NEURAL INFORMATION PROCESSING SYSTEMS
December 02-07, 2025

Qualcomm

OMNIDRAFT: A CROSS-VOCABULARY, ONLINE ADAPTIVE DRAFTER FOR ON-DEVICE SPECULATIVE DECODING

Ramchalam Kinattinkara Ramakrishnan, Zhaocong Yuan, Shaojie Zhuo, Chen Feng, Yicheng Lin, Chenzheng Su, Xiaopeng Zhang

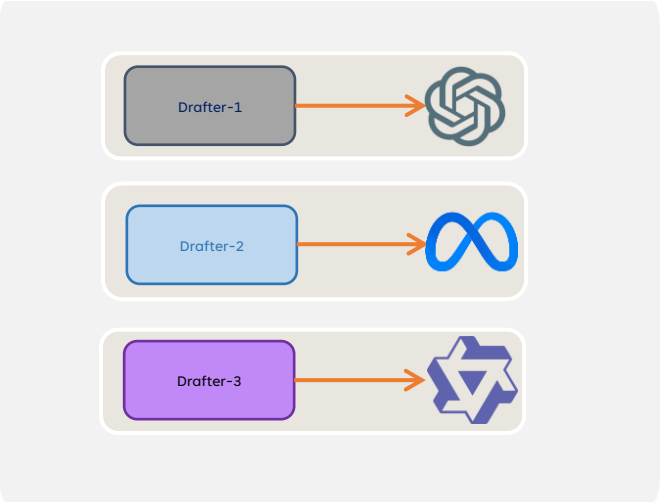
Qualcomm Canada ULC

The Goal: Fast AI on every single device

- We want powerful models to run on-device (privacy, speed, offline availability).
- Bottleneck → **models are huge**
- Solution→ ***Speculative Decoding (SpD)***

Problem: A "locked-in" system

- The draft and target models are tightly coupled → same tokenizer & vocabulary
- New target requires training new drafters → inefficient and not scalable

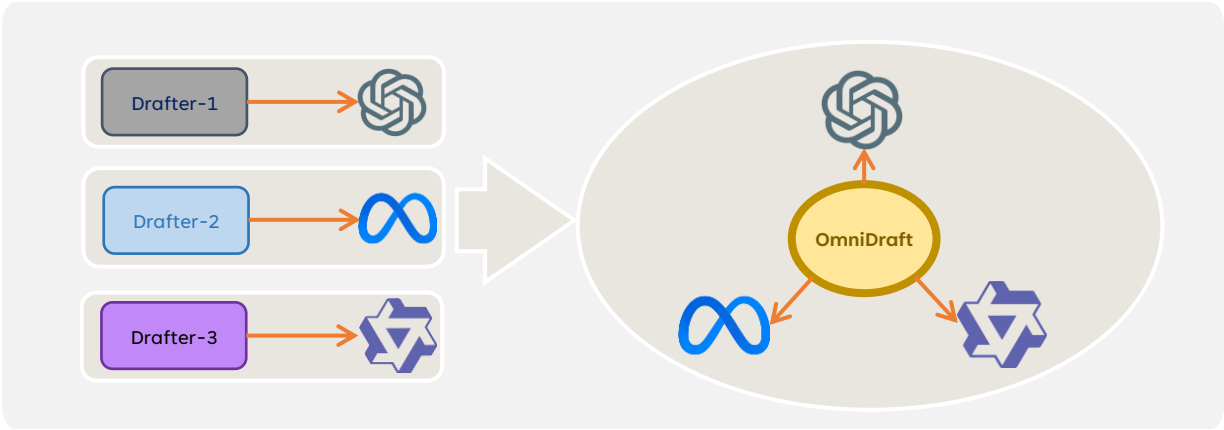


What if…

- Decouple the drafter and the target?
- A Single, universal drafter on-device to pair with **any target model?**
- A Single, universal drafter on-device to adapt to **any target application**?

Introducing: ***OmniDraft***

- *Plug and learn* drafter.
- Cross vocabulary, online adaptive to any target.

