OpenOmni

A Collaborative Open Source Tool for Building Future-Ready Multimodal Conversational Agents

Qiang Sun[1], Yuanyi Luo[2], Sirui Li[3], Wenxiao Zhang[1], Wei Liu[1]

[1]The University of Western Australia, Perth, WA, Australia
[2]Harbin Institute of Technology, Harbin, China
[3]Murdoch University, Perth, WA, Australia

Correspondence: pascal.sun@research.uwa.edu.au

November 22, 2024

Table of Contents

1. Motivation
2. Related Work
3. System Design
4. Demo
5. Conclusion

NTLP

Motivation

OpenAI released GPT-4o on 13 May 2024

Figure: Real-time demonstration[1]

[1] https://www.youtube.com/watch?v=DQacCB9tDaw

Google Project Astra on 14 May 2024

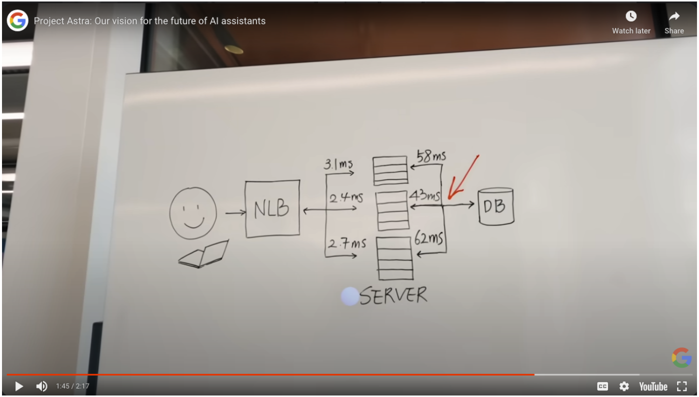


Figure: Google released demo video regarding Project Astra[2]

[2] https://deepmind.google/technologies/gemini/project-astra/

Telsa We Robot Launch on 10 Oct 2024

Figure: Telsa We Robot Launch[3]

[3] https://www.tesla.com/en_au/we-robot

Opportunities and Challenges

Opportunities

- Expanding research frontiers
- Diverse application areas

Challenges

- Performance optimization
- Domain-specific adaptation

Key Performance Metrics:

- **Latency:** Defining and achieving ‘real-time’ responses
- **Accuracy:** Ensuring domain-specific reliability and appropriateness

NLP

Related Work

Solution 1: Divide-and-Conquer

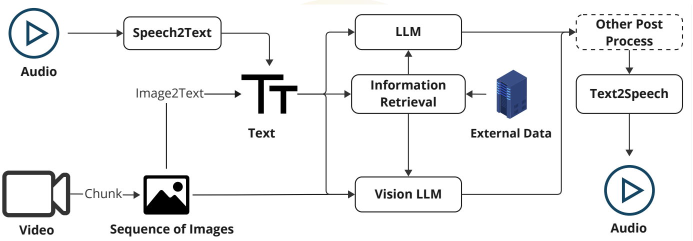


Figure: Traditional divide-and-conquer end-to-end multimodal conversation system

Solution 2: Fully End-to-End

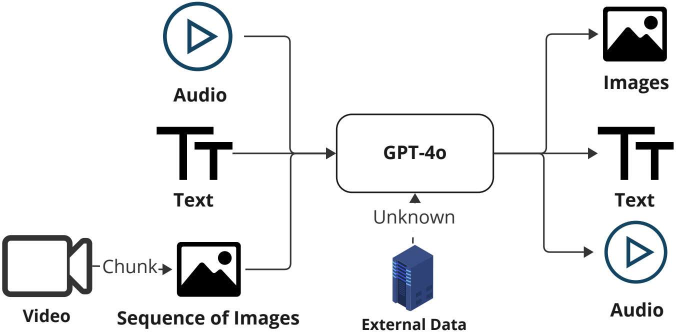


Figure: Our assumptions about how the fully end-toend model: GPT-4o works

Solution 3: Hybrid Solution

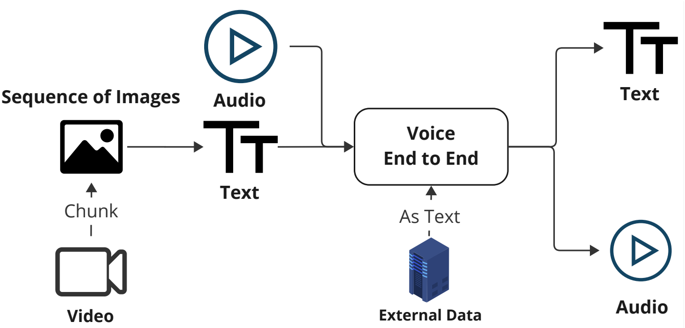


Figure: Hybrid solution via the combination of image2text and end-to-end voice model

