PinTok: Tokenizers Deserve Dedicated Pinned CPU Compute and Memory

Sean Choi[1*], Myungheon(Young) Chin[2*], Ernest K. Ryu[3]
*Equal contribution, [1] Santa Clara University, [2] Stanford University (currently at Google), [3] University of California, Los Angeles

ICML International Conference On Machine Learning

code available

1. Motivation: Tokenizer

- **Misconception**
  - Tokenization is the **first point of contact** for LLMs.
  - Tokenization performance is often **assumed to be negligible.**
  - Tokenizer usually runs as **slow, less optimized,** CPU process.
- **High Impact on Small Models**
  - As models become faster (e.g., embeddings, SLMs), tokenizers becomes the large proportion of total inference time (Fig. 1).
- **System-Level Overheads:** Conventional tokenizers suffer from following system inefficiencies:
  - **Hardware:** Translation Lookaside Buffer (TLB) misses and swaps..
  - **Operating System:** Process scheduling, context-switching.
  - **Networking:** CPU interrupts and redundant packet copies.

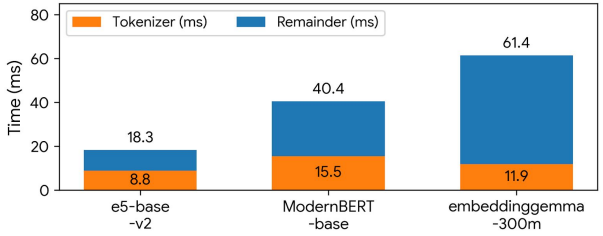


Figure 1. Proportion tokenization time becomes significant for small models.

Old

```
from pintok import pipeline
pipeline = PinTokPipeline("model")
pipeline.start()
... # Retrieve and use tokens
pipeline.stop()
```

PinTok

```
from transformers import pipeline
pipeline = PythonPipeline("model")
pipeline.start()
... # Retrieve and use tokens
pipeline.stop()
```

Plug & Play!

2. PinTok Overview

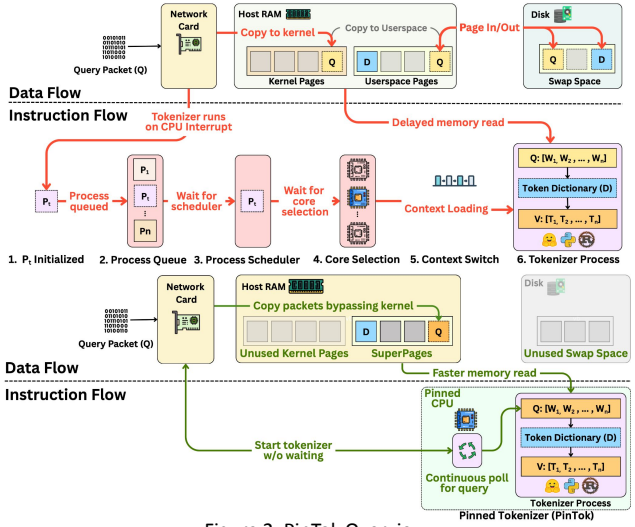
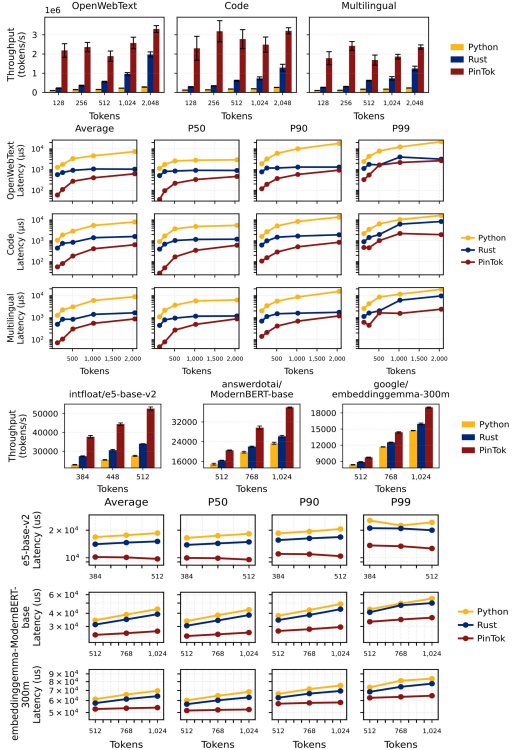


Figure 2. PinTok Overview

PinTok is an ***plug & play, algorithm-agnostic, software-only*** architecture:

- **Core & Memory Pinning (Hardware)**: Dedicates specific CPU cores to the tokenizer at boot time. Uses unswappable memory "superpages" to eliminate page-swap and maximize TLB hit rates.
- **Context Switch Avoidance (OS)**: Operates in a continuous polling mode (Run-to-Completion).
- **Zero-Copy Kernel Bypass (Networking)**: Packets are pulled directly from the NIC into user space, avoiding redundant copies and network processing delays.

3. Results



- **Throughput**: Up to **2,084%** higher throughput.
- **Latency**: Up to **95% avg. & 87% P99** latency reduction