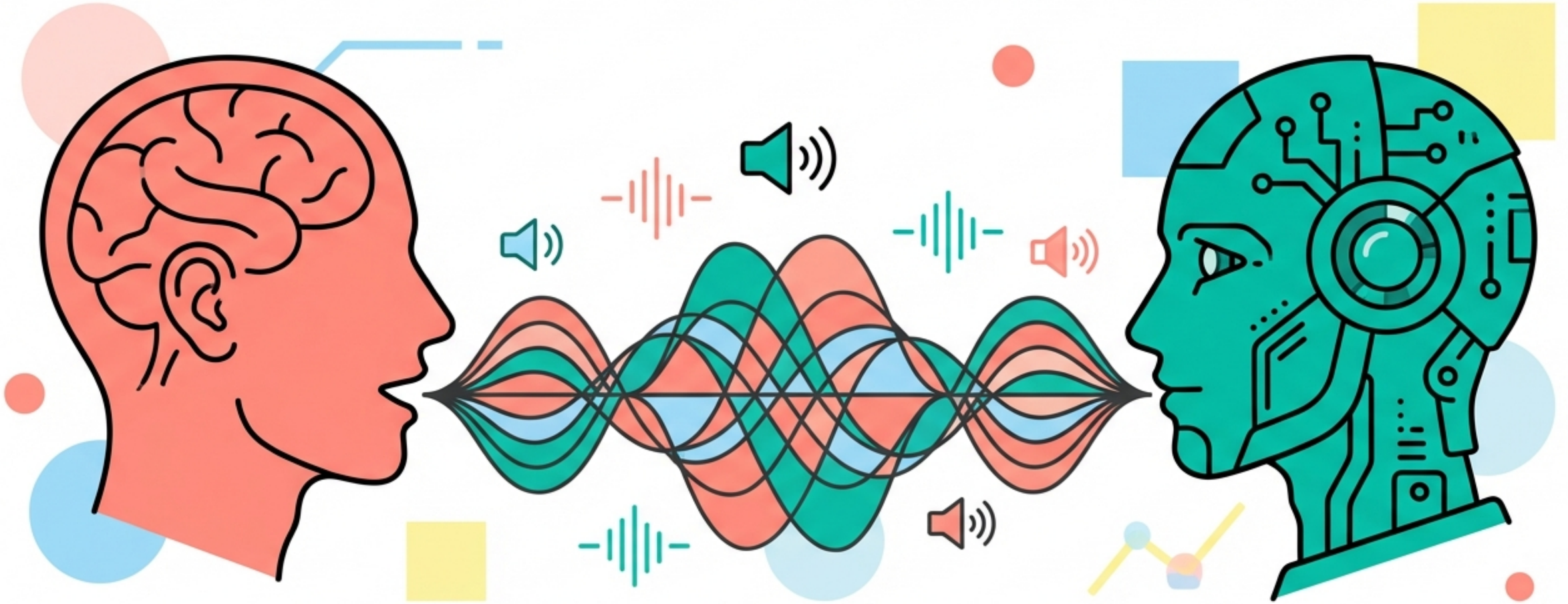


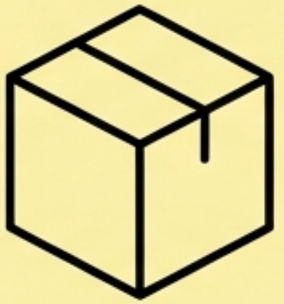
AIとの会話は、ついに「リアル」になる

NVIDIA PersonaPlex-7B：フルデュプレックス（全二重）がもたらす音声対話の革命

PersonaPlex-7B が変える4つの常識

フルデュプレックス (Full Duplex)

従来の「交互に話す」形式を撤廃。相手の話を「聞きながら同時に話す」ことが可能に。



単一モデル (Single Model)

