

# THE PASCAL FRONTIER

MODERN AI INFERENCE ON HARDWARE  
THAT APPARENTLY DIDN'T GET  
THE RETIREMENT MEMO.

IT'S NOT OLD.  
IT'S VINDICATED.

**MOSTLY FOR ENGINEERS.**

Translated for family, friends & curious humans.

**WHITEPAPER #1**

**vLLM vs. llama.cpp**

14B · 32B · 35B-CLASS MoE

UNSUPPORTED?  
Cute. ☺

# TABLE OF CONTENTS

THE JOURNEY

|    |                                                                                                  |    |
|----|--------------------------------------------------------------------------------------------------|----|
| 0  | HUMANS, AI, AND THE RECEIPTS<br>Authority, model lineage, and auditability                       | 03 |
| 1  | A LONG TIME AGO, IN A DATACENTER NOT SO FAR AWAY...<br>Family briefing on 04; engineering begins | 05 |
| 2  | YOU KEEP USING THAT WORD: "UNSUPPORTED"<br>Capability versus software support                    | 07 |
| 3  | ROADS? WHERE WE'RE GOING, WE NEED CUDA<br>The modern stack and its tripwires                     | 07 |
| 4  | FINE. WE'LL BUILD IT OURSELVES.<br>vLLM .019 / .021 / .023 qualification                         | 08 |
| 5  | HOUSTON, WE HAVE AN ENGINECORE PROBLEM<br>Failures, regressions, and boundaries                  | 12 |
| 6  | TWO ENGINES ENTER, ONE V100 LEAVES<br>Comparison method                                          | 13 |
| 7  | THE LITTLE ENGINE THAT APPARENTLY COULD<br>llama.cpp compatibility                               | 14 |
| 8  | RELEASE THE MODELS<br>14B, 32B, 35B-class workloads                                              | 15 |
| 9  | SHOW ME THE TOKENS<br>Measured results                                                           | 16 |
| 10 | WAIT... IT RAN WHAT?<br>Interpretation; Data Decoders follow                                     | 19 |
| 11 | TO BE CONTINUED...<br>WP2-WP4: engines, boundaries, and the final showdown                       | 23 |

APPENDICES    Commands · Evidence · Claim Ledger · Founder Review

Pacing authority: numbered color bands, technical panels, comic punctuation, and minimal dead air.

|                                                        |                                                   |                                                    |                                                                                       |
|--------------------------------------------------------|---------------------------------------------------|----------------------------------------------------|---------------------------------------------------------------------------------------|
| <p>THE MISSION</p> <p><i>Evidence before hype.</i></p> | <p>THE CAST</p> <p><i>Scott · Ren · Lurch</i></p> | <p>THE HARDWARE</p> <p><i>1 × V100 · 32 GB</i></p> |  |
|--------------------------------------------------------|---------------------------------------------------|----------------------------------------------------|---------------------------------------------------------------------------------------|



0

# Humans, AI, and the Receipts

The machine ran the benchmark. Humans retained authority.

BUILT FOR ENGINEERS.

TRANSLATED FOR  
EVERYONE ELSE.

The commands, measurements, failures, and evidence remain engineering-grade. Each major section ends with a short everyday explanation so family, friends, leaders, and curious readers can follow the plot without already knowing CUDA.

## HUMAN AUTHORITY

**J. Scott Henderson** directed the research, defined the engineering question, supplied and reviewed the evidence, and held final authority over every coding, editorial, publication, and claim decision.

- Reviewed each proposed coding change through PDS guardrails.
- Approved, rejected, or redirected publication decisions.
- Classified what the record supports—and what remains a gap.

## AI ROLE

**Ren-AI / ChatGPT and Codex** assisted with research, log analysis, code and document drafting, evidence collation, visual production, and quality checks.

**NO AUTONOMOUS APPROVAL** AI output was advisory. It did not independently authorize code, claims, releases, or publication.

| WORK PHASE                                    | CHATGPT / CODEX VERSION | ASSISTED ROLE                                                       |
|-----------------------------------------------|-------------------------|---------------------------------------------------------------------|
| Exploration, research, and working discussion | ChatGPT 5.5 Instant     | Research support, drafting, review assistance                       |
| Whitepaper visual and technical production    | Codex 5.6 Sol High      | Layout authority analysis, technical synthesis, creative direction  |
| Evidence collation and transparency revision  | Codex 5.6 Terra Medium  | Archive inventory, duplicate analysis, sanitization plan, manifests |
| Render QA and publication packaging           | Codex 5.6 Sol Light     | Final visual verification, packaging, and consistency checks        |

Version labels record the production workflow used for this Founder Review build. Models assisted; they did not replace the Founder’s judgment.

## AUDIT TRAIL

Technical claims are classified as **Observed, Analysis, Conclusion, Hypothesis, or Planned**. Supporting run evidence is mapped in the Claim / Evidence Ledger and will be released as a privacy-sanitized public projection with hashes, provenance, retained failures, and a redaction log.

Founder direction → AI-assisted analysis → Founder review  
Evidence record → claim classification → publication gate  
No evidence → recorded gap; never reconstructed fact



SCOTT / FOUNDER

“I approve the claim—or send it back.”

### THE GUARDRAIL

AI PROPOSES.  
THE EVIDENCE ANSWERS.  
THE HUMAN DECIDES.

CODE

CLAIMS

PUBLICATION



REN-AI / ADVISOR

“Receipts attached. Ego detached.”

Governance basis: PDS AI engagement, human authority, and Documentation & Editorial Standard. This page is a production disclosure—not a substitute for the underlying evidence.

HUMAN TRANSLATION · WHO IS DRIVING?

Think of AI as the navigation system: it can suggest the route and spot traffic, but Scott keeps both hands on the wheel and decides where the car goes.

BEFORE THE ENGINEERING STARTS...

# SO, WHAT ARE WE ACTUALLY DOING HERE?

*A two-minute briefing for the people who love an engineer—but do not want to become one just to follow the story.*



## 1 MEET LURCH

Lurch is Scott's home-lab computer. Its V100 graphics card was designed years ago for serious computing work.

MILLENNIUM FALCON RULE: old does not mean finished.

## 2 THE MYSTERY

Modern AI software increasingly says this older card is "unsupported." Scott wanted to know whether the hardware was truly incapable—or merely left behind by newer software.

THE TV STILL WORKS. THE APP STOPPED SUPPORTING IT.

## 3 THE TEST

We gave Lurch several AI models and ran them through two inference engines: software that turns a prompt into generated answers.

SAME CAR. TWO TRANSMISSIONS. REAL ROAD TEST.

