Capstone report plan

Tuesday, 13 May 2025 3:24 PM

0. Abstract: what is being done, why & what was achieved

— — — — —

1. Intro

bcg & motivation → prop

our contributions
scope
paper overview

2. Theoritical bcg
- → LLM & LLM context, sampler, etc
- → M1 SoC, GPU, NPU, Shared address space
- → zero copy reads (server data streams,)
- → quantization
- → data leakage

3 Related work
- → ollama & llamafile
- → relevant papers from proposal
- → ghotz rev eng
- → how coreml model uses NPU

4 methodology / design
- ↳ vecdump & mmap
- ↳ NPU hack
- ↳ RAG pipeline

5 Implementation & architecture
(UI, static lib, etc)
- ↳ context pool & model reuse

6 Experimentation
bertscore, models, quant, etc
Results & analysis

7 Conclusion & future work
- ↳ limitations
- ↳ NPU backend for llama