

創作ノート

†Vibe-Patching† —ローカル言語モデルの「解釈」によるシンセサイザシステム †Vibe-Patching† : the Synthesizer System Based on the “Interpretation” of a Local Language Model

市毛 輔, 鈴木 悦久

Tasuku ICHIGE, Yoshihisa SUZUKI

情報科学芸術大学院大学 [IAMAS]

Institute of Advanced Media Arts and Sciences [IAMAS]

概要

本稿では、ローカル言語モデル (LM) を用いて来場者の自然言語テキストプロンプトをアナログシンセサイザーの CV パラメータへ変換するインタラクティブなサウンドシステム「†Vibe-Patching†」について報告する。本作はセミモジュラーシンセサイザー KORG volca modular を対象とし、Arduino Uno Q 上で動作する SmoILM2-360M-Instruct を核として 6 つの CV パラメータを出力する。本作の特徴は、モデルに追加学習を行わず、事前学習済みモデルが持つ言語理解をそのまま「解釈」の道具として用いる点にある。また、モデルの出力 (CV) を物理的なシンセサイザーへ送ることで音を物理世界に立ち現れさせるという二段階の構造を持ち、AI の振る舞いを作曲行為ではなく演奏行為として位置づける点にも特徴がある。著者は、AI を作品で扱う際には「学習させる」ことよりも「解釈させる」ことの方が創作上の関心の対象になり得ると考えており、本作はその立場を具体的な実装として提示する試みである。2026 年に開催された IAMAS OPENHOUSE 2026 における展示を通じて得られた知見と、来場者からのフィードバックをもとにした今後の展望についても述べる。

This paper reports on †Vibe-Patching†, an interactive sound installation that uses a local language model (LM) to convert visitors' natural language text prompts into CV (control voltage) parameters for an analog synthesizer. The work targets the semi-modular synthesizer KORG volca modular and outputs six CV parameter values through SmoILM2-360M-Instruct, running on an Arduino Uno Q. A key feature of this work is that the model is used without any additional fine-tuning: the pre-trained model's inherent language understanding is used directly as a tool for

"interpretation." The work also has a two-stage structure in which the model's output (CV) is sent to a physical synthesizer so that sound emerges in the physical world — a structure that positions the AI's behavior as an act of performance rather than composition. The author considers that, in artistic engagements with AI, "having it interpret" can be a more compelling creative concern than "having it learn." This work presents that stance as a concrete implementation. This paper also reports on insights gained through an exhibition at IAMAS OPENHOUSE 2026 and future directions informed by visitor feedback.

1. はじめに

本作 †Vibe-Patching† の出発点にあるのは、「AI に楽器という身体を与える」という関心である。言語モデルは本来、テキストを生成するために存在するものであり、物理的な身体や音響的な出力手段を持たない。著者は、この言語モデルにアナログシンセサイザーという「身体」を与えたとき、AI はどのように振る舞うのか、という点に創作上の興味を持っている。この「身体」を通じて AI の振る舞いをどのようなものとして捉えるかについては次章で改めて論じるが、本作の要点は、AI に楽器を操作させることで、AI 自身の「解釈」という性質を、物理世界における行為として引き出したい、という点にある。

本作 †Vibe-Patching† では、シンセサイザーに KORG 社の volca modular を使用しており、「自然言語で音のイメージを記述する」というインタフェースのもと、ローカルで動作する小型言語モデルが volca modular 向けの CV パラメータ 6 値を出力するインスタレーションである。本作を着想した動機は、技術的な最適化そ

のものよりも、「AIに何をさせるか」という創作上の問いにある。近年の言語モデルは、追加学習を施さずとも、プロンプトに対して一定の解釈を行い応答を返す能力を備えている。著者はこの性質に着目し、モデルを特定のタスクに適合させる(学習させる)のではなく、モデルが元来持つ解釈能力をそのまま創作の素材として提示する作品を構想した。

