FT-LLM 2026 / Team031

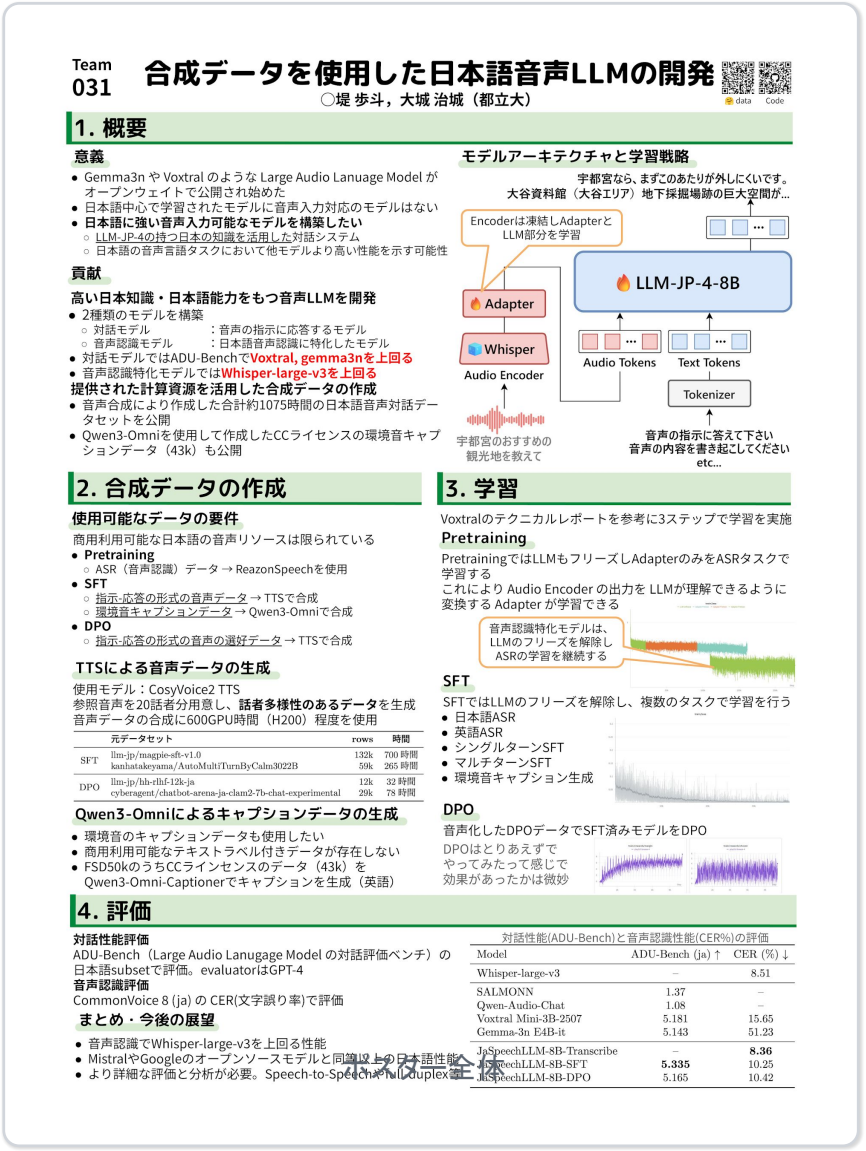
合成データを使用した日本語音声LLMの開発

堤 歩斗，大城 治城（都立大）

日本語合成音声データを作成し、音声理解可能なマルチモーダルLLMを構築

- 音声対話・音声認識・音声キャプション生成などの複数の音声タスクを学習
- 日本語の対話性能で既存オープンモデルを上回る
- 音声認識ではWhisper-large-v3を上回る

