University of Nottingham
UK | CHINA | MALAYSIA

Bangbang 2.0

Human-in-the-Loop Edge-Cloud LLM Decision Support for Inclusive Service Robots

Younghoon Chang[1], **Yanjie Zhou[2]**, Changshuai Liu[2], Siheng Ye[2], Yuechen Cao[2], Hoe Yin Khong[3], Yuhong Dai[4] and David Chieng[2]

[1] Nottingham University Business School China, University of Nottingham Ningbo China
[2] Department of Electrical and Electronic Engineering, University of Nottingham Ningbo China
[3] Global College, Shanghai Jiao Tong University
[4] Department of Electrical Engineering and Computer Science, University of California, Irvine

Presenter: Yanjie Zhou

IHCI 2026, oral presentation, Paper 99



Figure 1. Bangbang robot

Inclusive service scenario



Figure 2. Bangbang robot and remote operator

A visitor asks for help in Chinese or English

- The operator sees the visitor through the robot’s camera.
- The visitor can ask for help in Chinese or English.
- The operator reviews the request and chooses a reply.

Design goal

Help the operator respond while keeping control of the final reply.

Human control

AI support

- Draft replies during the conversation.
- Choose a local model or a cloud service.
- Type a request or use Chinese or English speech.

Operator control

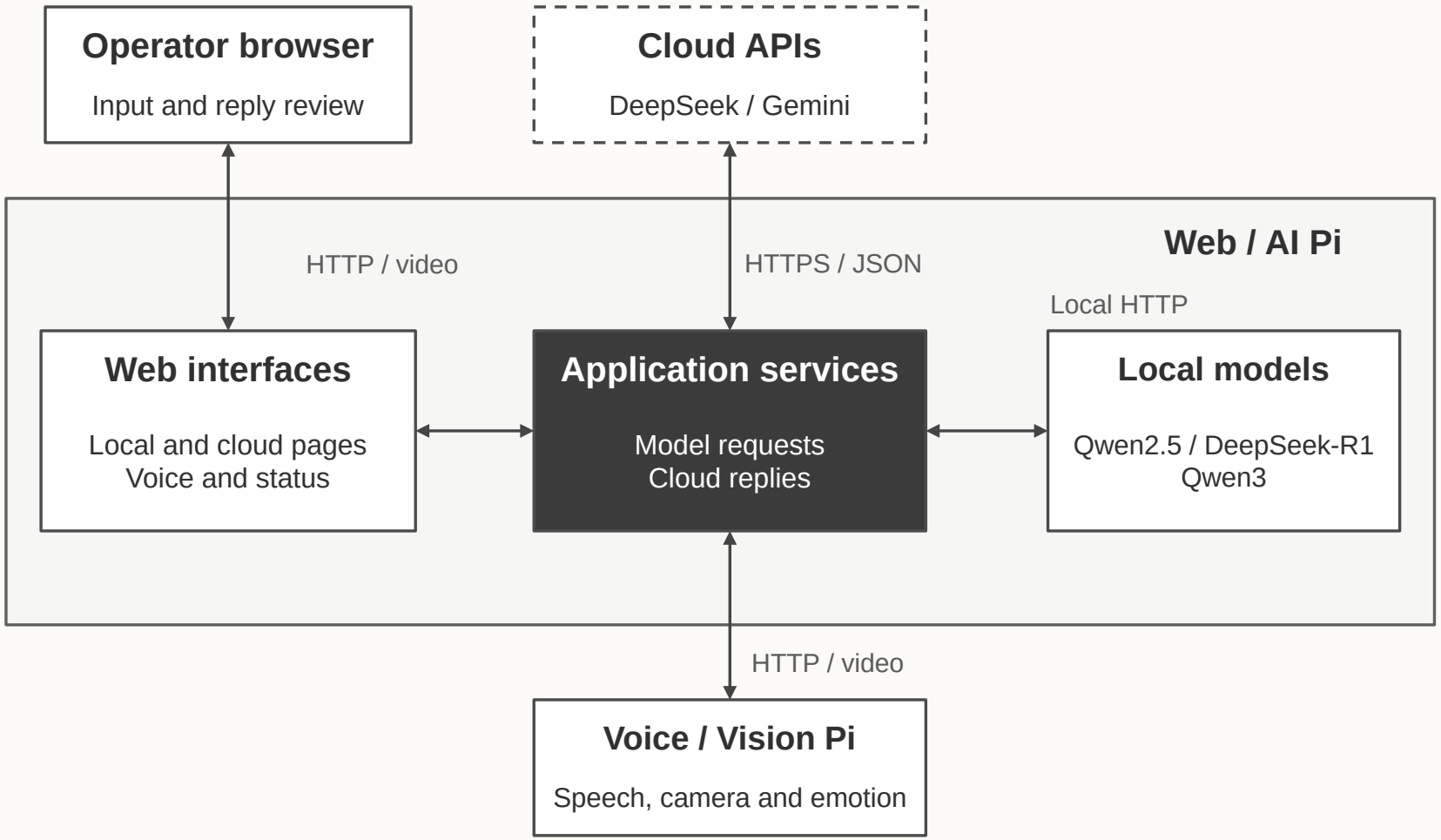
- Check the draft for errors or missing information.
- See which model produced the reply.
- Edit, accept or reject the draft before replying.

