Prompt & CoT Compression: LLMLingua-2 & TokenSkip

2026. 07. 30.

윤정호

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

Intro

***입력과 출력, 두 곳의 token을 줄인다**

- 토큰 효율의 두 병목
- 긴 prompt context: prefill 계산, 메모리, 문맥 잡음
- 긴 Chain-of-Thought: autoregressive decoding 지연

공통 질문
- 정답에 기여하지 않는 token은 무엇인가?
- 줄인 뒤에도 task fidelity를 유지할 수 있는가?

오늘의 논문

LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression

Zhuoshi Pan[1†], **Qianhui Wu**[2‡], **Huiqiang Jiang**[2], **Menglin Xia**[2], **Xufang Luo**[2], **Jue Zhang**[2], **Qingwei Lin**[2], **Victor Rühle**[2], **Yuqing Yang**[2], **Chin-Yew Lin**[2], **H. Vicky Zhao**[1], **Lili Qiu**[2], **Dongmei Zhang**[2]
[1] Tsinghua University, [2] Microsoft Corporation
{qianhuiwu, hjiang, xufang.luo}@microsoft.com

ACL 2024

TokenSkip: Controllable Chain-of-Thought Compression in LLMs

Heming Xia, **Chak Tou Leong**, **Wenjie Wang**, **Yongqi Li***, **Wenjie Li**
Department of Computing, The Hong Kong Polytechnic University
University of Science and Technology of China
{he-ming.xia, chak-tou.leong}@connect.polyu.hk

ENMLP 2025

LLMLingua-2

*Abstract & Introduction

- 긴 prompt는 비용뿐 아니라 정보 선택을 어렵게 함
- 기존 scoring 기반 압축은 causal LM 호출이 많아 느림
- task-aware 압축은 query마다 다시 최적화해야 함
- 목표: task-agnostic, fast, faithful extraction
- 새 단어·순서를 바꾸지 않고 필요한 token만 남김