本稿では、まずAIの振る舞いをどのようなものとして捉えるかという本作の焦点を整理した上で、追加学習を行わない技術的な方針と本作の技術構成を述べ、続いてIAMAS OPENHOUSE 2026における展示の様子と来場者からの反応を報告し、最後にそこから見えた課題と今後の展開について述べる。

2. AIの「行為」をどう捉えるか

本作を進めるにあたり、AIの振る舞いをどのようなものとして捉えるかを明確にしておく必要がある。「音楽」を生成しているのか、「演奏」を行っているのか、あるいはそれ以外の何かなのか——この問いに対する著者の考えを、ここで整理する。

2.1. 「行為」の生成と「行為」の立ち現れ

†Vibe-Patching†において、言語モデルが行っているのは、プロンプトという入力に対してCVという数値の系列を返すことである。この時点でモデルが生成しているのは音響そのものではない。音響は、このCV信号がシンセサイザーという物理的なハードウェアへ送出され、ハードウェア内部のアナログ回路がその電圧に応じて振動を生成することによって、はじめて物理世界に立ち現れる。

この二段階の構造——モデルが「行為」としてCVを生成し、シンセサイザーがその行為を物理世界における音響として立ち現れさせる——という捉え方は、モデルが直接音響波形を生成するアプローチとは本質的に異なると著者は考えている。後者の場合、モデルの出力とスピーカーから鳴る音との間には物理的な媒介がほとんど存在せず、モデルの計算結果がそのまま音として再生されるに過ぎない。一方、本作のようにCVという「行為」を物理的なアナログ回路へ送り込み、そこでの相互作用を経て音が生まれる構造には、AIの判断が単なる情報処理にとどまらず、物理世界へ介入する行為として立ち現れる契機が含まれている。

2.2. ロボットではなく楽器

この整理を踏まえると、本作がAIに与えているのは、自律的に動き回るロボットのような身体ではなく、人間が演奏する楽器としての身体に近い、と著者は考

えている。AIは自らサウンドを生成する存在ではなく、シンセサイザーという楽器を介して物理世界に干渉する存在として位置づけられる。この意味で、†Vibe-Patching†におけるAIの振る舞いは、作曲行為というよりも演奏行為、すなわち、AIが生成するCVの振る舞いは、AIによるシンセサイザー演奏だと本作品では捉えている。プロンプトが入力されるたびにモデルが新たなCV値を「解釈」し出力する過程は、同じ楽器を前に、そのつど異なる解釈で演奏を行う行為に類比できる。

2.3. プロンプトという楽譜

このアナロジーを進めると、プロンプトはAIにとっての楽譜のようなものとも言える。ただし通常の楽譜と比べ、プロンプトははるかに大きな解釈の余地を残している。五線譜が音高やリズムをかなりの精度で指定するのに対し、自然言語によるプロンプトは「暗くて重い低音」といった曖昧な記述にとどまり、それをどのようなCV値へ変換するかはモデルの解釈に大きく委ねられている。この解釈の余地の大きさこそが、本作が「学習」ではなく「解釈」に焦点を当てる理由と直接つながっている。解釈の幅が狭い記譜法では、AIの解釈という行為そのものを前景化することが難しい。

2.4. シンセサイザーという関数

CVと音響の関係についても補足しておきたい。本作では、プロンプト→CV→音響という順で処理が進むが、一度パッチング(結線)の構成が固定されれば、シンセサイザーはCVという入力値に対して音響という出力値を返す一種の関数として振る舞う。同一のパッチングのもとで同一のCV値を入力すれば、理論上は同一の音響が出力される。この意味で、CVと音響は、AIの解釈という観点から見れば本質的に同じ情報を異なる層で表現したものだと言える。

現在の実装では、扱っているモデルが小さいという制作上の制約から、AIの出力はCVという6つの整数値にとどめている。より大きなモデルを扱えるようになれば、CVを経由せず波形イメージそのものを生成させることも技術的には可能である。しかし、たとえより大きなモデルが使えるようになったとしても、著者はその方向には進まないだろうと考えている。