## 4 THE RECEIPTS

We measured completions, waiting time, speed, heat, power, and failures. We kept the bad runs along with the good ones.

NO "TRUST ME, BRO." THE LOGS GET A VOTE.



SCOTT

Designs the experiment and makes every final call.



REN-AI

Helps research, analyze, explain, and package the evidence.



LURCH

Runs the work, heats the room, and keeps the receipts.

## MOSTLY FOR ENGINEERS. TRANSLATED FOR EVERYONE ELSE.

Engineers get the full commands, tables, and evidence. Family, friends, and curious humans get bold plain-English translations and three Data Decoder intermissions. You can follow the plot without knowing CUDA, Linux, or what a token is.

THE QUESTION TO CARRY WITH YOU: *Did the old hardware fail—or did modern software simply stop making room for it?*

### YOUR ONLY JOB:

Follow the mystery. When a chart appears, look for **IN PLAIN ENGLISH**. Ren already fought the acronyms.





1

A Long Time Ago, in a Datacenter Not So Far Away...  
Meet Lurch—and the question modern software left behind



ONE OLD SERVER. ONE MODERN QUESTION.

Lurch is a home-lab engineering platform built around an NVIDIA Tesla V100 PCIe 32 GB accelerator: Volta, compute capability 7.0. This is the accelerator qualified throughout WP1.

**THE QUESTION** Does “unsupported” describe what the hardware cannot do—or what current software no longer chooses to support?

CANONICAL PLATFORM

| FIELD            | OBSERVED                      |
|------------------|-------------------------------|
| GPU              | Tesla V100 PCIe 32 GB         |
| Architecture     | Volta / SM70                  |
| Attention target | TRITON_ATTN                   |
| Role             | Local inference qualification |

WHY TEST IT

- Extend useful hardware life.
- Separate compatibility from capability.
- Build repeatable qualification methods.

WHAT WP1 PROVES

- Modern vLLM lanes can run on SM70.
- 14B and 32B workloads can qualify.
- llama.cpp crosses a 35B boundary.

WHAT IT DOES NOT

- V100 rivals current GPUs.
- Cheap purchase means cheap operation.
- One machine proves an industry thesis.

OBSERVED E-HW-01

Figure 1. NVIDIA Tesla V100 PCIe 32 GB installed in Lurch. Canonical Project Dominion Studios hardware reference.

HUMAN TRANSLATION · MEET THE MACHINE

Lurch is the Millennium Falcon of this story: older, unfashionable, and still perfectly capable of surprising newer ships when the right pilot and parts are involved.



LURCH / FIELD NOTE

Old hardware. *New tricks.* Keep up.

VOLTA

SM70

32 GB

LURCH // FIELD SYSTEM

1.1

Lurch / Canonical Field Dossier

ANTON.AOS HOME baseline joined to the WP1 qualification record



Figure 2. Founder-supplied Lurch character image. The 23.8 front-panel display remains visible; its unit and sensor placement are not proven by the photograph.

LURCH

HOME-LURCH-REV-A · CANONICAL BASELINE

| FIELD          | CANONICAL VALUE                                    | STATE                                                                                                |
|----------------|----------------------------------------------------|------------------------------------------------------------------------------------------------------|
| Host           | lurch                                              | Current formal hostname; repository evidence record needs synchronization                            |
| OS             | Ubuntu 24.04.4 LTS                                 | Verified repository snapshot                                                                         |
| Kernel         | 7.0.0-28-generic                                   | Verified snapshot; not a required pin                                                                |
| Board / BIOS   | ASUS ROG STRIX B450-F / 6001                       | Verified                                                                                             |
| CPU            | AMD Ryzen 5 3600 · 6C / 12T                        | Verified                                                                                             |
| Memory         | 64 GB Corsair Vengeance LPX DDR4                   | Four 16 GB modules founder-confirmed; family and DDR4 photo-verified                                 |
| Memory modules | 4 × 16 GB · CMK32GX4M2E3200C16 · 3200 MHz · 1.35 V | Two 32 GB (2 × 16 GB) kits; label details photo-verified                                             |
| Storage        | Samsung SSD 980 NVMe + 2 × 14 TB HDD               | SSD 980 installed model photo-verified; NVMe temperature verified; two 14 TB drives founder-reported |
| Display GPU    | AMD Radeon RX 580                                  | Verified; display-only role                                                                          |
| Compute GPU    | Tesla PG500-216 / V100 PCIe                        | Verified                                                                                             |
| VRAM / ISA     | 32,768 MiB HBM2 · Volta SM70                       | Verified                                                                                             |
| Driver         | NVIDIA 580.173.02                                  | Verified through nvidia-smi                                                                          |



SCOTT  
So... what OS is Lurch running?



REN-AI  
Ubuntu 24.04.4 LTS. Canonical record located. No guessing required.

LURCH: "RECEIPTS. I LIKE RECEIPTS."

VERIFIED HOST → GPU PATH

| LAYER                | CANONICAL STATE                    |
|----------------------|------------------------------------|
| Driver compatibility | CUDA 13.0 maximum reported         |
| Docker / Compose     | Functional; exact versions pending |
| NVIDIA toolkit       | 1.19.1-1 packages                  |
| Container GPU        | V100 visible in CUDA container     |
| CUDA smoke           | nbody completed on SM70            |

**DO NOT CONFUSE THE TWO** nvidia-smi CUDA 13.0 is driver compatibility —not proof of a CUDA 13 host toolkit. WP1 builds use the separate CUDA 12.6 qualification lane.

WP1 QUALIFICATION OVERLAY

| FIELD      | WP1 VALUE                           |
|------------|-------------------------------------|
| Python     | 3.12.3                              |
| Build CUDA | 12.6                                |
| Attention  | TRITON_ATTN                         |
| vLLM lanes | .019 / .021 / .023                  |
| Chassis    | Antec tower; exact model unverified |

**SOURCE BOUNDARY** Host facts: canonical ANTON.AOS Lurch baseline plus founder-confirmed hostname and installed memory count. Memory label details come from canonical repository photos. Benchmark-lane facts: WP1 evidence package.

OBSERVED PDS-KO-ANTON-LURCH-HARDWARE-0001 / E-COMPAT-REPORT-01

HUMAN TRANSLATION · WHAT IS UNDER THE HOOD?

This is Lurch's window sticker and service record: the exact engine, operating system, memory, and installed parts behind every claim in the paper.