METHODOLOGY

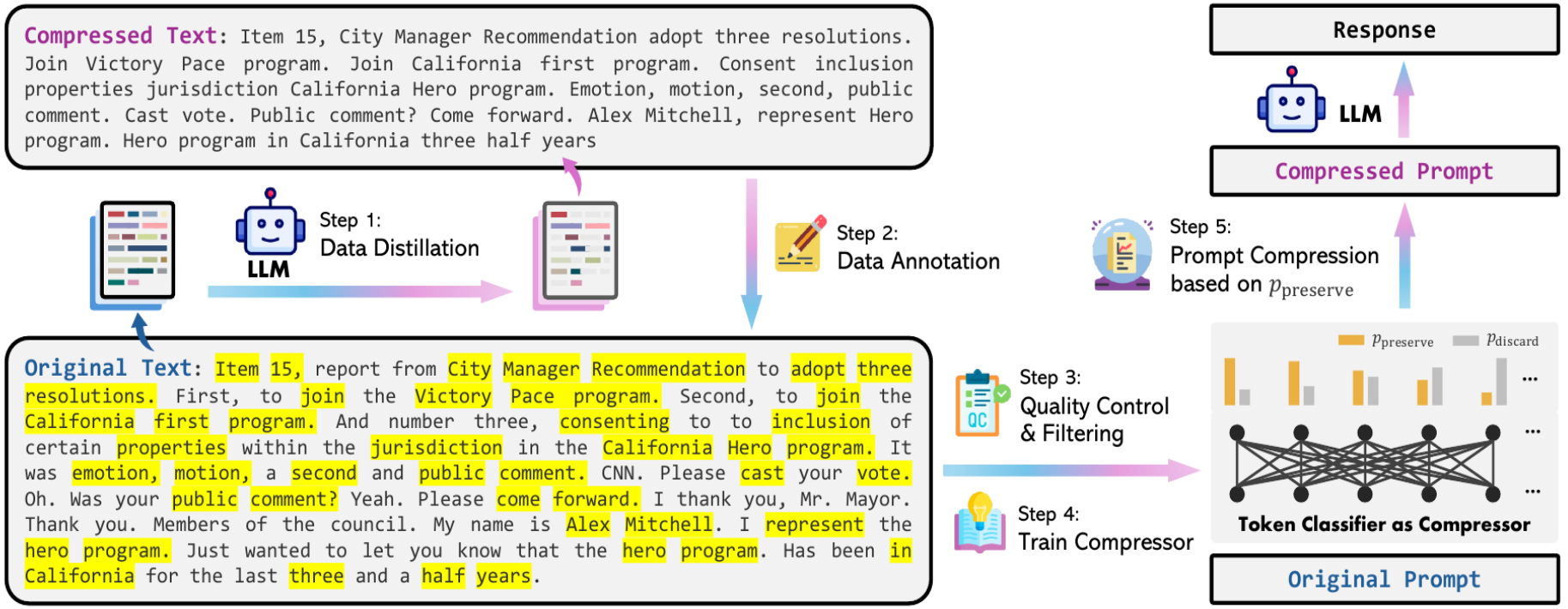


Figure 1: Overview of LLMLingua-2.

EXPERIMENTS

- Training data: MeetingBank 5,169 transcripts
- 41,746 chunks; GPT-4 deletion-only labels
- variation / alignment-gap quality filtering
- Compressor: XLM-RoBERTa-large 355M, mBERT 110M
- Evaluation: MeetingBank, LongBench, ZeroSCROLLS
- different target LLMs

Original Texts
Item 15, report from City Manager Recommendation to adopt three resolutions. First, to join the Victory Pace program. Second, to join the California first program. And number three, consenting to to inclusion of certain properties within the jurisdiction in the California Hero program.

Compressed Texts
City Manager Recommendation adopt three resolutions. Join California first program. Consent properties inclusion jurisdiction California Hero program.

Figure 5: Challenges in data annotation.
(i) Ambiguity: a word in the compressed texts may appear multiple times in the original content.
(ii) Variation: GPT-4 may modify the original words in tense, plural form, *etc*. during compression.
(iii) Reordering: The order of words may be changed after compression.

EXPERIMENTS

Methods	QA	Summary					Length	
	EM	BLEU	Rouge1	Rouge2	RougeL	BERTScore	Tokens	$1/\tau$
Selective-Context	66.28	10.83	39.21	18.73	27.67	84.48	1,222	2.5x
LLMLingua	67.52	8.94	37.98	14.08	26.58	86.42	1,176	2.5x
LLMLingua-2-small	85.82	**17.41**	48.33	**23.07**	**34.36**	**88.77**	984	3.0x
LLMLingua-2	**86.92**	17.37	**48.64**	22.96	34.24	88.27	970	3.1x
Original	87.75	22.34	47.28	26.66	35.15	88.96	3,003	1.0x

Table 1: In-domain evaluation of different methods on MeetingBank.

EXPERIMENTS

Methods	GSM8K						BBH					
	1-shot constraint			half-shot constraint			1-shot constraint			half-shot constraint		
	EM	Tokens	$1/\tau$	EM	Tokens	$1/\tau$	EM	Tokens	$1/\tau$	EM	Tokens	$1/\tau$
Selective-Context†	53.98	452	5x	52.99	218	11x	54.27	276	3x	54.02	155	5x
LLMLingua†	**79.08**	446	5x	77.41	171	14x	**70.11**	288	3x	61.60	171	5x
LLMLingua-2-small	78.92	437	5x	77.48	161	14x	69.54	263	3x	60.35	172	5x
LLMLingua-2	**79.08**	457	5x	**77.79**	178	14x	70.02	269	3x	**61.94**	176	5x
Full-Shot	78.85	2,366	-	78.85	2,366	-	70.07	774	-	70.07	774	-
Zero-Shot	48.75	11	215x	48.75	11	215x	32.32	16	48x	32.32	16	48x

Table 3: Out-of-domain evaluation on reasoning and in-context learning. †: numbers reported in Jiang et al. (2023b).