ポスターではデモも準備しています！



アプローチ

利用可能な日本語音声資源が少ないため、提供頂いた計算資源を活用し合成データを作成し使用

ねらい

- 日本語に強い音声入力対応モデルは少ない
- LLM-JP-4の日本知識を活かした
  音声対話を実現したい
- 対話モデルとASR特化モデルを構築

1075h
合成した日本語
音声対話データ

43k
音声
キャプション

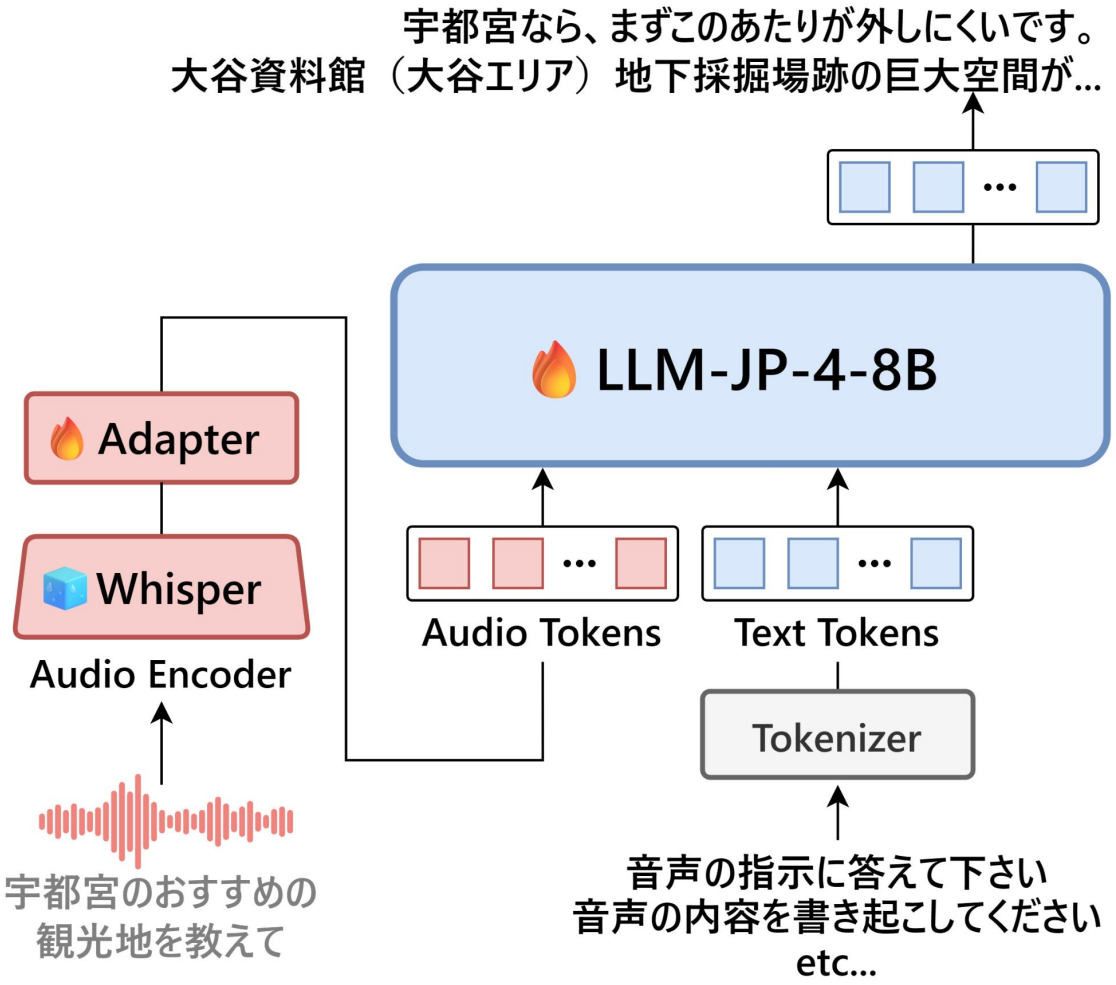
多話者
話者多様性のある合成音声

学習の流れ