System requirements

How can local and cloud AI help the operator while the operator controls the final reply?

Requirement	System design
Recover from AI failure	Retry, choose another model, or answer directly.
Keep a local option	Run three language models on the Web/AI Pi.
Keep the operator in control	Show the selected model and any errors. Keep replies editable.
Make deployment manageable	Use browser access, separate services and stable network settings.

System architecture



System roles

- Web/AI Pi: hosts the web pages and connects to local and cloud models.
- Voice/Vision Pi: provides speech recognition, video and expression labels.
- The operator selects a model and reviews its reply.
- Chinese speech recognition uses Baidu Cloud.

Local and cloud models

The operator can choose from three local models and two cloud services.

Model / service	Runs on	Role	Main constraint
Qwen2.5-0.5B	Local Pi	Small language model	Limited capacity
DeepSeek-R1-1.5B	Local Pi	Reasoning model	May spend its output limit on reasoning
Qwen3-4B	Local Pi	Larger language model	Needs more processing resources
DeepSeek	Cloud	Cloud chat service	Internet connection and account quota
Gemini 2.5 Flash	Cloud	Alternative cloud service	Internet connection and service access

Operator interface

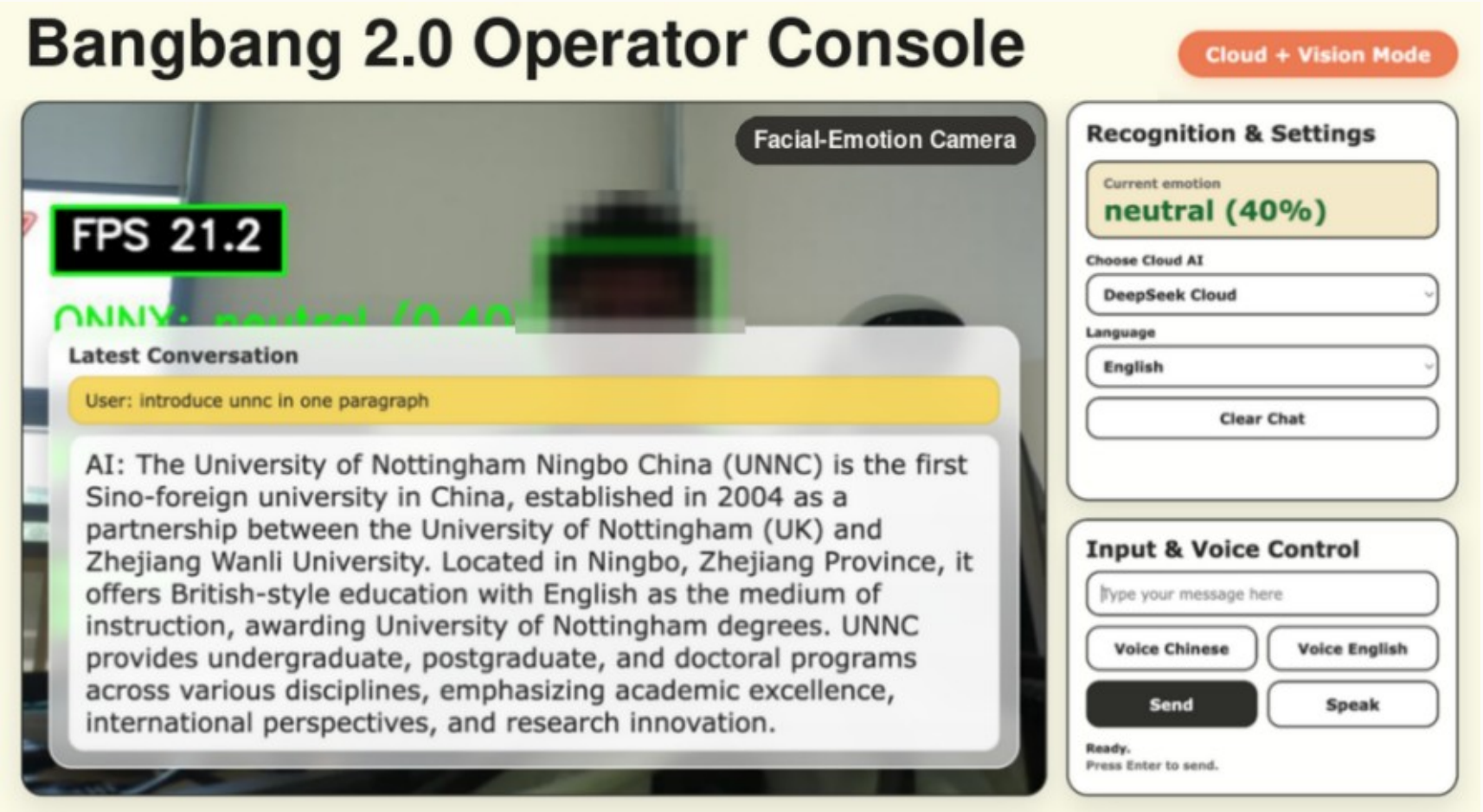
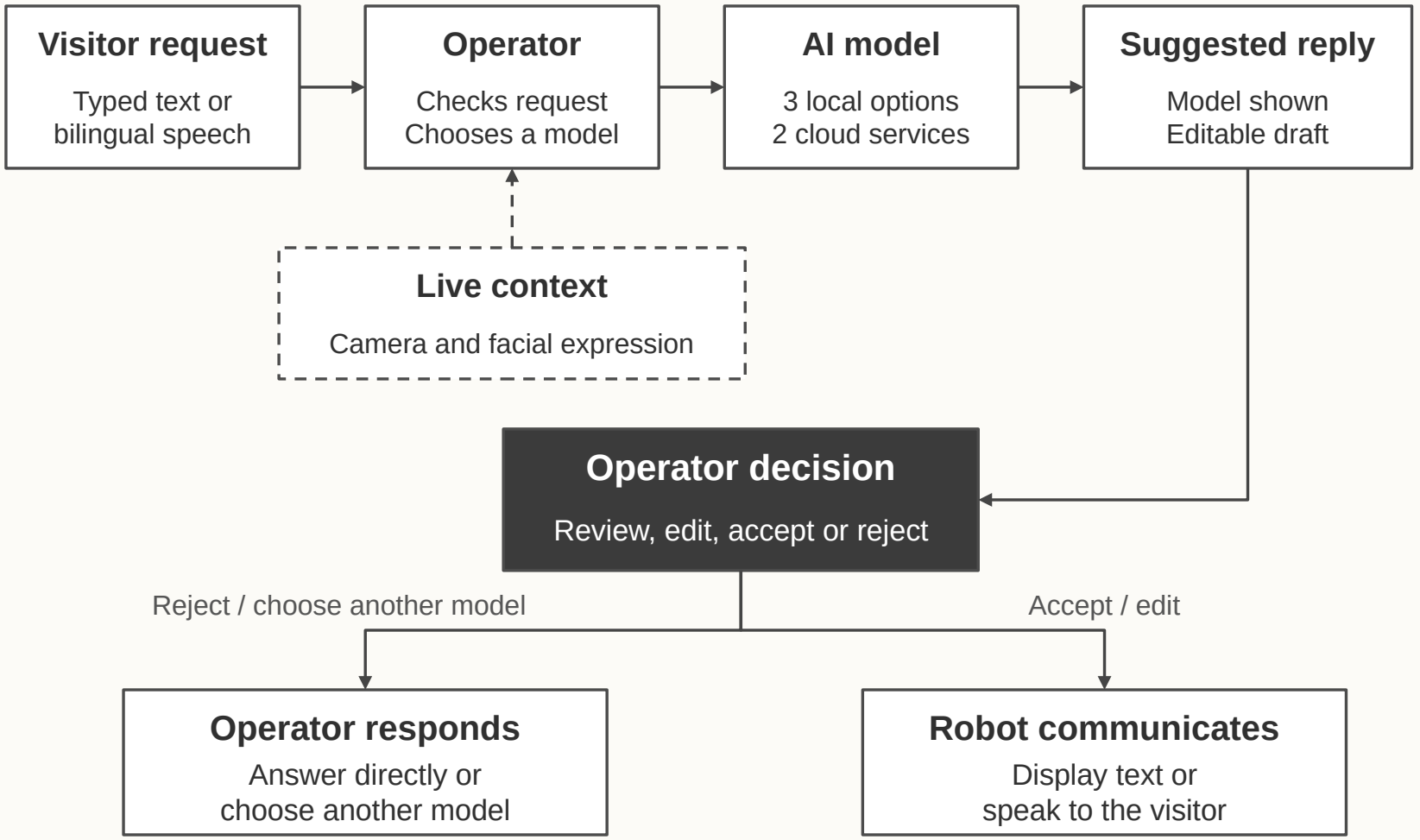