Constraint Triangle

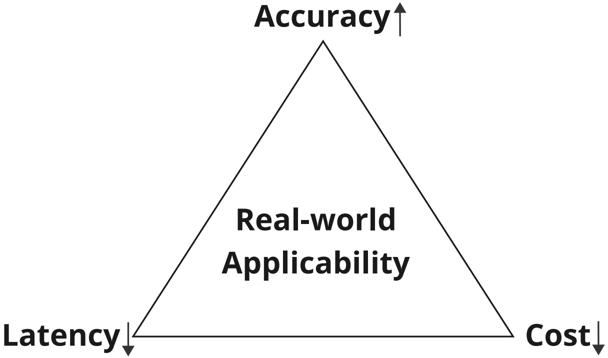


Figure: Constraint triangle for real-world applicability in multimodal conversational agent development

System Design

System Architecture

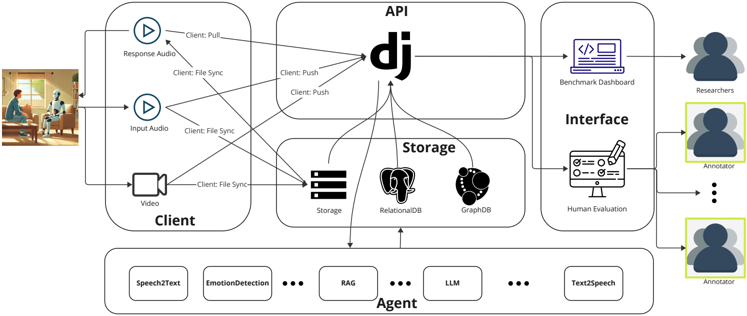


Figure: Architecture Design for OpenOmni Framework

Demo

Can “AI” be your president?

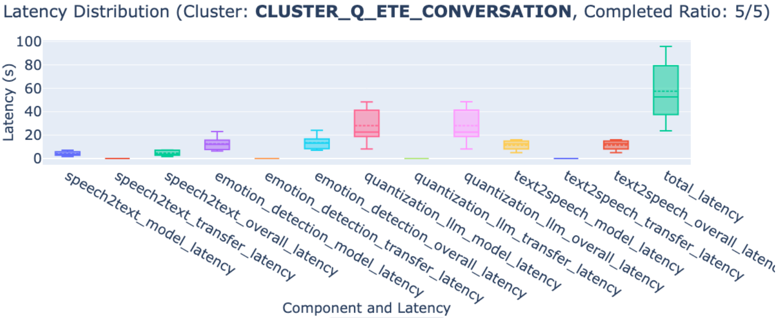
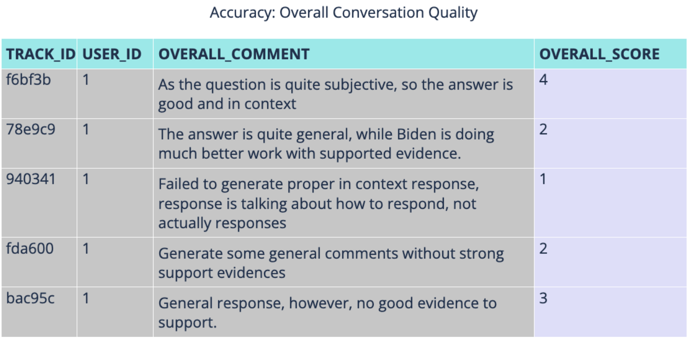


Figure: Screenshot of the end-to-end latency benchmark statistics for the setup: Local Whisper, Emotion Detection, Quantization LLM, and OpenAI Textto-Speech. This visualization is one example of the generated benchmark report; you can customize it or explore more details within our demo.

Can “AI” be your president?



Accuracy: Overall Conversation Quality

TRACK_ID	USER_ID	OVERALL_COMMENT	OVERALL_SCORE
f6bf3b	1	As the question is quite subjective, so the answer is good and in context	4
78e9c9	1	The answer is quite general, while Biden is doing much better work with supported evidence.	2
940341	1	Failed to generate proper in context response, response is talking about how to respond, not actually responses	1
fda600	1	Generate some general comments without strong support evidences	2
bac95c	1	General response, however, no good evidence to support.	3

Figure: Screenshot of annotated overall conversation accuracy statistics and comments for each conversation within GPT4O_ETE. Scores range from 0 to 5

Conclusion

Start of the Journey

- This is one of the early attempts toward developing multimodal conversational agents to benefit humans.
- Key challenges:
    - Latency remains an issue under current hardware constraints.
    - Accuracy shows promise but needs more work to be production ready.
- Our codebase aims to ease the burden for researchers and innovators, helping to bring advanced technology closer to practical applications.

Demo Links

Github Codebase

Demo Website

Documentation

The urls are: [4] [5] [6]

Demonstration Video is available: [7]

Presentation Video is available: [8]

[4] https://github.com/AI4WA/OpenOmniFramework

[5] https://openomni.ai4wa.com/

[6] https://openomnidocs.ai4wa.com/

[7] https://www.youtube.com/watch?v=zaSiT3clWqY

[8] https://youtu.be/R3IX24dKjw4?si=hnyNWOIxSOrfv2rV