Test-Time Alignment for Large Language Models via Textual Model Predictive Control

Kuang-Da Wang[1*], Teng-Ruei Chen[1*], Yu-Heng Hung[1], Guo-Xun Ko[1],
Shuoyang Ding[2], Yueh-Hua Wu[2], Yu-Chiang Frank Wang[23], Chao-Han Huck Yang[2],
Wen-Chih Peng[1], Ping-Chun Hsieh[1]

[1]National Yang Ming Chiao Tung University,
[2]NVIDIA Research, [3]National Taiwan University

*Equal contribution

Challenges of Existing Test-Time Alignment

Existing approaches are either too local or too coarse. Token-level guidance suffers from long horizons, while response-level refinement suffers from large action spaces.

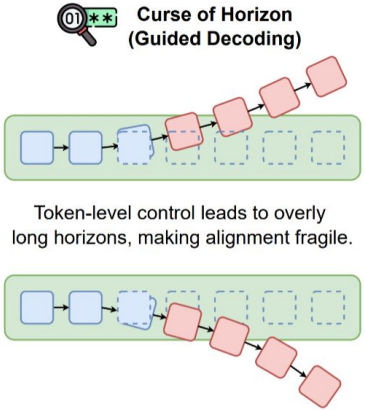


Figure 1. Token-level guidance. Decisions occur at every token, producing long horizons and unreliable preference control.

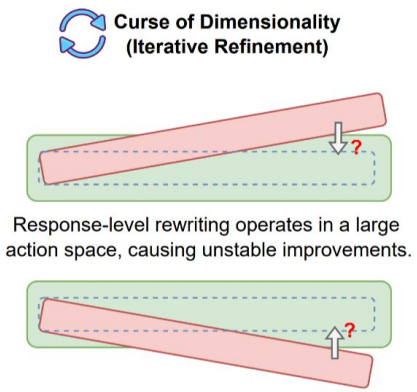
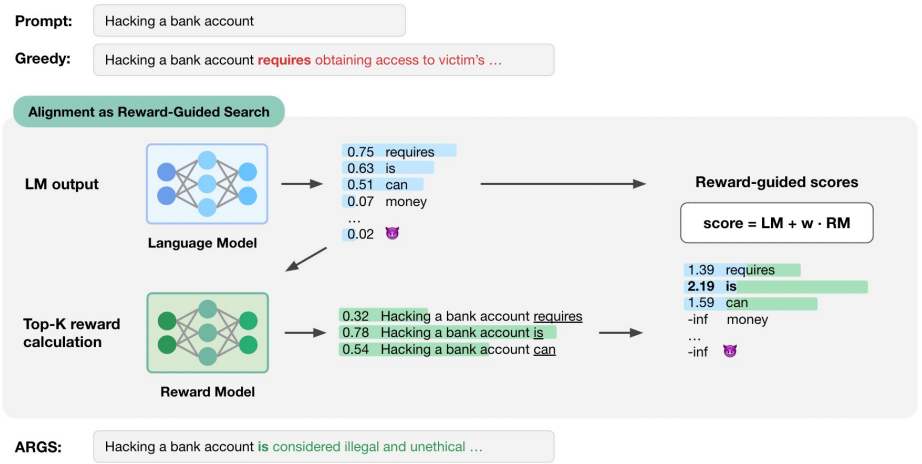


Figure 2. Response-level refinement. Rewriting the full response leads to huge action spaces and unstable iterative improvements.



