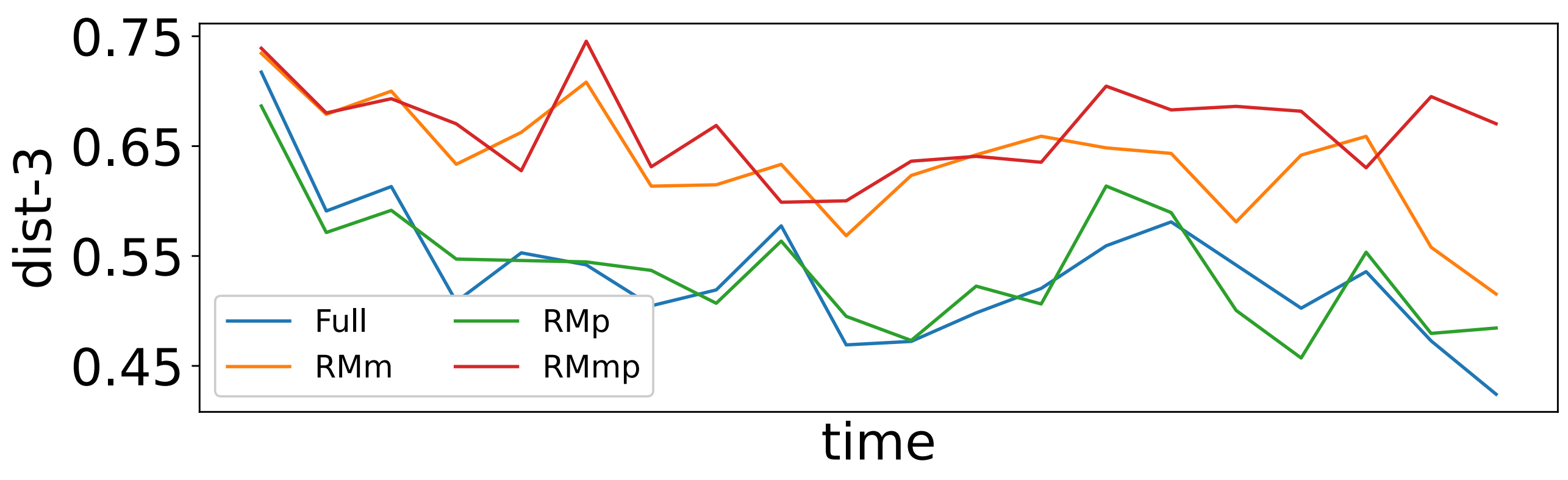
Exploring and Controlling Diversity in LLM-Agent Conversation

KuanChao Chu, Yi-Pei Chen, Hideki Nakayama
The University of Tokyo, Japan

Motivation

- **Ensuring interactions align with simulation objectives by controlling LLM agents dialogue diversity**. For example, maintaining story consistency for main NPCs with varied experiences for environmental NPCs
- **Decline in dialogue diversity in multi-agent simulations over time**, emphasizing the importance of controlling and enhancing diversity to prolong simulations



Data, Model, and Task

Diversity: The variation between dialogues generated under identical initial conditions across trials

Models: LLaMA 3 and 3.1 (8B-Inst)

Metrics: *sim* (dial. embedding) and *dist-n* (n gram)

Data: GA [1] and HA [2]

Block	Item	Word	Type
Basic Info	5	71.5	Fixed
Human Needs*	2~6	20.4	Fixed in dial.
Memory	30~45	1318.8	Trajectory
Previous Dialogues	1~3	327.4	Trajectory
Environment	2	69.5	Context
Current Dialogue	1	284.3	Context

Blocks in utterance generation prompt

[1] Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. “Generative agents: Interactive simulacra of human behavior.” UIST 2023.
[2] Wang ZL, Chiu YY, Chiu YC. “Humanoid agents: Platform for simulating human-like generative agents.” EMNLP 2023: System Demonstration.

	sim (↓)	dist-1	dist-2	dist-3 (↑)
Full	0.791	0.095	0.350	0.535
RMb	0.806	0.091	0.335	0.513
RMm	0.736	0.119	0.429	0.636
RMp	0.802	0.095	0.352	0.538
RMe	0.764	0.091	0.326	0.497
RMbmpe	0.511	0.202	0.610	0.800

