

# Lokális visszalépési mechanizmusok a generatív modellezésben

Molnár Máté Norbert  
PTI BSc

Témavezető:  
Dr. Tajti Tibor

2025

A nagy nyelvi modellek figyelemre méltó képességeket mutattak a természetes nyelv megértésében, azonban a szabványos dekódolás *visszafordíthatatlan*. Ez a korai hibák terjedéséhez, majd helytelen kimenetekhez vezethet. Kutatásunk egy feltételes visszalépési operátort vezet be a közelmúltbeli generálási lépések visszavonására közbenső minőségi értékek alapján.

## Definíciók

Legyen  $M$  egy előre betanított nagy nyelvi modell a Transformer architektúrán alapulva[2], amely tokenekből álló szókészlet  $\mathcal{V}$ -nek egy véges sorozatára visszaadja a logitok vektorát.  $M$  generálási folyamata  $t = 1, 2, \dots$  lépésekben:

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

ahol  $x_{1:t}$  a  $t$ . lépésig generált elemek sorozata.

## Probléma

- Ha  $x_t$  egyszer rögzítésre kerül, az minden további lépésnél a kontextusban marad.
- Korai hibák elronthatják a generálási folyamatot.

## Megoldás

Egy *feltételes visszalépési operátor* bevezetése, amely megadott lépésekben eltávolíthatja az utolsó  $n$  token, egy kritériumnak megfelelően. Hipotézisünk, hogy ez javíthatja a generálás minőségét.

# Feltételes visszalépés

Rögzítsünk egy ellenőrzési pont időszakot  $\kappa \in \mathbb{N}$ . Minden  $t$  lépésnél, ahol  $t \bmod \kappa = 0$  kiértékeljük a következő értéket:

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

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

ahol

- $\delta$  a visszalépés feltétele,
- $\Delta_{|\mathcal{V}|}$  az összes lehetséges valószínűségi vektor halmaza  $\mathcal{V}$  felett,
- $i_t$   $x_t$  indexe  $\mathcal{V}$ -ben,  $i_t \in \{0, 1, \dots, |\mathcal{V}| - 1\}$ ,
- $\Theta$  pedig összegyűjti a hiperparamétereket.

# Feltételes visszalépés

Ha  $n_t > 0$ , töröljük  $x_{t-n_t+1:t}$  útótagot, és folytatjuk a generálást  $x_{1:t-n_t}$ -ről. A belső Transformer állapotát ennek megfelelően szintén visszatekerjük<sup>1</sup>.

---

<sup>1</sup>Az összes tokenhez múltbeli kulcs- és érték-tenzorokat tartunk fenn, ezeket a szöveges szekvenciával együtt csonkítjuk.

Többfajta feltételes visszalépési operátort is implementáltunk:

- Valószínűség küszöbérték
- Entrópia küszöbérték
- Valószínűség határ
- Valószínűség csökkenés
- Valószínűség trend
- Ismétlés
- N-gram átfedés
- Logit küszöbérték

## Értékelés módja

- HumanEval benchmark[1],
- Qwen2.5-0.5B-Instruct[3],
- elsődleges mérőszám a helyesen generált programok százalékos aránya,
- minden kísérlet esetén a standard generálási paramétereket állandó értéken tartottuk.



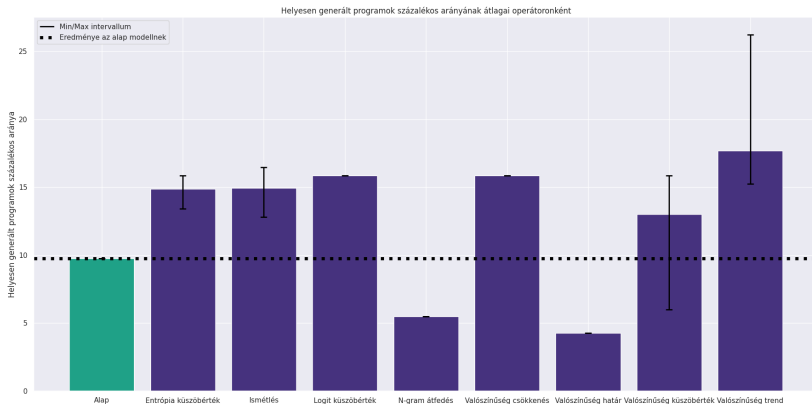
## HumanEval benchmark

```
def truncate_number(number: float) -> float:
    """ Given a positive floating point number, it can be decomposed into
    and integer part (largest integer smaller than given number) and decimals
    (leftover part always smaller than 1).

    Return the decimal part of the number.
    >>> truncate_number(3.5)
    0.5
    """
    ----
    return number - int(number)
```

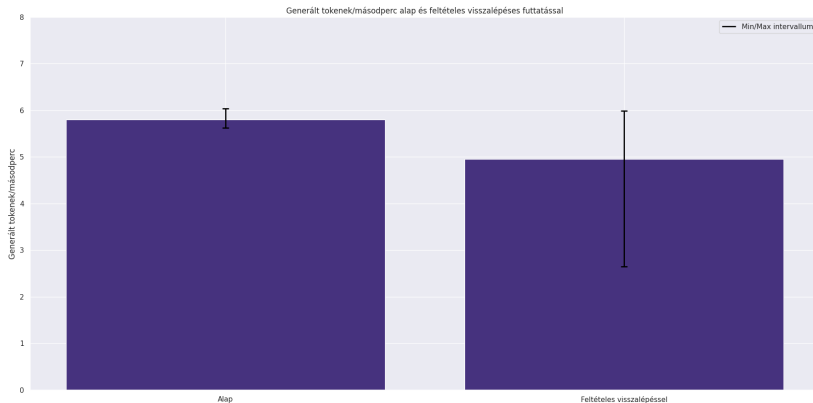
1. ábra. Egy eleme a benchmarknak

## Helyesen generált programok százalékos aránya



**2. ábra.** Helyesen generált programok százalékos arányának átlagai operátoronként

## Generált token/másodperc



**3. ábra.** Generált token/másodperc összehasonlítása standard és visszalépéses futtatással

- Nagyobb modellen kiértékelés.
- Optimális értékek megkeresése hiperparaméterekhez.
- Neurális háló felépítése és betanítása a visszalépési operátorra.

- [1] Mark Chen és tsai. *Evaluating Large Language Models Trained on Code*. 2021. arXiv: 2107.03374 [cs.LG]. URL: <https://arxiv.org/abs/2107.03374>.
- [2] Ashish Vaswani és tsai. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [3] An Yang és tsai. *Qwen2 Technical Report*. 2024. arXiv: 2407.10671 [cs.CL]. URL: <https://arxiv.org/abs/2407.10671>.

**Köszönöm a figyelmet!**