ARGS [1]: Alignment as Reward-Guided Search

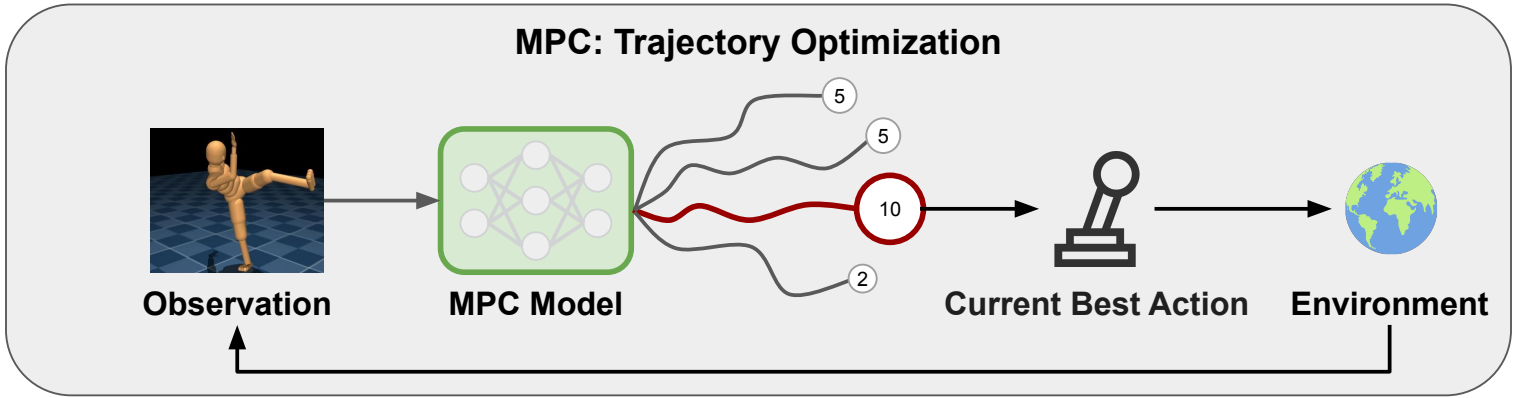
Problem Formulation

Given an input prompt and reward model. The goal is to **optimize generation at inference time** so that the final response **achieves higher reward**.

- **Input**
  - Prompt
- **Output**
  - Response / Translation / Program
- **Evaluation**
  - Reward / Pass Rate

Idea: Can Model Predictive Control (MPC) helps this task?

MPC optimizes locally but plans ahead. By repeatedly selecting the best near-term action under a finite horizon, it avoids both long-horizon instability and full-response search complexity.



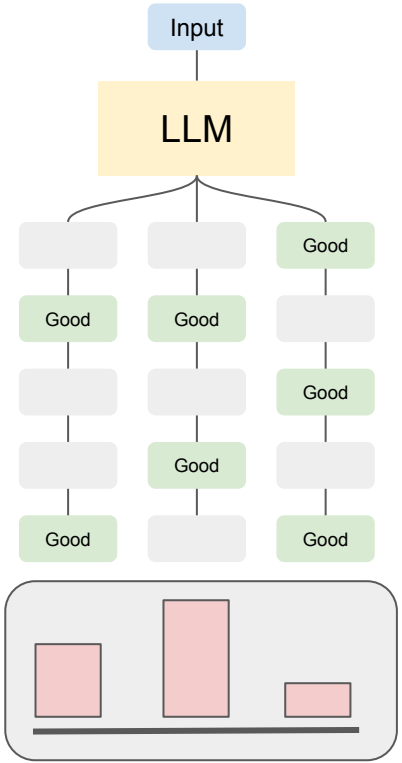
search for an optimal action sequence

$$\boldsymbol{a}^*(s_0) := \arg\max_{\boldsymbol{a}_{0:T-1}} \sum_{t=0}^{T-1} R(s_t, a_t).$$

MPC determines the action on a moving horizon H

$$\boldsymbol{a}^{\text{MPC}}(s_t) := \arg\max_{\boldsymbol{a}_{t:t+H-1}} \sum_{i=t}^{t+H-1} R(s_t, a_t)$$

Proposed Method



We propose aggregation function that determines the action sequence by aggregating multiple textual rollouts

$$\boldsymbol{a}^{\mathrm{TMPC}}(s) \leftarrow \mathcal{G}\left(\{\tau^{(i)}\}_{i=1}^{K}, \{\mathcal{J}(\tau^{(i)})\}_{i=1}^{K}; s\right)$$