2

# You Keep Using That Word: “Unsupported”

A support label is a stack trace, not a single fact

## THE SIX-LAYER PROBLEM

| LAYER         | FAILURE MODE      | WPI EXAMPLE             |
|---------------|-------------------|-------------------------|
| Hardware      | Missing feature   | No standard FA2         |
| Toolchain     | Target omitted    | SM70 excluded           |
| Dependency    | Backend removed   | xFormers drops V100     |
| Framework     | Fallback mismatch | Unusable attention path |
| Model loader  | Architecture gap  | 35B GGUF lane fails     |
| Runtime state | Contamination     | 27 GB stale VRAM        |

## THE QUICK VERSION



A GPU can remain computationally capable while one layer above it refuses to cooperate.

**LITERAL MECHANISM** vLLM 0.10.2 compiled, loaded, and initialized EngineCore on SM70—then failed at first inference because its selected attention path had no viable Volta operator.



The hardware didn't quit.  
The fallback did.

3

# Roads? Where We're Going, We Need CUDA

The route from prompt to token—and where Volta can fall off

| MODEL                            | FRAMEWORK         | ATTENTION                                    | PYTORCH + TRITON  | CUDA 12.6                        | DRIVER + GPU           |
|----------------------------------|-------------------|----------------------------------------------|-------------------|----------------------------------|------------------------|
| HF / GGUF · FP16 / GPTQ / Q4_K_M | vLLM or llama.cpp | TRITON_ATTN avoids the FA2/xFormers dead end | Kernel generation | SM70 compilation target retained | V100 executes the work |

```
$ export TORCH_CUDA_ARCH_LIST=7.0
$ export CUDA_HOME=/usr/local/cuda-12.6
$ export VLLM_ATTENTION_BACKEND=TRITON_ATTN
✓ make every compatibility decision explicit
```

**BOUNDARY** CUDA 13 removes Volta compilation and library support. The reproducible lane remains in CUDA 12.x.

**ANALYSIS** E-COMPAT-REPORT-01

**HUMAN TRANSLATION ·**  
“UNSUPPORTED” IS NOT “BROKEN”

Your television can still work after a streaming app stops supporting it. The screen did not fail; one layer of software stopped speaking its language.



LURCH / SUPPORT DESK

“Unsupported” sounds suspiciously like a software problem.

HARDWARE READY

STACK TRACE

FIND THE LAYER

LURCH // FIELD SYSTEM

4

Fine. We'll Build It Ourselves.

Three isolated vLLM lanes; one GPU; zero baseline vandalism

| vLLM       | PYTORCH      | CUDA | ATTENTION   | STATUS |
|------------|--------------|------|-------------|--------|
| 0.19.1.dev | 2.10.0+cu126 | 12.6 | TRITON_ATTN | ✓ PASS |
| 0.21.1.dev | 2.11.0+cu126 | 12.6 | TRITON_ATTN | ✓ PASS |
| 0.23.1.dev | 2.11.0+cu126 | 12.6 | TRITON_ATTN | ✓ PASS |

BUILD CONTRACT

- Separate versioned environments.
- Explicit SM70 binary-profile checks.
- Pinned CUDA/PyTorch combinations.
- Smoke before benchmark.
- Preserve failures and log paths.

WHY THREE LANES

.019 was the strongest community-backed baseline. .021 retained newer-model compatibility. .023 tested whether the same approach survived further upstream change.

ARCHITECTURAL DECISION

Maintain a reproducible Volta build profile first. Accept a dedicated fork only when features cannot remain isolated.

THE TRIPWIRES

1. New kernels assume newer GPU features.
2. Dependencies remove older operators.
3. Model loaders take engine-specific paths.
4. Dirty VRAM turns a test into fiction.

Marlin said  
"Nope."

SM80 ONLY?  
NOT TODAY.

RESULT All three lanes loaded models and generated non-empty output.



OBSERVED E-VLLM-SMOKE-XVER

HUMAN TRANSLATION · WHY BUILD THREE VERSIONS?

It is the same car on three carefully separated test tracks. If one setup fails, Scott can identify the changed part instead of blaming the whole vehicle.



LURCH / BUILD POLICY

Three lanes. One GPU. No baseline vandalism.

.019 .021 .023

LURCH // FIELD SYSTEM

41

# Build Once. Prove Twice.

Qualification gates and the small-model plumbing proof

## QUALIFICATION GATES

| GATE           | PASS CONDITION                |
|----------------|-------------------------------|
| Environment    | Capture exact versions        |
| Binary profile | Confirm SM70 path             |
| Idle-state     | Reject stale workloads        |
| Smoke          | Load + first generation       |
| Concurrency    | Exercise schedule/decode      |
| Telemetry      | Baseline, average, peak       |
| Cooldown       | Return near settled reference |

## PROM FROM THE LAB NOTEBOOK

```
STATE: SERVER STARTING
API server ready
STATE: CLIENT STARTING
CONCURRENT_RESULT: total=1000
completed=1000 | failures=0
✓ CONCURRENT_TEST_RESULT=PASS
```

Scott

Hit me with update, and fire away...

Ren

The server is fired up and ready to roll.

## OPT-125M · 100-user cross-version throughput

|      |                                                                                    |                |
|------|------------------------------------------------------------------------------------|----------------|
| .019 |  | 2,148.19 tok/s |
| .021 |  | 2,151.21 tok/s |
| .023 |  | 2,117.28 tok/s |

.019

**281.46** smoke out tok/s  
**0.2409 s** p95  
**36.3 W** avg

.021

**276.05** smoke out tok/s  
**0.2417 s** p95  
**42.7 W** avg

.023

**273.60** smoke out tok/s  
**0.2449 s** p95  
**44.1 W** avg

**READ THIS CORRECTLY** The small model validated the machinery. It was not promoted as the headline benchmark.

SMOKE

LOAD

CONCURRENCY

TELEMETRY

COOLDOWN

PASS

OBSERVED E-VLLM-CONCUR-SMALL

## HUMAN TRANSLATION · THE SMALL TEST MODEL

Before towing a camper across the country, you start the truck, test the brakes, and drive around the block. This page proves the test rig itself works.



