

AbletonBuddy: A Domain-Aware Natural Language Interface for Autonomous Control of Ableton Live via OSC

Pratham Vadhulas

Course: CS6235

Term: Fall 2025

November 23, 2025

Contents

1	Introduction	4
2	Motivation	4
3	Related Work	4
4	System Architecture & Methodology	6
4.1	High-Level Overview	6
4.2	The Tech Stack	6
4.3	The “Agentic Swarm” Pipeline	6
4.3.1	Step 1: Input & Disambiguation	6
4.3.2	Step 2: Classification & Task Extraction (The Router)	7
4.3.3	Step 3: Agent Assignment & Knowledge Retrieval	7
4.3.4	Step 4: Execution & Verification	7
4.3.5	Step 5: The Summary Agent	7
4.4	Visual Architecture Diagram	7
5	Implementation Steps	8
5.1	Phase 1: Core Architecture & Agent Framework	8
5.2	Phase 2: API Exposure	8
5.3	Phase 3: Frontend Integration	8
5.4	Phase 4: Domain Expansion (The Agent Swarm)	8
5.5	Phase 5: Multi-Agent Collaboration & Optimization	9
6	Experimental Setup & Results	9
6.1	Participants	9
6.2	Experimental Design	10
6.3	Task Design	10
6.4	Evaluation Metrics	10
6.4.1	Quantitative Metrics	11
6.4.2	Qualitative Metrics (NASA-TLX)	11
6.5	Hypothesis	11
6.6	Results	11
7	Meta Skills	11
7.1	Hybridizing Stochastic and Deterministic Logic	12
7.2	Constraint Engineering in Probabilistic Systems	12
7.3	Semantic Mapping of Abstract Domains	12

8	Conclusion & Future Work	12
8.1	Refinement & Stabilization	12
9	Deliverables	13

1 Introduction

This project presents a full-stack application designed to bridge the gap between natural language intent and digital music production. The core objective is to develop a chat interface that translates specific user requests into Open Sound Control (OSC) messages to control Ableton Live in real-time. Unlike standard mapping tools, this system utilizes a Large Language Model (LLM) backend to interpret semantic intent (e.g., “Make the drums punchier”) and execute the corresponding technical commands.

2 Motivation

The intersection of AI and creative tools is often hindered by the “translation gap” between human intent and software parameters.

The Efficiency Problem: Current state-of-the-art systems often rely heavily on generic protocols (such as MCP) or massive context injection to manage tool use. This approach is computationally inefficient, resulting in high latency and excessive token consumption, both of which are unacceptable in a real-time creative environment like music production.

The Workflow Solution: There is a critical need for a system that does not rely solely on brute-force context. Instead, we need a dynamic solution that leverages domain-specific heuristics, encoding the actual steps music producers take. By mimicking the workflow of a producer rather than just the API of the software, we can achieve a higher degree of autonomy with significantly lower computational overhead.

3 Related Work

The application of artificial intelligence in music production has seen significant growth, shifting from simple information retrieval and symbolic generation (MIDI) toward integrated, intelligent assistance within Digital Audio Workstations (DAWs) [2, 11]. Recent efforts have focused on transforming AI from a generator into a co-creative partner capable of navigating complex software environments [8, 3].

Current Landscape of Intelligent Music Assistants

Several prominent systems have emerged that attempt to bridge natural language and DAW control, each with distinct methodologies:

DAWZY: An open-source, voice- and text-driven assistant for the REAPER DAW [5]. It primarily utilizes code generation to execute low-level tasks, functioning effectively as a “scripting assistant” but lacking higher-level production nuance.

FilmComposer: Utilizes a multi-agent system operating a DAW to arrange and mix music specifically for silent films [9]. While innovative in its use of visual and rhythmic

inputs, its scope is limited to autonomous arrangement rather than granular control of existing production elements.

MixAssist: Introduces an audio-language dataset for training models to offer expert mixing advice [13]. It represents a step toward “expert systems” but focuses on providing advice rather than executing real-time control within the session.