Figure 3. Operator interface with AI support

Main controls

- Choose a model and input language.
- Review or edit the recognised request and AI reply.
- Observe live video and facial expression labels.
- Press Speak to send the reply to the visitor.

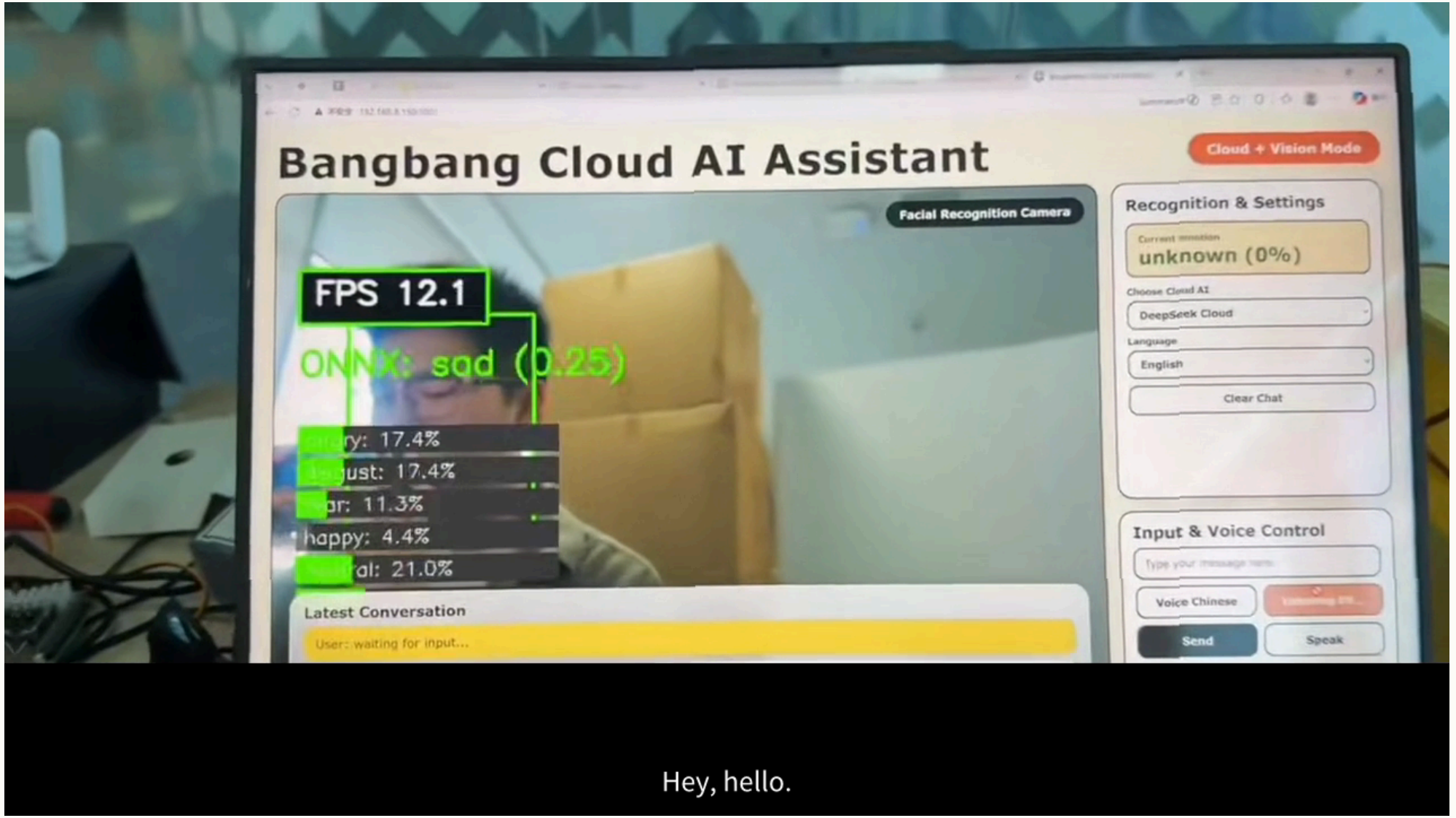
Interaction workflow



Operator actions

- Choose a local model or cloud service.
- Review the draft alongside the visitor's expression cues.
- Edit the reply, answer directly, or choose another model.
- Press Speak when ready to deliver the reply.

System demonstration



Demo: bilingual speech input, continuous facial expression cues and an editable AI reply.

System testing

Test setup: Raspberry Pi 5 (16 GB), 64-bit Raspberry Pi OS, shared local network

Function	How we checked it	Result
Web interfaces	Startup and browser records	Pages and services available
Local models	Saved model replies	All three local models returned replies
Cloud services	Provider and browser records	Both cloud services returned replies
Bilingual speech-to-text (STT)	Checks between the two Pis	Editable Chinese and English text
Facial expressions	Status and display records	Labels and confidence updated
Camera stream	Browser demonstration	Live video displayed
Speech output	Team confirmation	Audible output confirmed
Error handling	Code and error records	Visible errors and manual recovery

