Democratising LLM Inference: Optimising MoE Architectures on Dual RTX 3090 GPUs

Draft | Internal research | Technical report · version 0.2 · updated 28 November 2026

Openference Labs

Draft: results are preliminary

This is a working draft, not a finished report. 7 sections are waiting on experiments and contain no results.

- Related work: awaiting data
- Baseline implementations: awaiting data
- Performance analysis: awaiting data
- Cost analysis: awaiting data
- Model quality evaluation: awaiting data
- Discussion: awaiting data
- Conclusion: awaiting data

Abstract

A draft technical report on serving a 26B-parameter open mixture-of-experts model on two 24 GB RTX 3090 GPUs. It documents the changes needed to run the model correctly on Ampere, the measured effect of a lower-precision MoE all-reduce and of speculative decoding, and what remains unmeasured. All results are preliminary single-repeat runs on one rented host; this draft makes no claim that has not yet got a harness and an archived raw output behind it.

Keywords: mixture of experts, LLM inference, RTX 3090, SGLang, speculative decoding, consumer GPUs

Review status

Internal research. Not peer reviewed.

Project

OFL-001: Consumer GPU inference optimisation

Funding

Funded by Openference. No external grant or sponsorship is declared for this report.

Conflicts of interest

Openference Labs is the research division of Openference, which sells API access and serves a byte-identical copy of the studied checkpoint under the name OFL-26B-Uncensored. The checkpoint was edited by a third party; Labs did not train or edit it. The benchmark host is rented from a commercial marketplace with no commercial relationship to Openference beyond the rental.

> **Draft 0.1, internal.** Not peer reviewed. Sections marked *Awaiting data* are placeholders: they exist so that the structure is fixed, and they contain no results.

Introduction

Open-weight mixture-of-experts (MoE) models activate a small fraction of their parameters for each token, which makes them attractive for hardware with limited compute but real memory. Two second-hand 24 GB GeForce RTX 3090 cards cost far less than a single data-centre accelerator. This report asks what such a pair can do as a serving platform for a recent open MoE model, and where the gaps are.

The model studied is Inkling 2 Mini (Thinking Machines Lab, Apache 2.0), in a reduced-refusal derivative made by OpenUncensored. Labs did not train or edit it; see the model database entry for the full lineage.

Problem statement

The model's reference serving path targets GPUs newer than Ampere: it assumes native NVFP4 execution, hardware FP8 conversion, a specific attention kernel and NVLink multicast. None of these exist on an RTX 3090 (compute capability 8.6). The questions are whether the model can run correctly on two such cards, how fast, at what concurrency, and which changes help.

Related work

Awaiting data. This section will cite prior work on MoE serving, speculative decoding and low-precision inference. No citations are listed yet and none are implied.

Research objectives

1. Characterise what two 24 GB Ampere GPUs can serve for a 26B-parameter MoE model, publishing every configuration and measurement.
2. Close the gaps between the engine's assumptions and Ampere, with a numerical-equivalence check for each change.
3. Measure cost and energy per token on affordable hardware.

Methodology

Benchmarks are closed-loop concurrency sweeps against a warm server using the engine's own serving benchmark, with fixed seeds, a recorded launch command and environment, and an unrecorded warm-up pass before each measurement. Numerical changes are gated by comparing prompt log-probabilities and greedy continuations between configurations on a deterministic server. Every run is stored with its provenance (measured, estimated or external) and maturity (preliminary or validated). Comparisons are only made between runs that share a methodology comparability group.

Hardware configuration

Two GeForce RTX 3090 cards (24 GB GDDR6X, 936 GB/s each, compute capability 8.6) on a rented single node with no NVLink bridge. The chipset does not support peer-to-peer between the GPUs (topology `NODE`), so NCCL traffic goes through host memory at about 5.1 GB/s. CPU model, system memory, PCIe link generation and width, power limits, clocks and driver version **were not recorded**; they are marked as not recorded in the dataset and a `collect-env` script will capture them before any result is validated. See the hardware configuration `hw-2x3090-pcie-host-a`.