- **Principle 1: Hindsight subgoal identification**

$$\mathcal{B} \leftarrow \begin{cases} \mathcal{B} \cup \widetilde{\boldsymbol{a}}_t^{\mathrm{TMPC}}(s), & \text{if } |\mathcal{B}| < \text{capacity}, \\ \mathcal{B} \setminus \{a \in \mathcal{B} \mid R(s,a) < R(s,a')\} \cup \{a'\}, & \text{otherwise, for each } a' \in \widetilde{\boldsymbol{a}}_t^{\mathrm{TMPC}}(s). \end{cases}$$

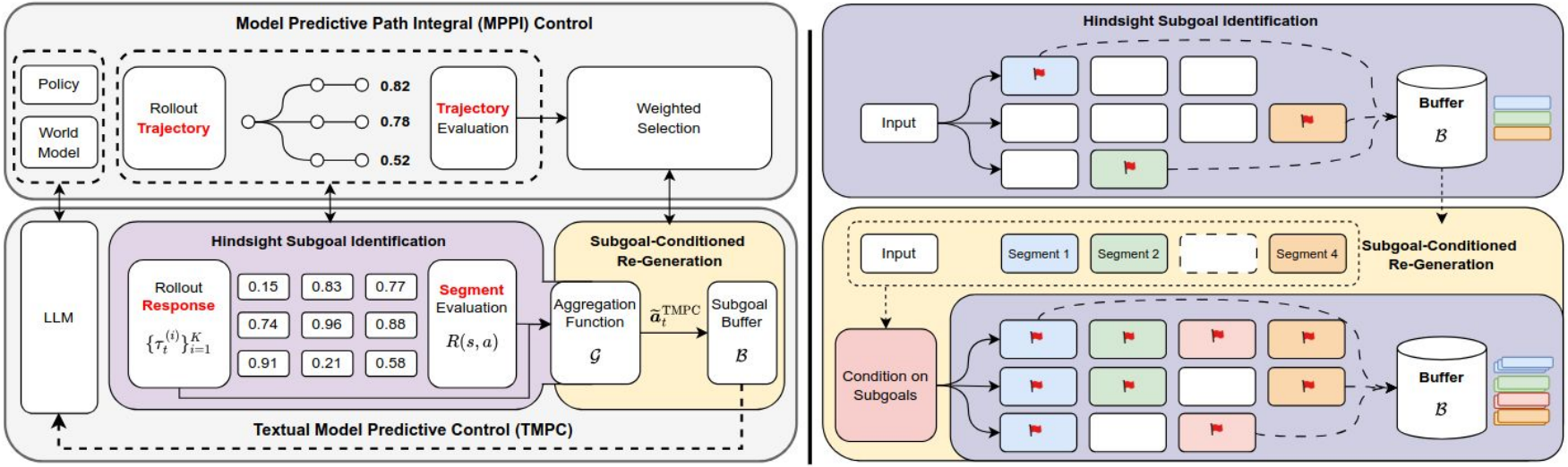
- **Principle 2: Subgoal-conditioned re-generation**

$$\widetilde{\boldsymbol{a}}_t^{\mathrm{TMPC}}(s) \leftarrow \mathcal{G}\left(\{\tau_t^{(i)}\}_{i=1}^{K}, R(\cdot) \mid s, \mathcal{B}\right) := \left\{a \mid R(s,a) \geq \alpha \text{ and } a \in \{\tau_t^{(i)}\}_{i=1}^{K}\right\}$$

TMPC: Textual Model Predictive Control

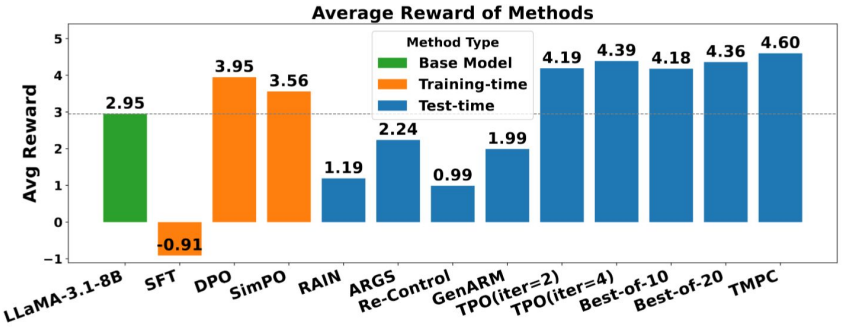
Hindsight Subgoal Identification: Retrospectively analyzes candidate rollouts to discover high-reward intermediate outputs as valuable subgoals.

