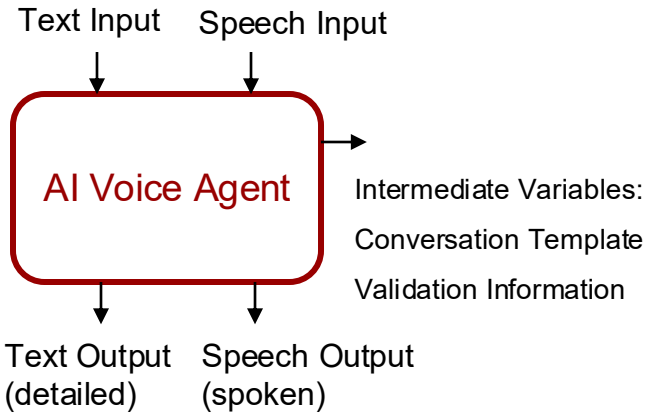
AI Voice Agent:
Modular Voice-Controlled Browser Automation

Presenter: Ruizhen Shen Jiazhuang Chen
Supervisor: Sonny Vu Johanna Björklund

Introduction

Goal

An interactive assistant designed to automate browser-based tasks through a graphical user interface.



Used Tools

- Speech Recognition

(Transfer speech to text)

- LLM-powered reasoning

(Via api)

- BCM

(Browser Control Module)

- TTS

(Transfer text to speech)

Completed Work

- UI

(To interact with user)

- Task Classification

(Using conversation template)