ASR（音声認識）→ LLM → TTS（音声合成）の複雑なパイプラインを排除。単一の脳で全てを処理。

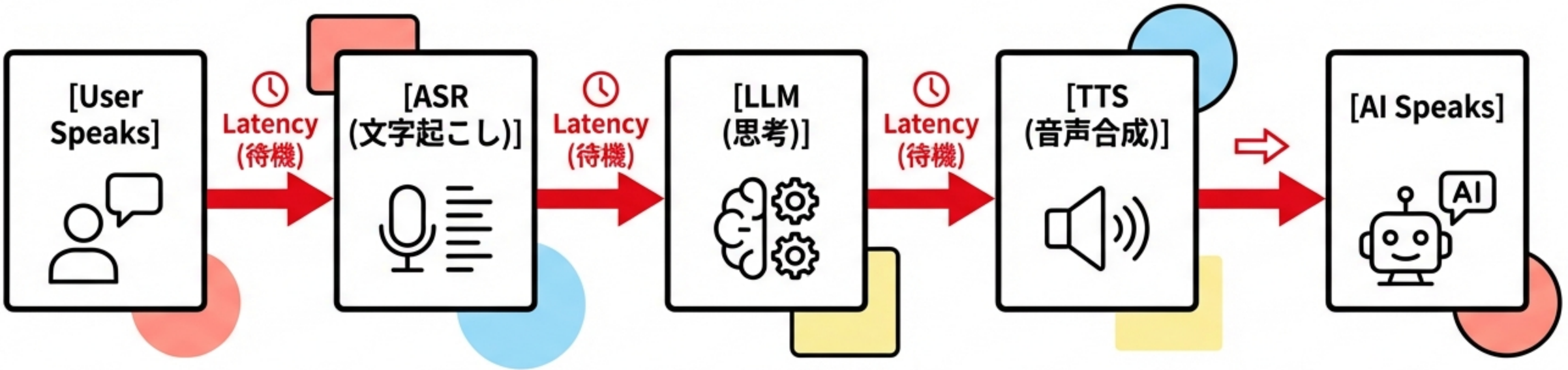
超低遅延 (Sub-second Latency)