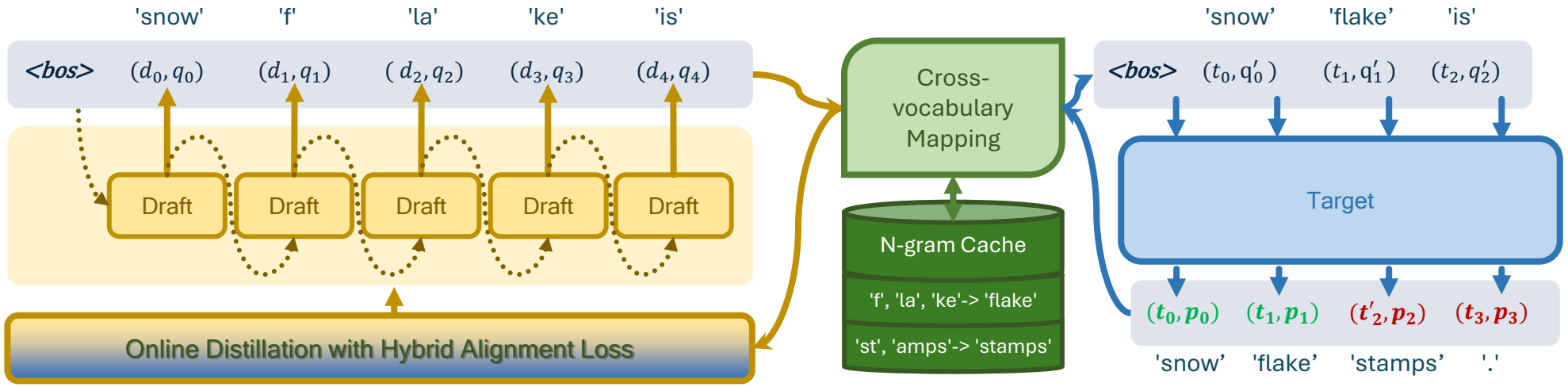


OMNIDRAFT

Contributions:

1. **Cross-vocabulary speculative decoding** via online n-gram cache
2. **Online knowledge distillation** with hybrid alignment loss

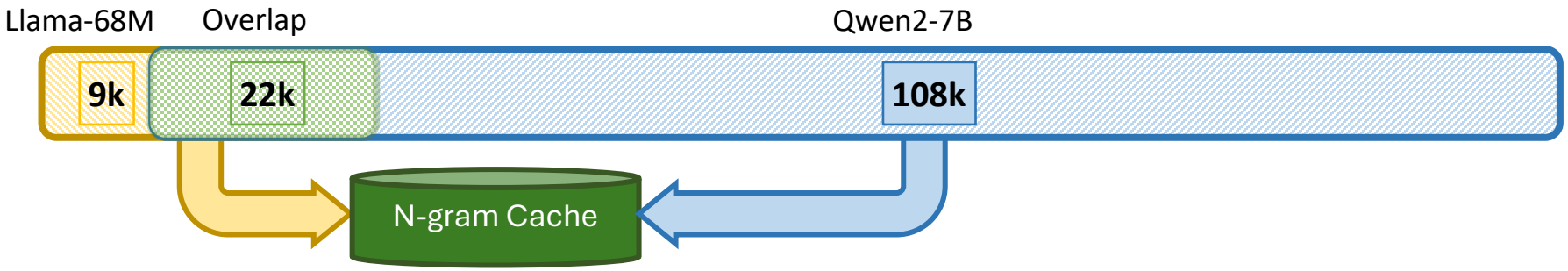
As an example, the proposal ("snowflake is") is tokenized differently by drafter and target. Cross-vocabulary SpD enables a mapping on tokens and their logits (['snow', 'f', 'la', 'ke', 'is'] → ['snow', 'flake', 'is']) to support rejection sampling between two vocabularies.



CROSS-VOCABULARY SPD VIA ONLINE **N-GRAM CACHE**

What is the n-gram cache?