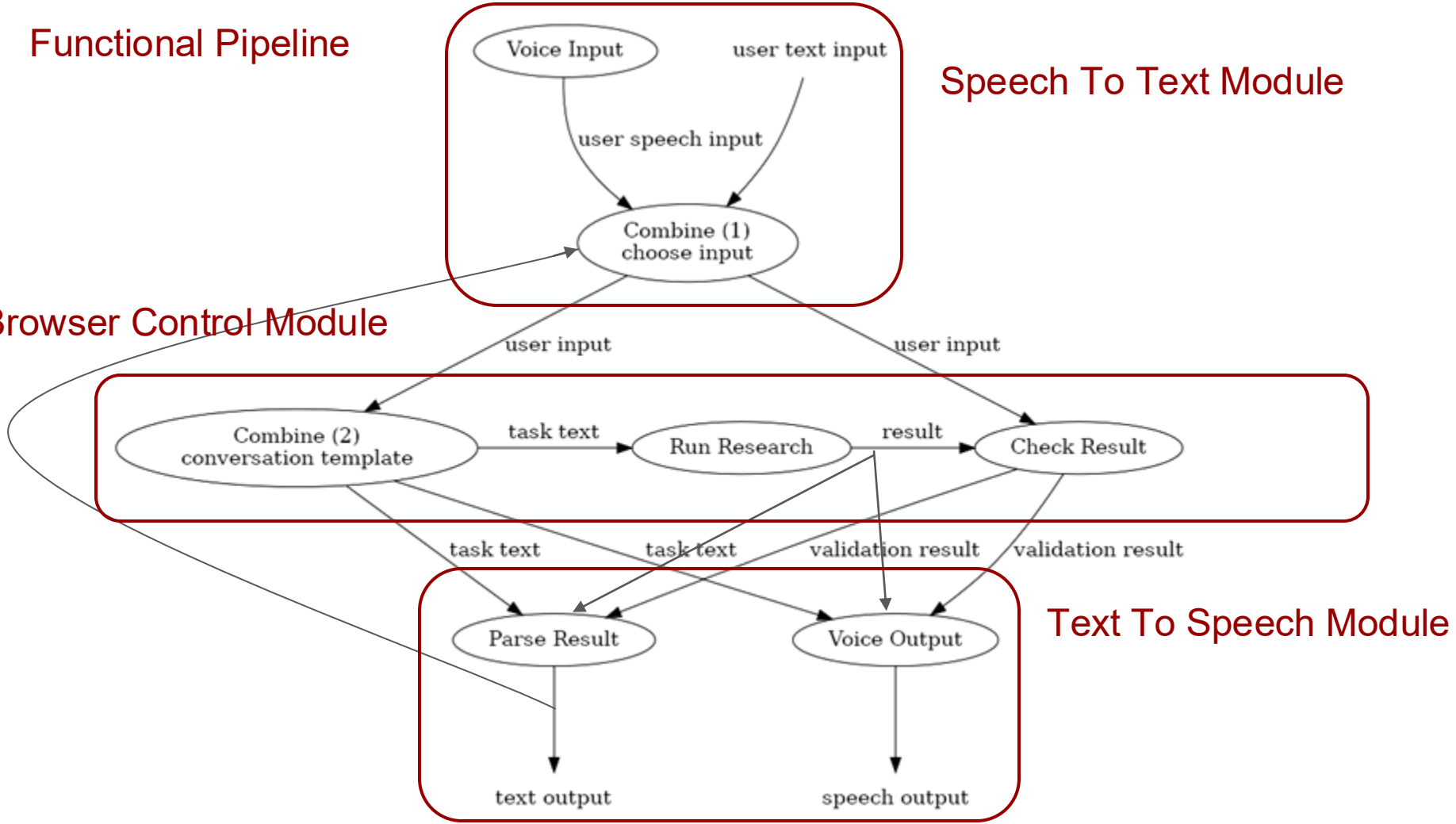
- BCM and TTS Pipeline

(All functionalities are encapsulated in a modular interface that supports fluid, looped interaction.)

- Result Validation Check

(Check the LLM output)

Functional Pipeline



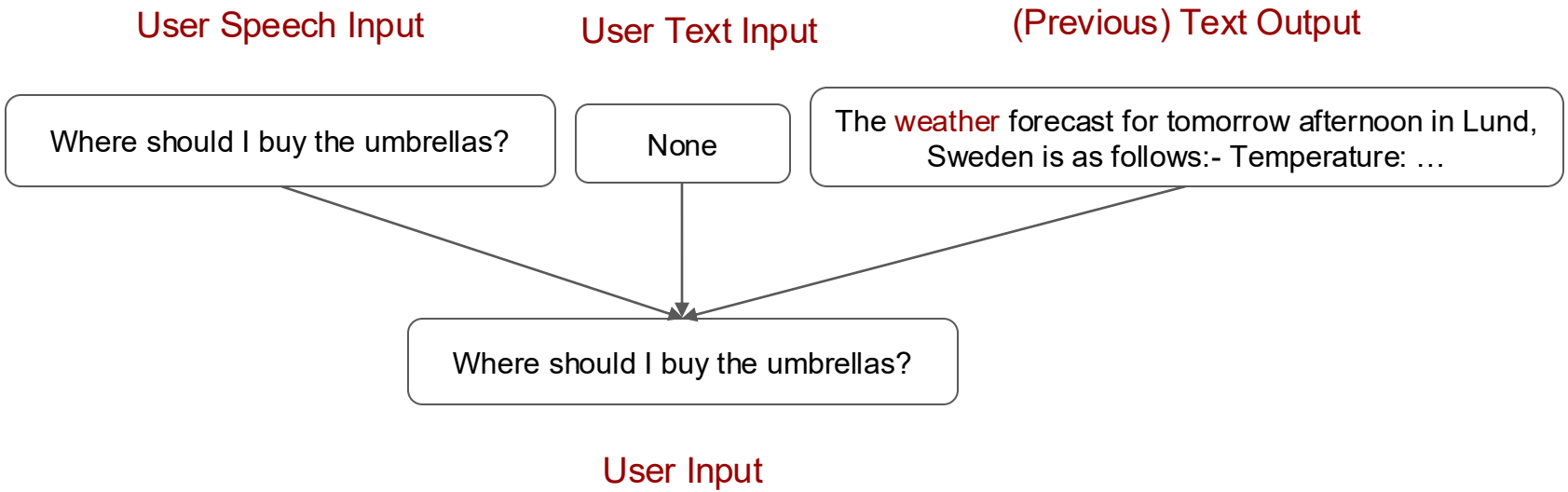
Voice Input

(Input Audio → User Speech Input)

User Speech Input → Where should I buy the umbrellas?