Software configuration

SGLang `main` at commit `37a52b4427` plus ten Labs patches on branch `ofl-sm8x` (and one experimental patch for speculative decoding), PyTorch 2.14.1 for CUDA 13.2, nvcc 12.9, Python 3.12. Tensor parallelism 2, Triton attention, page size 128, 32,768-

token context, BF16 KV cache. The checkpoint is loaded as released (expert weights NVFP4 and MXFP8, remainder BF16 and FP32) and experts run through Marlin weight-only dequantisation. See the software configurations `sw-sglang-sm8x-e02-*` and `sw-sglang-sm8x-e03-*`.

Dataset and workload description

Two workloads are used. The fixed-length workload sends random token prompts of exactly 512 tokens and generates exactly 256 tokens; it isolates the engine from content but has no linguistic structure. The chat workload samples ShareGPT conversations with a 256-token output cap and is the only workload on which speculative decoding can be assessed. Both run at concurrency 1, 4, 16 and 64 with at least 16 requests per level.

Experimental setup

Each configuration is started fresh, warmed with an unrecorded pass at every concurrency level, then measured once. Server and client run on the same host over loopback. The same launch command is used for every configuration except the single factor under test.

Baseline implementations

Awaiting data. A comparison with a second serving engine on the same weights and host exists in the Labs notes but has not yet been archived as a dataset run, so it is not reported here.

Results

Preliminary. The runs behind these statements are in the dataset under project OFL-001, experiments E02 and E03.

- **E02, bf16 MoE-combine all-reduce.** On the fixed-length workload, aggregate output throughput changed by about −0.3% at concurrency 4 (214.4 to 213.8 tok/s), rose 16% at concurrency 16 (393.8 to 458.6) and 5.5% at concurrency 64 (493.5 to 520.4). Single-stream throughput and inter-token latency are unchanged.
- **E03, DSpark speculative decoding.** On the chat workload, single-stream throughput rose 19% (75.0 to 89.2 tok/s). At concurrency 4, 16 and 64 aggregate throughput fell (214.9 to 187.3, 482.6 to 300.8 and 383.8 to 323.7 tok/s). Mean accept length was 2.41 tokens per verification step.
- **E04, KV pool sizing.** On the chat workload without speculation, enlarging the sliding-window KV pool from about 29 thousand to between 42 and 94 thousand tokens left throughput at 64 concurrent requests at 395 to 404 tok/s, against 384 tok/s at the default pool. The KV pool was not the bottleneck. The engine had captured decode CUDA graphs only up to batch size 48, so decode at 64 requests ran eagerly. A follow-up with a higher cap is recorded in the notebook at 828 tok/s at 64 requests; its raw output is not archived, so it is not reported here as a result.

Each figure is a single run. No spread is available yet. The E02 and E03 results at 64 concurrent requests, and possibly the speculative-decoding results at four or more, were measured with the default decode CUDA-graph cap and may change when it is raised.

Experiment E02, concurrency sweep. Host: 2x GeForce RTX 3090 24 GB, PCIe only (no NVLink bridge), rented combine all-reduce. treatment: bf16 MoE-combine all-reduce.Each point is a single run (1 repeat).

Concurrency	Output tok/s baseline	Output tok/s treatment	Change treatment vs baseline	Total tok/s baseline
1	73.0	73.2	+0.3%	219.0
4	214.4	213.8	−0.3%	643.2
16	393.8	458.6	+16.5%	1,181.5
64	493.5	520.4	+5.5%	1,480.4

Measured · Preliminary · Click a throughput figure for the full run, its limitations and its raw output.

Equivalence check: Prompt-logprob equivalence gate, 3 fixed prompts. Run details.

