DEEP DREAM

FALCON BARISTA

FALCON LLM HACKATHON

The Problem

- Major challenge for restaurant industry

High Labor Costs

- Solution

Let AI solve it

The Solution

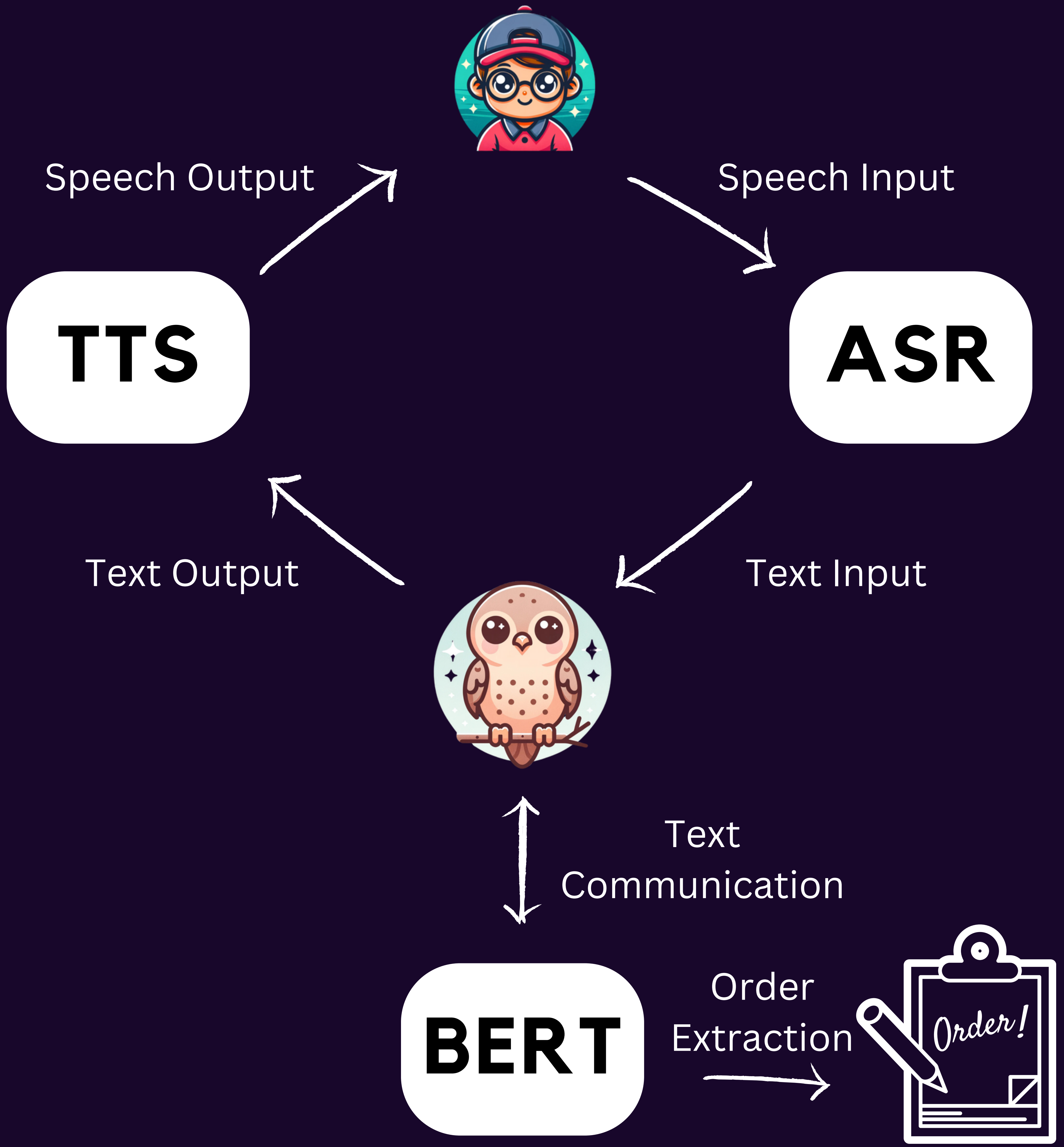
- An order taking bot with speech capabilities