Ableton Copilot & Loop Copilot: These systems represent the current state-of-the-art in session management [1, 12]. Ableton Copilot is built on the Model Context Protocol (MCP) [14] to manage granular MIDI operations and device control. Similarly, Loop Copilot integrates LLMs with a Global Attribute Table (GAT) to maintain continuity during iterative editing.

Limitations of Current Approaches (The “Context” Gap)

While the systems above demonstrate the viability of AI in DAWs, they generally rely on the LLM’s ability to translate natural language directly into actionable commands via the Model Context Protocol (MCP) or massive context injection [7, 15].

This approach presents two critical inefficiencies:

Computational Inefficiency: Relying on MCP requires the system to constantly “read” and maintain the entire session state to maintain coherence. This “brute force” context management burns tokens rapidly and introduces latency, which is detrimental to the real-time flow of music production [4].

The “Theory vs. Production” Gap: Many existing models are trained on symbolic representations (music theory/MIDI) rather than audio engineering physics [16, 10]. They struggle to bridge the gap between high-level creative intent (e.g., “Make it punchy”) and the nuanced signal processing required to achieve it (e.g., specific compression ratios and EQ curves) [6].

Our Contribution: Domain-Specific Knowledge Priors

Our system addresses these limitations by moving away from generic MCP-based state recording. Instead, we propose a workflow that incorporates explicit, structured music producer knowledge priors.

Rather than relying on an LLM to “figure out” the session state from scratch every time, our system utilizes a domain-aware pipeline:

Intent Parsing: Decoupling the user’s creative goal from the technical implementation.

Priors Consultation: referencing a static knowledge base of expert signal chains (e.g., standard vocal processing chains) rather than dynamically generating them via token-heavy reasoning.

Heuristic Execution: Using lightweight Open Sound Control (OSC) messages [17] mapped to these expert priors.

This approach allows the system to provide expert-level guidance that aligns with professional mixing philosophies while significantly reducing the token overhead and latency associated with standard MCP agents.

4 System Architecture & Methodology

4.1 High-Level Overview

The system is designed as a modular, event-driven application that decouples the user’s natural language intent from the low-level technical execution. The architecture follows a Human-in-the-Loop (HITL) design, ensuring that generative AI suggestions are validated against structured “Knowledge Priors” before interacting with the live production environment.

4.2 The Tech Stack

Component	Technology	Role
Frontend	React.js	Provides the chat interface and real-time state visualization for the user.
Backend API	Flask (Python)	Orchestrates the application logic and manages websocket connections.
AI Orchestration	Marvin + Pydantic	Handles intent classification, entity extraction, and agentic workflow management using structured outputs.
Control Layer	AbletonOSC	The translation layer that converts Python commands into Open Sound Control (OSC) messages compatible with Ableton Live.

Table 1: Technology Stack Components

4.3 The “Agentic Swarm” Pipeline

Unlike standard “Single Shot” LLM calls, our backend utilizes a multi-step reasoning pipeline to ensure safety and accuracy in the creative workflow.

4.3.1 Step 1: Input & Disambiguation

Raw user input (e.g., “The drums sound weak”) is first passed to a Disambiguation Module. This module checks for vagueness. If the request lacks context (e.g., “Fix it”), the system halts and requests clarification. If clear, it passes to the classifier.

4.3.2 Step 2: Classification & Task Extraction (The Router)

We utilize Marvin to classify the intent into specific domains.

Input: “Create a hiphop melody.”

Classification: Composition_Task (vs. Song_Task or Track_Task).

Extraction: The system extracts structured entities using Pydantic models: {Task: “Create”, Genre: “Hiphop”, Element: “Melody”}.

4.3.3 Step 3: Agent Assignment & Knowledge Retrieval

Once classified, the task is assigned to a specialized Domain Agent (e.g., The Composition-Agent).

The Critical Difference: Instead of hallucinating a solution, the Agent retrieves Knowledge Priors—pre-validated heuristics (e.g., “Hiphop Melody = Minor Pentatonic Scale + 808 Bass + Syncopated Rhythm”).