Combine(1) Choose Input

(User Speech Input or User Text Input + (Previous) Text Output→ User Input)

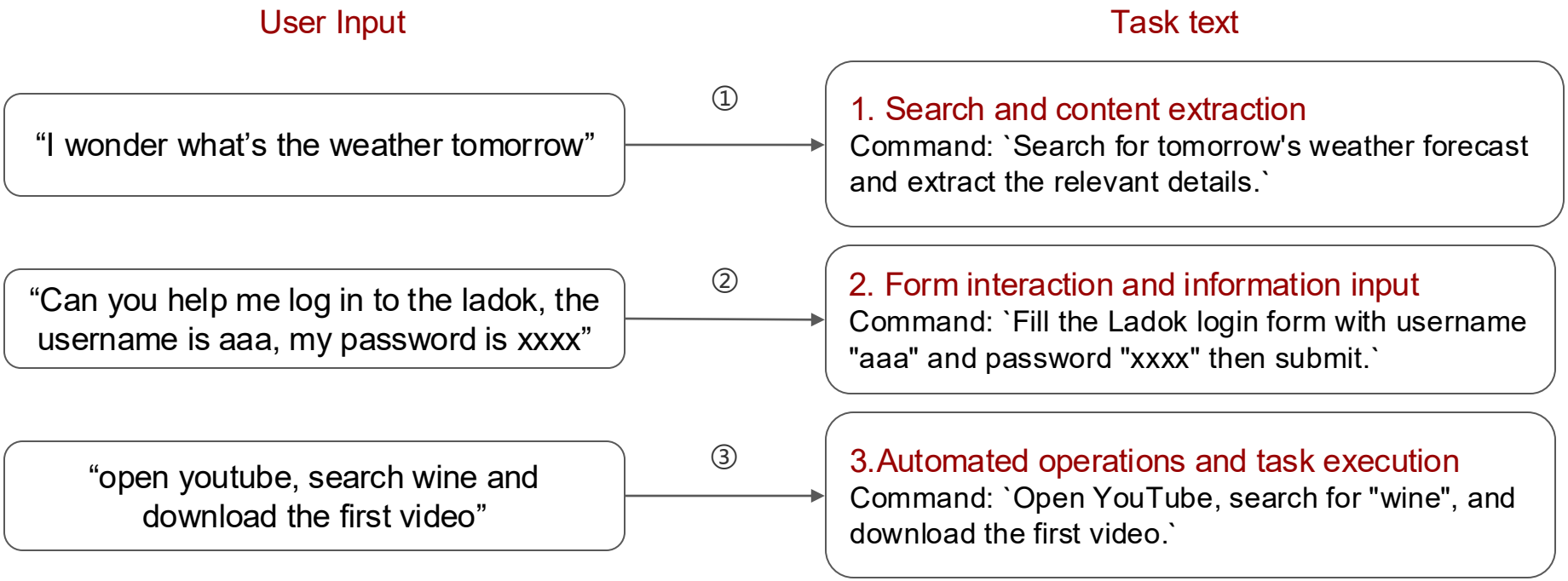


Combine(2) Conversation Template

(User Input → Task Text)

3 different conversation templates:

1. Search and content extraction
2. Form interaction and information input
3. Automated operations and task execution



Run Research

(Task Text → BCM Result)