EXPERIMENTS

Methods	MeetingBank				LongBench-SingleDoc					
	QA	Summ.	Tokens	$1/\tau$	2,000-token cons.	Tokens	$1/\tau$	3,000-token cons.	Tokens	$1/\tau$
Selective-Context	58.13	26.84	1,222	2.5x	22.0	2,038	7.1x	26.0	3,075	4.7x
LLMLingua	50.45	23.63	1,176	2.5x	19.5	2,054	7.1x	20.8	3,076	4.7x
LLMLingua-2-small	75.97	29.93	984	3.0x	25.3	1,949	7.4x	**27.9**	2,888	5.0x
LLMLingua-2	**76.22**	**30.18**	970	3.0x	**26.8**	1,967	7.4x	27.3	2,853	5.1x
Original Prompt	66.95	26.26	3,003	-	24.5	14,511	-	24.5	14,511	-

Table 4: Evaluation with Mistral-7B as the Target LLM on MeetingBank and LongBench single doc QA task. We report Rouge1(Lin, 2004) for summary.

$1/\tau$	1x	2x	3x	5x
End2End w/o Compression		14.9		
End2End w/ LLMLingua-2	-	9.4 (1.6x)	7.5 (2.1x)	5.2 (2.9x)
Selective-Context	-	15.9	15.6	15.5
LLMLingua	-	2.9	2.1	1.5
LLMLingua-2	-	**0.5**	**0.4**	**0.4**

Table 5: Latency (s) comparison on MeetingBank.

LLMLingua-2: Conclusion

- 355M encoder가 prompt-level importance를 빠르게 예측
- MeetingBank 3.1x 압축에도 QA EM 86.92, 원문 87.75에 근접
- 5x 압축에서 end-to-end latency 14.9s → 5.2s

***한계**

- high-quality teacher label과 alignment filtering에 의존
- task-agnostic은 query-aware 방법보다 OOD에서 항상 우세하지 않음
- 입력을 줄여도 출력 CoT의 decode 비용은 남아 있음

TokenSkip

*Abstract & Introduction

- CoT는 reasoning accuracy를 높이지만 output token마다 decode latency가 누적
- CoT 안의 token 중요도는 균일하지 않음
- 수식·숫자·결정 단계는 보존
- 연결어·반복 표현은 압축 가능
- 질문: 짧은 CoT를 '명령'이 아니라 모델이 생성하게 할 수 있는가?

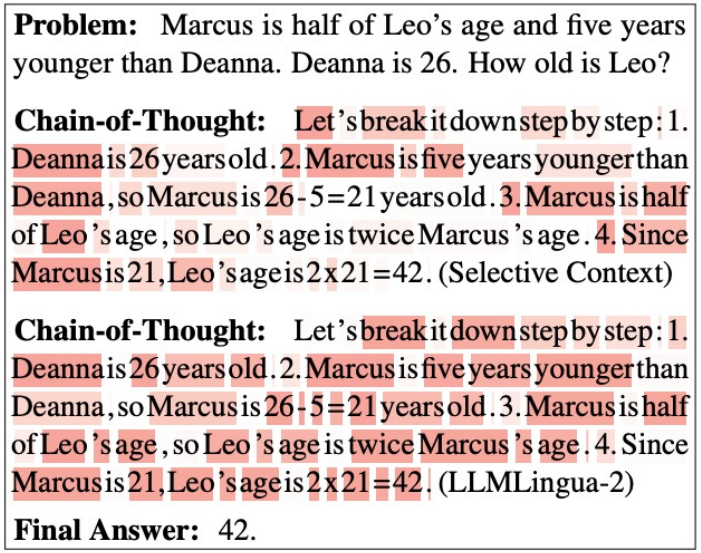


Figure 2: Visualization of token importance within a CoT sequence, with darker colors indicating higher values. This figure compares two token importance measurements: Selective Context and LLMLingua-2.