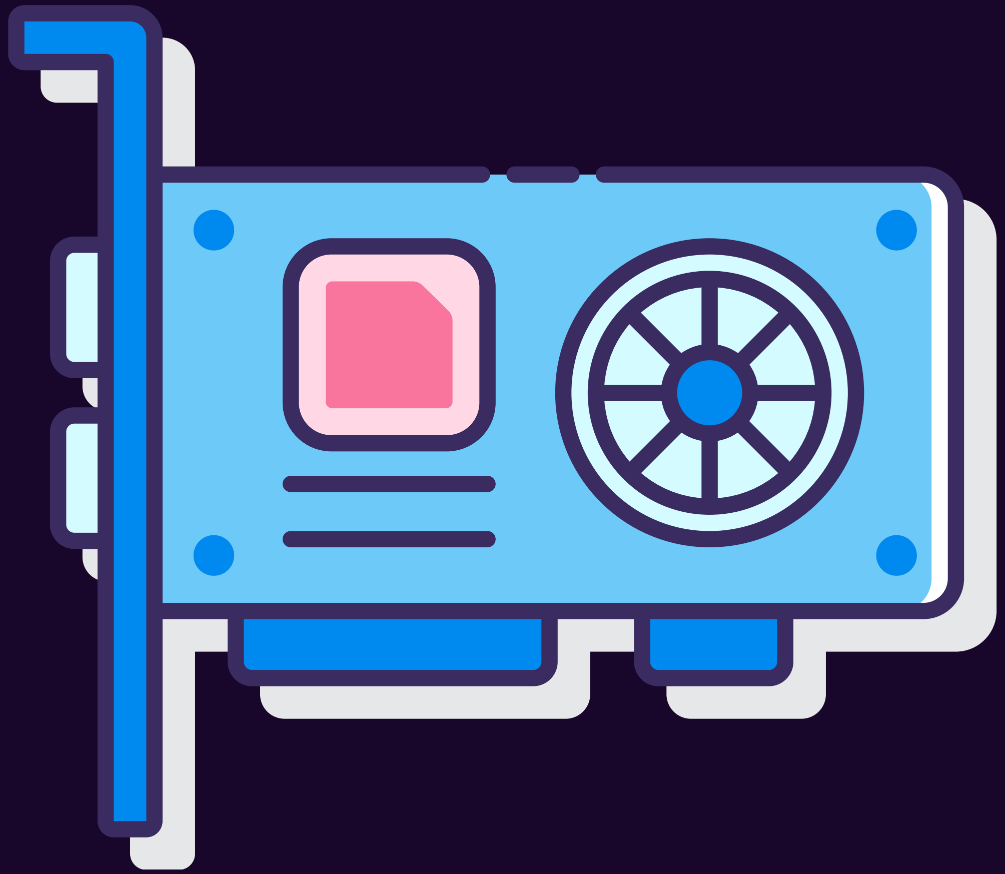
FALCON BARISTA

- No manpower required for order taking
- For order taking on counters, drive-throughs, and over the phone

Salient Features

- Has ability to hear and speak
- Falcon-7B LLM for conversation
- Fine-tuned BERT for information extraction





Works on single
16 GB RAM GPU

Our Innovation

- Falcon Barista has minimal compute needs.
- Most chatbots use LLMs with over 100 billion parameters.
- Falcon Barista uses the fine-tuned Falcon-7B LLM with just 7 billion parameters along with smaller fine-tuned BERT models to give best performance / compute ratio