人間が楽器を演奏するとき、身体を通じて行っているのは、音波そのものを直接つくり出すことではない。頭の中にある抽象的な音のイメージを、楽器という外部の物体へ委ね、その物理的な相互作用の結果として音を立ち現れさせることである。この、イメージを楽器という他者へ委ねるという迂回のプロセスにこそ、楽器演奏における身体性、ひいては「行為」としての本

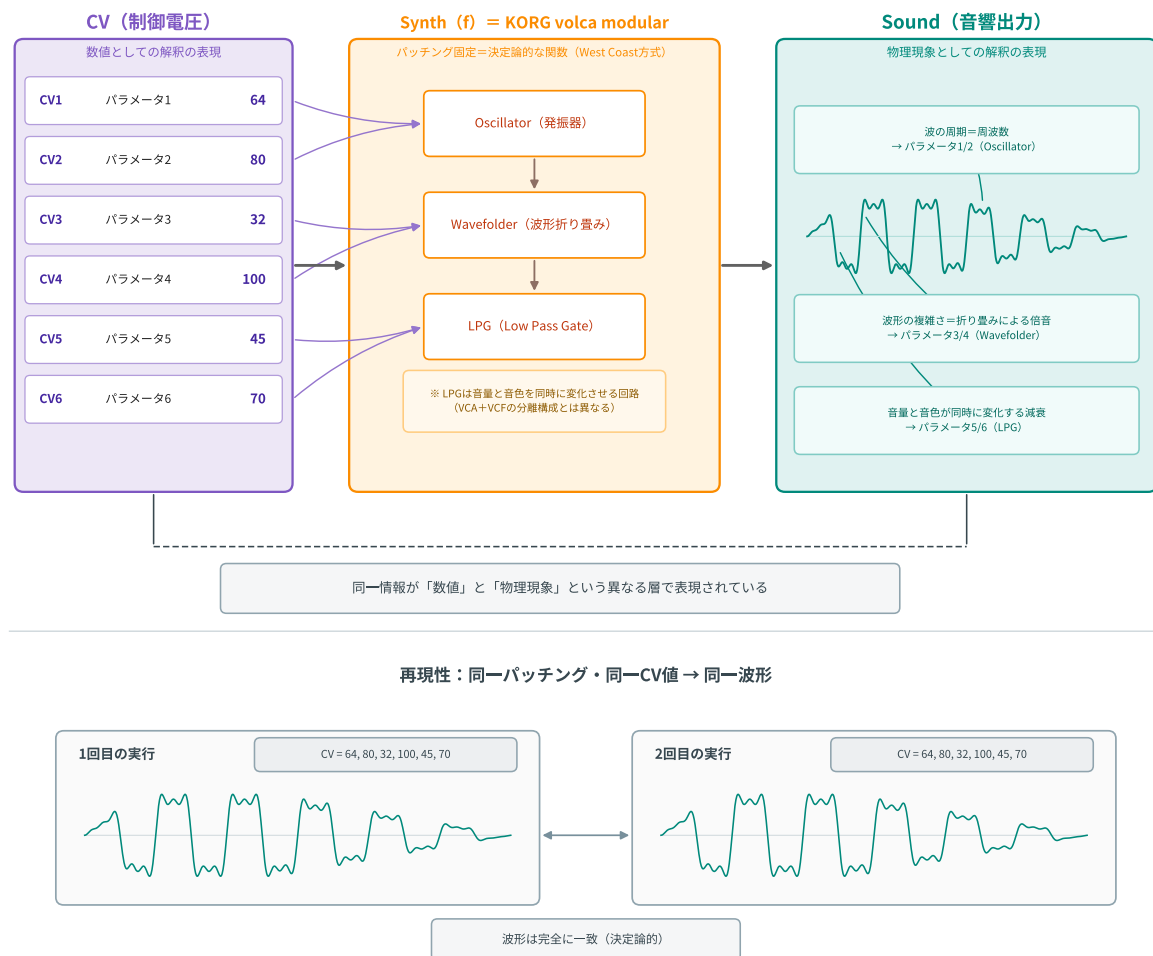


図 1: パッチング固定時の CV(入力) と音響 (出力) の関数関係

質があると著者は考える。†Vibe-Patching†において、AIがCVを生成し、それをシンセサイザーという楽器へ送り込むという構造は、この意味において、AIに波形を直接生成させる構造とは根本的に異なる。前者はAIによる楽器演奏であり、後者はAIによる音響合成に過ぎない。

