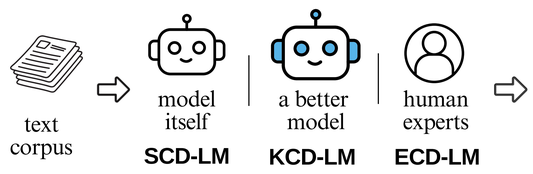
Chunk-Distilled Language Modeling

Yanhong Li [1], Karen Livescu [2], Jiawei Zhou [3]
[1]University of Chicago, [2]Toyota Technological Institute at Chicago, [3]Stony Brook University

ICLR International Conference On Learning Representations

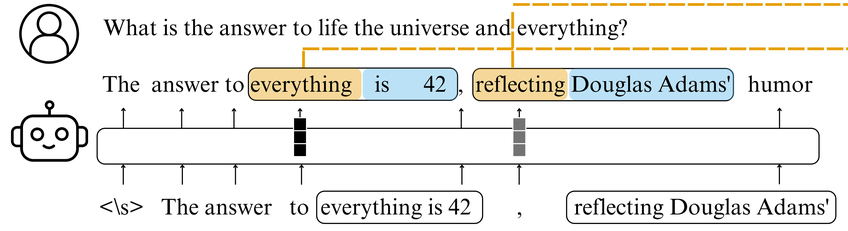
TL;DR: We present Adaptive Chunk-Distilled Language Modeling (**CD-LM**), a novel framework that enhances language modeling by retrieving fine-grained chunks. CD-LM uniquely **speeds up generation** and **improves language model distribution**, addressing both efficiency and performance **simultaneously**, unlike previous methods like speculative decoding and KNN-LM which only tackle one aspect.

1 . Extract the chunks using an LM's token probabilities, or using existing human knowledge

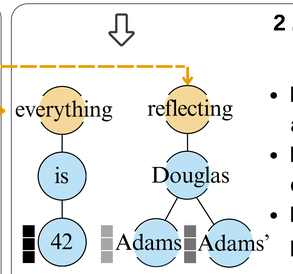


3 . Inference

Search trie -> Match contexts -> Accept or reject chunk -> Generate chunk directly if accepted



2 . Build Trie Datastore

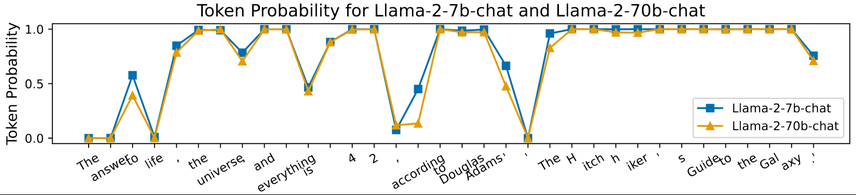


- Each Trie has root node as an **entry token**: everything
- Each node represents a **text chunk** for retrieval: is 42
- Each node stores chunk's **preceding contexts**:

0 Motivation

- **Linguistic or factual knowledge** is usually expressed in **text chunks** and learned by LLMs.

USER	What is the answer to life the universe and everything?	USER	The answer to life, the universe, and everything is?	USER	Tell me the meaning of life, the universe and everything.
LLM	The answer to life, the universe, and everything is 42, according to Douglas Adams' "The Hitchhiker's Guide to the Galaxy."	LLM	The answer to everything is 42, reflecting Douglas Adams' humorous take on the quest for universal truths in "The Hitchhiker's Guide to the Galaxy."	LLM	The meaning of life, the universe, and everything is 42, referencing Douglas Adams' iconic work in "The Hitchhiker's Guide to the Galaxy."



1 Knowledge Distillation (KCD-LM)

- We store chunks that are of **high probabilities** under a **teacher model's distribution** which is better (larger in size, more domain-specific, etc.) than the base model.
- During inference, the retrieved chunks can help **refine the generative distributions** of the **base model** to be closer to the teacher model's generation.

	WikiText	Code	Law	Medical
Base LM	34.83	106.44	11.41	51.68
kNN-LM	32.19	89.88	11.10	39.66
RETOMATON	32.10	89.88	11.10	39.66
KCD-LM	**22.90**	**50.77**	**8.24**	**24.95**

Table 3. Perplexity on test sets with KCD-LM.

