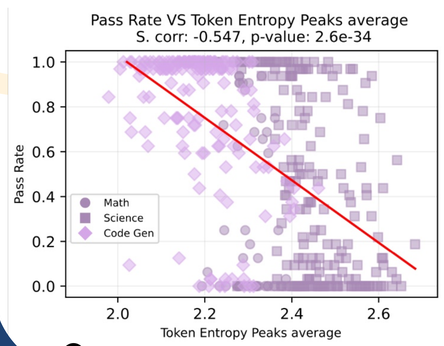




Entropy Peak Lookout

Token-level entropy spikes during generation correlate with lower Pass Rate, making peak entropy a cheap, real-time proxy for problem difficulty.

- Average peak entropy (99.9th percentile) negatively correlates with Pass Rate ($\rho \approx -0.55$, $p<1e\text{-}30$);
- High-entropy moments → genuine decision points worth exploring;
- Low entropy → redundant continuation.



The Branch Switch

Branch a sequence into two candidates only when token entropy exceeds threshold θ.

- Branch = pick top-2 most likely tokens, spawn new candidate continuation;
- Shared prefix tokens are never regenerated → direct token savings.

Token Savings Platform

EAGer-init alone uses less than half the tokens of Full Parallel sampling across all models and benchmarks, before any performance gain is even counted.

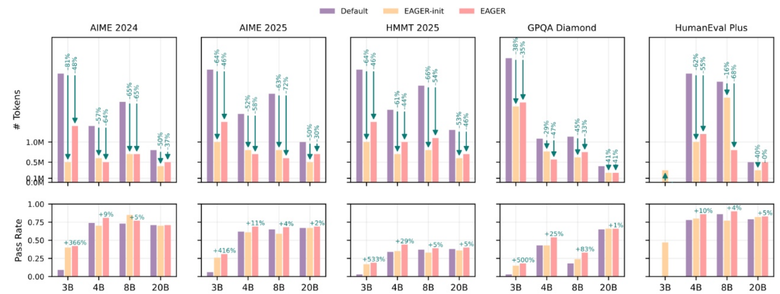


Figure 3. Compute and performance trade-offs of EAGer-init and EAGer. Across all benchmarks and model size, the efficiency of EAGer-init and EAGer consistently outperforms Full Parallel sampling, requiring only half as many tokens in most cases (top). In addition, they achieve higher pass rate accuracy (bottom). For issues specific to the smallest 3B model, see Appendix D.

Benchmark Depot

Evaluated across 4 open models (3B–20B) and 5 reasoning benchmarks spanning math, science, and code.

Reallocation Interchange

Unused budget from easy prompts (capped at 2M) **is redirected to harder prompts** — either saturating prompts (label-free: EAGER-adapt) or prompts with Pass@k=0 (label-available: full EAGER).

- Total token usage still stays below fixed-budget FULL PARALLEL baseline.

Accuracy Terminal

Pass Rate, Pass@k, and Cons@k:
- up to +37% Pass@k with labels,
- +12% without,

while cutting tokens by up to 64%.

Pareto Express

As budget M grows, EAGER shifts the entire Pareto frontier outward.

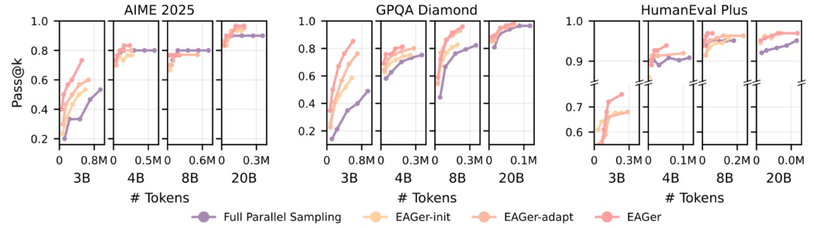


Figure 4. Performance comparison with scaling the total allowed sequences for generating ($M \in \{1, 4, 8, 16, 24, 32\}$). As M increases (line's markers), EAGer consistently improves Pass@k (y-axis) while reducing the number of tokens needed to find the correct solution (x-axis), further shifting the Pareto frontier of the performance–efficiency trade-off.

Terminus

EAGer consistently beats Full Parallel sampling requires no fine-tuning, generalizes across model families/sizes and task domains, and needs only a single threshold θ calibrated from 10 examples.