4.3.4 Step 4: Execution & Verification

The Agent attempts to execute the task via the OSC Layer.

Success: The state change is confirmed by Ableton (via an OSC reply).

Failure: If the track doesn’t exist or the device is missing, an error flag is raised.

4.3.5 Step 5: The Summary Agent

Regardless of success or failure, the Summary Agent aggregates the logs. It translates the technical outcome back into natural language for the user (e.g., “I created a hiphop melody on track 1 using a minor pentatonic scale in C minor with syncopated rhythms.” or “I couldn’t create the melody—no MIDI track was available. Should I create a new track?”).

4.4 Visual Architecture Diagram

Figure 1 illustrates the system architecture. The diagram shows the flow from the React frontend through the Flask/Marvin core, which includes the Disambiguator Node and Router that splits into specialized agents (Mixing Agent, Arrangement Agent). These agents consult the Knowledge Priors database before executing commands via the OSC protocol to Ableton Live. The feedback loop completes through OSC responses processed by the Summary Agent, which returns natural language feedback to the React UI.

Figure 1: System Architecture Diagram

5 Implementation Steps

The development of the system followed an iterative, incremental methodology. We began with a minimal viable command-line architecture to validate the core logic before progressively expanding into a full-stack application with a multi-agent swarm. This process was divided into five distinct phases.

5.1 Phase 1: Core Architecture & Agent Framework

The initial phase focused on establishing the foundational logic without the overhead of web protocols. We developed the core orchestration framework, which included the bi-directional Open Sound Control (OSC) layer for communicating with Ableton Live and the memory management system.

During this stage, the primary multi-stage pipeline—*Disambiguation* → *Classification* → *Extraction* → *Execution* → *Summarization*—was implemented and validated using a single “Song Agent.” This isolated environment allowed us to perfect the conversation management (creation, resumption, and deletion) and task execution logic before introducing the complexity of a full agent swarm.

5.2 Phase 2: API Exposure

Once the CLI workflow was stable, we encapsulated the core logic within a Flask-based RESTful API. This phase required mapping the internal Python agent methods to exposed HTTP endpoints, establishing a standard for request handling, error propagation, and status codes. Rigorous integration testing was conducted using Postman to ensure that the API layer correctly translated external HTTP requests into the underlying agent operations, effectively preparing the system for client-side integration.

5.3 Phase 3: Frontend Integration

To bridge the gap between the backend logic and the end-user, we developed a React.js frontend. This phase focused on User Experience (UX), specifically the creation of a chat interface capable of capturing natural language input and rendering the system’s real-time state. The frontend was tightly coupled with the Flask endpoints, requiring end-to-end validation to ensure that asynchronous agent responses were accurately visualized in the browser, thereby completing the full-stack loop.

5.4 Phase 4: Domain Expansion (The Agent Swarm)

With the full-stack architecture in place, the system was systematically scaled from a single agent to a comprehensive suite of eight specialized domain agents. We adopted a standardized expansion pattern for each new domain: defining the specific context, implementing the agent instructions, developing the OSC tools, and finally performing integration testing.

The resulting swarm covers the full breadth of DAW operations:

- **Structural Management:** The *Track*, *Scene*, *View*, and *Application Agents* manage the session layout and global settings.
- **Sonic Manipulation:** The *Device*, *Clip*, *Clip Slot*, and *Song Agents* handle the direct modification of audio and MIDI data.

By adhering to a modular implementation strategy, we ensured that adding complex capabilities (such as device control) did not disrupt the stability of the existing architecture.

5.5 Phase 5: Multi-Agent Collaboration & Optimization

The introduction of multiple agents necessitated a dedicated optimization phase. Initial stress testing with complex, multi-step prompts (e.g., “Create a new track and add a reverb”) revealed bottlenecks in state synchronization and error propagation across agent boundaries.

To address this, we refined the coordination mechanisms to ensure memory consistency. This involved optimizing the task execution pipeline to handle sequential dependencies between agents and implementing robust recovery strategies for partial failures. This phase was critical in transforming a collection of individual tools into a cohesive, collaborative system.