Remove (RM) block x: these information collectively plays a constraining role

Adaptive Prompt Pruning (APP)

Calculate an attention score for each item

⬇

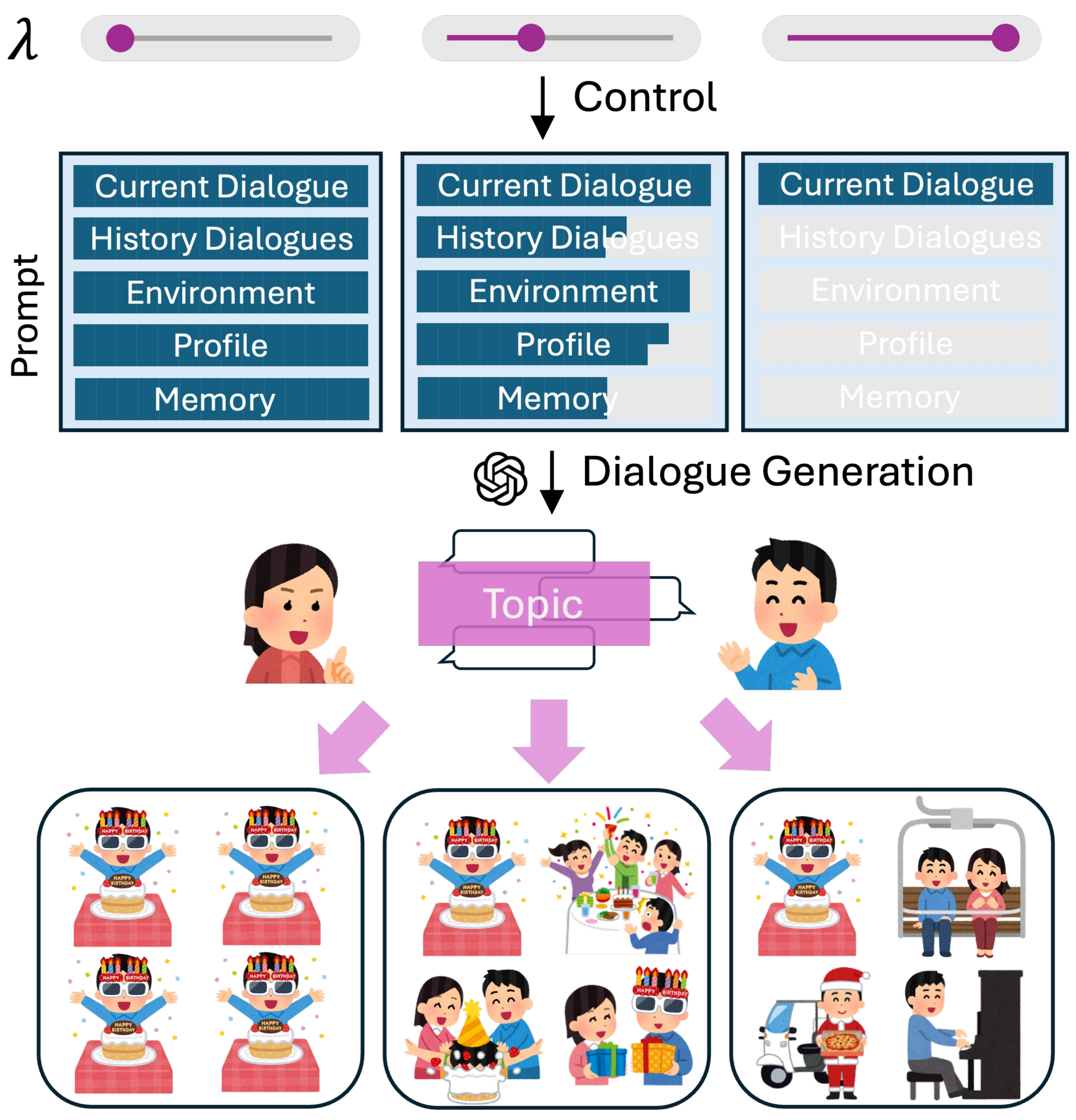
Sort the items by the attention scores in descending order

