Al 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)



(Input Audio \rightarrow User Speech Input)

Voice Input

User Speech Input

Combine(1) Choose Input

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



User Input

Combine(2) Conversation Template

(User Input \rightarrow Task Text)

User Input

3 different conversation templates:

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

Task text



Run Research

(Task Text \rightarrow BCM Result)

Task Text

Category: Search and content extraction Search for umbrella stores near Lund, Sweden. ThES [agent] & Starting task: 1. Search and content extraction Search for a restaurant in New York City on Yelp. INFO [browser] . Launching local browser driver*playwright channel*chrc INFO [agent] * Step 1: Evaluating page with 0 interactive elements on: INFO [agent] 🧠 LLM call => ChatOpenAI [🖂 7 msg, ~2363 tk, 5777 char] + [agent] ? Eval: Unknown - The current page is blank, and no intera INFO T\$F0 [agent] @ Memory: Starting the task: Search for a restaurant in Ne [agent] @ Next goal: Navigate to Yelp's website to begin the searc INFO [controller] @ Navigated to https://www.yelp.com INFO INFO [agent] S Executed action 1/1: go_to_url INFO [agent] * Step 2: Ran 1 actions in 15.99s: 5 1 INFO [agent] . Step 2: Evaluating page with 123 interactive elements on TNED [agent] @ LLM call => ChatOpenAI [||| 10 msg. -4121 tk. 10548 char. INFO [agent] . Eval: Success - Successfully navigated to Yelp's mebsite INFO [agent] @ Memory: Currently on Yelp's nomepage. The next step is t INFO [agent] @ Next goal: Input 'New York City' in the location search INFO [controller] = Input New York City into index 4 INF-0 [agent] Z Executed action 1/2: input_text INFO [agent] Something new appeared after action 1 / 2 INFO [agent] * Step 3: Ran 2 actions in 18.75s: 5 2 **IAFO** [agent] * Step 3: Evaluating page with 125 interactive elements or INFO [agent] @ LLM call => ChatOpenA1 [14 msg, ~4308 tk, 10541 char **TNFO** [agent] . Eval: Failed - The input action redirected to a Yelp pag INFO [agent] @ Memory: Attempted to search for 'New York City' Google Graph Rad Income State State

BCM Result

Found several umbrella stores near Lund. Sweden: 1. Lund & Lund https://lundochlund.com/collections/umb rellas 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-inthe-auditorium-of-the-university-bui.html 4. Lund University Shop....

Check Result

(User Input + BCM Result→ Validation Result)

Validation Result



User Input "What's the weather tomorrow?"
BCM Result "It's sunny tomorrow."

Yes,

the BMC output "It's sunny tomorrow" is sufficiently relevant to the user input..... The inferred context is that the user is asking about the weather forecast...

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.

Parse Result

(BCM Result + Validation Result + Task Text \rightarrow Text Output)

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)

	Voice AlOperator –	×
Speech Output		
	You: What's the weather tomorrow AI: Tomorrow's weather forecast for Lund, Skane, Sweden: Date: May 28, High Temperature: 16°C, Low Temperature: 9°C, Description: Rain in the morning; clearing, Precipitation Chance: 70%	•
	🖉 Play — 🗆 🗙	
	Text Output	
	Playing Voice	
	Play Voice	-
	Hold to Talk Type your sentence here 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!