割り込み検知から反応までわずか0.240秒。人間と区別がつかないリアルタイム性。

オープン＆商用利用可

NVIDIA Open Model License (CC-BY-4.0)。誰でも無償でビジネスに組み込める7Bモデル。

従来の音声AI：「トランシーバー」の時代

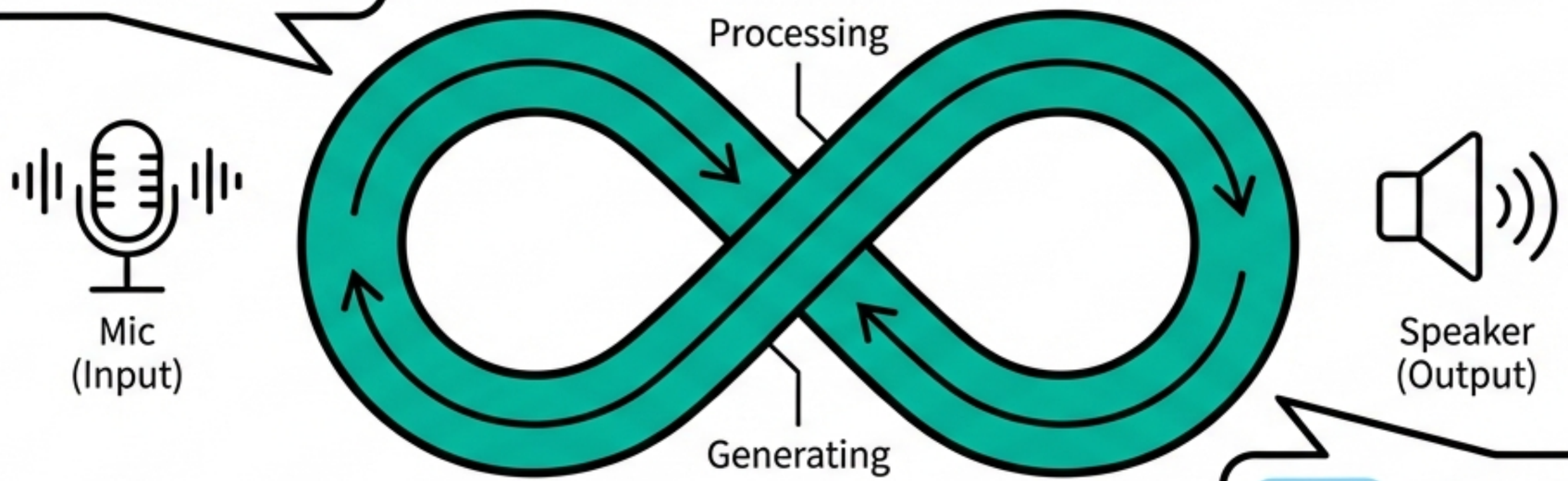


これまでの音声AIは、処理が直列（パイプライン）でした。ユーザーが話し終わるのを待ち、文字に起こし、考え、音声を作るまで、どうしても数秒の「不自然な沈黙」が発生します。

最大の問題点: AIが話している間、ユーザーは割り込めません。
まるでトランシーバーを使っているような硬直した体験です。

PersonaPlexの解決策：「電話」の時代へ

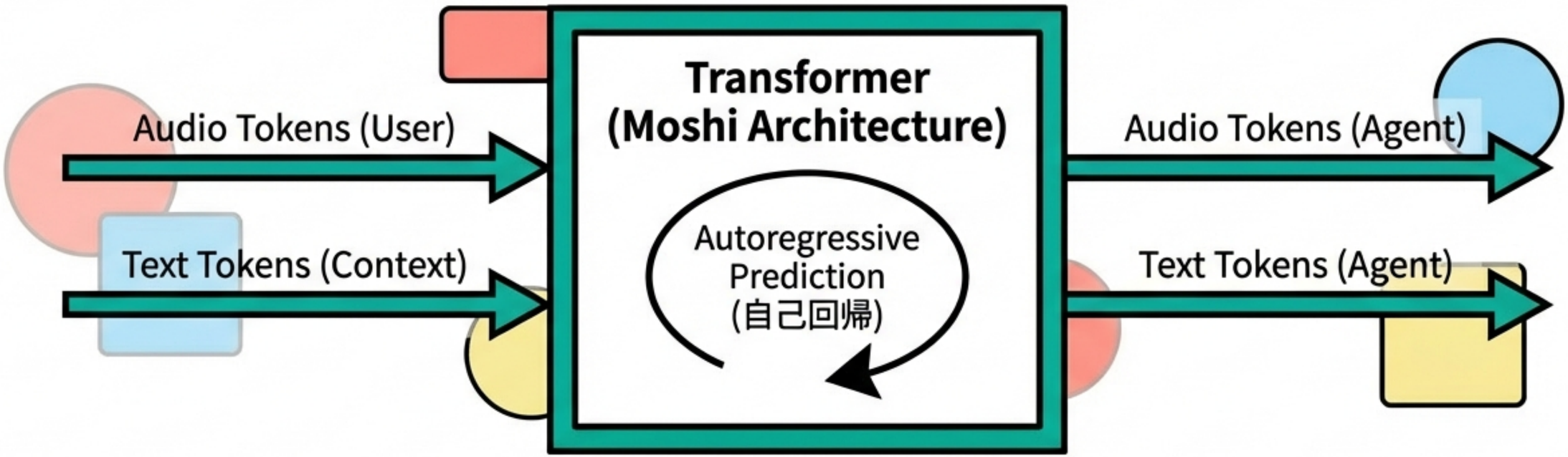
バージイン (Barge-in)：
話の途中で遮っても即座に対応。



バックチャネル:「うんうん」「へぇ」といった相槌をリアルタイムで生成。

PersonaPlexは、入力と出力を同時に処理する「デュアルストリーム」構成を採用。
AIは自分の発話中も、ユーザーの声を常に聞いています。

技術の核心：Moshiアーキテクチャ



PersonaPlex-7Bは、Kyutaiの「Moshi」をベースに開発されました。

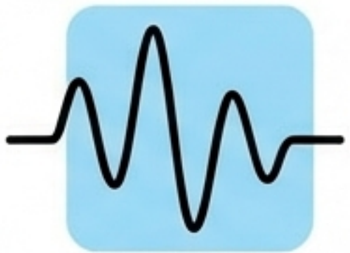
- **Speech-to-Speech: 音声トークンとテキストトークンを同時予測。 中間の「文字起こし」プロセスに依存しません。**
- パラメータ数: 7B (70億)
- 音声入出力: 24kHz (WAV/WebAudio)
- コンテキスト: 音声とテキストの両方の文脈を理解