TokenSkip

*From importance scoring to controllable CoT

- TokenSkip은 transformer layer를 동적으로 skip하는 기법이 아님
- LLMLingua-2의 p_preserve로 CoT token을 prune
- 압축된 CoT를 새로운 supervision으로 사용
- 목표 compression ratio γ를 조건으로 제공
- 하나의 모델이 여러 길이의 reasoning을 생성하도록 학습

METHODOLOGY

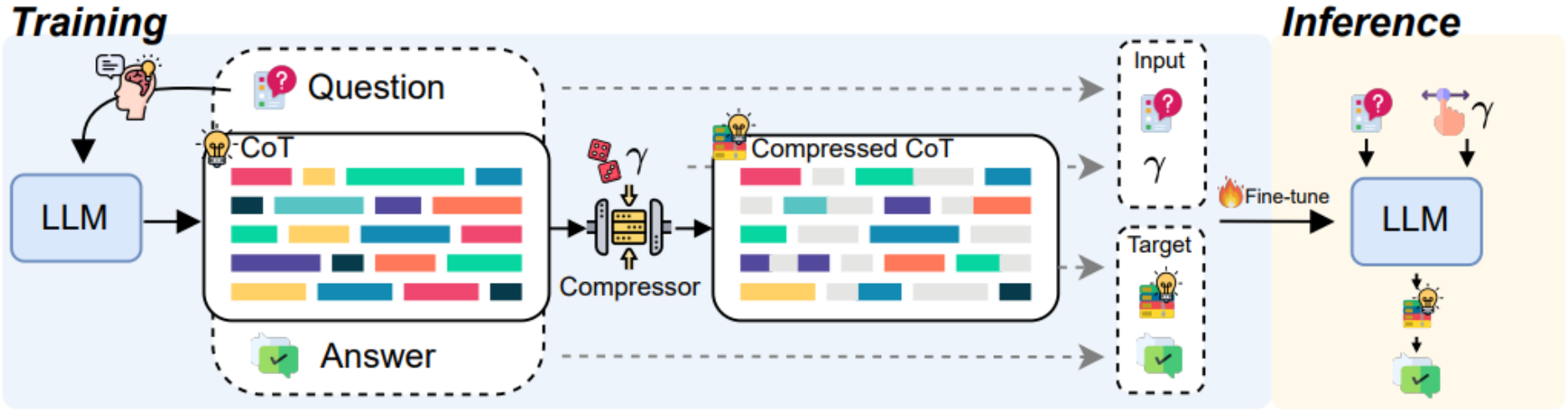
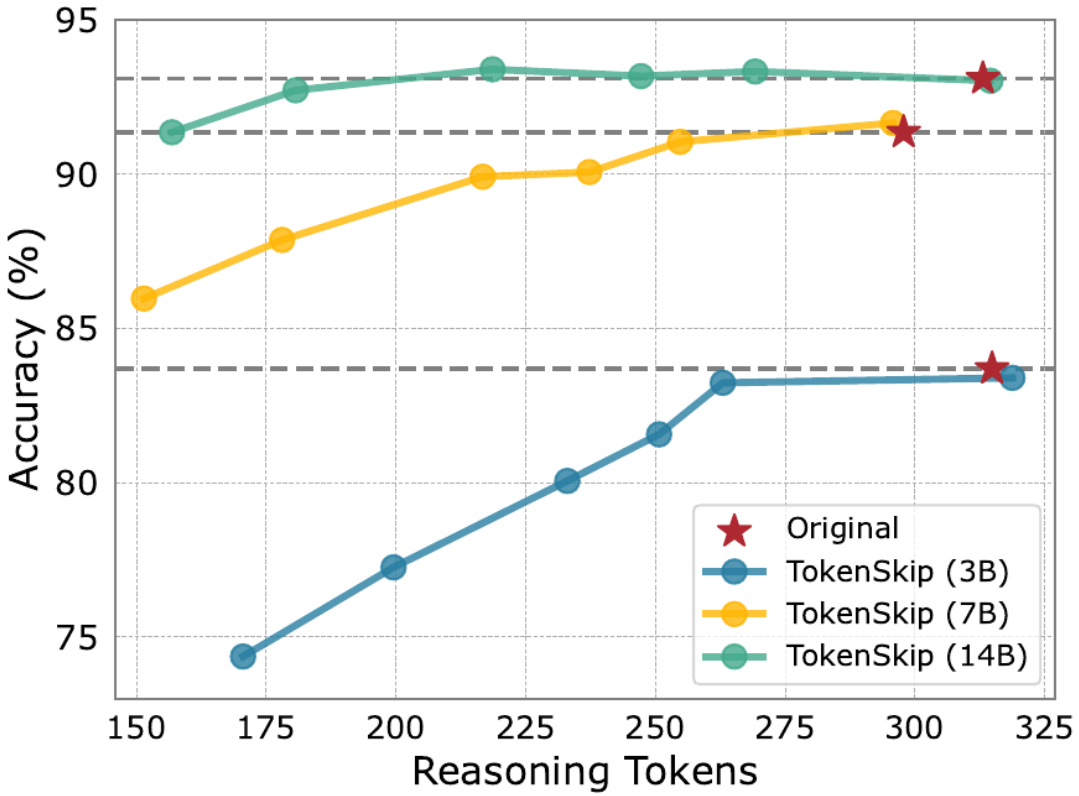


