International Conference on Formal Methods
and Foundations of Artificial Intelligence
Eszterházy Károly Catholic University
Eger, Hungary, June 5–7, 2025

Local backtracking mechanisms in generative modeling

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

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

[b]University of Debrecen

Abstract

Large language models have demonstrated remarkable capabilities in natural language understanding and generation, however, standard decoding is *irreversible*, once x_t is committed it remains in the context for all subsequent steps. This can lead to the propagation of early errors, resulting in wrong outputs such as repetitions, incoherence, divergence from the initial prompt's intent, or generation of low-quality or undesirable content. Our research introduces a conditional backtracking mechanism to retract recent generation steps based on intermediate quality assessments, which could potentially mitigate these issues.

Backtrackable decoding. The core problem addressed is the lack of adaptive control and revision capability within the standard autoregressive LLM generation. The irreversibility of token selection can lock the generation process into suboptimal trajectories. We hypothesize that introducing a *local backtracking operator* into the generation loop which, at selected checkpoints, may remove the last $n \geq 0$ tokens according to a task-specific quality criterion, can improve overall generation quality, coherence and adherence to desired properties. The challenge lies in defining effective criteria for triggering backtracking and determining the extent of revision without incurring excessive computational overhead or destabilizing the generation process. Research focusing on allowing similar operations to large language models has been conducted before, either for reasoning performance improvements [6][7], or for safety reasons [8][3].

Large language models. Let $\mathcal{V}$ denote a finite vocabulary and $\Delta_{|\mathcal{V}|}$ the set of all possible probability vectors over $\mathcal{V}$. Given a pre-trained autoregressive large language model based on the Transformer architecture [4] which computes a vector of logits $M(s) \in \mathbb{R}^{|\mathcal{V}|}$ for an input sequence $s \in \mathcal{V}^*$. Generation proceeds by iterating for $t = 1, 2, \dots$

$$z_t \; = \; M(x_{1:t-1}), \qquad p_t \; = \; softmax(z_t), \qquad x_t \sim p_t$$

where $x_{1:t-1}$ denotes the sequence of tokens generated up to step $t-1$. This produces a sequence $(x_t)_{t \geq 1}$, $x_t \in \mathcal{V}$.

Conditional backtracking. Fix a call period $\kappa \in \mathbb{N}$. At every time step t with $t \bmod \kappa = 0$ we evaluate

$$\delta : \mathbb{R}^{|\mathcal{V}|} \times \Delta_{|\mathcal{V}|} \times \mathcal{V} \times \{0, 1, \dots, |\mathcal{V}| - 1\} \longrightarrow \{0, 1, \dots, t\},$$

$$n_t \; = \; \delta\big(z_t, p_t, x_t, i_t; \Theta\big),$$

where

- δ is the decision function,
- Θ collects hyper-parameters,
- i_t is the index of x_t, $i_t \in \{0, 1, \dots, |\mathcal{V}| - 1\}$

If $n_t > 0$ we delete the suffix $x_{t-n_t+1:t}$ and resume decoding from $x_{1:t-n_t}$. The internal transformer state is correspondingly rewound.[1]
We implemented a variety of decision functions exploring different heuristics:

- Probability-based
- Distribution-based
- Content-based
- Logit-based

Evaluation Protocol We evaluate the models on the HumanEval benchmark [1], which assesses functional correctness for synthesizing programs from docstrings. We used the Qwen models, specifically Qwen-2.5-Coder-7B-Instruct [2] and Qwen2.5-0.5B-Instruct [5]. The primary metric reported is the percentage of correctly synthesized programs (often referred to as Pass@k, where here we used $k = 1$).

Our evaluation encompasses the following experimental setups:

- **Baseline**: We established a baseline performance using standard autoregressive generation for the target LLM on the HumanEval test set.

[1] We maintain past-key tensors for all prefix tokens and truncate them jointly with the textual sequence.

- **Decision Function Comparison**: We conducted comparative evaluations where different implementations of the decision function were employed within our framework, using default hyperparameters for each decision function and a fixed frequency. Performance metrics on HumanEval are compared across different decision functions and against the baseline.
- **Hyperparameter Optimization**: To assess the sensitivity and potential of the framework, we implemented a random search procedure. This search explored the space of decision-function specific parameters and the backtracking frequency to identify configurations that yield the best performance on the HumanEval benchmark for a given implementation.

All experiments were run on an identical hardware setup. Standard generation parameters (e.g., temperature, top-p) were kept constant across all evaluations.

Results. Experimental evaluation on the Qwen2.5-0.5B-Instruct model shows that the proposed framework yields pronounced quality gains. The method achieves a relative Pass@1 improvement of up to 168%, with a mean gain of approximately 62 % over the conventional autoregressive decoding baseline. Crucially, these gains are obtained with negligible computational overhead: generation throughput decreases by only about 1% when backtracking is enabled.

References

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

[2] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, K. Dang, Y. Fan, Y. Zhang, A. Yang, R. Men, F. Huang, B. Zheng, Y. Miao, S. Quan, Y. Feng, X. Ren, X. Ren, J. Zhou, J. Lin: *Qwen2.5-Coder Technical Report*, 2024, arXiv: 2409.12186 [cs.CL], URL: https://arxiv.org/abs/2409.12186.

[3] B. Sel, D. Li, P. Wallis, V. Keshava, M. Jin, S. R. Jonnalagadda: *Backtracking for Safety*, 2025, arXiv: 2503.08919 [cs.CL], URL: https://arxiv.org/abs/2503.08919.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin: *Attention Is All You Need*, 2023, arXiv: 1706.03762 [cs.CL], URL: https://arxiv.org/abs/1706.03762.

[5] A. Yang et al.: *Qwen2 Technical Report*, 2024, arXiv: 2407.10671 [cs.CL], URL: https://arxiv.org/abs/2407.10671.

[6] X.-W. Yang, X.-Y. Zhu, W.-D. Wei, D.-C. Zhang, J.-J. Shao, Z. Zhou, L.-Z. Guo, Y.-F. Li: *Step Back to Leap Forward: Self-Backtracking for Boosting Reasoning of Language Models*, 2025, arXiv: 2502.04404 [cs.CL], URL: https://arxiv.org/abs/2502.04404.

[7] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan: *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, 2023, arXiv: 2305.10601 [cs.CL], URL: https://arxiv.org/abs/2305.10601.

[8] Y. Zhang, J. Chi, H. Nguyen, K. Upasani, D. M. Bikel, J. Weston, E. M. Smith: *Backtracking Improves Generation Safety*, 2024, arXiv: 2409.14586 [cs.LG], URL: https://arxiv.org/abs/2409.14586.