Challenges

Reduce Latency

Currently ~10s

BERT Gets Confused

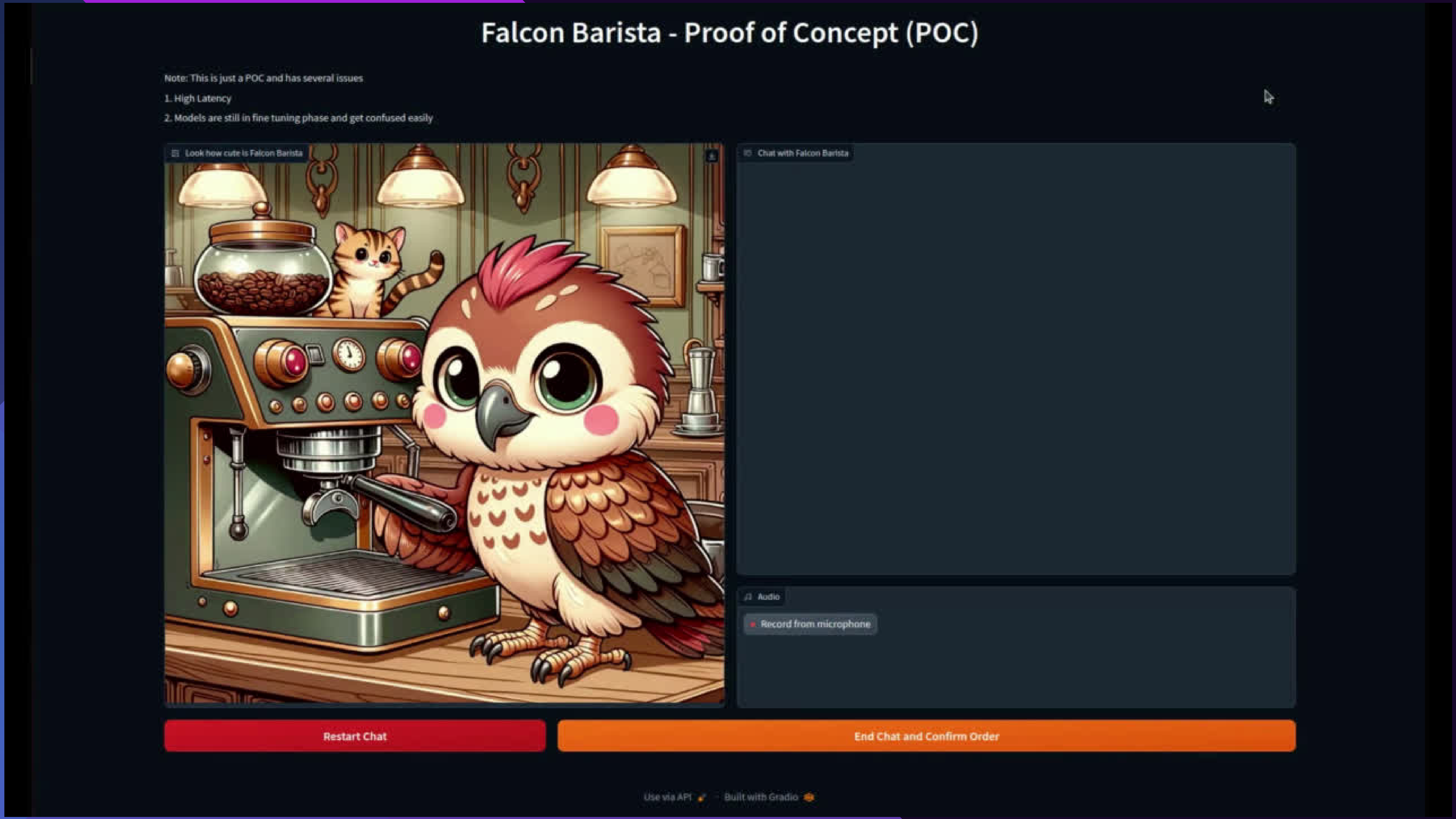
Needs further Finetuning

Falcon Hallucinates

Needs Finetuning

Way Ahead? Need more Compute!

DEMO VIDEO



Falcon Barista - Proof of Concept (POC)

Note: This is just a POC and has several issues

1. High Latency
2. Models are still in fine tuning phase and get confused easily

Video link

FALCON BARISTA

THANK YOU