

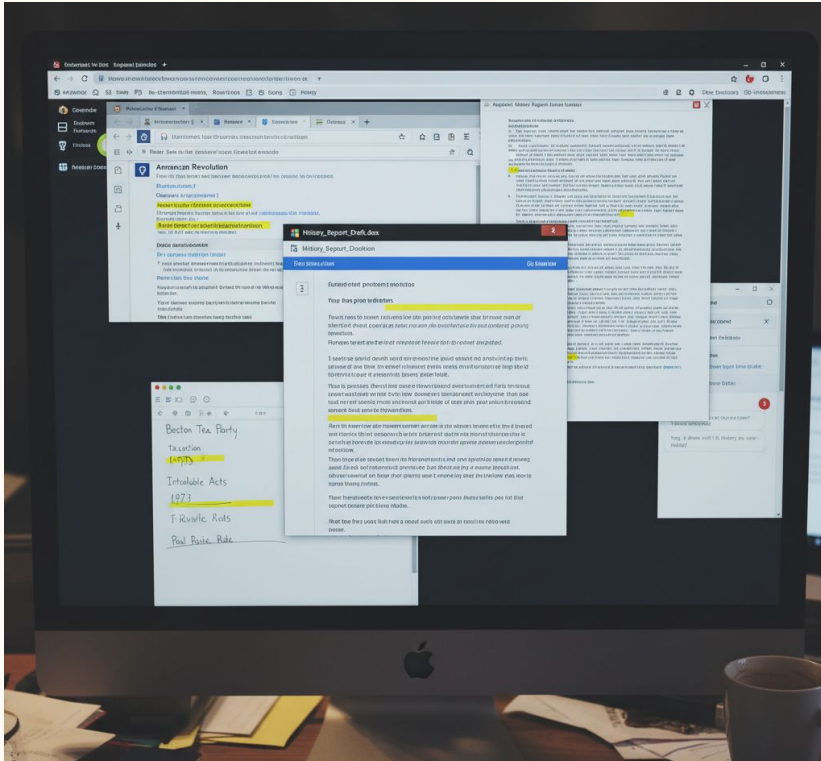
→ Team C4

SPARK

Design Review

Tatyana Cruz
Sida Cheng
Leonard Cheng

Use Case



S

SPARK
AI Desktop Assistant

LIVE CONTEXT

Safari Active
Craft Team Interest

Slack
Team Design Channel

Chrome
Google Form - CMU Survey

LIVE CAPTURE

● Live

[14:23] Ronnie: Hey team, just wanted to follow up on the project timeline. Are we still on track for the demo next week?

SUGGESTED ACTIONS

Draft Slack Reply
Generate response to Ronnie

Add to Exam Study Guide
Save key concepts

Summarize Thread
Quick overview of discussion

Create Calendar Event
From mentioned dates

Use Case Functionality

Action Button	Use Case Example
Context Summary	Collect your input and the context from your last 10 window switches into a high-level briefing of your recent workflow.
Key Word Search	Highlights a term to generate a blurb listing all recent files and pages where that topic appears.
Text Formatting	Instantly cleans up or professionally rewrites a highlighted snippet of text for better clarity.
Responses	Drafts a 1–3 sentence reply to the most recent Slack or email based on the current context.

Use-Case Requirements

Privacy: 100% local data processing, zero cloud dependency

Ease of Use: Universal USB compatibility, <2s per atomic action

Autocomplete Delay: < 500ms

Grammar correction: >90% correctness rate

~~Learning: 70% acceptance rate after 10K tokens of personalized input~~

Context retrieval and storage : able to capture the status and context of Apps (Slack, Word, iMessage) for a 1-hour work session.