## LURCH / QUALIFICATION ORDER

Load it. Stress it. Cool it. *Then believe it.*

SMOKE

1,000 REQUESTS

PASS

LURCH // FIELD SYSTEM

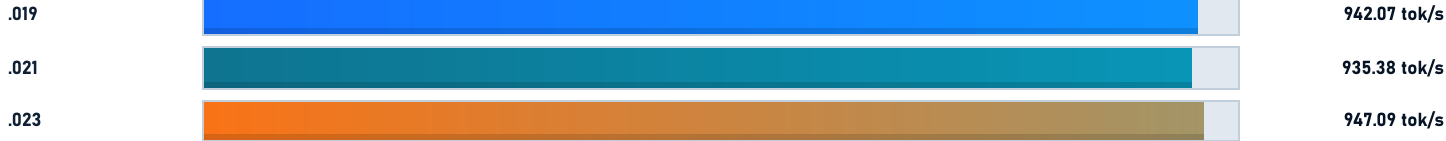


4.2

# Fifty Users Walk Into a V100...

Qwen2.5-14B cross-version evidence

## TOTAL TOKEN THROUGHPUT · 50 USERS × 5 REQUESTS



| LANE | IN TOK/S | OUT TOK/S | TOTAL  | P50 S  | P95 S   | AVG W | SCORE |
|------|----------|-----------|--------|--------|---------|-------|-------|
| .019 | 791.87   | 150.20    | 942.07 | 8.3007 | 12.8486 | 147.2 | 99.44 |
| .021 | 786.25   | 149.13    | 935.38 | 8.4669 | 12.6442 | 154.1 | 98.22 |
| .023 | 796.09   | 151.00    | 947.09 | 8.2507 | 12.7924 | 152.3 | 99.44 |

## PDS RECOMMENDED

.023

Highest total throughput and best median latency. .019 tied the composite score and used less average power; the tie-break favored the newer validated lane.

HONEST FOOTNOTE "Recommended" means best fit for this tested configuration—not universal superiority.

## WORKLOAD

250 total requests  
random 512–1,536 input  
64–256 output

## RESULT

250 / 250 each lane  
zero observed failures  
all lanes PASS

## WHAT CHANGED

Software versions changed.  
GPU and workload contract  
did not.

## THE VERDICT

.023 TAKES THE LANE.  
THE OTHER TWO  
REFUSE TO DIE.

Three lanes enter.  
All three work.

OBSERVED E-VLLM-14-XVER

HUMAN TRANSLATION · FIFTY USERS AT ONCE Picture fifty restaurant tables ordering in waves. All three kitchens served every order; version .023 finished the combined rush by a small margin.



## LURCH / BENCHMARK VERDICT

All three lanes work. .023 takes the chair.

942.07 935.38 947.09 TOK/S

LURCH // FIELD SYSTEM

4.3

# Big Mama Finally Made Lurch Work

Practical 14B load—and the 32B capability proof

## 14B · PRACTICAL SWEET SPOT

| METRIC          | OBSERVED                   |
|-----------------|----------------------------|
| Completed       | 40 / 40                    |
| Throughput      | 483.94 tok/s               |
| p50 / p95       | 14.1341 / 20.0607 s        |
| GPU utilization | 71.6% avg / 100% peak      |
| Power           | 117.5 W avg / 242.8 W peak |
| Temperature     | 42.9°C avg / 56.0°C peak   |

**RESULT** Sustained qualification, not merely a first-token demo.

## 32B · CAPABILITY PROOF

| METRIC          | OBSERVED                   |
|-----------------|----------------------------|
| Model           | Qwen2.5-32B Q4_K_M         |
| Completed       | 40 / 40                    |
| Throughput      | 89.35 tok/s                |
| p50 / p95       | 83.0012 / 119.5556 s       |
| GPU utilization | 80% avg / 100% peak        |
| Power           | 181.9 W avg / 257.0 W peak |
| Temperature     | 66.2°C avg / 81.0°C peak   |

**TRADEOFF** Reliable service—with steep latency and thermal cost.

14B

THE OPERATIONAL ANSWER

32B

THE ENGINEERING ANSWER

40 REQUESTS

FULL GPU

ZERO FAILURES

**LURCH  
THREW DOWN  
THE GAUNTLET.**

**BOTTOM LINE** 14B was the practical sweet spot. 32B was the capability proof. Both statements can be true at the same time.

INSIDE LURCH

**THE V100 HITS 100%.  
THE THERMALS  
START NEGOTIATING.**



**OBSERVED** E-VLLM-14-LOAD / E-VLLM-32-GGUF

**HUMAN TRANSLATION · 14B VERSUS 32B**

The 14B model is the dependable daily pickup. The 32B model is the overloaded moving truck: it completed the trip, but slowly, hotter, and with the engine working much harder.

LURCH / OPERATING ENVELOPE

14B for work. 32B for proof.

40/40

FULL GPU

ZERO FAILURES

LURCH // FIELD SYSTEM

5

# Houston, We Have an EngineCore Problem

Failures, regressions, fallback paths, and boundaries

I

## FAILED TESTS ARE PART OF THE RESULT.

| FAILURE                | BOUNDARY                                    | ENGINEERING RESPONSE            |
|------------------------|---------------------------------------------|---------------------------------|
| 0.10.2 first inference | Unusable Volta attention fallback           | Move to Triton lane             |
| 32B GPTQ start         | EngineCore failed; excerpt lacks root cause | Do not invent a deeper cause    |
| 32B GGUF dirty state   | ~27 GB stale VRAM                           | Add idle gate; rerun clean      |
| 35B-class under vLLM   | Loading path did not qualify                | Dedicated WP3 lane              |
| 35B llama.cpp, 4K      | 68/80; context overflow                     | Raise context; preserve failure |

Initial reference: VRAM=27003 MiB  
Pre-lane state: VRAM=32404 MiB  
FAIL: API server did not become ready  
✓ test environment was not idle

### THE RULE

A retry can correct test state. It cannot erase the failed run.

### THE TEMPTATION

"It passed later" is not permission to quietly delete why it failed first.

THE LOG  
REMEMBERS.

**EVIDENCE DISCIPLINE** Preserve the failure, change one condition, record the new result, and explain the boundary.

FAILURE IS DATA

THE RUN ENDED.  
THE EVIDENCE  
DID NOT.

HUMAN TRANSLATION · WHY KEEP  
THE FAILURES?

A failed run is the skid mark at an accident scene. Cleaning it away might look nicer, but it destroys the clue that explains what happened.

LURCH / BLACK BOX

Failure is data. The log remembers.

PRESERVE

CHANGE ONE THING

RERUN

LURCH // FIELD SYSTEM



## PART B

# TWO ENGINES ENTER. ONE V100 LEAVES.

## INITIAL vLLM vs. llama.cpp COMPARISON

### RULE OF ENGAGEMENT

Compare like with like only when model, workload, context, concurrency, and measurement contract align. Otherwise say why they are unlike.

*“Same GPU” is not the same thing as  
“same experiment.”*

#### vLLM

High-throughput serving

#### llama.cpp

Hardware + model-format reach

HUMAN TRANSLATION · TWO  
ENGINES, ONE CAR

vLLM and llama.cpp are two different transmissions bolted to the same vehicle. The fair question is not simply “which is faster?” but “which one fits this particular job?”