自在なコントロール：ペルソナと声質

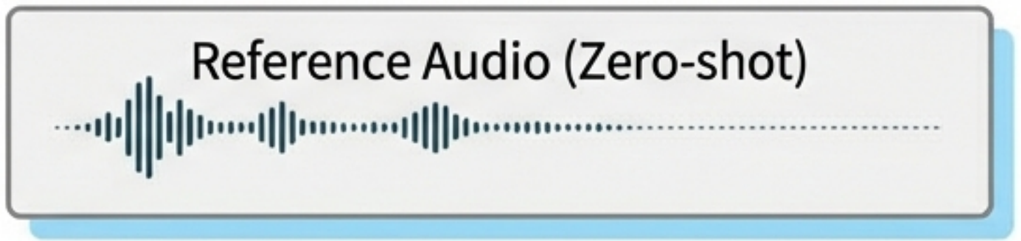
1. Text Prompt（役割設定）

```
Role: Sarcastic Astronaut
Context: Mars mission panic
```

背景、性格、シチュエーションをテキストで指示。「皮肉屋の宇宙飛行士」や「親切な教師」など、詳細なロールプレイが可能。

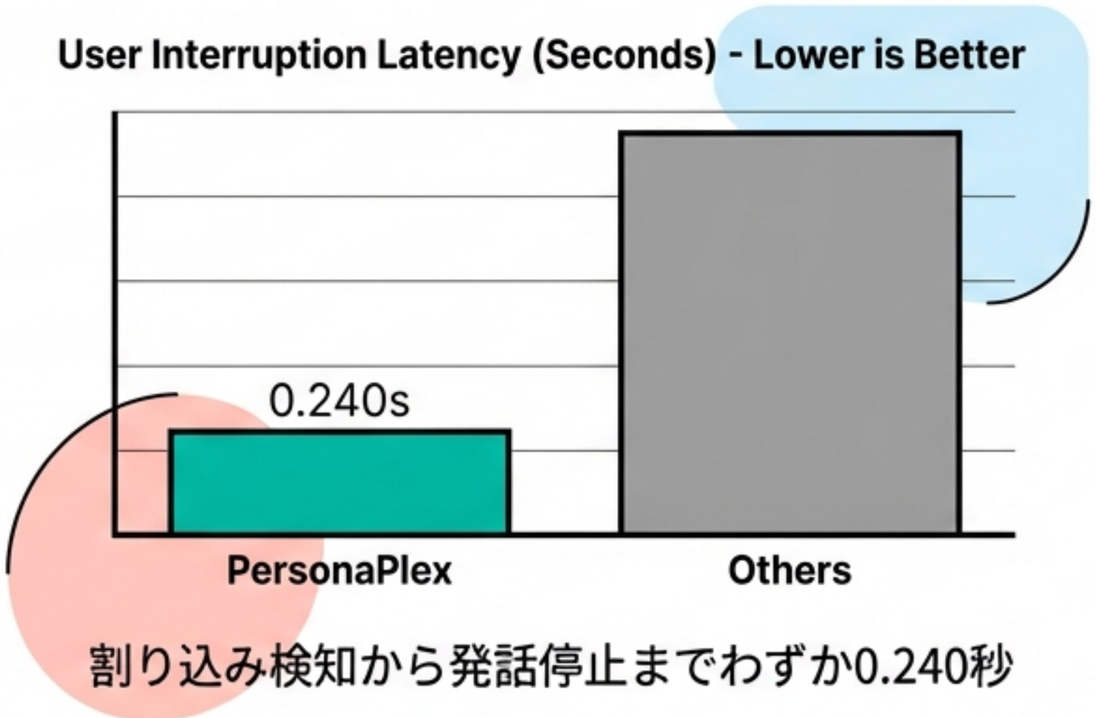


2. Voice Prompt（ボイスクローニング）

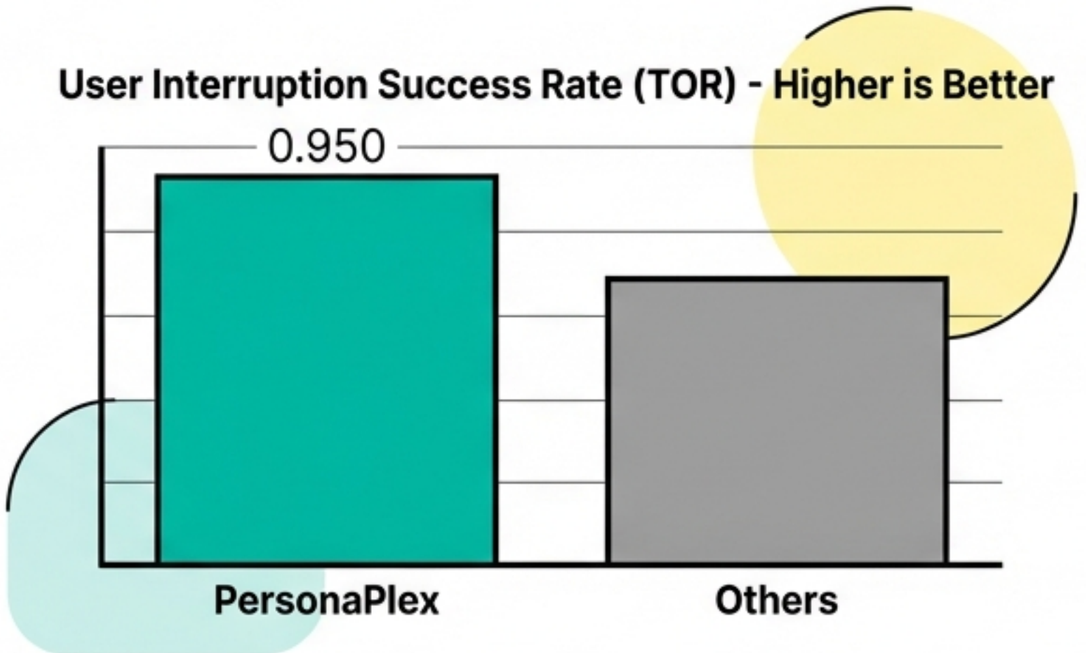


数秒の参照音声を与えるだけで、その声質や話し方を模倣（Zero-shot Voice Cloning）。追加の学習は不要です。

ベンチマークが証明する圧倒的な性能



割り込み検知から発話停止までわずか0.240秒



会話の主導権交代成功率 95%

公開ベンチマーク「FullDuplexBench」における評価結果。GPT-4oと比較しても、割り込み時の反応速度と自然さにおいて優れた結果を示しています。

ビジネスとエンターテインメントへの応用

次世代カスタマーサポート

顧客の言い直しや割り込みに即座に対応し、イライラ（呼損率）を低減。

ゲーム内NPC

プレイヤーの発言に食い気味に反応する、没入感の高いキャラクター対話。

音声アシスタント

「コマンド」ではなく「会話」でタスクをこなす真のパートナー。

トレーニング・面接練習

予測不能なタイミングで話しかけるリアルな対話シミュレーション。

導入による戦略的インパクト

顧客体験（CX）の向上

不自然な「間」を排除することで、CSAT（顧客満足度）とNPS（ネット・プロモーター・スコア）を改善。

業務効率化

言い直しや修正がスムーズなため、一件あたりの解決時間（AHT）を短縮し、初回解決率（FCR）を向上。

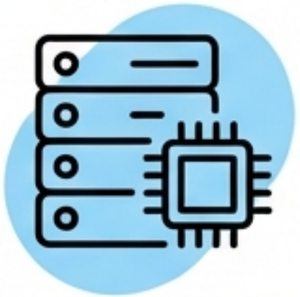
業務効率化

言い直しや修正がスムーズなため、一件あたりの解決時間（AHT）を短縮し、初回解決率（FCR）を向上。

ブランドの一貫性

独自のボイスプロンプトにより、企業の「ブランドボイス」を統一したキャラクターで展開可能。

実行環境とハードウェア要件

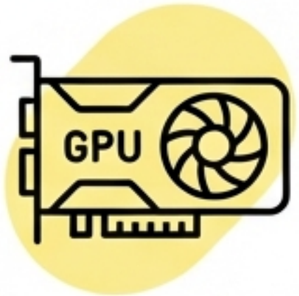


Ideal (Datacenter)

NVIDIA A100 / H100

