

# Reproduction of Lightning-AI/lightning-thunder: Hugging Face LLM generation, eager vs Thunder

K-Veritas Team

*Submitted by K-Veritas Team*

kv:2610.00036v1 · Published 2026-10-11

Independent reproduction · Benchmark or leaderboard entry · Programming languages and compilers

## Abstract

Ran the quickstart examples/quickstart/hf\_llm.py at the pinned commit: greedy generation of 100 tokens with a static cache, timed with Thunder's benchmark\_n over two runs, once in PyTorch eager mode and once compiled with Thunder's hf-transformers recipe. The script's default model, meta-llama/Llama-3.2-1B, is gated, so the first public model from its own list was used instead (deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B, bf16). PyTorch 2.8 with CUDA 12.8 to match the pinned nvFuser build. One NVIDIA RTX 5060 Ti.

### About this record

Title, authors and abstract were provided by the authors and reviewed for completeness, not for scientific soundness. Every number in this record is copied from the signed data of the sealed reports [2], which remain authoritative.

This cover page is not signed. Check it against the SHA-256 listed at <https://kveritas.org/records/2610.00036v1>, and verify each report with its code bundle: `kveritas verify <report.pdf> --bundle <bundle.zip>`.

## 1 Sealed reports

The experiments were sealed with K-Veritas [1], which binds each result to the code, hardware and time that produced it.

| Report | Label                       | Sealed     | Runs | Invocations | Signer                         |
|--------|-----------------------------|------------|------|-------------|--------------------------------|
| r1 [2] | Eager vs Thunder generation | 2026-10-10 | 1    | 1           | Signed by the K-Veritas server |

## 2 Results

### r1: Eager vs Thunder generation

| Run | Metric              | Final value | Step |
|-----|---------------------|-------------|------|
| 1   | eager_generate_ms   | 1221.03     |      |
| 1   | thunder_generate_ms | 2554.79     |      |

| Run | Inline claim        | Value   | Phase |
|-----|---------------------|---------|-------|
| 1   | eager_generate_ms   | 1221.03 |       |
| 1   | thunder_generate_ms | 2554.79 |       |

### 3 Setup

| Report | Item             | Value                                           |
|--------|------------------|-------------------------------------------------|
| r1     | Operating system | linux / amd64                                   |
| r1     | CPU              | Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz, 16 cores |
| r1     | GPU              | NVIDIA GeForce RTX 5060 Ti, 16311 MiB           |
| r1     | Memory           | 31.0 GB                                         |
| r1     | Total run time   | 423.4 s                                         |

### 4 Files

Each report and code bundle must hash to these values. A bundle belongs to the report whose signed data names its hash.

| File            | SHA-256                                                          | Size   |
|-----------------|------------------------------------------------------------------|--------|
| r1.pdf          | cf2dddc7b738d11c1f9c7530b54fb9fbe163bde11ccddbe3275d5acf0bd1e8e7 | 3.7 MB |
| r1.kvbundle.zip | bd016a21116e9c54890828a70a7014367d9cb7cee9e4f75c498541741a8816b8 | 2.2 MB |

### 5 Links and licence

Reproduction of: <https://github.com/Lightning-AI/lightning-thunder/tree/62dfb0b8daeb240fd3a5615966ff0b9bb7efd710>

Record: CC-BY 4.0. Code: Apache-2.0, as declared by the authors. Tags: compiler, pytorch, llm-inference, thunder.

Listed, not endorsed. Review checks the artifact, never the science. Reports are never edited; changing them makes a new version. Other fields are corrected in place, and every correction is logged in metadata.yaml.

### References

[1] M. K. Keita and C. Homan. Computer Science Conferences Should Require Nonrepudiable Experimental Results. NeurIPS 2026 Position Paper Track. arXiv:2605.08586.

[2] Sealed experiment report r1 (Eager vs Thunder generation), kv:2610.00036v1/r1. Data hash d5f0b916cca0952bbef89c86fb15264e33d54a4fbc9b0a4b3e48186de56ffeff.

<https://kveritas.org/records/2610.00036v1/r1.pdf>