6

# Two Engines Enter, One V100 Leaves

One platform; two runtime philosophies

COMPARISON CONTRACT

| DIMENSION        | vLLM                         | llama.cpp                      |
|------------------|------------------------------|--------------------------------|
| Primary strength | High-throughput serving      | Portable local inference       |
| Model ecosystem  | HF-native + selected GGUF    | GGUF-native                    |
| Volta work       | SM70 builds + attention lane | No target-model source patches |
| WP1 evidence     | 14B x-version; 14B/32B .023  | 14B/32B/35B demonstrations     |
| Direct race?     | Only when workload matches   | Many runs intentionally unlike |

METHOD

Use shared workload parameters where possible. Preserve engine-specific configuration. Label unmatched data **not directly comparable**.

MODEL   CONTEXT   REQUESTS   OUTPUT   ENGINE

**EDITORIAL RULE** A graph can be numerically correct and still be dishonest.

7

# The Little Engine That Apparently Could

llama.cpp on V100 without source compatibility patches

```
$ llama-server -m <MODEL.gguf> -ngl 999 -c 8192 --reasoning off
API: /v1/chat/completions
✓ 14B PASS · 32B PASS · 35B-class MoE PASS
```

WHAT "NO PATCHES" MEANS

- No llama.cpp source changes for these target runs.
- Not zero configuration.
- Not universal compatibility.

WHY IT MATTERS

- A different engine crossed the boundary.
- 35B separated hardware capability from the tested vLLM loading path.

OBSERVED E-LLAMA-COMPAT-01

HUMAN TRANSLATION · THE FAIR-FIGHT RULE

Two race cars sharing a track are not comparable if one carries passengers and the other runs empty. Match the load and rules before declaring a winner.



LURCH / RUNTIME DIPLOMACY

Different engine. Same V100. *New boundary.*

vLLM   llama.cpp   NO FAKE HORSE RACE

LURCH // FIELD SYSTEM

8

Release the Models

The evidence inventory—not a universal winner board

| MODEL               | RUNTIME             | WORKLOAD  | RESULT   | ROLE                |
|---------------------|---------------------|-----------|----------|---------------------|
| Qwen2.5-14B         | vLLM .019/.021/.023 | 50×5      | PASS all | Cross-version proof |
| Qwen2.5-14B         | vLLM .023           | 8×5       | 40/40    | Practical load      |
| Qwen2.5-32B Q4_K_M  | vLLM .023           | 8×5       | 40/40    | Qualification       |
| Qwen2.5-14B Q4_K_M  | llama.cpp           | 8×10 / 8K | 80/80    | Compatibility       |
| Qwen2.5-32B Q4_K_M  | llama.cpp           | 8×10 / 8K | 80/80    | Compatibility       |
| Hermes 35B-A3B GGUF | llama.cpp           | 8×10 / 8K | 80/80    | MoE proof           |

**IMPORTANT** Model artifacts, request counts, and runtime contracts differ. The matrix tells you what was demonstrated; it does not manufacture a horse race.

14B

THE PRACTICAL LANE

Fastest useful balance in WP1

32B

THE PRESSURE TEST

Qualified, slower, hotter

35B

THE PLOT TWIST

llama.cpp passed; vLLM becomes WP3

RELEASE  
THE MODELS.

OBSERVED E-WORKLOAD-MATRIX-01

HUMAN TRANSLATION · THREE  
MODEL SIZES

Think compact car, pickup, and tractor-trailer. Bigger can carry more capability, but it also demands more room, fuel, cooling, and patience.



LURCH / MODEL ROLL CALL

Fourteen. Thirty-two. Thirty-five. **Release them.**

PRACTICAL

PRESSURE TEST

PLOT TWIST

LURCH // FIELD SYSTEM



9

# Show Me the Tokens

Measured llama.cpp results

## DEMONSTRATED TOTAL TOKEN THROUGHPUT



| MODEL         | DONE  | TOK/S  | P50 S   | P95 S   | P99 S   | PEAK °C |
|---------------|-------|--------|---------|---------|---------|---------|
| 14B Q4_K_M    | 80/80 | 449.20 | 15.4620 | 22.5791 | 25.1043 | 68      |
| 32B Q4_K_M    | 80/80 | 259.34 | 24.3756 | 41.3963 | 43.6957 | 75      |
| 35B-class MoE | 80/80 | 434.16 | 16.8933 | 21.2228 | 24.0966 | 52      |

14B

Input 63,665  
Output 9,622  
Avg power 177.4 W

32B

Input 63,665  
Output 7,576  
Avg power 203.4 W

35B MoE

Input 62,305  
Output 13,140  
Avg power 106.5 W

WAIT...  
IT RAN THE 35B?

**COMPARISON WARNING** Same envelope, different models and output totals. Treat these as demonstrations, not a universal ranking.

**HUMAN TRANSLATION · TOKENS ARE** Tokens are small pieces of language. Counting tokens per second is like counting meals leaving a kitchen—but every order still has to arrive correctly.



LURCH / BENCHMARK REQUEST

Enough adjectives. *Show me the tokens.*

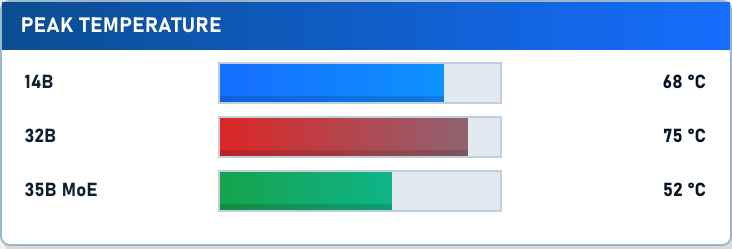
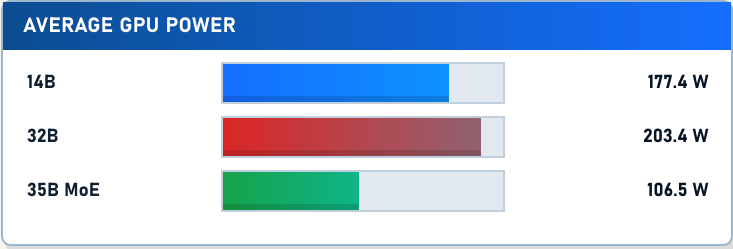
449.20 259.34 434.16 TOK/S

LURCH // FIELD SYSTEM

# 9.1

## Watts, Heat, and Useful Work

Telemetry without the convenient bullshit



| MODEL   | AVG W | PEAK W | AVG °C | PEAK °C | AVG UTIL |
|---------|-------|--------|--------|---------|----------|
| 14B     | 177.4 | 232.9  | 57.1   | 68      | 82.0%    |
| 32B     | 203.4 | 251.2  | 65.0   | 75      | 87.3%    |
| 35B MoE | 106.5 | 163.6  | 45.9   | 52      | 74.5%    |

**OBSERVED**

The 35B-class MoE run drew less average power than the dense 14B and 32B runs.