Response time observations

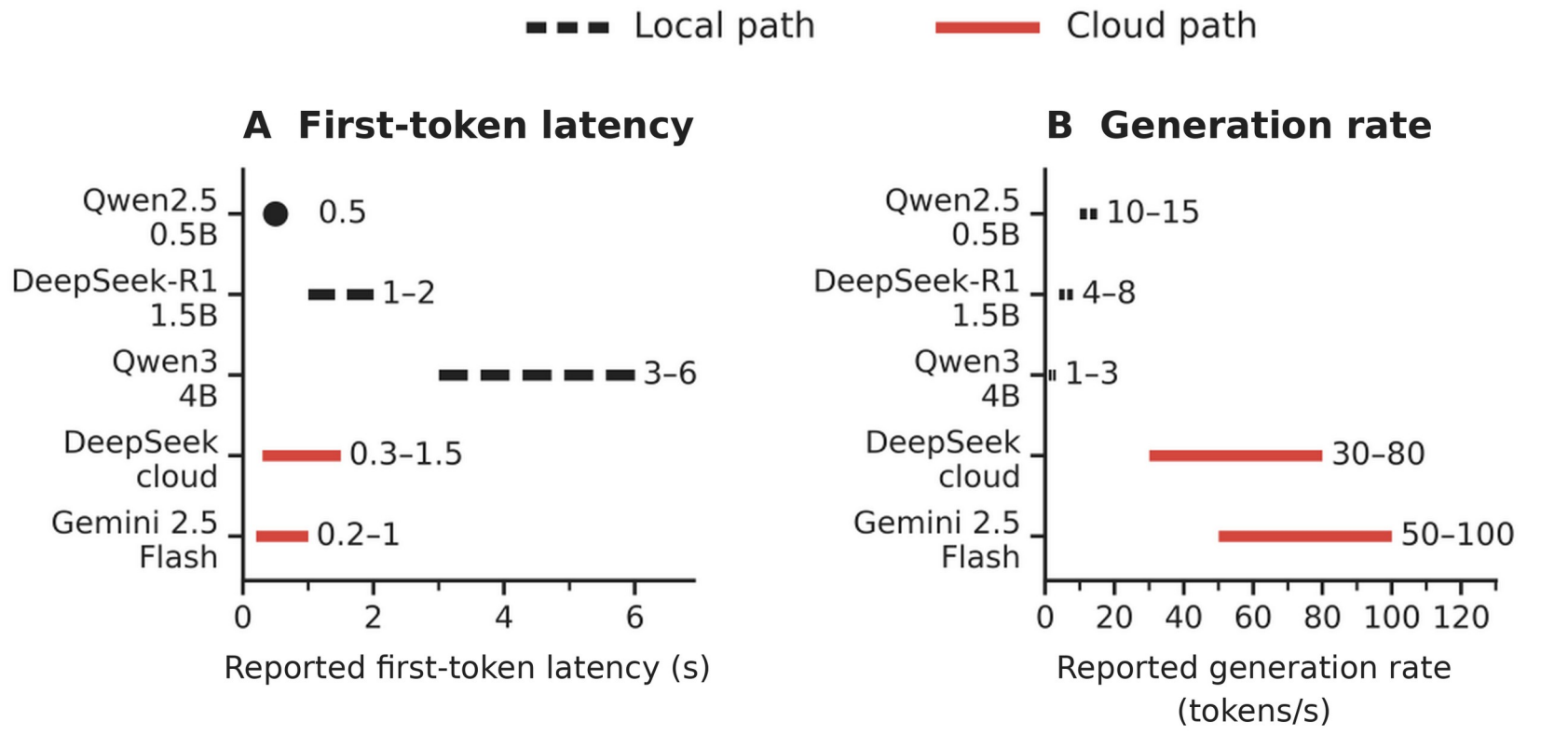


Figure 4. Response time observations [Human design, drawn by AI]

Development observations

- Recorded during laboratory development.
- Prompts and test settings varied between models.
- Timing and token counts were not measured consistently.
- Cloud values are not browser completion times.

Problems encountered and changes made

Issue	What we observed	What we changed
Chinese input and remote display	The desktop interface caused input and display problems.	Moved the operator interface to the browser.
Reasoning output	A model could reach its output limit before giving a usable answer.	Showed response status and whether an answer was available.
Incorrect replies	A local model gave an incorrect explanation of Bangbang.	Kept human review and the option to answer directly.

Future Work

Current limitations

- Testing so far covers system operation rather than user benefits.
- Reply accuracy and workload need further evaluation.
- Facial expression labels are visual cues, not a measure of internal emotion.
- Choosing a model manually may add work for the operator.

Planned work

- Test bilingual requests against verified answers.
- Add trusted information sources and show references.
- Compare direct answering, fixed models, manual selection and automatic selection.
- Measure accuracy, response time, workload and usability with users.

Conclusions

Local and cloud AI

Three local models and two cloud services connected through two Raspberry Pis.

Integrated interaction

Bilingual input, live video, expression labels and robot speech in one system.

Operator control

The selected model stays visible, and the operator controls the final reply.

Working prototype

The team demonstrated the workflow and documented deployment and recovery.

University of Nottingham
UK | CHINA | MALAYSIA

Thank you

Questions?

Yanjie Zhou, IHCI 2026, Paper 99

Bangbang 2.0