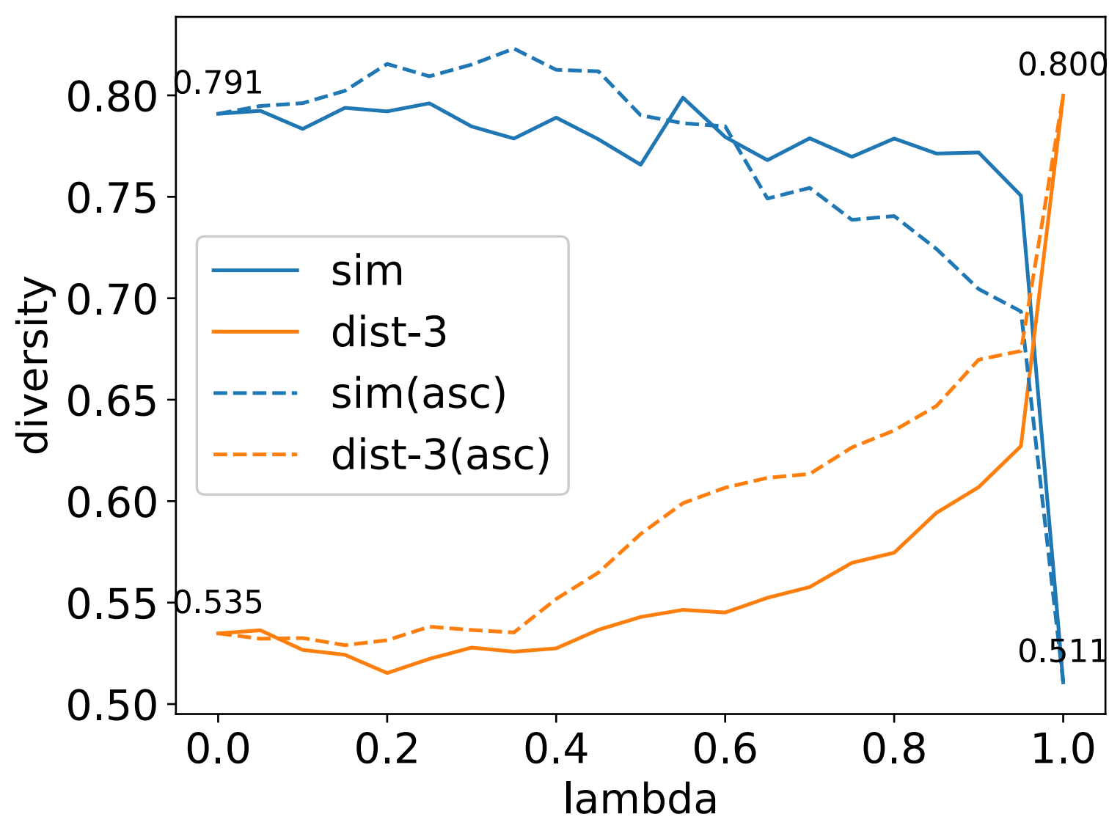
⬇

$\lambda \in [0,1]$ determines the items to remove (cumulative score)