**NOT YET PROVEN**

That lower average watts equals higher efficiency or better operational value.

**DO NOT MISREAD THIS** Normalize completed useful work, model architecture, token mix, quality, and time before claiming efficiency.

LOWER WATTS  
≠  
FREE LUNCH.

**HUMAN TRANSLATION · POWER IS THE FUEL BILL** A machine being busy is not automatically efficient. A delivery truck can burn more fuel because it carried more packages—or because it took a terrible route.



LURCH / TELEMETRY

Watts are evidence. *Not a verdict.*

POWER THERMALS USEFUL WORK

LURCH // FIELD SYSTEM

## 9.2

## The Run That Failed for the Right Reason

35B-class MoE context boundary

4K CONTEXT

68 / 80

12 HTTP 500 failures  
CONTEXT EXCEEDED

8K CONTEXT

80 / 80

0 failures · 434.16 tok/s  
PASS

## WHAT FAILED

```
requested context exceeded n_ctx=4096
HTTP 500 | 12 requests
CONCURRENT_TEST_RESULT=FAIL
```

The randomized multi-turn workload exceeded the configured 4K window.

## WHAT CHANGED

```
-c 8192
completed=80 | failures=0
CONCURRENT_TEST_RESULT=PASS
```

One parameter changed. The repeated test crossed the boundary.

**WHY KEEP THE FAILURE** Because “we changed one parameter and it passed” is reproducible only when the original condition and failure remain visible.

FAIL

CHANGE ONE THING

RE-RUN

PRESERVE  
BOTH.

OBSERVED E-LLAMA-35-4K / E-LLAMA-35-8K

## HUMAN TRANSLATION · THE 4K FAILURE

*Twelve pieces of luggage did not fit in the smaller trunk. Moving to the larger trunk let every bag travel; the luggage and car stayed the same.*



## LURCH / INCIDENT REVIEW

4K said no. 8K said yes. *Write down both.*

FAIL 68/80

CHANGE ONE THING

PASS 80/80

LURCH // FIELD SYSTEM



10

## Wait... It Ran WHAT?

Sanitized lab evidence and the interpretation it supports

## FROM THE LAB NOTEBOOK

```
[2026-08-07T20:35:31-07:00]
users=8 requests/user=10 context=8192
model=/home/<USER>/models/...35B-A3B...gguf
completed=80 | failures=0
Concurrent total tokens/sec: 434.16
peak=52.0 C | avg=106.5 W | peak=163.6 W
✓ CONCURRENT_TEST_RESULT=PASS
```

Username and home-directory identity normalized;  
technical measurements retained.

## THE PLOT TWIST

## THE 35B RESULT CHANGED THE QUESTION.

The tested vLLM configuration did not qualify the 35B-class GGUF model. llama.cpp loaded it, served it, and completed the 80-request workload on the same V100.

**CONCLUSION** The vLLM failure did not establish a hardware capability limit. It established a boundary in the tested software/model-loading path.

*The GPU  
had receipts.*

## WHAT WE KNOW

- Hardware executed the 35B-class MoE workload.
- 8K run completed 80/80 with telemetry.
- Known-good vLLM lanes remain untouched.

## WHAT WE DO NOT KNOW

- Whether bounded vLLM work can serve it.
- Which blocker appears next.
- Whether success would be operationally useful.

CONCLUSION E-INTERPRET-35

## HUMAN TRANSLATION · WHY 35B MATTERS

*One doorway was too narrow, but another doorway let the same couch through. That proves the couch fits inside the house—the first doorway was the limitation.*



## LURCH / RECEIPTS DEPARTMENT

The hardware had receipts. The software had questions.

SAME GPU

35B EXECUTED

BOUNDARY MOVED



SCOTT + REN / LAB DESK

D1

# How to Read the Speed Tests

The numbers without the engineering decoder ring

## THE BIG IDEA

*Speed is useful only when requests finish reliably and people are not left waiting forever.*

### TOKENS / SECOND

How quickly the system processes and generates pieces of text. Higher usually feels faster.

#### IN PLAIN ENGLISH

*Like dishes leaving a restaurant kitchen each minute.*

### P50 LATENCY

The middle wait time: half the requests finished faster and half finished slower.

#### IN PLAIN ENGLISH

*The wait most customers are likely to recognize.*

### P95 LATENCY

A near-worst-case wait: 95% finished within this time, while 5% took longer.

#### IN PLAIN ENGLISH

*The slow table during the dinner rush—not the average table.*

### 40/40 · 80/80 · 250/250

Completed requests divided by attempted requests. Matching numbers mean none failed.

#### IN PLAIN ENGLISH

*Every passenger who boarded reached the destination.*

### THROUGHPUT VS. LATENCY

Throughput measures total work; latency measures an individual wait. A system can excel at one and disappoint at the other.

#### IN PLAIN ENGLISH

*A busy checkout can scan many carts per hour while one shopper still waits too long.*

### WHY THE BARS LOOK CLOSE

Small visual differences mean the lanes performed similarly. The shared PASS matters more than a tiny lead.

#### IN PLAIN ENGLISH

*Three runners cross together; do not invent a photo finish championship.*

**READER TAKEAWAY** Ask three questions: Did every request finish? How much work was completed? How long did people wait?



LURCH:

*“Fast is good. Finished is better.”*

# How to Read Heat, Power, and Size

Running it is only half the story

## THE BIG IDEA

*A model can run successfully and still be too slow, hot, or power-hungry for the job you actually need.*

### 14B · 32B · 35B-CLASS

These labels describe model scale. Larger does not automatically mean better for every task.

#### IN PLAIN ENGLISH

*A pickup, moving truck, and tractor-trailer solve different jobs.*

### AVERAGE POWER

The normal electrical draw across the measured workload.

#### IN PLAIN ENGLISH

*The usual fuel burn during the trip.*

### PEAK POWER

The highest momentary draw recorded during the run.

#### IN PLAIN ENGLISH

*The surge while climbing the steepest hill.*

### AVERAGE TEMPERATURE

The normal operating heat across the run.

#### IN PLAIN ENGLISH

*The oven temperature through most of the bake.*

### PEAK TEMPERATURE

The hottest recorded moment. It helps identify cooling limits and risk.

#### IN PLAIN ENGLISH

*The hottest minute when every burner is on.*

### GPU UTILIZATION

How busy the accelerator was. High use means it was working hard—not necessarily working efficiently.

