13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Local Backtracking Mechanisms in Generative Modelling*

Máté Norbert Molnár[a], **Tibor Tajti**[ab]

[a]Eszterházy Károly Catholic University
molnar.mate.norbert@uni-eszterhazy.hu
tajti.tibor@uni-eszterhazy.hu

[b]University of Debrecen, Doctoral School of Informatics

Abstract

Autoregressive models generate text token-by-token in a process where the set of generated tokens typically only grows. This creates an *irreversibility problem*: if a model samples a suboptimal token—such as a hallucinated variable in a Python function—that error becomes a permanent part of the context. While global search methods like Beam Search [4] or Tree of Thoughts [7] address this by exploring multiple candidates, they incur high computational and memory costs. We propose a lightweight, local alternative: a mechanism to detect potential errors immediately after they occur and backtrack by deleting the recent suffix of the sequence. Unlike tree-search methods that fork the generation history, our method maintains a single active candidate sequence but allows for "self-correction" via a learned operator.

To implement this efficiently, we developed a framework integrating with the Hugging Face Transformers ecosystem [5]. To avoid the prohibitive cost of re-generating the prefix, our system performs a synchronized rollback of the token IDs and the Transformer's Key-Value (KV) Cache. When a backtrack is triggered, the KV tensors are cropped to the target position, and the internal state of the decision operator is reverted.

We define a *Backtracking Operator* as a function that evaluates the current generation step and returns an integer $k \geq 0$, representing the number of tokens to discard. We investigated two distinct strategies for this operator. First, we utilized

*This research was supported by the EKÖP-25 University Research Fellowship Program of the Ministry for Culture and Innovation from the Source of the National Research, Development and Innovation Fund.

Heuristic Operators based on output logits, identifying the "Probability Trend" heuristic (triggering a backtrack if probability drops below a moving average) as the most robust static method. Second, to enable adaptation, we trained a *Reinforcement Learning (RL) Agent*. We utilized a Recurrent Neural Network (LSTM) [2] trained via Proximal Policy Optimization (PPO) [3]. To overcome the sparsity of unit test signals, we employed large-scale open-weights (GPT-OSS-20B) teacher model as a judge to provide dense reward feedback on partial generation quality, without access to reference solutions. Crucially, the agent observes a lightweight feature vector (confidence, entropy, token frequency) rather than raw text to minimize computational overhead.

We evaluated our approach on the HumanEval dataset [1] in a zero-shot setting using the base versions of the Qwen2.5 model family [6] (0.5B and 3B, and the 7B model with 8-bit quantization). To ensure statistical significance under stochastic generation (temperature=0.5, top_k=50), the reported results represent the mean Pass@1 accuracy aggregated over 100 independent runs. Variance across runs was low.

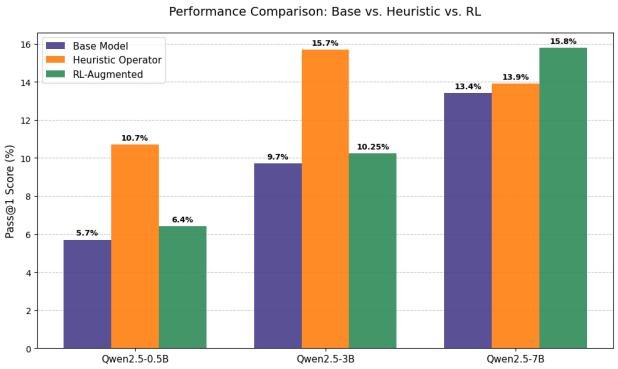

Figure 1. Pass@1 accuracy comparison (Mean of 100 runs): Base Model vs. Backtracking Model. The RL agent demonstrates improved robustness at larger model scales.

Our results (Figure 1) demonstrate a distinct scaling behaviour. On small to mid-sized models (0.5B and 3B), simple heuristics were highly effective, significantly outperforming the RL agent. However, for the larger (7B) model, static heuristics plateaued, rendering the RL agent essential for performance gains, improving Pass@1 accuracy from 13.4% to 15.8%.

We observe that heuristic backtracking outperforms RL control at smaller model scales. We attribute this to increased noise and reduced representational capacity in smaller models, which makes learned credit assignment unstable, while simple confidence- and entropy-based heuristics reliably capture first-order error signals. As model scale increases, RL benefits from richer internal representations and consistently surpasses heuristic control.

This reliability comes with a computational trade-off. We measured a decrease

in generation speed from ≈5.8 tokens/sec (Base) to ≈4.9 tokens/sec (Backtracking). This 15% latency cost is significantly lower than the ≈100% overhead of a standard Beam Search ($N = 2$), making it a practical trade-off for higher reliability in offline coding tasks. By bridging the gap between greedy decoding and expensive global search, this framework offers a robust method for generative self-correction.

References

[1] M. Chen et al.: *Evaluating Large Language Models Trained on Code*, CoRR abs/2107.03374 (2021), arXiv: 2107.03374, URL: https://arxiv.org/abs/2107.03374.

[2] S. Hochreiter, J. Schmidhuber: *Long Short-Term Memory*, Neural Computation 9.8 (1997), pp. 1735–1780, URL: https://www.bioinf.jku.at/publications/older/2604.pdf.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov: *Proximal Policy Optimization Algorithms*, CoRR abs/1707.06347 (2017), arXiv: 1707.06347, URL: http://arxiv.org/abs/1707.06347.

[4] I. Sutskever, O. Vinyals, Q. V. Le: *Sequence to Sequence Learning with Neural Networks*, in: Advances in Neural Information Processing Systems, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger, vol. 27, Curran Associates, Inc., 2014, URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5a18e133cbf9f257297f410bb7eca942-Paper.pdf.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush: *Transformers: State-of-the-Art Natural Language Processing*, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ed. by Q. Liu, D. Schlangen, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45, DOI: 10.18653/v1/2020.emnlp-demos.6, URL: https://aclanthology.org/2020.emnlp-demos.6/.

[6] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu: *Qwen2.5 Technical Report*, 2025, arXiv: 2412.15115 [cs.CL], URL: https://arxiv.org/abs/2412.15115.

[7] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan: *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, in: Advances in Neural Information Processing Systems, ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, vol. 36, Curran Associates, Inc., 2023, pp. 11809–11822, URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.