Quantitative Design Requirements

<i>User Requirement</i>	<i>Design Requirement</i>
Data Privacy	Hardware can host local LLM and database
Universal USB	Hardware has native USB support
LLM Autocomplete Delay	<500ms $\approx$ 10ms(USB) + 500ms(LLM Reading speed) For a 20-word sentence: 60+ token/s Reading speed
<2s / Atomic Action	For a 20-word sentence: 20+ token/s Output speed
Grammar Correction	>90% correctness $\approx$ LLM with 3B+ parameters
Context Capture	Access to app info, Max Context Length > 4000 tokens (~3000 words)

Solution Approach

Two-processor design:

- Jetson AGX Xavier (AI Core)
- RP2040 Pico (peripherals/USB)

Software + Hardware Integrated Solution:

- PC-side daemon for context extraction (active window, selected text)
- Physical panel with controls and display

Hardware and LLM: AGX Xavier(32GB RAM) + Qwen3-4B / Phi-3 mini

Inference engine: llama.cpp, llama-server(API calls)

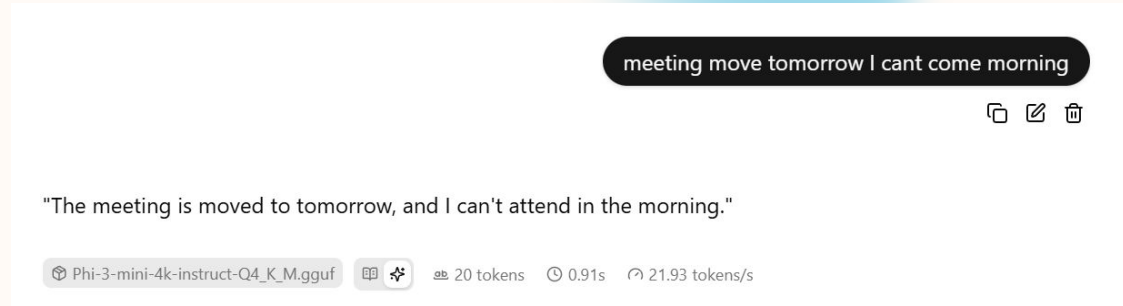
Social Impact

Mental Health: open-source models and programs

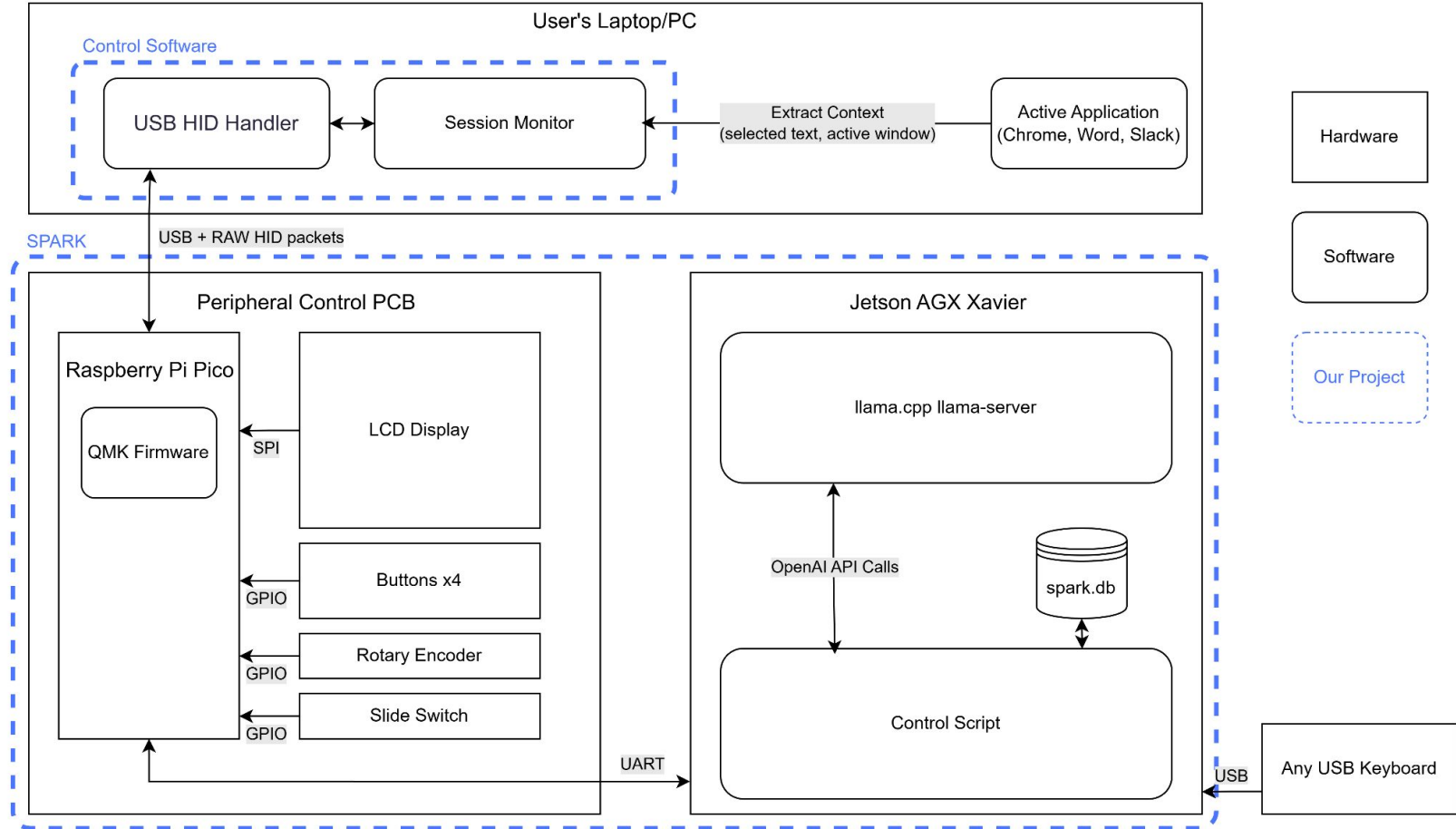
Environment: energy-efficiency of small models