6 Experimental Setup & Results

To validate the efficacy of the system, we designed a comparative Human-Computer Interaction (HCI) study. The primary objective was to quantify the "Efficiency Lift" provided by the semantic interface compared to traditional manual workflows in Ableton Live. The study focuses on three core metrics: Time to Task Completion (TTC), Cognitive Load (NASA-TLX), and Semantic Accuracy.

6.1 Participants

The study design calls for $N = 10$ participants with varying levels of music production expertise, stratified into two groups:

- **Novices (n=5):** Less than 1 year of experience; familiar with concepts but struggle with technical implementation.
- **Experts (n=5):** 5+ years of experience; highly proficient with Ableton Live shortcuts and signal flow.

This stratification allows us to test the hypothesis that the system democratizes complex production techniques for novices while accelerating workflow for experts.

6.2 Experimental Design

We utilized a **Within-Subjects Design**, where each participant performs a set of standardized production tasks using two distinct methods. The order of methods is counterbalanced to prevent learning effects.

1. **Control Condition (Manual):** Participants execute tasks using the standard mouse, keyboard, and browser search (for technique lookup) within Ableton Live.
2. **Experimental Condition (Semantic-Live):** Participants execute the same goals using the natural language chat interface.

6.3 Task Design

The experiment consists of three tasks increasing in complexity and abstraction, designed to stress-test specific architectural components:

Task A: Explicit Command (Structural)

"Create three Audio tracks, rename them to 'Vox', 'Gtr', and 'Bass', and group them."

- *Objective:* Tests the **Track Agent's** ability to handle multi-step batch operations.

Task B: Technical Configuration (Parameter)

"Add a Compressor to the Bass track and sidechain it to the Kick drum."

- *Objective:* Tests the **Device Agent's** routing capabilities. This is a high-friction task manually (requires clicking multiple sub-menus).

Task C: Abstract Semantic Request (Knowledge Prior)

"The vocals sound too thin. Make them warmer and more intimate."

- *Objective:* Tests the **Translation Layer**. In the Manual condition, the user must know *which* tools create "warmth" (e.g., EQ low-mid boost, saturation). In the Experimental condition, the system uses its Knowledge Priors to execute the chain automatically.

6.4 Evaluation Metrics

We capture both quantitative system logs and qualitative user feedback:

6.4.1 Quantitative Metrics

- **Time to Task Completion (TTC):** Measured in seconds from the start of the prompt/action to the final state verification.
- **Interaction Steps:** The total number of discrete clicks or keystrokes required. This measures physical effort.
- **Token Efficiency:** For the Experimental condition, we measure the total tokens consumed per task to validate our claim that the "No-MCP" architecture is cost-effective.

6.4.2 Qualitative Metrics (NASA-TLX)

After each condition, participants complete a simplified NASA Task Load Index (NASA-TLX) survey to rate:

- **Mental Demand:** "How much thinking was required?"
- **Frustration:** "How insecure or discouraged did you feel?"
- **Perceived Success:** "How satisfied were you with the sonic result?"

6.5 Hypothesis

We hypothesize that:

- **H1:** The Experimental condition will show a $> 50\%$ reduction in TTC for Task C (Abstract Request) across all user groups.
- **H2:** Novices will report significantly lower Mental Demand using the system compared to manual configuration.
- **H3:** The system will maintain a sub-200ms latency for command acknowledgement, validating the architectural decoupling strategy.

6.6 Results

The execution of the experimental study and collection of results is deferred to future work. The experimental framework described above provides a rigorous methodology for validating the system's efficacy once the study is conducted with participants.

7 Meta Skills

This project required moving beyond standard software engineering practices into the realm of experimental AI system design. The challenges encountered—specifically regarding latency, hallucinations, and protocol translation—necessitated the development of several key meta-skills.

7.1 Hybridizing Stochastic and Deterministic Logic

A primary learning outcome was realizing the limitations of “pure” Generative AI in real-time systems. Initially, the temptation was to allow the LLM to manage the entire state via the Model Context Protocol (MCP). However, we identified that this approach introduced unacceptable latency and token costs (“burning tokens”) for a creative workflow.