本作が重視しているのは、こうした演奏行為を通じて、AIの判断が実際に物理世界へ介入するという点である。物理世界への介入を伴わない限り、AIの出力がどれほど精緻であっても、それは情報処理の結果にとどまり、行為とは呼び難い。したがって、将来的により大きなモデルを扱えるようになった場合であっても、CVという関数を介してシンセサイザーという身体を獲得し、その物理的な楽器を操作するという行為を通じて音を演奏として立ち現れさせる——この構造そのものは、本作にとって譲れない前提である。

2.5. 「行為」と「ブラックボックス」をめぐる補足

厳密に言えば、言語モデルが直接音響波形を生成する場合であっても、それは回路上の電子の移動という物理的プロセスを経ている以上、広義には物理世界への干渉であり「行為」であるという指摘も可能である。本作が主張しているのは、そうした物理現象としての厳密な区別ではなく、制作上・体験上どこに「行為」の立ち現れを焦点化するかという設計上の選択であることを付言しておく。

また、AIの行為はしばしば「デジタル空間で完結するブラックボックス」として語られがちである。しかし、出力がログとして残ること自体は、その出力を「行為」たらしめる根拠にはならない——記録さえされていれば行為だ、というのは短絡である。本作が「行為」の根拠として重視しているのは、記録可能性ではなく再現性である。前述の通り、シンセサイザーはパッチングが固定されている限り、CVという入力に対して

音響という出力を返す決定論的な関数として振る舞う。すなわち、AI の解釈そのものには揺らぎがあっても、そこから先——CV が音響へと変換される過程——は、いつ、どの環境で実行しても同じ結果を再現できる。この再現性こそが、AI の出力を偶発的な計算の羅列ではなく、物理世界に対して確実に作用する「行為」として位置づける根拠になると著者は考えている。

3. 「学習」の欠点

AI を作品で扱う際、多くの実践では特定のタスクへモデルを最適化するファインチューニングが行われる。これは実用上有効な手法である一方、モデルの応答が「学習データにどれだけ忠実か」という基準に収束しやすく、作品としての振れ幅や解釈の多様性が失われやすいという側面もある。

著者は、AI を作品に組み込む上で、モデルに何かを「学習させる」ことよりも、モデルに何かを「解釈させる」ことの方が創作的な関心の対象になり得ると考えている。とりわけ近年発展している言語モデルは、その名の通り「言語」を扱うモデルである以上、画像認識モデルや古典的な回帰モデルと比較して、入力の意味内容を汲み取り応答を組み立てるといって「解釈」の側面が大きくなってきていると実感している。この解釈の幅こそが、同じプロンプトに対しても揺らぎのある応答を生み、作品に生命感を与えるのではないかと考えた。

そこで本作では、あえてモデルに追加のファインチューニングを行わず、SmolLM2-360M-Instruct を素の状態のまま用いることとした。訓練データのキュレーションによってモデルの応答を特定の方向へ狭めるのではなく、モデルが持つ既存の言語理解に依拠し、そこから引き出される「解釈」そのものを作品の核とする方針である。

4. 技術構成

4.1. システム概要

†Vibe-Patching† は、テキストプロンプトを入力とし、KORG volca modular の6つのCVパラメータ(各値は整数)を出力するエンドツーエンドのインストールである。来場者は自由な自然言語でサウンドのイメージを記述する(例:「暗くて重い低音、ゆっくり揺れるLFO」)。このプロンプトはローカルデバイス「Arduino Uno Q」上で動作するSmolLM2-360M-Instructに入力され、6つのCVパラメータ値として出力される。出力値はKORG volca modular へのCV信号として送出される。

