LLM Data/Applied ML (survey parsing automation).

Core asset: JSON extraction app with lite monitoring.

Stack: llama.cpp, LoRA adapter, Streamlit, Python.

Highlights:

- Fine-tuned LLAMA2 with domain specific dataset, Quantization and GGUF conversion.
- Built a prompt + few-shot parser with single-JSON guarantees and {} for non-questions.
- Deterministic decode: Temperature 0.0, top_k 0, top_p 1.0, fixed stops.
- Balanced single-JSON guard and minimal validators: Latency-safe (Median latency ~60s CPU).
- Monitoring lite: runs.csv with Valid%, median_ms, avg_tokens, items + alert.
- Repro notebook

Results:

- Accuracy: 95% Valid JSON: 100% on checked set. Stable latency.
- Reproducibility: Same stops/decoding in app and notebook.

Links:

https://github.com/M-Ramkumar20/SP-Auto-Demo