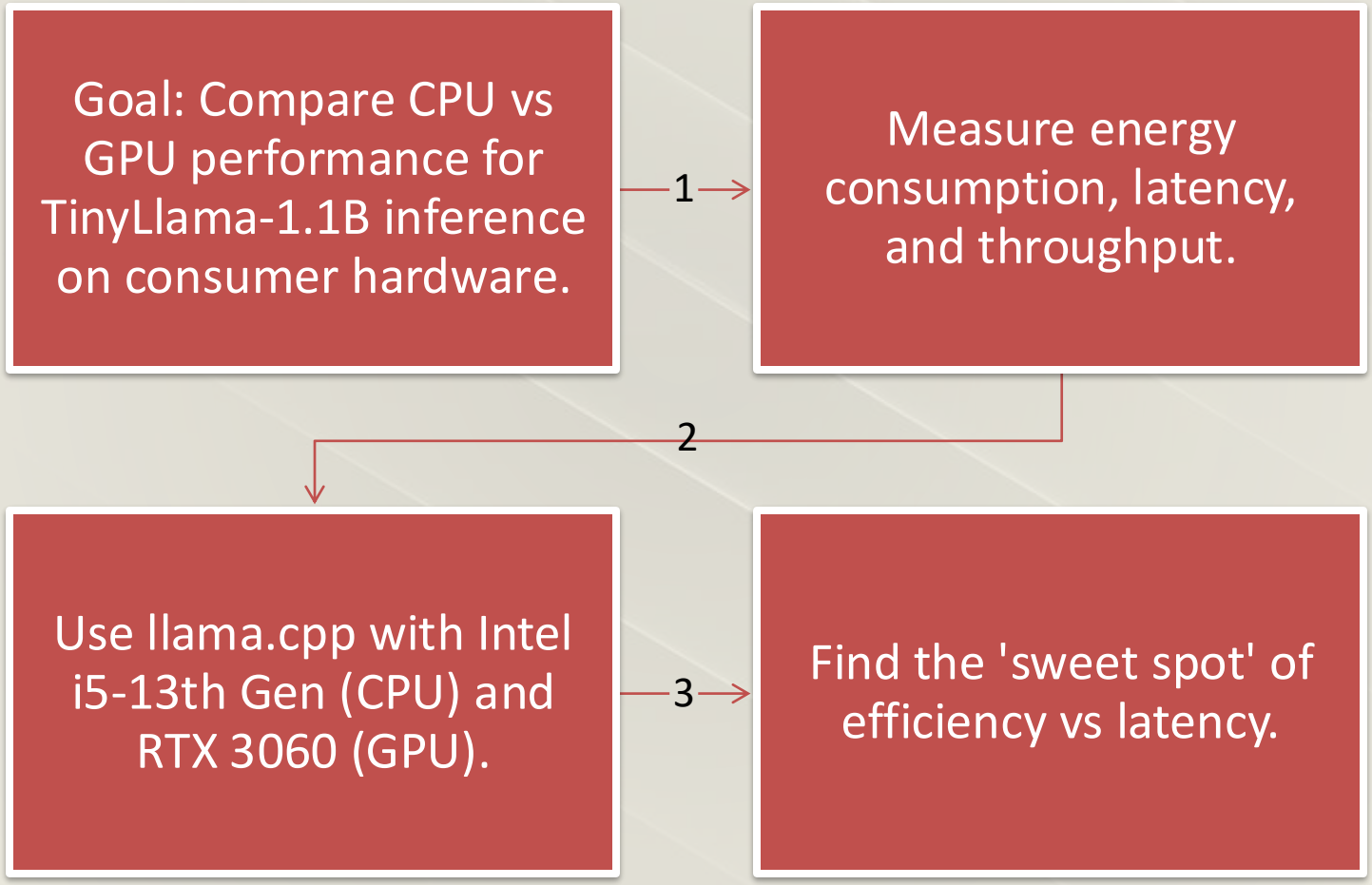
Milestone 2

Energy-Aware LLM Inference on Consumer Hardware

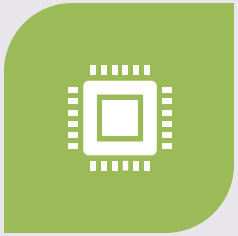
CSCE 585

Recap of Proposal



Accomplishments

SET UP LLAMA.CPP LOCALLY WITH TINYLLAMA MODEL.



CPU INFERENCE PIPELINE WORKING END-TO-END.

TELEMETRY LOGGING INTEGRATED (LATENCY + ENERGY).

RAW CPU POWER LOGS COLLECTED AND GRAPHED.

CPU Telemetry Output

Proof of working CPU pipeline:

Automatic CSV logging from Intel Power Gadget.

Energy consumption recorded per run.

Example CPU power graph generated (see below).

[Insert CPU Graph + CSV Snippet Here]

Proof of working CPU pipeline

Output: Runs inference on CPU and logs telemetry.