Task Text

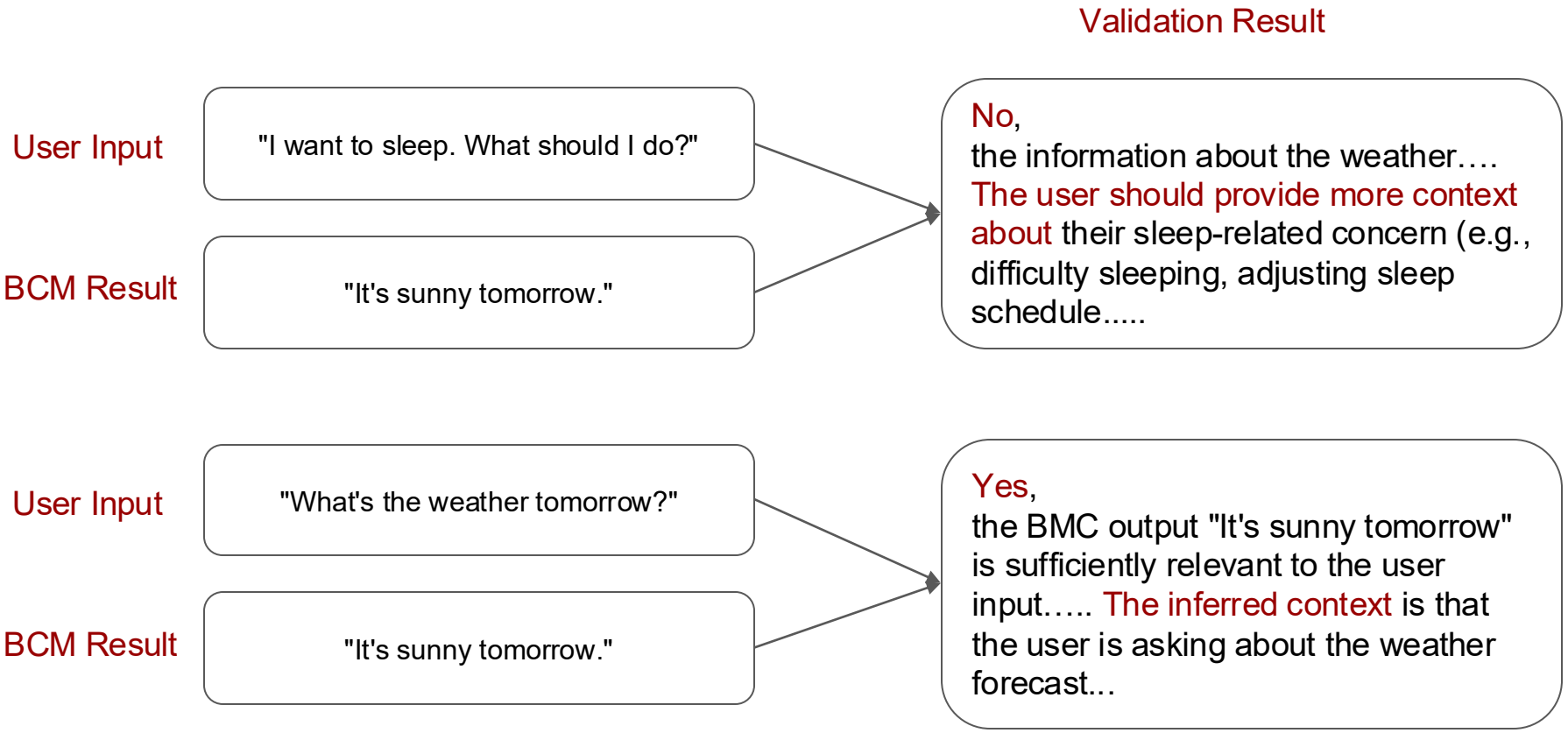
Category: Search and content extraction
Search for umbrella stores near Lund, Sweden.

BCM Result

Found several umbrella stores near Lund, Sweden:
1. Lund & Lund - https://lundochlund.com/collections/umbrellas
2. Lund University Shop - https://extern.shop.lu.se/en/group-4122447/collapsible-umbrella.html
3. Lund University Shop (Umbrella with the canopy in the auditorium of the university building) - https://extern.shop.lu.se/en/group-4122447/umbrella-with-the-canopy-in-the-auditorium-of-the-university-bui.html
4. Lund University Shop.…

Check Result

(User Input + BCM Result→ Validation Result)



Parse Result

(BCM Result + Validation Result + Task Text → Text Output)

BCM Result

Found several umbrella stores near Lund,Sweden:
1. Lund & Lund -
https://lundochlund.com/collections/umbrellas 2…

Validation Result

"Yes, the BCM output is relevant as it provides …

Task Text

Category: Search and content extraction;
Search for umbrella stores near Lund, Sweden.