Economic Factors

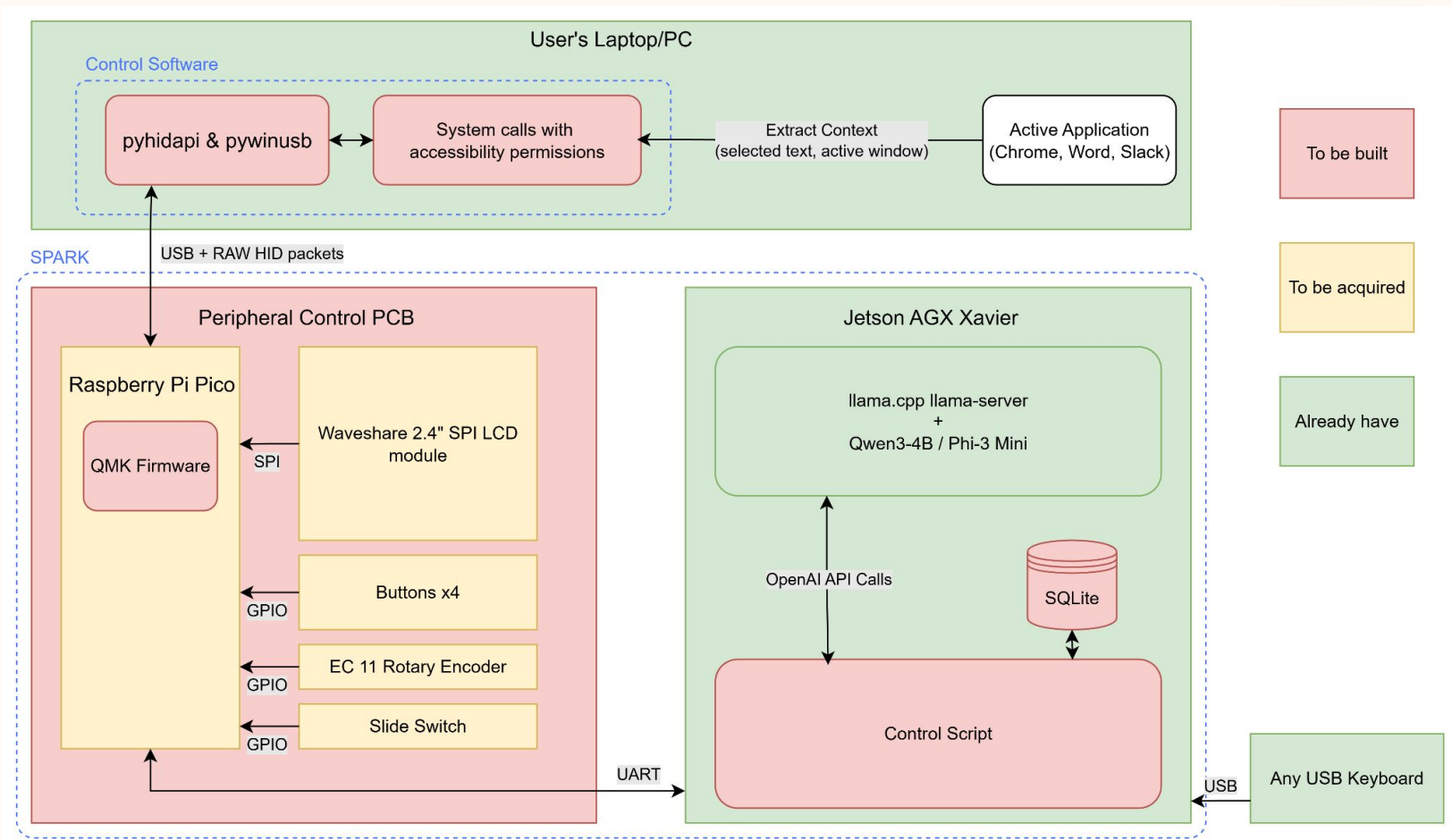
Financially reasonable local LLM application