Figure 4: Illustration of TokenSkip. During training, TokenSkip first generates CoT trajectories from the target LLM. These CoTs are then compressed to various ratios sampled from the ratio set. TokenSkip fine-tunes the LLM using compressed CoTs with mixed ratios, enabling controllable CoT inference at any desired $\gamma \in \{\gamma_0, \ldots, \gamma_z\}$.

EXPERIMENTS

- Model: LLaMA-3.1-8B; Qwen2.5 3B / 7B / 14B
- Training: LoRA (r=8, α=16), 3 epochs
- $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- answer token은 압축하지 않음
- Benchmark: GSM8K, MATH, MATH-500
- Metric: accuracy, CoT tokens, latency, actual ratio

EXPERIMENTS



EXPERIMENTS

*Main results

Methods	Ratio	GSM8K				MATH-500			
		Accuracy ↑	Tokens ↓	Latency (s) ↓	*Act*Ratio	Accuracy ↑	Tokens ↓	Latency (s) ↓	*Act*Ratio
Original	-	86.2$_{(0.0\downarrow)}$	213.17	5.96$_{1.0\times}$	-	48.6$_{(0.0\downarrow)}$	502.60	16.37$_{1.0\times}$	-
BeConcise	-	82.9$_{(3.3\downarrow)}$	161.32	4.73$_{1.3\times}$	0.76	47.4$_{(1.2\downarrow)}$	471.34	15.54$_{1.1\times}$	0.94
OnlyNumbers	-	83.2$_{(3.0\downarrow)}$	165.27	4.95$_{1.2\times}$	0.78	46.4$_{(2.2\downarrow)}$	487.00	15.93$_{1.0\times}$	0.97
AbbreWords	-	83.7$_{(2.5\downarrow)}$	170.33	5.15$_{1.2\times}$	0.80	47.6$_{(1.0\downarrow)}$	489.07	15.94$_{1.0\times}$	0.97
LC-Prompt	0.9	84.1$_{(2.1\downarrow)}$	226.37	6.12$_{1.0\times}$	1.06	48.6$_{(0.0\downarrow)}$	468.04	15.39$_{1.1\times}$	0.93
	0.7	84.9$_{(1.3\downarrow)}$	209.39	5.51$_{1.1\times}$	0.98	48.4$_{(0.4\downarrow)}$	472.13	15.55$_{1.1\times}$	0.94
	0.5	83.7$_{(2.5\downarrow)}$	188.82	4.97$_{1.2\times}$	0.89	47.8$_{(0.4\downarrow)}$	471.11	15.48$_{1.1\times}$	0.94
Truncation	0.9	70.2$_{(26.0\downarrow)}$	202.06	5.29$_{1.1\times}$	0.95	47.8$_{(0.8\downarrow)}$	440.33	14.56$_{1.1\times}$	0.88
	0.7	25.9$_{(60.3\downarrow)}$	149.99	3.97$_{1.5\times}$	0.70	45.0$_{(3.6\downarrow)}$	386.89	12.85$_{1.3\times}$	0.77
	0.5	7.0$_{(79.2\downarrow)}$	103.69	2.95$_{2.0\times}$	0.49	27.4$_{(21.2\downarrow)}$	283.70	9.40$_{1.7\times}$	0.56
TokenSkip	1.0	86.7$_{(0.5\uparrow)}$	213.60	5.98$_{1.0\times}$	1.00	48.2$_{(0.4\downarrow)}$	504.79	16.43$_{1.0\times}$	1.00
	0.9	86.1$_{(0.1\downarrow)}$	198.01	5.65$_{1.1\times}$	0.93	47.8$_{(0.8\downarrow)}$	448.31	15.26$_{1.1\times}$	0.89
	0.8	84.3$_{(1.9\downarrow)}$	169.89	5.13$_{1.2\times}$	0.80	47.3$_{(1.3\downarrow)}$	398.94	13.39$_{1.2\times}$	0.79
	0.7	82.5$_{(3.7\downarrow)}$	150.12	4.36$_{1.4\times}$	0.70	46.7$_{(1.9\downarrow)}$	349.13	11.55$_{1.4\times}$	0.69
	0.6	81.1$_{(5.1\downarrow)}$	129.38	3.81$_{1.6\times}$	0.61	42.0$_{(6.6\downarrow)}$	318.36	10.58$_{1.6\times}$	0.63
	0.5	78.2$_{(8.0\downarrow)}$	113.05	3.40$_{1.8\times}$	0.53	40.2$_{(8.4\downarrow)}$	292.17	9.67$_{1.7\times}$	0.58