Text Output

Found several umbrella stores near Lund, Sweden:
1. Lund & Lund -
https://lundochlund.com/collections/umbrellas
2. Lund University Shop.....

Speech Output

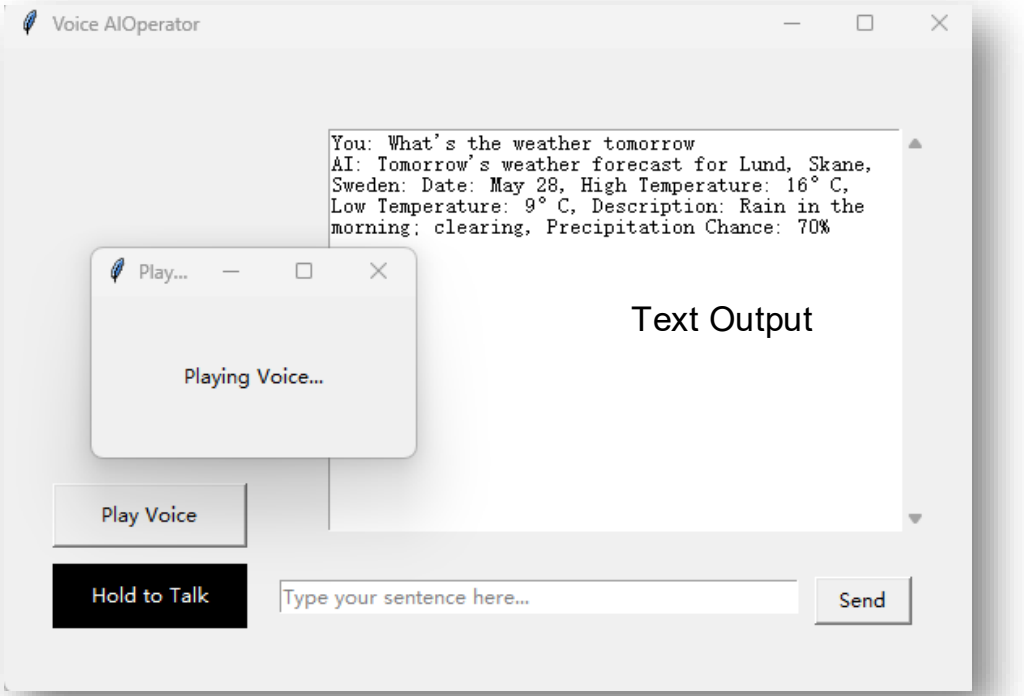
I found several places to buy umbrellas in Lund, Sweden...

Voice Output

(BCM Result +Validation Result + Task Text → Speech Output)