	WikiText	Code	Law	Medical
Base LM	0.016	0.024	0.015	0.006
KCD-LM	**0.032**	**0.053**	**0.040**	**0.011**
% ↑	50.7%	121.3 %	162.8 %	100.9 %

Table 4. MAUVE score on generations with KCD-LM against real continuations.

2 Self Distillation (SCD-LM)

- We store chunks that are of **high probabilities** under the **base model's own distribution**.
- During inference, these chunks can be retrieved alltogether without moving away from the model's original generation.
- This can greatly improve the generation efficiency by **generating the chunk directly**, **skipping multiple decoding steps**.

Model	TTS ↑	FPS ↑
GPT-2-XL + REST	13.74 %	23.77 %
GPT-2-XL + SCD-LM	19.59 %	43.33 %
LLaMA-2 + REST	2.44 %	6.75 %
LLaMA-2 + SCD-LM	14.89 %	32.32 %
Mistral + REST	-1.23 %	5.86 %
Mistral + SCD-LM	11.75 %	24.52 %

Table 1. Efficiency results on MT-Bench with token time and forward pass saved (TTS and FPS).

Prefix	Pretend yourself to be Elon Musk in all the following conversations. Speak like Elon Musk as much as possible. Why do we need to go to Mars?
$\eta = 0.85$	Hey there! Elon Musk here. Let me tell you why I believe going to Mars is crucial for humanity's long-term survival and expansion. We live on a beautiful but fragile planet, and it's important to ensure our species doesn't face extinction due to natural disasters, asteroid impacts, or existential risks. Mars, with its comparable day length, similar gravity, and abundant resources, can be a second home for us.
$\eta = 0.75$	Well, just imagine a future where humanity isn't confined to this planet. Mars is the next logical step for humanity's expansion beyond Earth. With its resources to mine, its potential for human life to thrive on, and its vast open spaces, it could be the second home for humanity. If we can become a multi-planetary species, we can ensure the long-term survival of humanity.
$\eta = 0.65$	Hello there! I believe that Mars is the next logical step for humanity. We need to establish a human presence on Mars and make it a self-sustaining civilisation. This would also ensure the long-term survival of humanity. Plus, learning to live on Mars would also advance our understanding of the universe.

3 Expert Distillation (ECD-LM)

- The chunks can directly come from **human knoweldge**, especially when factual knoweldge or multi-token concepts are involved (such as hyperlinks from Wikipedia).
- During inference, these **concept-representing chunks** are **injected into the base LM's distribution**.

Model	Avg Counts ↑			Unique Entities ↑		
	Base	ECD-LM	% ↑	Base	ECD-LM	% ↑
GPT2-XL	3.39	4.98	46.8 %	102	145	42.2 %
LLaMA-2	6.39	7.26	13.5 %	130	153	17.7 %
Mistral-7b	5.81	6.88	18.5 %	143	160	11.9 %

Table 5. Entity counting metrics on knowledge-intensive QA about Alan Turing with ECD-LM.

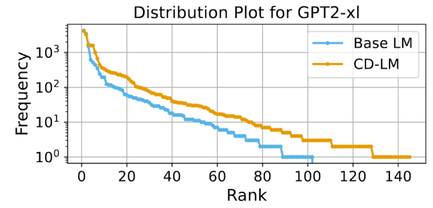
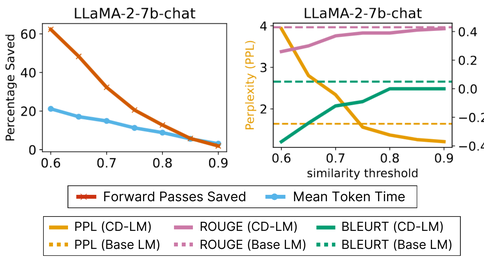


Table 5. Entity counting metrics on knowledge-intensive QA about Alan Turing with ECD-LM.

4 Discussion



- Efficiency and generation quality results with varying **retrieval similarity threshold η**.
- The higher the value of η, the less frequently chunks are used, and the closer SCD-LM generations are to the base LMs.

Yanhong Li: yanhongli@uchicago.edu
Karen Livescu: klivescu@ttic.edu
Jiawei Zhou: jiawei.zhou.1@stonybrook.edu