Measure	Value	Unit
Prompt 1: prompt tokens	2,061	tokens
Prompt 1: mean absolute prompt-logprob difference	0.053	nats
Prompt 1: maximum absolute prompt-logprob difference	1.444	nats
Prompt 1: top-1 token agreement	0.980	fraction
Prompt 1: perplexity, fp32 combine	2.499	ppl
Prompt 1: perplexity, bf16 combine	2.486	ppl
Prompt 1: greedy continuation identical (1 = yes)	0	bool
Prompt 2: prompt tokens	1,794	tokens
Prompt 2: mean absolute prompt-logprob difference	0.062	nats
Prompt 2: maximum absolute prompt-logprob difference	1.898	nats
Prompt 2: top-1 token agreement	0.966	fraction
Prompt 2: perplexity, fp32 combine	2.240	ppl
Prompt 2: perplexity, bf16 combine	2.247	ppl
Prompt 2: greedy continuation identical (1 = yes)	1	bool
Prompt 3: prompt tokens	607	tokens
Prompt 3: mean absolute prompt-logprob difference	0.010	nats
Prompt 3: maximum absolute prompt-logprob difference	0.522	nats
Prompt 3: top-1 token agreement	1	fraction
Prompt 3: perplexity, fp32 combine	1.171	ppl
Prompt 3: perplexity, bf16 combine	1.172	ppl
Prompt 3: greedy continuation identical (1 = yes)	1	bool

Experiment E03, concurrency sweep. Host: 2x GeForce RTX 3090 24 GB, PCIe only (no NVLink bridge), rented
new tokens cap, closed loop. baseline: no speculation. treatment: DSpark speculative decoding with 8 draft t

Concurrency	Output tok/s baseline	Output tok/s treatment	Change treatment vs baseline	TTFT mean ms baseline
1	75.0	89.2	+19.0%	198
4	214.9	187.3	-12.9%	328
16	482.6	300.8	-37.7%	1,424
64	383.8	323.7	-15.7%	6,223

Measured Preliminary Click a throughput figure for the full run, its limitations and its raw output.

Equivalence check: Greedy continuation agreement, 6 prompts x 256 tokens. Run details.

Measure	Value	Unit
Continuations identical to the non-speculative run (of 6, 256 tokens each)	2	prompts
Prompt 1: first divergence	1	token index
Prompt 1: top-1 minus top-2 margin at divergence, non-speculative	0.028	nats
Prompt 1: top-1 minus top-2 margin at divergence, speculative	0.200	nats
Prompt 2: first divergence	11	token index
Prompt 2: top-1 minus top-2 margin at divergence, non-speculative	0.044	nats
Prompt 2: top-1 minus top-2 margin at divergence, speculative	0.085	nats
Prompt 3: first divergence	1	token index
Prompt 3: top-1 minus top-2 margin at divergence, non-speculative	0.107	nats
Prompt 3: top-1 minus top-2 margin at divergence, speculative	0.044	nats
Prompt 5: first divergence	166	token index
Prompt 5: top-1 minus top-2 margin at divergence, non-speculative	0.078	nats
Prompt 5: top-1 minus top-2 margin at divergence, speculative	0.139	nats
Maximum prompt-logprob difference, speculative server vs non-speculative server (3 prompts)	0	nats

Experiment E04, concurrency sweep. Host: 2x GeForce RTX 3090 24 GB, PCIe only (no NVLink bridge), rented speculation, larger KV pool (--mem-fraction-static 0.85). m85r30: no speculation, larger KV pool (--mem-fract swa-full-tokens-ratio 0.6).Each point is a single run (1 repeat).

Concurrency	Output tok/s m85	Output tok/s m85r30	Output tok/s m85r60	TTFT mean ms m85	T
16	471.4	471.0	470.5	1,611	
64	394.5	400.9	403.9	5,890	

Measured · Preliminary · Click a throughput figure for the full run, its limitations and its raw output.