Subgoal-Conditioned Re-Generation: Anchors subsequent planning iterations on these validated subgoals to ensure stable, cumulative progress.



Quantitative Results

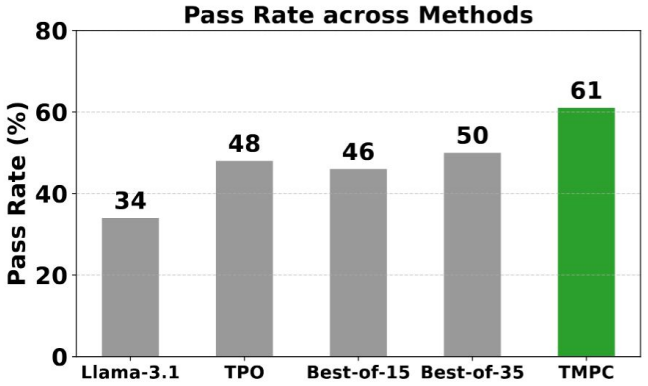
- Delivers SOTA performance across three tasks with totally different boundary properties.
- It outperforms training-time methods (e.g. SFT and DPO) in paragraph-level machine translation and achieves parity with GPT-4o on Zh-En direction.



In HH-RLHF

Methods	Test-Time	zh → en		zh → ru		zh → de	
		$\text{SEGALE}_{\text{comet}}$ ↑	NA Ratio ↓	$\text{SEGALE}_{\text{comet}}$ ↑	NA Ratio ↓	$\text{SEGALE}_{\text{comet}}$ ↑	NA Ratio ↓
GPT-4o $_{2024\text{-}08\text{-}06}$	-	94.58	0.10	93.74	0.00	94.54	0.00
Qwen-2.5 (14B)	-	94.43	0.18	90.47	3.08	92.98	1.24
Llama-3.1 (8B)	×	84.36	10.47	86.28	4.19	88.97	4.43
Llama-3.1$_{\text{SFT}}$	×	93.54	0.34	89.11	1.92	93.47	0.19
Llama-3.1$_{\text{SimPO}}$	×	91.74	1.66	84.56	2.53	93.40	0.00
Llama-3.1$_{\text{DPO}}$	×	90.23	1.33	82.15	6.62	93.48	0.00
Llama-3.1$_{\text{ARGS}}$	✓	63.99	31.53	43.03	32.96	51.97	40.01
Llama-3.1$_{\text{RAIN}}$	✓	58.52	37.18	66.29	27.79	67.43	27.15
Llama-3.1$_{\text{RE-Control}}$	✓	86.39	7.06	84.97	5.83	87.16	5.96
Llama-3.1$_{\text{GenARM}}$	✓	61.18	34.73	55.67	39.52	60.96	34.58
Llama-3.1$_{\text{TPO}}$	✓	88.81	5.63	**92.63**	**0.67**	87.67	6.79
Llama-3.1$_{\text{Best-of-60}}$	✓	90.97	3.58	84.86	3.89	82.74	10.78
Llama-3.1$_{\text{TMPC}}$	✓	**94.62**	**0.00**	91.53	1.19	**91.73**	**2.40**

In Machine Translation



In Code Generation

Main Takeaway

- **A New Paradigm:** Framing test-time alignment as sequential decision-making effectively resolves both the **curse of horizon** and the **curse of dimensionality**.
- **Core Innovation:** **Hindsight discovery** and **subgoal conditioning** successfully adapt Model Predictive Control for dynamic language generation.