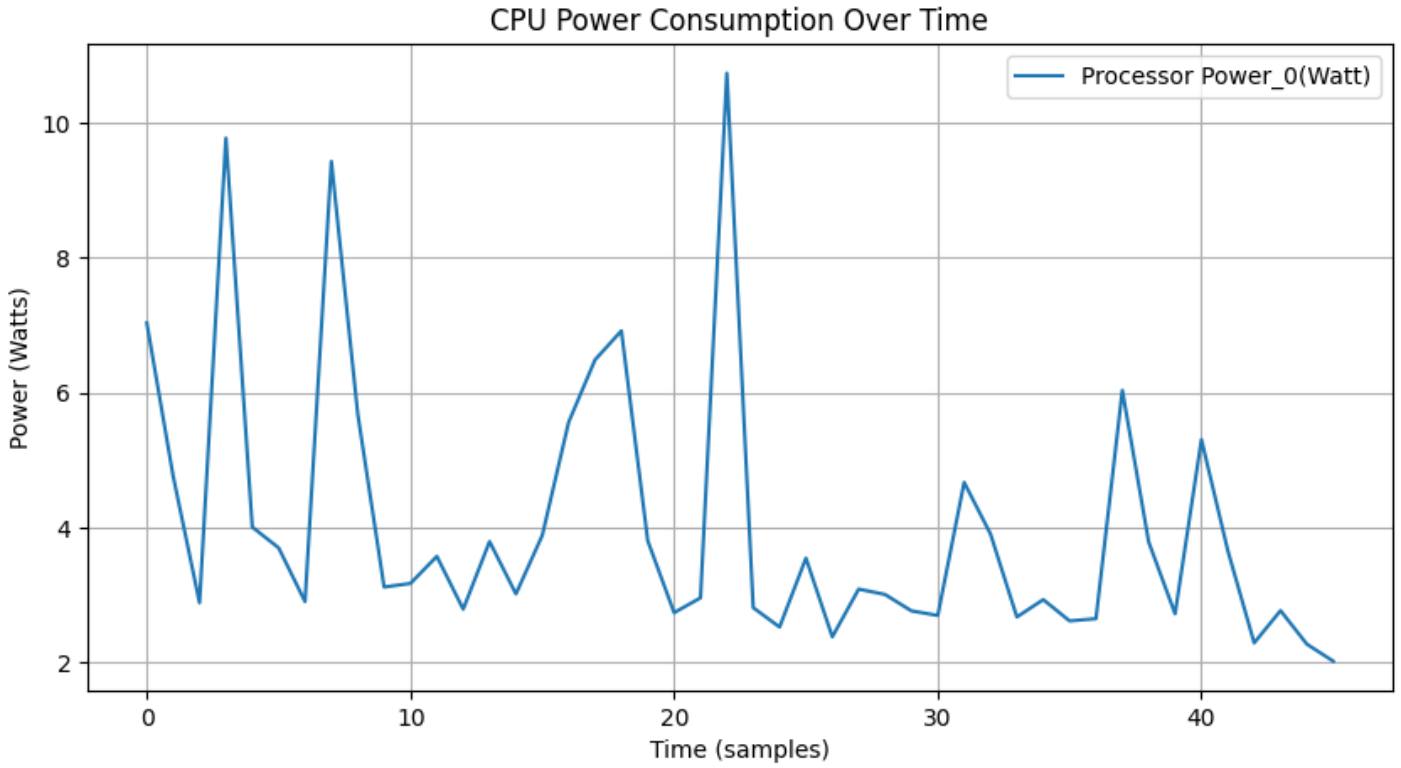
```
PS C:\Users\tysup\Documents\Energy-Aware-LLM-Inference-on-Consumer-Hardware> python src
build\bin\Release\llama-cli.exe"
>>
error: failed to initialize MDH_Context
Max Temp = 100
number of nodes = 1
TDP(mWh)_0 = 65.00
Base Frequency = 2500.00(MHz)
Logging...Done
☑ CPU power logged: 52.58 J (raw CSV saved to data\raw_cpu_power_20251002_010315.csv)
```

timestamp	backend	energy_joules	notes
2025-10-01T22:52:04.560	cpu	48.239673913043475	prompt=baseline-001
2025-10-01T23:13:45.566	cpu	20.220108695652176	prompt=baseline-001

Automatic CSV logging (Intel Power Gadget)

Energy consumption per run

CSV snippet for slides


CPU Power Consumption Over Time
Processor Power_0(Watt)
Power (Watts)
Time (samples)
2
4
6
8
10
0
10
20
30
40

Challenges

GPU pipeline not yet working (CUDA init failed).

No GPU energy telemetry from NVML yet.

Haven’t generated CPU vs GPU comparison graphs yet.

Next Steps

- Fix GPU pipeline (resolve CUDA + NVML logging).
- Run multiple trials with varying batch size and context.
- Generate CPU vs GPU latency vs energy graphs.
- Deliver full comparison and analysis for next milestone.