Final Summary – Comparative A/B Test Consolidation

Weekly Objective

This week focused primarily on the comparative evaluation of the performance of two AI music generation models — **Suno v4.5 (Model A)** and **ElevenLabs Music (Model B)** — using three prompts (cinematic orchestral, meditative ambient, and vocal pop ballad). The goal was to identify qualitative differences, adherence to the brief, and structural coherence, evaluating each output according to artistic, technical, and usability criteria.

Methodology

For each prompt (P01–P03), two distinct outputs were generated while keeping all parameters fixed and the brief unchanged, in order to conduct an impartial A/B test.
The tracks were analyzed using an evaluation grid (1–5 scale) structured across ten criteria: structure and cohesion, melody and harmony, rhythm and timing, orchestration, narrative/emotional arc, creativity/originality, audio quality and mix, accessibility, usability, and flow.
Each track was accompanied by qualitative critical notes, and the results were collected in a comparative Excel file.

In addition to the evaluation spreadsheet, a structured JSON metadata file was created for each generated audio output.

Each JSON file includes technical and contextual fields such as prompt ID, AI model and mode, generation date, versioning, local file reference, audio link, and frozen parameters (target duration, target BPM range, vocal presence, key or tonal center when specified, genre/style, and core instrumentation).

The metadata also records the original brief, the exact prompt text used, actual output duration, and qualitative evaluation notes.

This structure mirrors real-world AI music training and human-in-the-loop evaluation workflows, enabling traceability, reproducibility, and scalable dataset analysis.

Key Results

- **P01 (Epic Orchestral Trailer):** Suno adhered more closely to the brief and instrumental requirements (strings, brass, percussion), delivering a more clearly defined climax. ElevenLabs showed cleaner audio but an overly short climax and missing brass section.
- **P02 (Calm Ambient Meditation):** Both models were coherent with the meditative style: Suno was simpler but slightly too active, while ElevenLabs was more faithful to the calm mood but included unrequested instrumentation and a less smoothly faded ending.

- **P03 (Female Pop Ballad):** Suno provided a more recognizable structure and a mood consistent with the genre; ElevenLabs delivered natural-sounding vocals and an excellent mix, but with poorly differentiated sections, a missing climax, and an abrupt ending.

Conclusion and Implications for Future Work

Overall, **Suno v4.5 (Model A)** demonstrated greater structural reliability and stronger adherence to briefs. However, it shows limited accuracy in respecting BPM and key, and track duration is highly variable and difficult to control. Some endings feel truncated, and harmonic language is often simple and linear, but requested instrumentation is almost always respected and overall audio quality is solid.
Suno v4.5 (Model A) proves superior in the creation of pop songs in this test set: vocals sound more natural and less artificial, partly thanks to the ability to customize them through the “+Persona” section; additionally, song sections are more clearly defined and recognizable.

ElevenLabs (Model B), on the other hand, allows greater control over BPM through manual input and appears to offer higher audio quality with a well-balanced mix. Unlike Suno, track duration is more controllable and reliable. However, it exhibits similar issues with abrupt endings and sometimes introduces unrequested instruments while omitting others.
In vocal pop songs, ElevenLabs proves inferior to Suno: manually specified sections (verse, chorus, bridge) are not always recognized, tend to sound too similar to one another, and the vocal performance appears somewhat artificial and monotonous. In addition, some lyrics are sung too quickly, with overlapping syllables.
ElevenLabs excels instead in freely generated songs, where the AI is allowed to autonomously define the song sections.