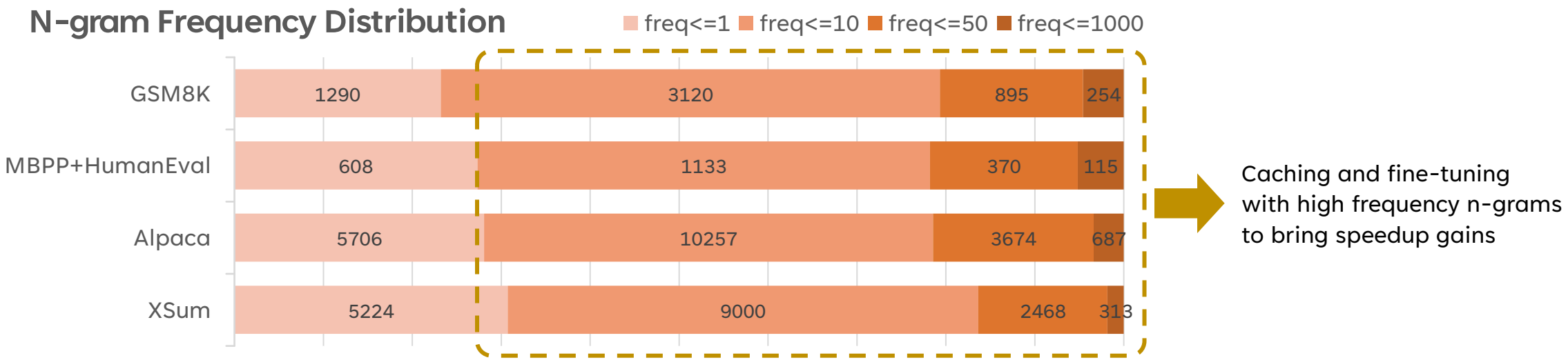
- For non-coupled draft-target pair, the vocabulary overlap is generally small.
- A (non-overlapping) target token can be detokenized as several draft tokens like an n-gram.
  - For example, (target): '**flake'** → (draft): **'f', 'la', 'ke'**



CROSS-VOCABULARY SPD VIA ONLINE **N-GRAM CACHE**

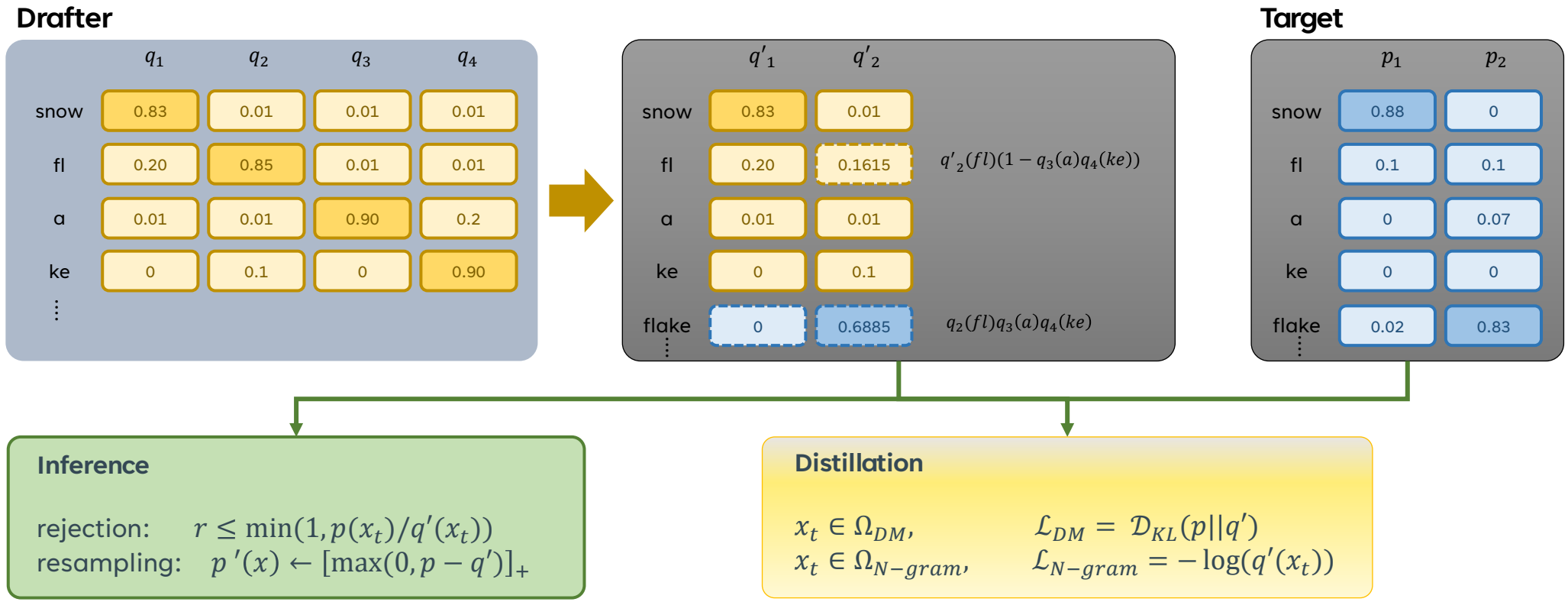
What do we gain from the cache?