#### IN PLAIN ENGLISH

*A worker can be 100% busy doing useful work or carrying boxes in circles.*

**READER TAKEAWAY** “It ran” proves capability. Practical value also depends on waiting time, stability, heat, power, output quality, and cost.



LURCH:

*“Check the gauges before declaring victory.”*

# How to Read Failures and Evidence

The receipts department, translated

## THE BIG IDEA

A trustworthy experiment shows what failed, what changed, what passed later, and where the proof lives.

### 4K → 8K CONTEXT

Context is the model's working space. The 4K setting was too small for twelve requests; 8K gave them room.

#### IN PLAIN ENGLISH

*A king-size mattress needs a wider doorway.*

### HTTP 500

The serving program encountered an internal error while handling a request.

#### IN PLAIN ENGLISH

*The kitchen could not complete the order; the customer did not cause it.*

### OBSERVED

Directly recorded by commands, logs, summaries, or telemetry.

#### IN PLAIN ENGLISH

*The security camera footage.*

### ANALYSIS / CONCLUSION

Reasoning based on observations. It must remain distinguishable from raw measurement.

#### IN PLAIN ENGLISH

*The detective explains what the footage supports.*

### HYPOTHESIS / PLANNED

A possible explanation or future test—not a finished result.

#### IN PLAIN ENGLISH

*The next episode preview, not tonight's ending.*

### EVIDENCE ID + HASH

The ID points to the supporting record; the hash helps prove a file did not silently change.

#### IN PLAIN ENGLISH

*A library call number plus a tamper-evident seal.*

**READER TAKEAWAY** A later PASS does not erase the earlier FAIL. Keeping both is how someone else can understand and reproduce the engineering decision.



LURCH:

*"No retcons. Keep both logs."*



11

## To Be Continued...

The Pascal Frontier series roadmap

# ONE MACHINE. FOUR MISSIONS. EACH OUTRO HANDS THE STORY TO THE NEXT.

WHITEPAPER #2

### SGLANG ENTERS THE ARENA

Lurch carries the same model dossiers into a third inference engine and a new evidence-first qualification round.

WP1 → WP2 · SAME HARDWARE · NEW ENGINE

WHITEPAPER #3

### SAVE vLLM. RUN THE 35B.

Scott, Ren, and Lurch return to determine whether disciplined engineering can carry vLLM across the 35B-class loading boundary.

WP2 → WP3 · TEAM MISSION · NO VICTORY ASSUMED

WHITEPAPER #4

### THE FINAL SHOWDOWN

About 80 participants perform comparable real-world assignments through different models under one evaluation contract.

WP3 → WP4 · CAPABILITY · ACCURACY · USEFULNESS

Same work. Same prompt. Different models.

*The proof is in the output.*

**THE RULE:** Preserve every prompt, output, failure, and scoring decision. Measure what the models help people accomplish—not merely peak concurrency.

WP1 → WP2

#### ENGINE HANDOFF

vLLM → llama.cpp → SGLang

The same hardware and model dossiers meet a new engine.

WP2 → WP3

#### THE RESCUE MISSION

SCOTT REN LURCH →

35B + vLLM

The team attempts the unresolved loading boundary.

WP3 → WP4

#### ANTON.AOS INCUBATION

BENCHMARK UTILITY →

ANTON.AOS / TBC

Internal benchmark components may become part of Anton.

HUMAN TRANSLATION · WHAT HAPPENS NEXT?

WP1 proved the veteran can still fight. The sequels bring in another engine, attempt a rescue mission, and finally give every model the same assignment.

# ENGINEERING TAKEAWAYS

|   |                                                                                   |            |
|---|-----------------------------------------------------------------------------------|------------|
| 1 | Modern vLLM .019/.021/.023 ran on SM70 through explicit CUDA 12.6 + Triton lanes. | OBSERVED   |
| 2 | 14B cross-version randomized concurrency passed in every lane.                    | OBSERVED   |
| 3 | vLLM .023 served 32B Q4_K_M reliably—with major latency and thermal tradeoffs.    | OBSERVED   |
| 4 | llama.cpp needed no source compatibility patch for the three target families.     | OBSERVED   |
| 5 | 35B llama.cpp 80/80 separates hardware capability from the tested vLLM path.      | CONCLUSION |
| 6 | Power conclusions require normalization by completed useful work.                 | ANALYSIS   |

**THEY STOPPED SUPPORTING LURCH.  
THEY DID NOT STOP LURCH.**

## CAPABILITY

Can the hardware execute the work? **Yes, within demonstrated boundaries.**

## VALUE

Is it operationally or economically worthwhile? **Requires a separate normalized decision.**

## WP1 · EVIDENCE BEFORE HYPE

### HUMAN TRANSLATION · THE BUSINESS-SIZED LESSON

*Do not replace an entire factory because one supplier stopped making a compatible part. First identify whether a safe replacement part or different process restores useful work.*



### LURCH / ENGINEERING PAYOFF

*They stopped supporting me. They did not stop me.*

CAPABILITY

COMPATIBILITY

EVIDENCE

LURCH // FIELD SYSTEM

# EVIDENCE HANDOFF / ARCHIVE INTAKE

FOUNDER-SUPPLIED PACKAGE

111

files safely extracted

OUTER ARCHIVE

120

entries inventoried

TRANSCRIPTS

86

session records

NESTED ZIP CONTENT

75

additional files inspected

## WHAT ARRIVED

| CATEGORY         | COUNT / STATUS |
|------------------|----------------|
| Primary reports  | 3              |
| Benchmark tools  | 5              |
| Reference images | 3              |
| Prior drafts     | 2              |
| Raw archives     | 5              |
| Integrity ledger | SHA256SUMS     |

## WHAT REN COULD VERIFY

- Outer ZIP extracted without path traversal.
- Two nested ZIP packages extracted and inventoried.
- The 35B llama.cpp summary/log family is present.
- Original files were not modified.
- The supplied SHA256 ledger is retained.

## RECOVERED 35B EVIDENCE

```
llamacpp-35b / summary-...203526.txt
llamacpp-35b / summary-...202952.txt
8K context: 80/80 completed
4K context: 68/80 completed
✓ both pass and failure artifacts preserved
```

**INTERPRETATION** The supplied handoff converts the 35B locator from “session-only” to an inspectable archive artifact. It does not make unmatched engine runs directly comparable.

## OPAQUE BUT PRESERVED

Three RAR archives remain intact but were not opened in this workspace because a compatible extractor was unavailable. Their presence is recorded; their contents are not guessed.