UI

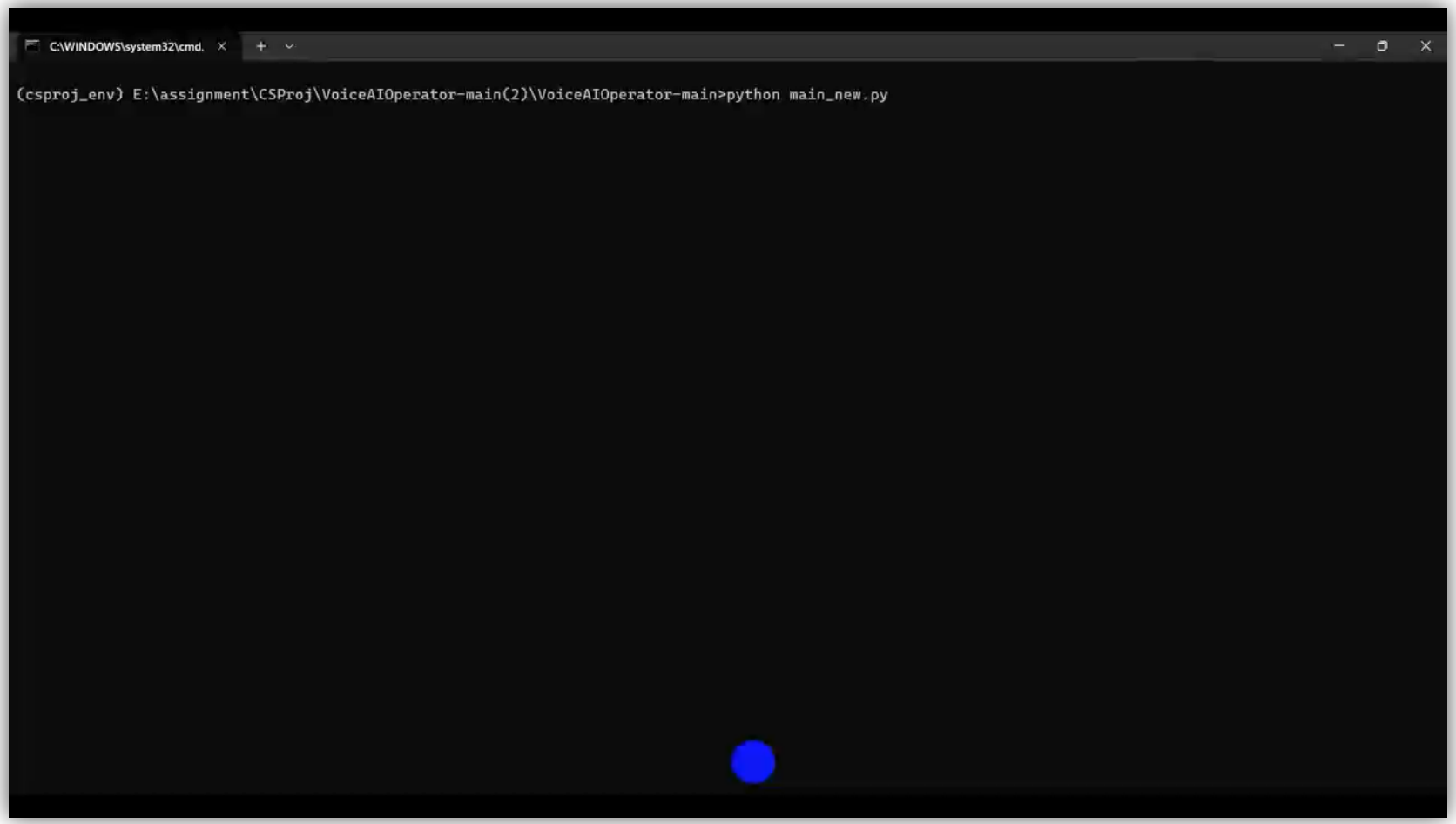
Speech Output



Hold the button to talk Type in sentence and click send

Instead of Avatar we now temporarily use this UI to demonstrate our functions.