- With training-free, it enables **dynamic token merging** during cross-vocabulary SpD.
- With fine-tuning, it gives **speedup boost** since drafter and target are more aligned on n-gram token semantics and distribution .



ONLINE DISTILLATION WITH **HYBRID ALIGNMENT LOSS**

A distribution mapping transformation between the drafter and target vocabularies, converting the token and logits to a compatible space.



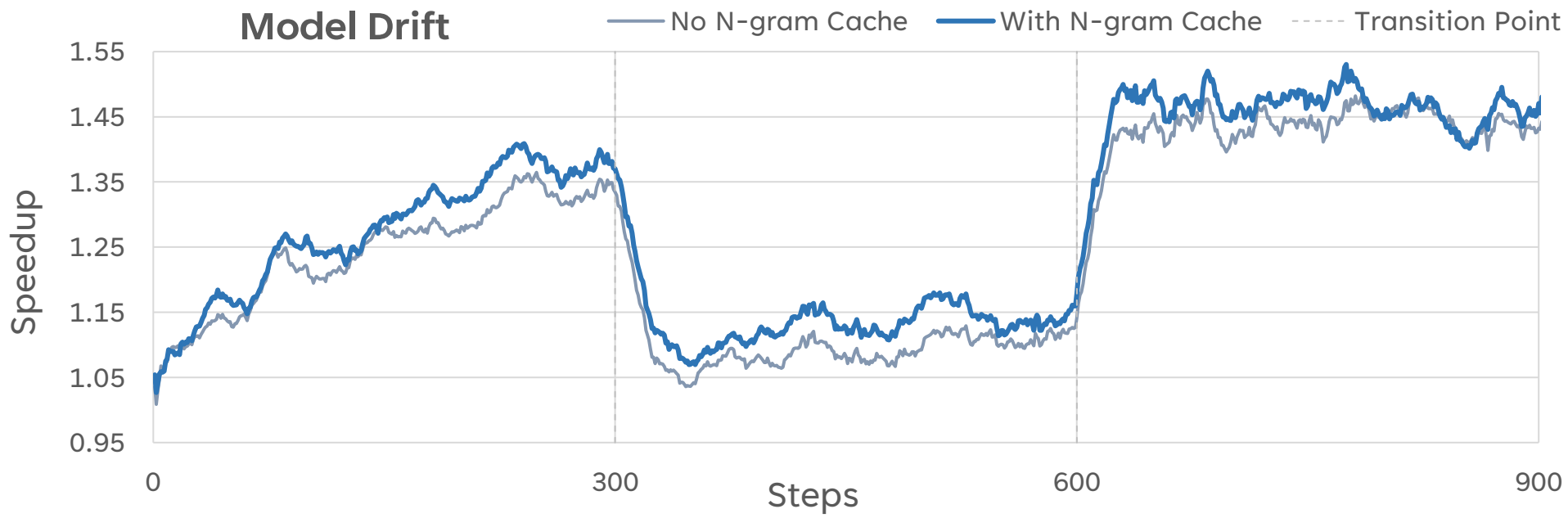
RESULTS

Speedup performances over math, coding and instruction following tasks using the Llama-68M drafter.
Additional results scaling target model to 14B, 32B achieve up to **1.5 – 2x boost** in inference speed.