System Block Diagram



Implementation Plan



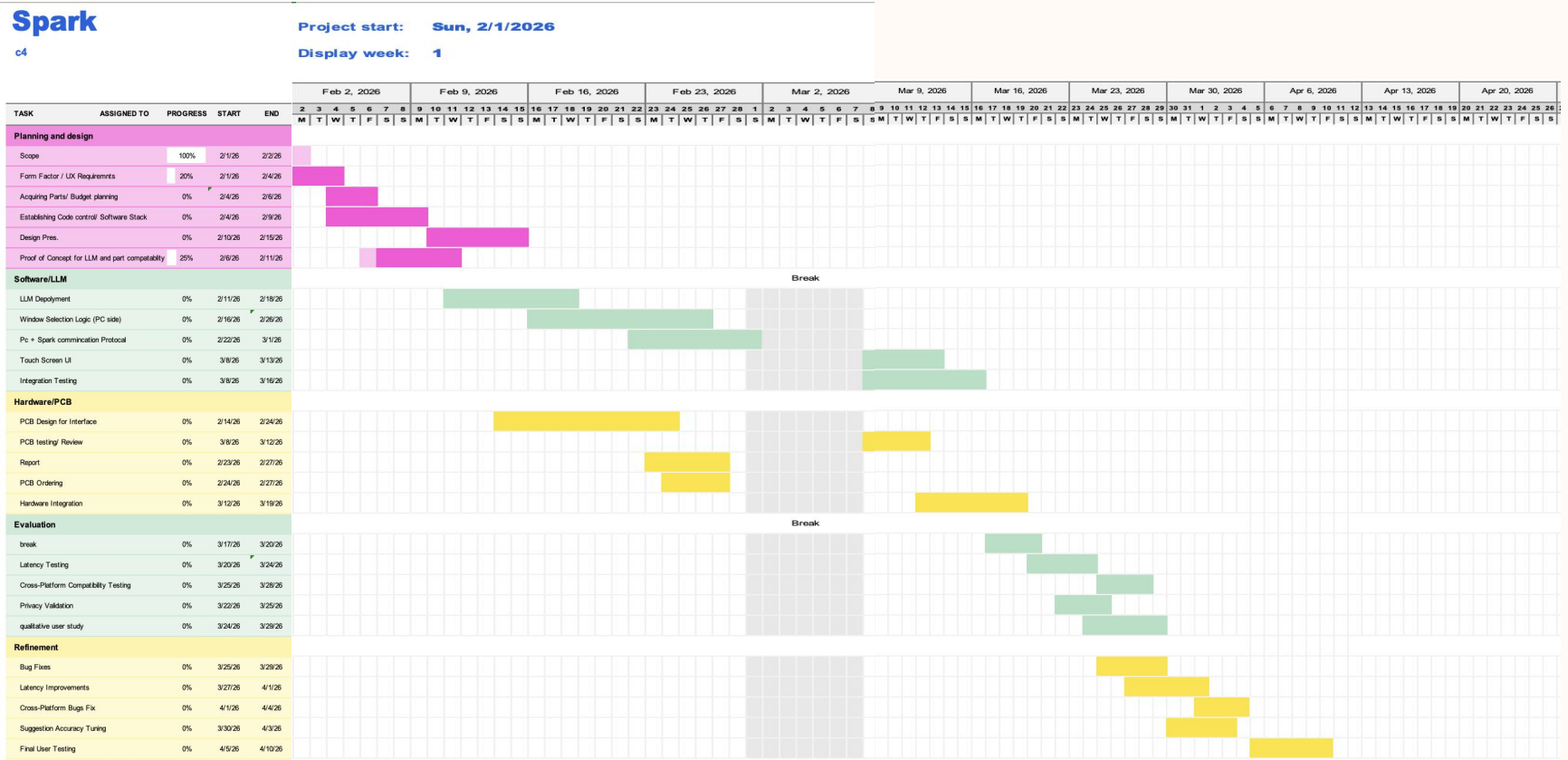
Testing, Verification, and Metrics

Requirement	Standardized Test Method	Measured Metric	Passing Criteria
Autocomplete	20-word controlled sentence × 50 trials	Avg latency	≤ 500 ms
Grammar Correction	100 paragraphs × 5 labeled errors	% corrected errors	≥ 90% accuracy
Context Capture	Simulated 1-hour workflow log	Max tokens processed	≥ 4000 tokens
Atomic Action	Button → Full response cycle × 50 trials	Completion time	≤ 2 s

Testing, Verification, and Metrics (cont.)

Use Case	Standardized Input	Validation
Context Summary	Pre-recorded 10 window switches with key events	Key Events Mentioned / Total Event Points $\geq 80\%$
Keyword Search	10 timestamped database entries	Correct file returned ≤ 2 s
Text Formatting	Highlighted 10 test snippets	Errors Corrected / Total Known Errors $\geq 90\%$
Response Drafting	20 Slack/email threads with key points	Key Points Mentioned / Total Key Points $\geq 95\%$

Project Management - Schedule



Project Management - Work Distribution

Core Responsibilities

- **Tatyana (Software/PM):** Project management, pattern recognition logic, and host-side context detection.
- **Sida (Software/Firmware):** System architecture, SPARK communication protocol, and touch screen UI.
- **Leonard (Hardware/Testing):** PCB design/ordering/assembly, enclosure fabrication, and hardware validation.

Workload & Workflow Strategy

- **Parallel Development:** The Client Side Software Application (Tatyana), LLM Pathway (Sida), and Hardware Development (Leonard) can be done in Parallel
- **Conflict points**
 - a. **Software Sync:** Tatyana & Sida integrate LLM and logic.
 - b. **Full System:** Full team collaboration
- **Balanced Scheduling:** Gantt chart milestones are staggered to prevent crunch periods and account for mid-semester deadlines.