VRAM 40GB+

推奨
(Recommended)



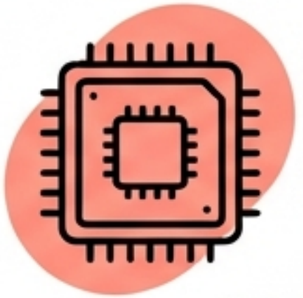
Consumer (High-End)

RTX 4090 / 3090

VRAM 24GB

可能
(Available)

Requires `--cpu-offload` flag



Low-End / Integrated

Ryzen 780M / RTX 4060

VRAM < 16GB

困難
(Not Viable)

注意: 統合GPU（Ryzen 780Mなど）ではリアルタイム動作は困難です。クラウドGPU（RunPod, Vast.ai, Colab Pro）の利用を推奨します。

エンジニア向け実装ガイド

```
# Install Opus codec first
sudo apt install libopus-dev  # or vcpkg install opus

# Clone and Install
git clone https://github.com/NVIDIA/personaplex
cd personaplex
pip install moshi/

# Run Server (with offload if needed)
python -m moshi.server --cpu-offload
```

Python 3.10+ と PyTorch 環境で動作します。

必須: Opusコーデックのインストールが必要です。

デプロイ: `moshi.server` モジュールを使えば、すぐにWebUIで対話デモを立ち上げられます。

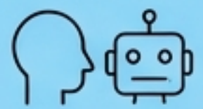
オープンソースという優位性

- **ライセンス: NVIDIA Open Model License Agreement (CC-BY-4.0)**
  商用利用: 条件を遵守すれば、商用プロダクトへの組み込みが可能です。
- **コスト効率**
  API従量課金ではなく、自社インフラ（または安価なクラウドGPU）で運用できるため、スケール時のコスト予測が容易です。
- **透明性とセキュリティ**
  ブラックボックスなAPIとは異なり、モデルの挙動やセキュリティを自社で管理できます。

コミュニティの声：「ロボット音声の死」

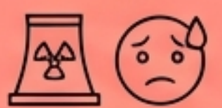
ロボットボイスエージェントは
永遠に死んだ。感情、専門用語、
緊急性すべてが完璧だ。