Final Demonstration



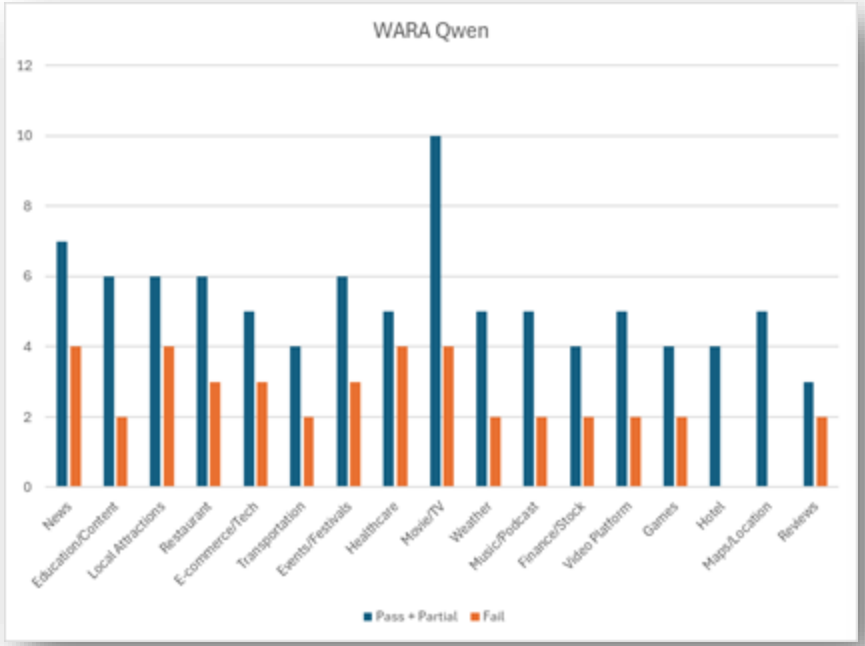
Overview of Key Improvements

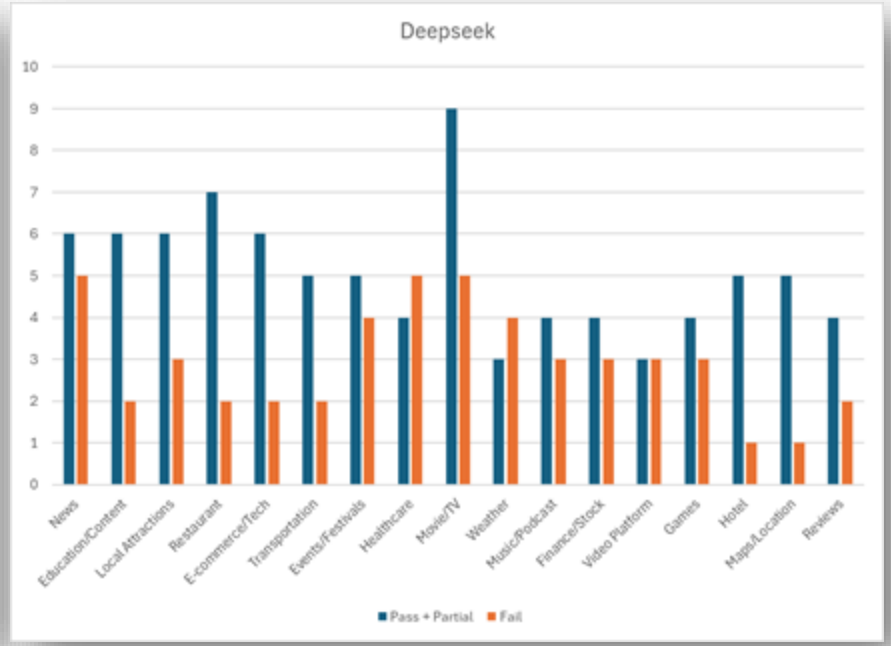
Achievements	Technical Key Point
Improved Execution Efficiency	Introduced a step limit mechanism to control the Agent's execution flow.
Enhanced Speech Interaction: TTS and LLM Output Optimization	Added a processing step to convert LLM output into concise, conversational responses.
Multi-Model Compatibility	Modularized support for custom LLMs

Technical Fixes and Enhancements:

- ✓ Fixed misordered chat history by refactoring update logic.
- ✓ Improved validation results with additional constraints.
- ✓ Resolved .wav permission issue by switching to click-to-play mode.
- ✓ Updated dynamic URL support to ensure consistency across modules.

Test Result Graph





Enhancement

Improving Large Model Adaptability While Ensuring Privacy

- Constructed closed-form prompts to avoid uploading raw audio or sensitive data.
- Used local proxy to control data flow and prevent leaks.
- Post-processed model output (e.g., clean_output) to ensure compliance and clarity.

Speed Optimization

- Used asynchronous execution to reduce blocking.
- Applied audio locking to avoid conflicts and playback delays.

Avatar Adaptation

- Structured output for compatibility with avatar narration or facial expression rendering.

Acknowledgments

We thank **WARA Media & Language** for providing API support and infrastructure.

We gratefully acknowledge the use of the **browseruse** tool from GitHub.

Special thanks to **Assist. Prof. Sonny Vu, Assoc. Prof. Johanna Björklund** for their invaluable guidance and support throughout the project.

We also thank **Quan** for providing key API-related materials that greatly facilitated our development.

Thank you for listening!