EXPERIMENTS

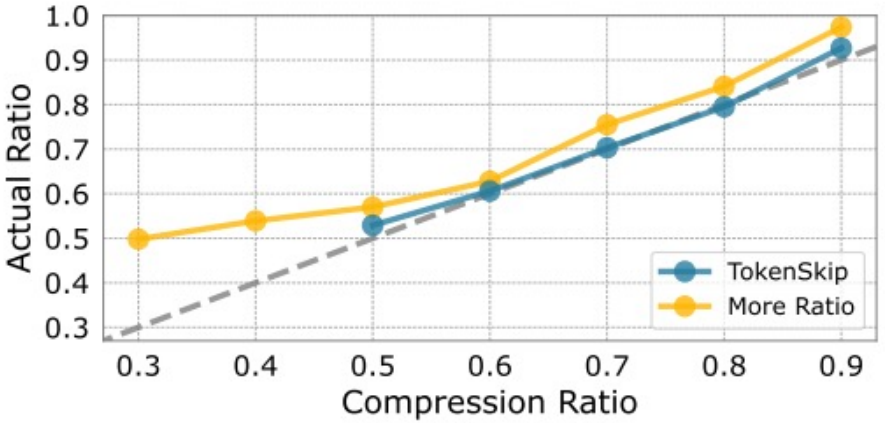


Figure 6: Comparison of ratio adherence across different compression ratio settings. The experimental results are obtained with LLaMA-3.1-8B-Instruct on GSM8K.