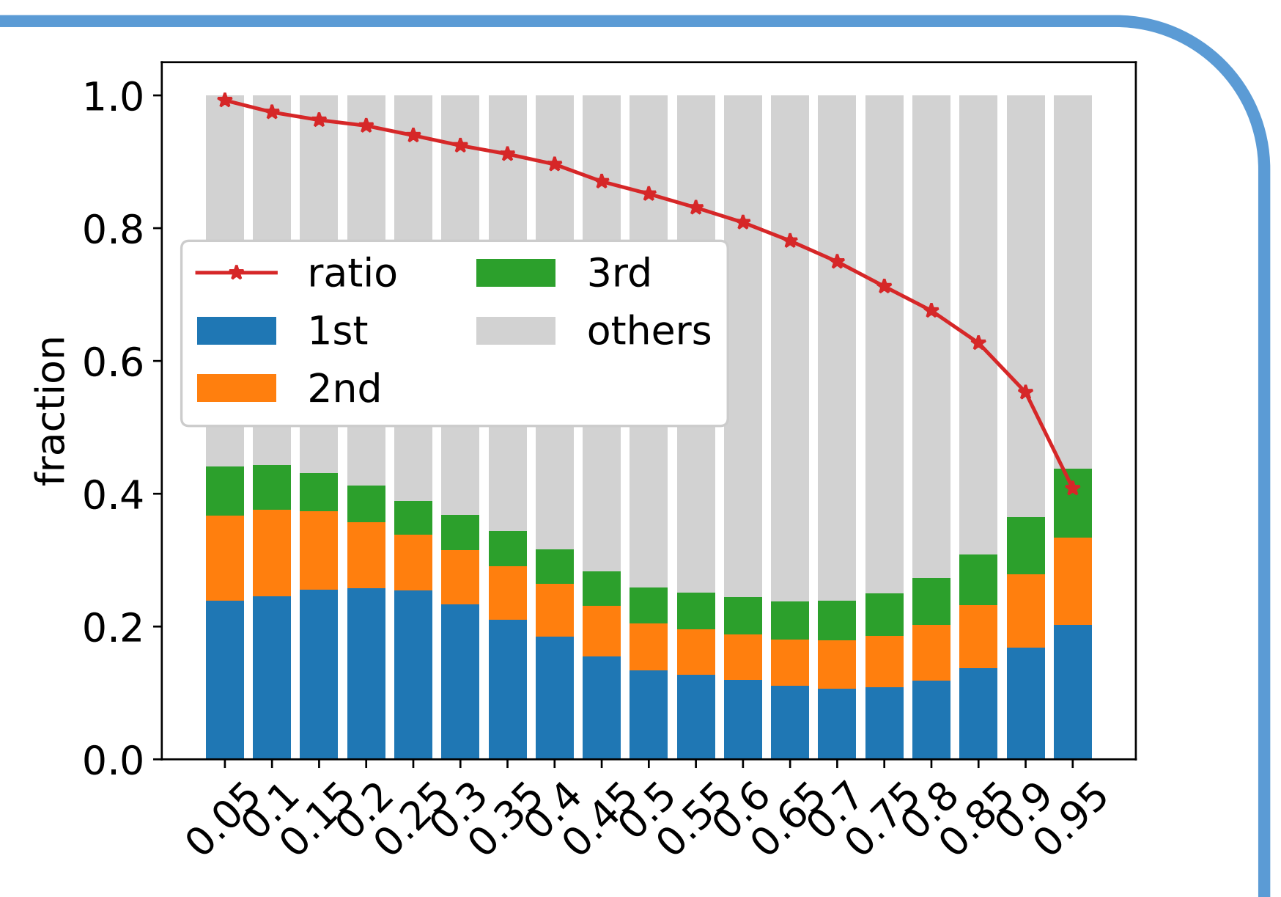
⬇

Generate using the pruned prompt

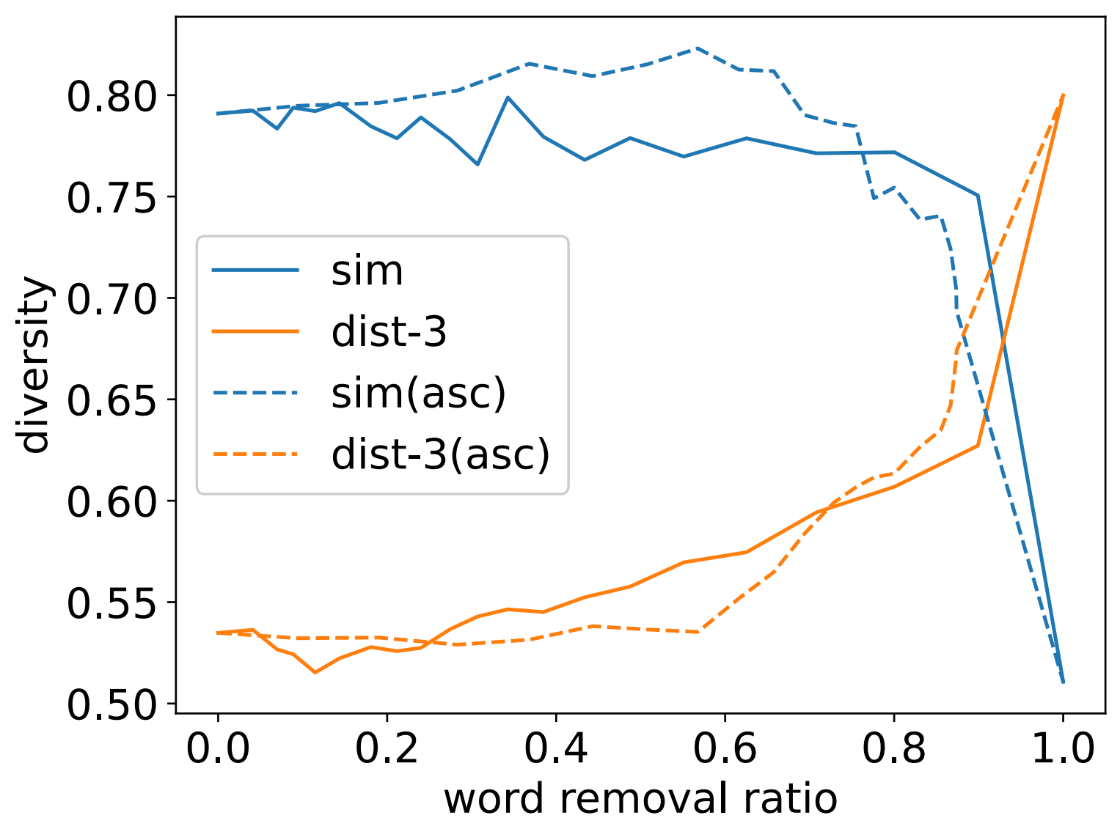




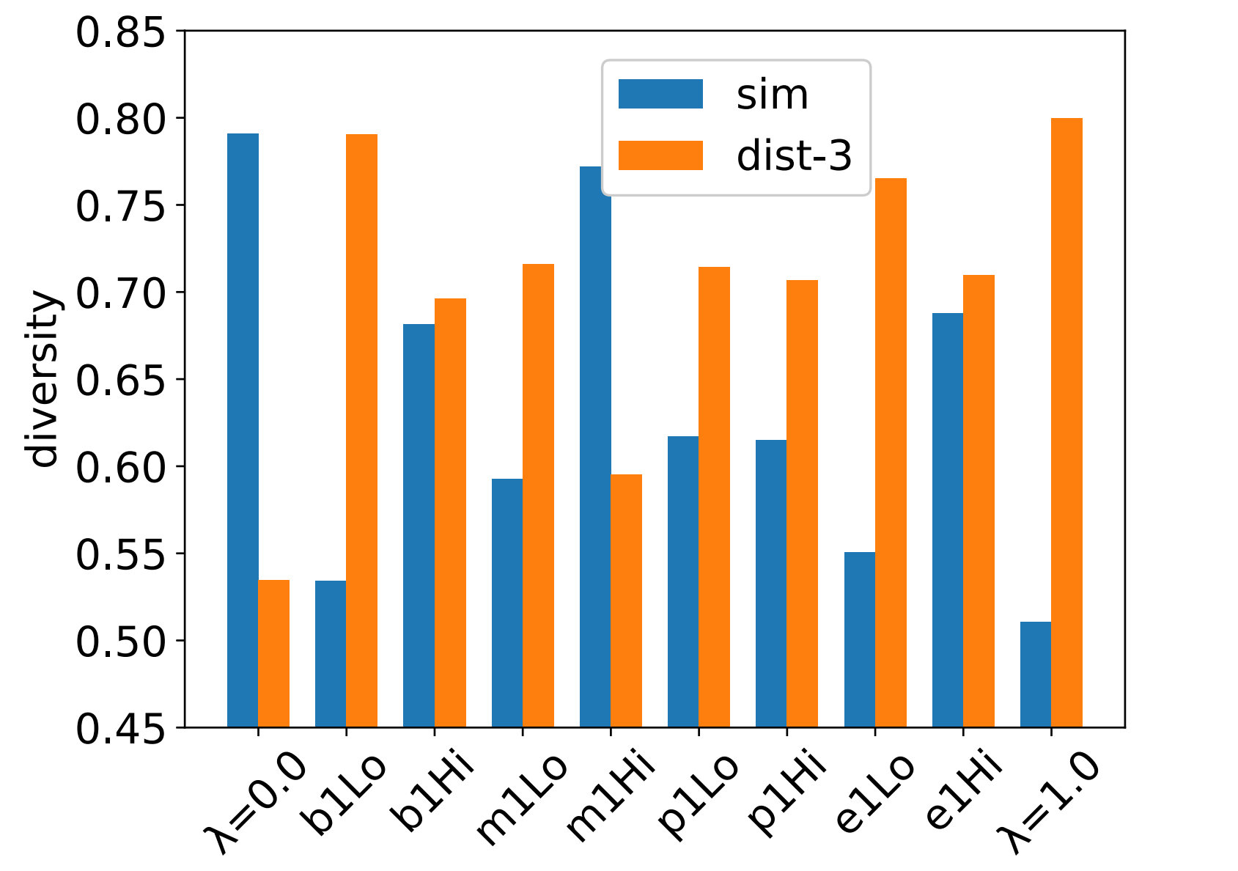
LLaMA 3, GA



Post-removal attention scores

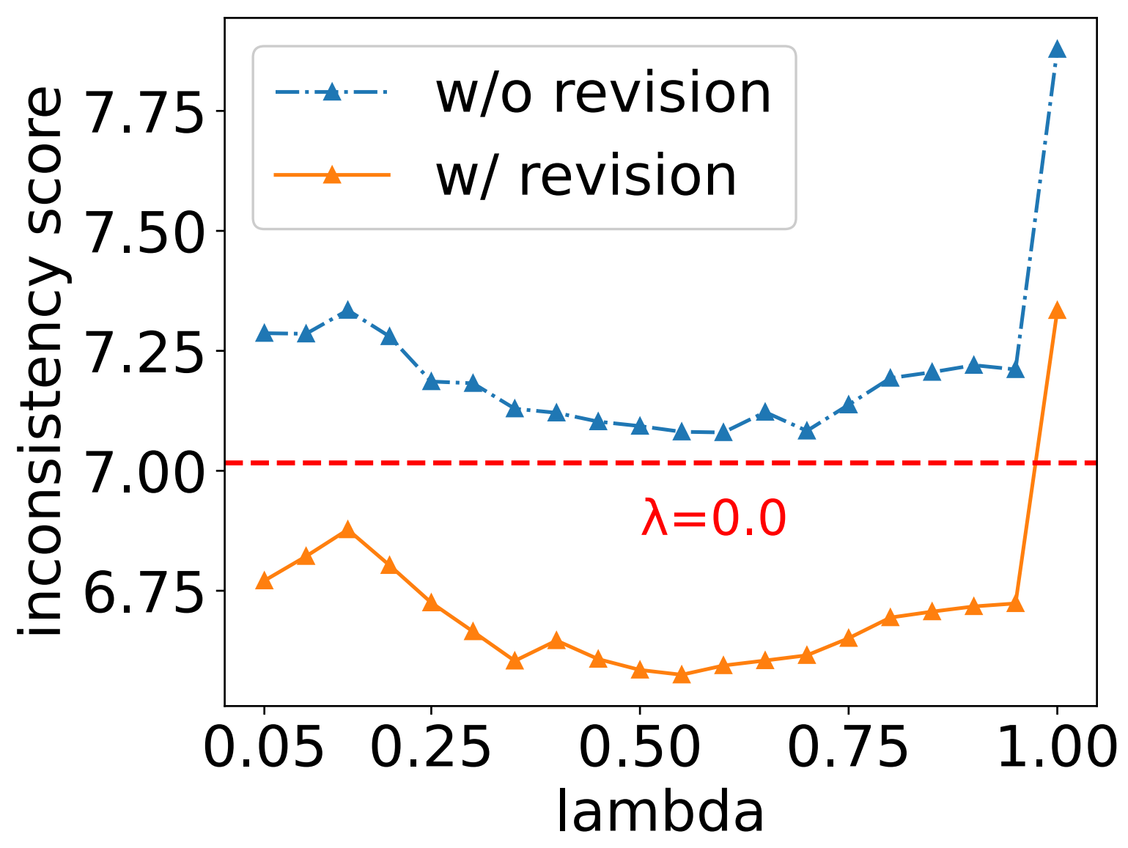


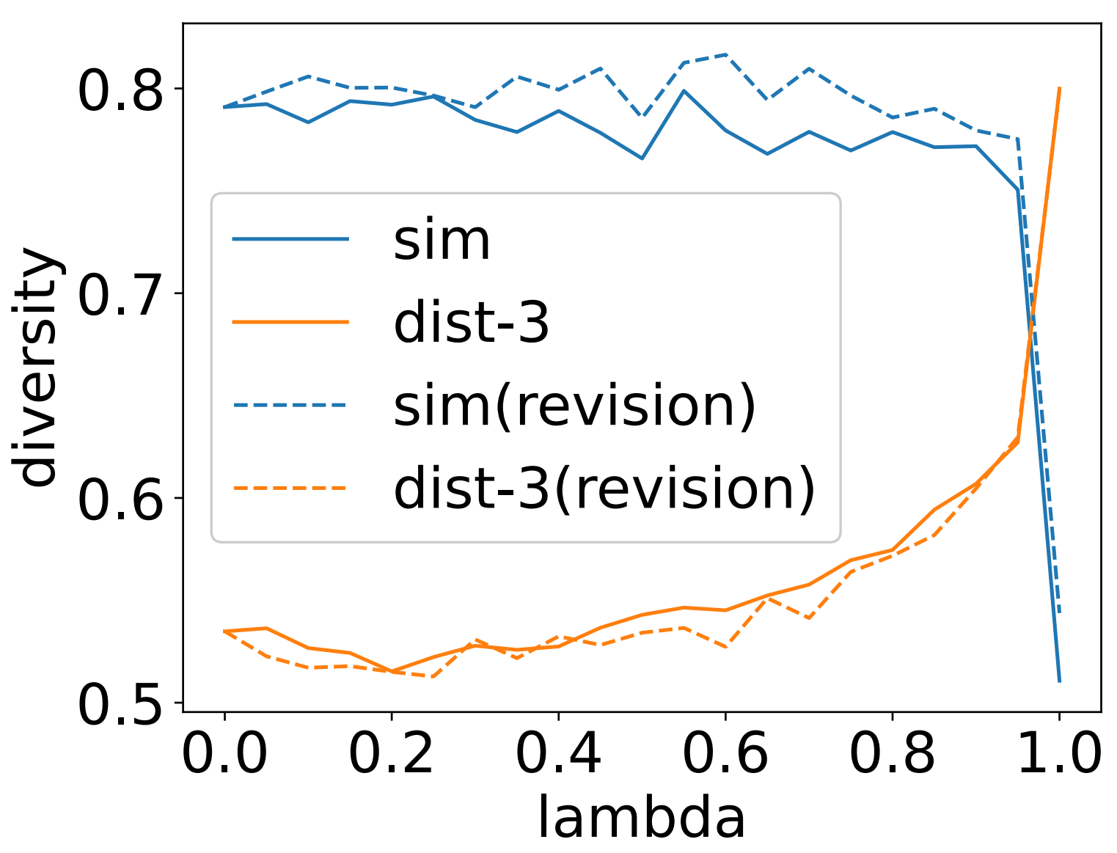
Word-efficient



Retain only one unit (Hi/Lo)

Discussion





Balancing Trade-off

Generated response may conflict with pruned information