Target	Method	GSM8K	MBPP+HumanEval	Alpaca	XSUM
Llama3-8B	SpD_{DM}	0.94x	1.03x	0.96x	0.91x
	$\mathcal{L}_{DM}$	1.58x	1.26x	1.25x	1.20x
	$\mathcal{L}_{DM} + \lambda\mathcal{L}_{N-gram}$	**1.70x**	**1.33x**	**1.30x**	**1.24x**
Qwen2-7B	SpD_{DM}	1.04x	0.91x	1.01x	0.96x
	$\mathcal{L}_{DM}$	1.50x	1.29x	1.25x	1.16x
	$\mathcal{L}_{DM} + \lambda\mathcal{L}_{N-gram}$	**1.61x**	**1.36x**	**1.30x**	**1.22x**

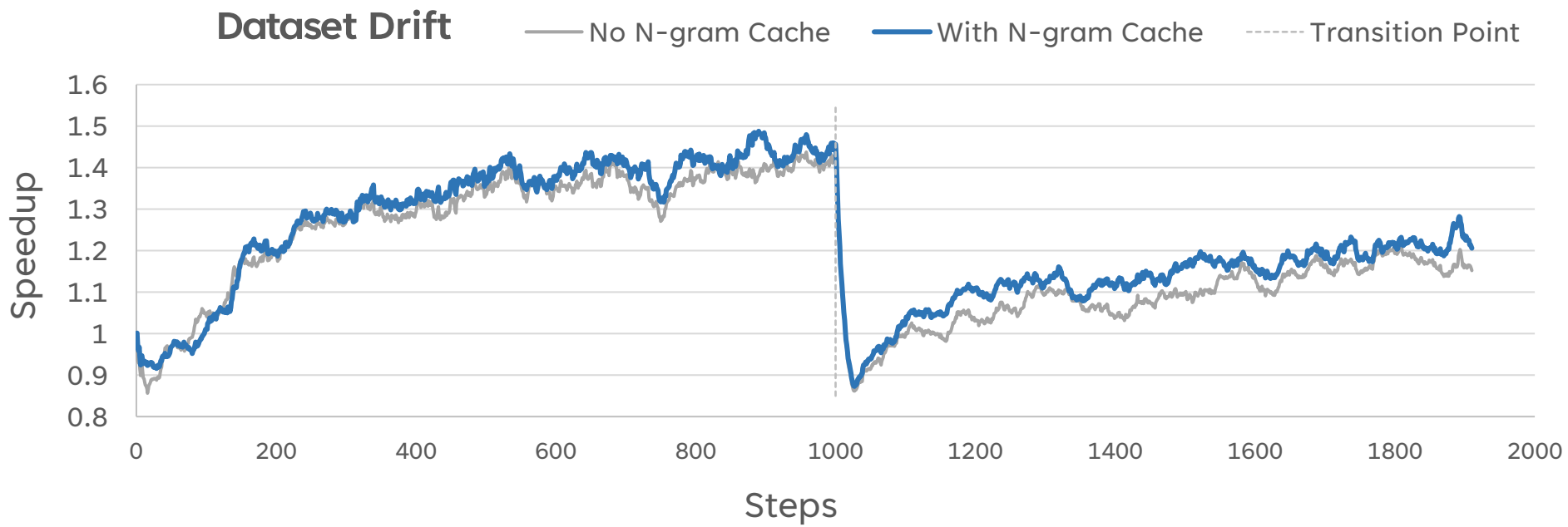
DISTRIBUTION DRIFT

Cross-vocabulary SpD is robust against distribution drifts in either **change of target model** or change in task data, recovery rate is also better due to the presence of the online cache.



DISTRIBUTION DRIFT

Cross-vocabulary SpD is robust against distribution drifts in either change of target model or **change in task data**, recovery rate is also better due to the presence of the online cache.



CACHE MEMORY

Memory footprint of the n-gram cache after online learning is **minimal compared to model sizes**.
Additional results using cache eviction policies (LRU, LFU) also show robust speedup performances.