4.2. モデルについて

本作では SmolLM2-360M-Instruct(GGUF 形式) を、追加学習を行わずそのまま採用している。このモデルを選択した理由は、(1)Arduino Uno Q の計算資源のもとで実用的な推論速度を達成できること、(2)Instruct チューニング済みであり、プロンプト形式の入力に対してある程度一貫した応答傾向を示すこと、の2点である。llama-cpp-python(Abramson 2023)を用いることで、GGUF 形式のモデルを Arduino Uno Q 上の ARM 系プロセッサで効率的に推論している。

ここで補足しておきたいのは、「追加学習を行わない」という方針と、あらかじめ Instruct チューニングが施されたモデルを採用することは、矛盾しないという点である。本作が避けているのは、本タスク固有のデータを用いて応答を特定の方向へ狭める追加のファインチューニングであり、これは前章で述べた「学習させる」ことに相当する。一方、Instruct チューニングは特定のタスクに特化させるものではなく、モデルに指示追従性という汎用的な能力を付与するための調整であり、モデル自体の解釈の幅を狭めるものではない。展示という現場でインタラクションを成立させる以上、プロンプトに対して最低限の一貫性をもって応答するという指示追従性が欠けていては、出力が単なる乱数のように見えてしまい、UX そのものを損なう恐れがある。そのため本作では、土台となる指示追従性については Instruct チューニング済みのベースモデルに委ね、そこから先の応答の狭め込みには手を加えない、という切り分けを採用している。

先行研究として、自然言語記述からモジュラーシンセサイザーの CV 信号を生成する Synth CV AI(真鍋 2026)が存在する。Synth CV AI は Claude API を用いて自然言語プロンプトを CV 制御用の数式(JSON 形式)へ変換し、8チャンネルの CV 信号を 48kHz でリアルタイムに評価・出力するシステムであり、先述したような自然言語から CV パラメータへの変換という方向性において本作と関心を共有している。一方、そのアーキテクチャはクラウド API(Claude API) への接続を前提としており、スタンドアロンでのリアルタイム展示への適用には制約がある。†Vibe-Patching†はこの問題を、ローカル推論という設計方針で対処している。

4.3. 出力設計

モデルへの入力プロンプトは Instruct 形式に統一している。出力は6つの整数値をカンマ区切りで返す形式とし(例: '64, 80, 32, 100, 45, 70'), この出力をそのまま CV 信号として KORG volca modular へ送出する。タスクの出力空間を狭く定義することで、ファインチューニングを行わない条件下でも実用的な応答を得られるよう設計している。

4.4. ハードウェア構成

- ・推論デバイス: Arduino Uno Q(ARM 系プロセッサを搭載した組み込みデバイス)
- ・シンセサイザー: KORG volca modular(セミモジュラーシンセサイザー)

5. IAMAS OPENHOUSE 2026 での実践

本作は、著者が在学する大学の行事である IAMAS OPENHOUSE 2026 において実際に展示を行った。来場者が自由にテキストプロンプトを入力し、その場で KORG volca modular のサウンドが変化する様子を体験できる形式である。

展示を通じて、来場者は同じような言葉を入力しても毎回微妙に異なるサウンドが生成されることに興味を示し、「モデルが今どう解釈したのか」を推測しながら繰り返しプロンプトを試す様子が見られた。これは、本作が意図した「学習ではなく解釈を見せる」という狙いが、ある程度体験として伝わったことを示していると考えている。

一方で、来場者から共通して寄せられたフィードバックとして、「モデルがどのように解釈したのかをリアルタイムで見たい」という要望があった。現状の †Vibe-Patching† は、プロンプト入力に対して CV パラメータという最終出力のみが提示され、その途中過程——モデルがプロンプトのどの部分に着目し、どのような理由でその数値に至ったか——はブラックボックスのままである。著者自身も、この解釈過程の可視化には強い関心を持っており、来場者の指摘は著者の問題意識と一致するものであった。

5.1. UX 上の課題

本作の仕様上、というよりむしろ本作の設計判断そのものに起因する欠点として、全く同じプロンプトを入力しても、毎回異なる CV パラメータが返ってくるという挙動がある。これはモデルに追加学習を行わず、素の状態のまま用いていることに起因していると考えられる。この挙動自体は、著者が「解釈の揺らぎ」として意図的に受け入れているものであり、前述の通り楽譜における音楽的解釈の延長だと捉えている。