ボットと人間の区別がつかない
レベルに近づいている。

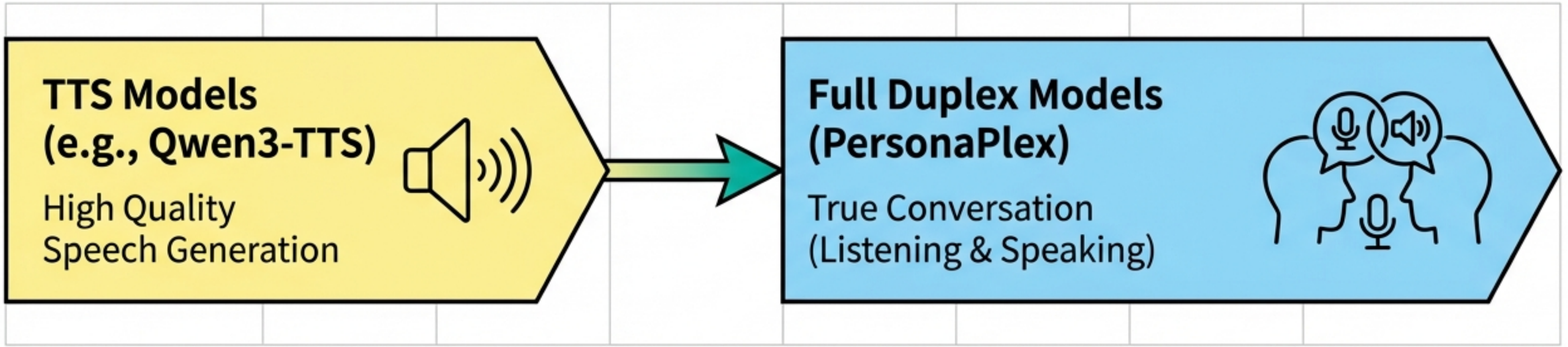


聞いて話す、ただそれだけだ
が、トランシーバー方式とは
次元が違う。

ストレスを感じるほどのリアリティ
（宇宙炉のデモを見て）。



音声AIの現在地と未来

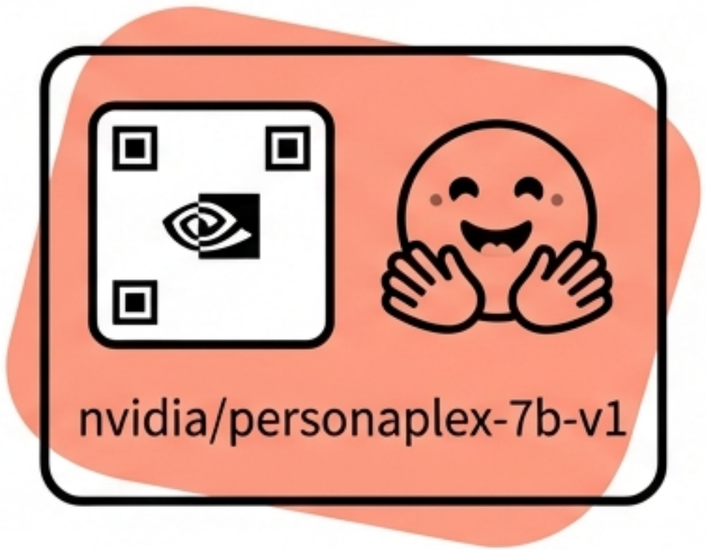


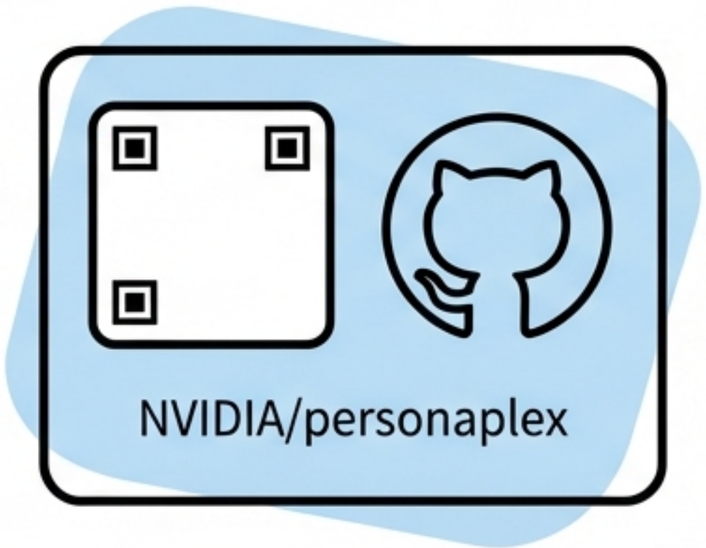
Qwen3-TTSのような高品質なTTSモデルも登場していますが、PersonaPlexの真価は「対話（Conversation)」にあります。

「相手の意図を汲み取り、適切なタイミングで言葉を発する」能力において、PersonaPlexは独自のポジションを築いています。

待つ時間は終わりました。
今すぐ会話を始めましょう。

PersonaPlex-7Bは、AIとのインタラクションにおける「摩擦」を取り除きました。





あなたのプロジェクトで、真のリアルタイム対話を実装してください。