Pretraining
Adapterを
音声認識で学習

SFT
対話・音声キャプション等を学習

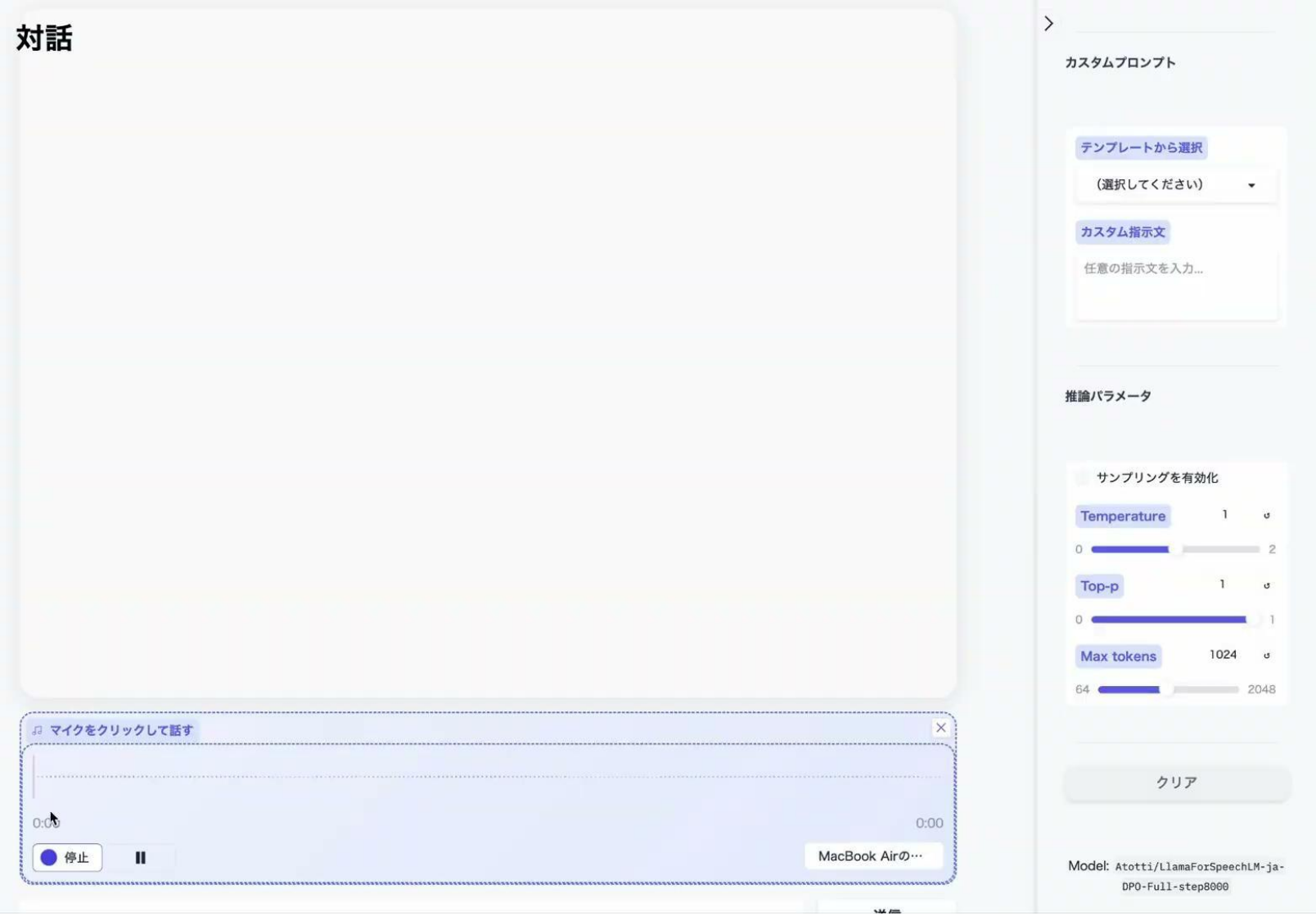
DPO
音声化した選好データで調整



AdapterとLLM-JP-4-8Bをフルで学習

デモ

A5000で推論しているデモ



見どころ

- 日本語の音声指示に有益な応答ができる
- LLM-JP-4由来の日本知識を活用した応答
- 1つのモデルでASRと対話のタスクの双方が可能

データセット公開済み
モデルはApache2.0で公開予定