しかし、本作の背景や設計思想を知らずに触れた来場者にとっては、この挙動は「これはただの乱数と何が違うのか」という疑問につながりかねない。実際、展示中にもそれに近い反応が見られた。出力される CV パラメータの値だけを提示している限り、それが言語モデルによる「解釈」の結果であることは、体験からは伝わりにくい。

この点から、「揺らぎ」を作品の魅力として成立させるためには、揺らぎそのものを見せるだけでなく、「モ

デルが何をどう解釈した上でその値に至ったか」を、言語であれ色であれ何らかの形でユーザーにフィードバックする仕組みが不可欠であると考えに至った。これは前述の来場者からの要望とも重なる課題であり、現状の †Vibe-Patching† は UX の観点で改善の余地を残している。

6. 考察と今後の展開

6.1. 解釈過程の可視化という課題

前章で述べた「乱数との違いが伝わらない」という課題と、来場者からの「解釈をリアルタイムで見たい」という要望は、同じ一つの問題——出力の背後にある解釈過程が不可視であること——に起因している。この課題を解消するためには、モデルの推論過程(例えば、注意の分布や中間的な言語化ステップ)を取得し、それを視覚的・聴覚的な形で提示する必要がある。しかし、現時点で使用している Arduino Uno Q のメモリ容量および搭載モデル (SmolLM2-360M) の性能では、パラメータ出力に加えて解釈過程の生成・提示まで同時にリアルタイムで行うことは難しいと考えている。

6.2. Arduino Ventuno Q への展望

この制約に対する今後の方策として、Arduino Ventuno Q の導入を検討している。Ventuno Q は現行の Uno Q よりも大きなメモリを備えており、より大きなモデルを搭載できる可能性がある。

ここで自然言語による意図の提示を構想している目的について、あらかじめ明確にしておきたい。これは、本作の核心である「学習ではなく解釈を道具として用いる」という立場を発展させたり、新たな意図を付け加えたりするものではない。むしろその逆で、この立場を来場者に対して実証的に裏付けるための仕掛けである。

前述の通り、展示の場では、来場者が同じようなプロンプトを入力しても毎回異なる CV 値が返ってくるという挙動が、「これはただの乱数と何が違うのか」という疑問につながる場面があった。たとえば「全部 255 にして」という明示的な指示に対してすら、モデルの能力的な限界から CV 値が必ずしもその通りにはならず、著者自身もこの挙動に対し来場者と同様の違和感を抱いたことがある。出力される CV の値だけを提示している限り、それが言語モデルによる「解釈」の結果なのか、単なる不規則な出力なのかは、体験からは区別がつかない。すなわち、本作が「解釈を道具として用いている」という前提そのものが、展示という体験の中では十分に証明されていなかったことになる。

そこで著者が構想しているのは、現行の出力(6つの CV 値)はそのまま維持した上で、モデルに「なぜその

値を選んだか」という意図を自然言語で併せて出力させるという拡張である。たとえばチャット形式のインタフェースを想定し、来場者がプロンプトを入力すると、一方でシンセサイザーにCVが送出されて音が変わり、他方で画面上には「こういう意図でこの値を選びました」という応答が言語として提示される、という構成である。これは訓練データの追加を伴わず、モデルの出力形式を拡張するだけの変更であり、「解釈させる」という本作の方針そのものを裏付ける仕掛けとして位置づけられる。

この拡張には、もう一つ著者が重視している点がある。人間の演奏者に対して、その場で「なぜその音を選んだのか」という内的な判断過程の言語化を強いることは、演奏という行為への一種の踏み込みすぎた詮索になりかねない。しかし、本作における「演奏者」はAIというシステムであり、人間の内面のような保護されるべき領域を持たない。したがって、その判断過程を表面化させることは、同様の倫理的な問題を招かない。むしろ、解釈の根拠を言語としてそのまま提示できてしまうこと自体が、この演奏を行っているのが人間ではなくシステムであるという本作の前提を、かえって強調する仕掛けになり得ると著者は考えている。