Performance analysis

Awaiting data. Interpretation requires repeats and a profile of the speculative path.

Cost analysis

Awaiting data. The economics model is published, but no power was sampled during load, so no energy or cost-per-token figure exists for these runs.

Model quality evaluation

Awaiting data. Only numerical-equivalence gates exist: three prompts for the bf16 combine (mean absolute prompt-logprob difference 0.01 to 0.06 nats, top-1 agreement 96.6% to 100%) and six prompts for speculation (two of six greedy continuations identical, the rest diverging at near-ties). These do not constitute a quality evaluation.

Limitations

- One rented host with incomplete metadata; no second machine.
- One run per point; spread unknown; nothing is validated.
- Random-token workload for engine changes; one natural-text workload for speculation.

- Quality gates are weak: three and six prompts, no task evaluation.
- One unexplained start-up crash, not reproduced; a soak test is planned.
- E02 and E03 used the default decode CUDA-graph cap (batch size 48). E04 showed this makes 64-request decode run eagerly, so their large-batch results may change; the re-run is planned (experiment G01).
- The follow-up that raises the cap has no archived raw output and is not reported as a result.

Discussion

Awaiting data.

Future work

Planned experiments include re-running E02 and E03 with decode CUDA graphs covering the tested batch sizes, a soak test, a task-level quality evaluation against a reference engine on the same weights, prefill cost reduction, KV cache options and context length, an NVLink-bridge comparison and power and cost measurement. See the project page for status.

Conclusion

Awaiting data. No conclusion is drawn from preliminary single-run results.

References

1. Thinking Machines Lab. *Inkling 2 Mini* model card. https://huggingface.co/thinkingmachines/Inkling-2-Mini (https://huggingface.co/thinkingmachines/Inkling-2-Mini) (retrieved 2026-11-28).
2. OpenUncensored. *Inkling 2 Mini Uncensored* model card. https://huggingface.co/OpenUncensored/Inkling-2-Mini-Uncensored (https://huggingface.co/OpenUncensored/Inkling-2-Mini-Uncensored) (retrieved 2026-11-28).

3. NVIDIA. *NVIDIA Ampere GA102 GPU Architecture* whitepaper v2. https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf (https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf) (retrieved 2026-11-28).
4. NVIDIA. *GeForce RTX 3090* product page. https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3090/ (https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3090/) (retrieved 2026-11-28).
5. SGLang project. https://github.com/sgl-project/sglang (https://github.com/sgl-project/sglang)

Reproducibility instructions

The exact launch arguments, environment variables, patch list and commit are in the software configurations. Benchmark commands and filtered raw output are archived with each run. The harness scripts and the patch series are not yet published; until they are, these results cannot be reproduced from this site alone.

Code and dataset availability

Dataset: the benchmark dataset, with a versioned export and checksums, on the Benchmarks page. Code: the patch series and harness are planned for release under the open-source programme; no repository is public yet. The dataset licence is pending a decision.

Conflict-of-interest and funding disclosures

Openference Labs is the research division of Openference, which sells API access and serves a byte-identical copy of the studied checkpoint under the name OFL-26B-Uncensored. Labs did not train or edit the checkpoint. The benchmark host is rented from a commercial marketplace. Funded by Openference; no external funding is declared.

Cite this publication

Openference Labs (2026). Democratising LLM Inference: Optimising MoE Architectures on Dual RTX 3090 GPUs. Openference Labs, draft, version 0.2. https://labs.openference.com/publications/democratising-llm-inference-moe-dual-rtx-3090/

```
@techreport{labs2026democratising,
  title       = {Democratising LLM Inference: Optimising MoE Architectures on Dual RTX
  author      = {Openference Labs},
  institution = {Openference Labs},
  year        = {2026},
  month       = {11},
  url         = {https://labs.openference.com/publications/democratising-llm-inference-
  note        = {Draft. Version 0.2. Not peer reviewed}
}
```