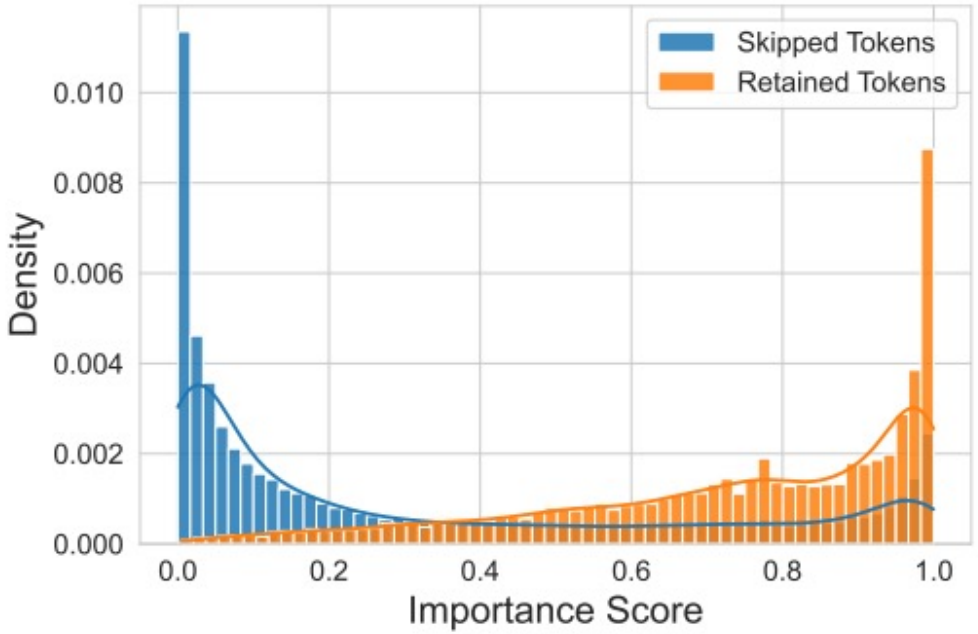


Figure 7: Distribution of token importance for skipped versus retained tokens. The LLM effectively learns to skip low-importance tokens and retain critical ones.

EXPERIMENTS

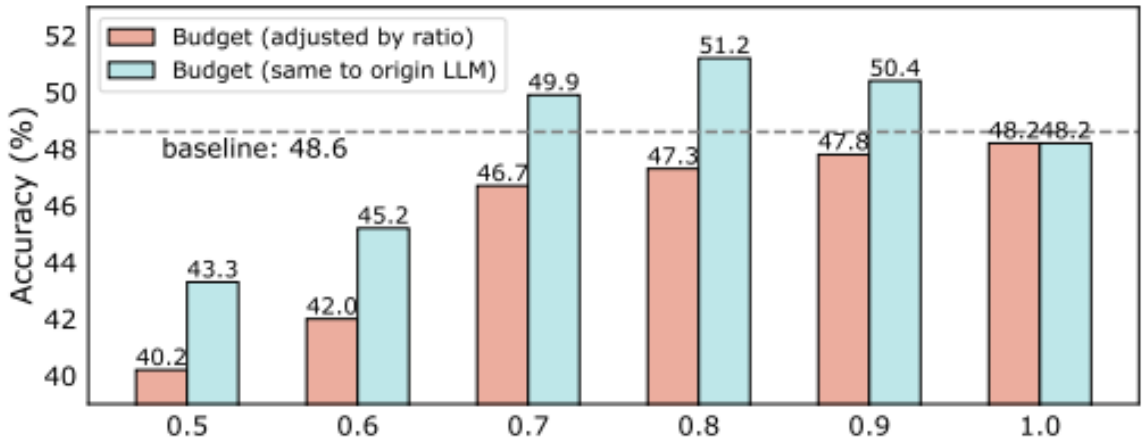


Figure 9: Performance comparison of TokenSkip with varying maximum length constraints, evaluated with LLaMA-3.1-8B-Instruct on the MATH-500 dataset.

TokenSkip: Conclusion

- CoT의 모든 토큰이 같은 역할을 하지 않음.
- 중요도 낮은 토큰을 줄여도 핵심 reasoning을 상당부분 보존할 수 있음
- 과한 압축이 아닐 땐 정확도 손실을 제한하면서도 Token 수와 latency를 줄일 수 있음

***한계**

- 중요도는 따로 학습한 것이 아닌 LLMLingua-2가 예측한 preserve 확률임
- 수식, 숫자 도메인에 최적화된 Score가 아님
- API 모델이나, 학습 없이 진행할 수 있는 방법이 아님

Thank you