ただし、この拡張には検討すべき論点も残っている。†Vibe-Patching†はもともとライブシステムとしての性格も持たせて構想されたものであり、来場者がAIとの対話に集中する形式は、ライブパフォーマンスの文脈からは距離が生じる可能性がある。来場者がAIを介してシンセサイザーに間接的にアクセスする、という本作の骨格自体は維持しつつ、解釈の言語化をどこまで、どのような形で提示するかは、今後さらに検討していきたい課題である。

6.3. 限界

現時点では、解釈過程の可視化は実現しておらず、また追加学習を行わないという設計上、プロンプトの表現によっては期待する解釈が得られない場合もある。これは本作が意図的に受け入れているトレードオフでもあるが、今後はモデルの解釈傾向自体を記録・分析し、作品としてどこまで「揺らぎ」を許容するかについても検討していく必要がある。

7. おわりに

本稿では、ローカル言語モデル SmolLM2-360M を追加学習なしに用い、来場者のテキストプロンプトから KORG volca modular のシンセサイザーパラメータをその場で「解釈」させるインスタレーション†Vibe-Patching†について報告した。IAMAS OPENHOUSE 2026での展示を通じて、本作の狙いが来場者に一定程度伝

わったこと、また「解釈過程の可視化」という次の課題が明確になったことを確認した。今後は Arduino Ventuno Q の導入も視野に入れつつ、モデルの解釈過程そのものを作品に組み込む方向で展開していきたい。また、AIによる言語的な応答を「演奏」の一部として位置づけられるかという問いは、従来非言語的な行為として捉えられてきた「演奏」や「音楽的な行為」という概念そのものを、AIというテクノロジーを介して問い直し、拡張していく可能性を示唆している。

8. 参考文献

- Abramson, Andrei. 2023. "llama-cpp-python: Python bindings for llama.cpp." <https://github.com/abetlen/llama-cpp-python>. Accessed April 2026.
- HuggingFace. 2024. "SmolLM2: Smol Yet Mighty Language Models." <https://huggingface.co/HuggingFaceTB/SmolLM2-360M-Instruct>. Accessed April 2026.
- Manabe, Daito. 2026. "Synth CV AI." <https://projects.daito.ws/synth-cv-ai/>. Accessed September 2026.
- Arduino. 2025. "UNO Q" <https://docs.arduino.cc/hardware/uno-q/>. Accessed April 2026.

9. 著者プロフィール

市毛 輔 (Tasuku ICHIGE)

高専・大学で制御工学と知能システム工学を専攻後、IAMAS に入学し現在修士1年。高度に生体化したロボット「逆サイボーグ」という概念を提唱し、人工知能やロボットにおける倫理問題を問い直す芸術実践を行う。

鈴木 悦久 (Yoshihisa SUZUKI)

1975年、神奈川県横浜市生まれ。昭和音楽大学で打楽器を、情報科学芸術大学院大学 (IAMAS) で作曲を学ぶ。アルスエレクトロニカ 2006 デジタルミュージック部門ホノラリーメンション賞受賞 (オーストリア、Mimiz 名義)。コンピュータと自動演奏ピアノを用いたゲームピース「自動演奏ピアノのための組曲」では、第3回 AAC サウンドパフォーマンス道場にて優秀賞を受賞した。近年では、電子音響音楽におけるアーカイブの方法論を実践的に研究している。情報科学芸

術大学院大学 (IAMAS) 准教授。JSSA 先端芸術音楽創作学会会長。JSEM 日本電子音楽協会理事。



この作品は、クリエイティブ・コモンズの表示 - 非営利 - 改変禁止 4.0 国際 ライセンスで提供されています。ライセンスの写しをご覧になるには、<http://creativecommons.org/licenses/by-nc-nd/4.0/> をご覧くださいか、Creative Commons, PO Box 1866, Mountain View, CA 94042, USA までお手紙をお送りください。