Additional step for revision: an inconsistency score rated by the LLM

	config	sim (↓)	dist-1	dist-2	dist-3 (↑)	len
Full	default	0.791	0.095	0.350	0.535	39.9
APP	default	0.771	0.107	0.393	0.594	38.4
Full	T=1.0	0.791	0.103	0.381	0.578	40.1
APP	T=1.0	0.778	0.113	0.419	0.634	38.7
Full	p=0.99	0.800	0.102	0.375	0.569	40.0
APP	p=0.99	0.776	0.111	0.414	0.624	38.4
Full	sequential	0.634	0.197	0.524	0.695	21.9
APP	sequential	0.645	0.216	0.563	0.740	21.3

Comparison

APP is more effective (T, p) or avoids coherence issue in sequential generation

Can further enhance diversity when combined

	sim (↓)	dist-1	dist-2	dist-3 (↑)
Full	0.791	0.095	0.350	0.535
		Order		
bpmec	0.789	0.098	0.352	0.535
bmepc	0.787	0.094	0.339	0.514
bmecp	0.761	0.081	0.276	0.413
cepmb	0.744	0.053	0.145	0.206
cempb	0.747	0.050	0.135	0.191
		Frequency		
HPSS	0.828	0.093	0.337	0.518
RMbmp	0.693	0.143	0.495	0.706
HPSS+RMbmp	0.693	0.176	0.553	0.761
TLCS+RMbmp	0.733	0.143	0.501	0.713

Other Factors Affecting Diversity

Block order critically affects diversity: negative patterns (e.g., c first and b last)

Frequent names can enhance diversity as parametric knowledge is amplified (**H**arry **P**otter is 1,000x more than **T**ifa **L**ockhart in C4)