We developed the skill of **Architectural Decoupling**: recognizing which parts of a system require creative reasoning (the “What”) and which require rigid, hard-coded execution (the “How”). By implementing “Knowledge Priors” (deterministic look-up tables) alongside the LLM, we learned to optimize for computational efficiency without sacrificing the flexibility of natural language, effectively bridging the gap between static automation and dynamic AI agents.

7.2 Constraint Engineering in Probabilistic Systems

Working with Large Language Models introduces non-determinism—a critical risk when controlling live audio software where a hallucinated command could corrupt a session. This project shifted our focus from simple “Prompt Engineering” (trying to persuade the model) to **Constraint Engineering** (forcing the model to obey syntax).

By utilizing libraries like Marvin and Pydantic to enforce strict schema validation *before* any command reaches the OSC layer, we learned to treat the LLM not as a magic box, but as an untrusted component that requires rigorous sanitation. This meta-skill involves designing “guardrails” that allow for creative expression while mathematically guaranteeing system stability.

7.3 Semantic Mapping of Abstract Domains

A significant challenge was translating vague human artistic intent (e.g., “make it punchier”) into precise numerical parameters (e.g., “set Compressor Ratio to 4:1, Attack to 10ms”). This required developing the skill of **Domain-Specific Abstraction**.

We had to codify the tacit knowledge of a music producer into a structured logic that a machine could parse. This went beyond coding; it required analyzing the “physics” of music production and creating a translation layer that maps qualitative adjectives to quantitative signal processing commands. This experience highlighted the difficulty and necessity of grounding AI reasoning in domain-specific reality rather than general text patterns.

8 Conclusion & Future Work

8.1 Refinement & Stabilization

The current development phase focuses on system hardening. While the functional feature set is complete, continuous refinement is applied to improve response latency and handle

edge cases in user intent. This stage prioritizes the resolution of bugs discovered during real-world usage scenarios and the optimization of the OSC message throughput to ensure a seamless, real-time user experience.

9 Deliverables

References

- [1] Ableton Copilot Project. Ableton copilot: Mcp-based session management for ableton live, 2024. Project Documentation.
- [2] Giorgio Agostini, Emmanouil Benetos, and Tillman Weyde. Deep learning for music information retrieval: A survey. *International Journal of Multimedia Information Retrieval*, 10(2):99–127, 2021.
- [3] Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka, and Frédéric Bevilacqua. Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems*, 4(4):1–34, 2014.
- [4] Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt’s behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.
- [5] DAWZY Project. Dawzy: Voice and text-driven assistant for reaper. <https://github.com/your-username/dawzy>, 2024. GitHub Repository.
- [6] Brecht De Man, Joshua D. Reiss, and Ryan Stables. Ten years of automatic mixing. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-13)*, 2013.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [8] Rebecca Fiebrink, Perry R. Cook, and Daniel Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 147–156, 2011.
- [9] FilmComposer Project. Filmcomposer: Multi-agent system for silent film music arrangement, 2024. Project Documentation.
- [10] Qiaoqiao Huang, Aren Jansen, Andy Lee, Rithesh Ganti, Yuzi Li, Daniel P. W. Ellis, et al. Musicgen: Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

- [11] Qiaoqiao Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Qingsong Chen, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
- [12] Loop Copilot Project. Loop copilot: Llm integration with global attribute table for iterative editing, 2024. Project Documentation.
- [13] MixAssist Project. Mixassist: Audio-language dataset for expert mixing advice, 2024. Project Documentation.
- [14] Model Context Protocol. Model context protocol specification. <https://modelcontextprotocol.io/>, 2024.
- [15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- [16] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.
- [17] Matthew Wright, Adrian Freed, and Ali Momeni. Open sound control: State of the art 2003. In *Proceedings of the 2003 Conference on New Interfaces for Musical Expression (NIME-03)*, pages 153–159, 2003.