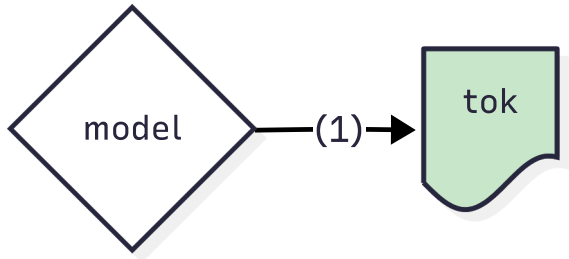
Accelerating LLM Inference with Lossless Speculative Decoding Algorithms for Heterogeneous Vocabularies

Nadav Timor [w], Jonathan Mamou[i], Daniel Korat[i], Moshe Berchansky[i], Gaurav Jain[d], Oren Pereg[i], Moshe Wasserblat[i], David Harel[w]

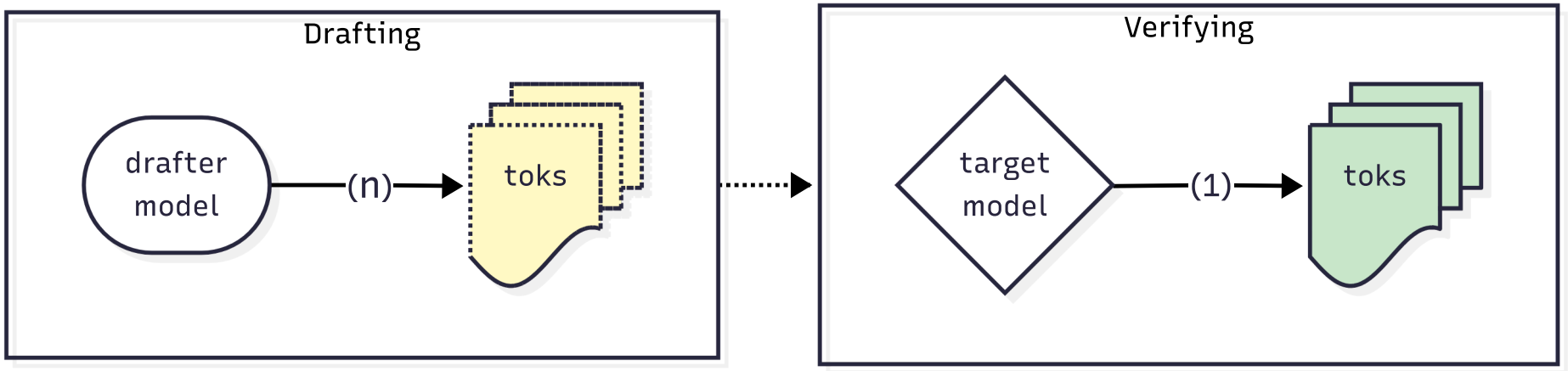
[w] Weizmann Institute of Science, [i] Intel Labs, [d] d-Matrix

Speculative decoding [Leviathan et al., 2023; Chen et al.,

- Autoregressive decoding



- Speculative decoding



  - up to 3x faster (↓ latency, ↑ throughput)
  - lossless

Contribution: Removing the shared-vocab constraint

Current limitation: drafter must share the same vocab as the target

- 💸 Training drafters from scratch:
  - No family (e.g., `DeepSeek-R1`, `phi-4`, `Mixtral-8x22B`, `CodeLlama`)
  - In-family is too slow (e.g., `DeepSeek-R1-Distill-*`, `Llama 3.1`, `gemma-2`)
  - No reuse

Our contribution: removing this limitation (& remaining lossless)

- 🆓 No training:
  - Any off-the-shelf drafter
  - Reuse
- Up to 2.8x faster (than autoregressive)
- Default in 🤗 Transformers (since Oct '24 + Feb '25)

Usage example

```
from transformers import pipeline

pipe = pipeline(
    "text-generation",
    model="google/gemma-2-9b-it",
-   assistant_model="google/gemma-2-2b-it"
+   assistant_model="double7/vicuna-68m"         # 1.5x lossless speedup!
)
out = pipe("Summarize this article...")
```

Our 3 algos

- Speculative decoding is undefined for heterogeneous vocabs

1. TLI, Token-level intersection 🤗
    - vocab pruning
2. SLEM, String-level exact match 🤗
    - back-and-forth tokenization + heuristic
3. SLRS, String-level rejection sampling
    - probs on strings

- How to choose?

Theoretical guarantees

- Lossless
- Acceptance rate (expected)
- Acceptance rate is higher than baseline

Empirical speedups

- **Up to 2.8× toks/sec**
- Various hardware
- Tasks:
  - summarization
  - coding
  - long-context understanding
- Independent evaluation by Hugging Face

Summary: free-lunch for everyone

- Speculative decoding with any off-the-shelf drafter
- Unlocks **lossless** speedups that previously required training
- Default in 🤗 (388k repos + 6k libs)

Qs?

Poster session, **4:30-7:00 pm** (📍East Exhibition Hall A-B #E-2810)

> **Summary:** free-lunch for everyone
>
> 1. Speculative decoding with any off-the-shelf drafter
> 2. Unlocks lossless speedups that previously required training
> 3. Default in 🤗 (388k repos + 6k libs)

Thank you!

Poster session, **4:30-7:00 pm** (📍East Exhibition Hall A-B #E-2810)

> **Summary:** free-lunch for everyone
>
> 1. Speculative decoding with any off-the-shelf drafter
> 2. Unlocks lossless speedups that previously required training
> 3. Default in 🤗 (388k repos + 6k libs)