## PRIVACY / RELEASE GATE

Raw archives and transcripts are private engineering evidence. Public release requires sanitization, secret scanning, licensing review, and Founder approval.

## OUTER ZIP SHA-256

```
ade0a04baa4050f7fb6f98cbb792076dabc2bab9dba5585785f3af2769029934
```

OBSERVED E - HANDOFF - V2 - 20260811

## HUMAN TRANSLATION · WHY 111 FILES MATTER

*This is the evidence warehouse inventory. A claim without its log is like a court case whose evidence box vanished between the crime scene and the trial.*



## LURCH / EVIDENCE LOCKER

One hundred eleven files walked in. *The claims brought ID.*

2 NESTED ZIPS

35B LOGS RECOVERED

3 RARS SEALED

LURCH // FIELD SYSTEM

# REPRODUCIBILITY MAP

| RUN ID         | RUNTIME / WORKLOAD      | SOURCE LOCATOR        |
|----------------|-------------------------|-----------------------|
| E-VLLM-14-XVER | vLLM 14B · 50×5         | summary-...032332.txt |
| E-VLLM-14-LOAD | vLLM .023 14B · 8×5     | summary-...034315.txt |
| E-VLLM-32-GGUF | vLLM .023 32B · 8×5     | summary-...160833.txt |
| E-LLAMA-14     | llama.cpp 14B · 8×10/8K | summary-...210220.txt |
| E-LLAMA-32     | llama.cpp 32B · 8×10/8K | summary-...204521.txt |
| E-LLAMA-35-8K  | llama.cpp 35B · 8×10/8K | summary-...203526.txt |
| E-LLAMA-35-4K  | llama.cpp 35B · 8×10/4K | summary-...202952.txt |

**ARCHIVE STATUS** The Founder-supplied v2 handoff restored inspectable raw ZIP evidence, including the 35B llama.cpp summaries. Three RAR containers remain preserved but opaque pending a compatible extractor.

## BUILD & TEST COMMANDS

```
CUDA_HOME=/usr/local/cuda-12.6
TORCH_CUDA_ARCH_LIST=7.0
VLLM_ATTENTION_BACKEND=TRITON_ATTN
python -m vllm.entrypoints.openai.api_server --model <MODEL> ...
pds-v100-llm-benchmark.sh --runtime <vllm|llamacpp> --users N ...
✓ capture versions · workload · result · telemetry · source path
```

### PUBLICATION TOOLING

- Historical scripts remain tied to historical runs.
- Clean scripts must not rewrite evidence.
- Before release: document, scan, validate, test, review licensing.

### SECURITY

Public service examples should bind locally or require authentication. The wildcard-CORS/no-key warning remains relevant.

### HUMAN TRANSLATION · HOW SOMEONE REPEATS THE TEST

*A recipe is reproducible only when it lists the ingredients, oven setting, timing, and result—not merely a photograph of the finished cake.*



### LURCH / REPRODUCIBILITY CONVOY

Capture → hash → rerun → preserve. *No mystery meat.*

SOURCE

WORKLOAD

RESULT

TELEMETRY

LURCH // FIELD SYSTEM



# CLAIM / EVIDENCE LEDGER

| CLAIM                                                       | CLASS      | EVIDENCE ID       | CONFIDENCE |
|-------------------------------------------------------------|------------|-------------------|------------|
| Modern vLLM lanes generated on SM70                         | Observed   | E-VLLM-SMOKE-XVER | HIGH       |
| 14B cross-version workload passed                           | Observed   | E-VLLM-14-XVER    | HIGH       |
| 32B GGUF served 40/40 in .023                               | Observed   | E-VLLM-32-GGUF    | HIGH       |
| llama.cpp served all target families without source patches | Observed   | E-LLAMA-COMPAT-01 | HIGH       |
| 35B result separates hardware from vLLM path                | Conclusion | E-INTERPRET-35    | HIGH       |
| Second-life GPUs may retain engineering value               | Hypothesis | H-LIFECYCLE-01    | OPEN       |
| vLLM can be extended to 35B-class model                     | Planned    | P-WP3-35B         | UNKNOWN    |

FOUNDER REVIEW COPY

VERSION 2.1 · 2026-08-12

AUTHORSHIP

**Engineering, experimentation, benchmarking, and primary authorship:**  
J. Scott Henderson, Founder, Project Dominion Studios.

**AI-assisted research, debugging, analysis, documentation, and editorial support:**  
Ren-AI, an OpenAI/ChatGPT-based AI advisor operating through ChatGPT and Codex.

HUMAN AUTHORITY / GUARDRAILS

**Scott remained the final human decision-maker.** AI work was advisory and derived. Scott analyzed each proposed coding or publication change and approved or rejected it under Studio/main:

- AI-ENGAGEMENT.md
- management/governance/ai/transparency-and-human-authority.md
- standards/knowledge/DOCUMENTATION-EDITORIAL-STANDARD.md

| WORK PHASE              | MODEL / MODE          | ROLE                                                             |
|-------------------------|-----------------------|------------------------------------------------------------------|
| Broader working process | ChatGPT 5.5 Instant   | Research, discussion, drafting, review support                   |
| Whitepaper production   | Codex 5.6 · Sol High  | Evidence judgment, visual authority, layout, technical synthesis |
| Publishing / QA         | Codex 5.6 · Sol Light | Final mechanical packaging, rendering, validation, and manifests |

Founder-confirmed transition to Codex 5.6 Sol Light occurred before final packaging and QA.

**REVIEW SCOPE** Approve or reject WP1’s publication system before it becomes the WP2–WP4 template. The series does not advance automatically.

“Sci-fi is the dream. Engineering is how we get there.”  
— J. Scott Henderson

Missing raw artifacts remain gaps—not reconstructed facts. AI systems did not hold approval or publishing authority.

HUMAN TRANSLATION · HOW TO READ THE RECEIPTS

Each claim has a claim-check ticket pointing to its supporting record. If the ticket is missing, the coat does not leave the evidence room.

THE PASCAL FRONTIER · NEXT ENGINE

# A NEW CHALLENGER ENTERS THE ARENA.

*SAME LURCH. SAME MODEL DOSSIERS.  
A DIFFERENT ENGINE GETS THE KEYS.*



WHITEPAPER #2

**SGLANG ENTERS THE ARENA.**



WP1 · QUALIFIED LANES  
**vLLM**



WP1 · HARDWARE WITNESS  
**llama.cpp**



WP2 · NEW ENGINE  
**SGLang**

HUMAN TRANSLATION · NEXT:  
SGLANG

*The same driver and racetrack remain; a new engine gets installed. That lets the next paper test the engine instead of quietly changing the whole car.*