	GSM8K	MBPP+HumanEval	Alpaca	XSUM
Training Samples	7473	910	8000	4000
Cache Size (#n-grams)	5569	2238	20339	17013
Cache Memory (MB)	1.372	0.501	4.569	3.924

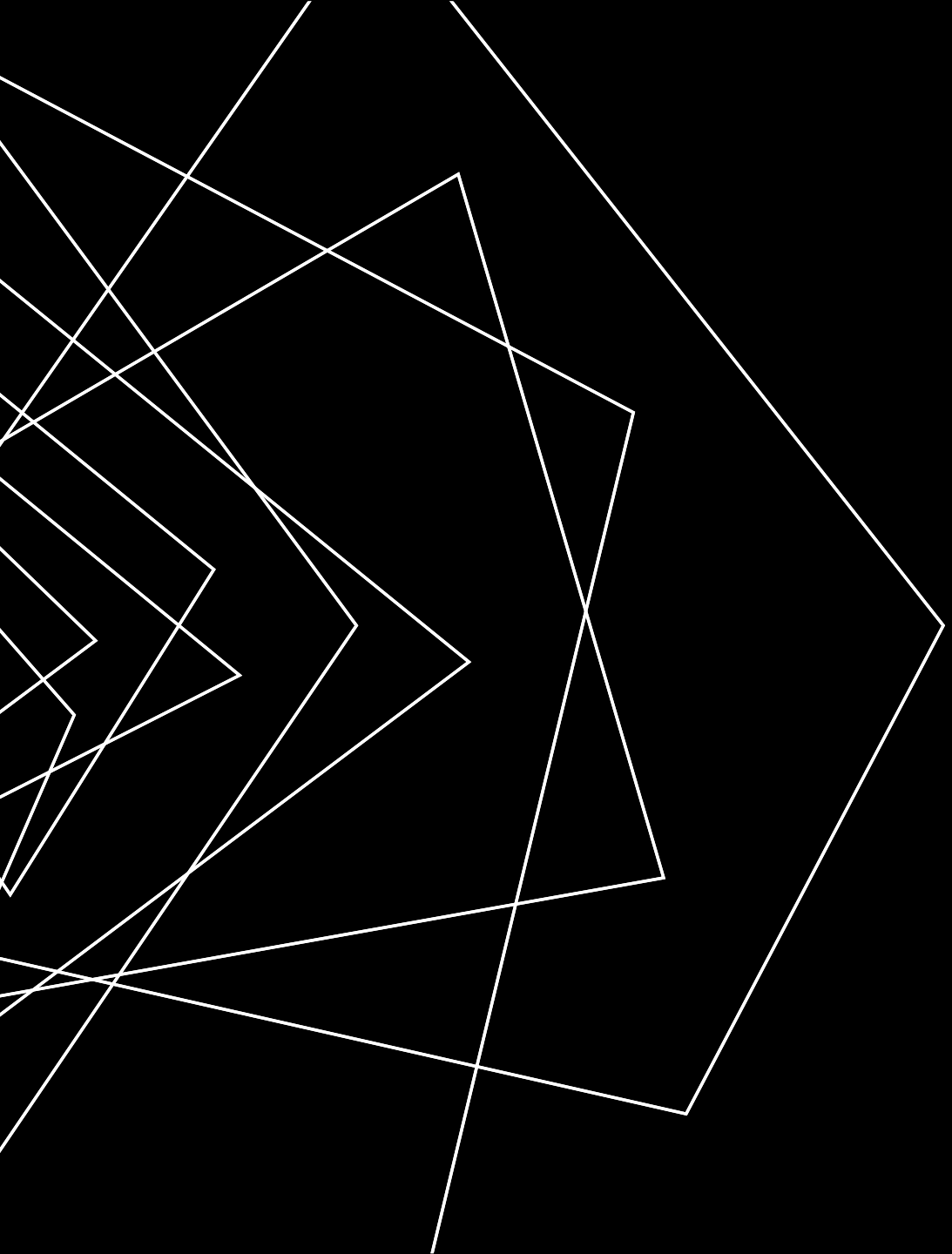
FUTURE WORK

Current limitations

- Special token handling (‘<think>’, ‘<image>’, …)
- Data efficiency

Future work

- Cache optimization
- Multi-modal extension (encoder-decoder, VLMs)
- Drafter customization



THANK YOU