Week 4 Report: Ứng dụng LLM trong Voice Assistants(VA)

1. Vấn đề đặt ra

Các trợ lý giọng nói(Voice Assistants - VA) trước đây dựa trên:

- ASR(Automatic Speech Recognition) pipeline dựa trên các model truyền thống như model **Hidden Markov Models (HMMs)** kèm theo **Gaussian Mixture Models (GMMs)** cho acoustic modeling, sau đó dùng HMM để ghép thành các phoneme và từ (e.g., CMU Sphinx)
- Kết hợp với **Statistical Language Models**: Chủ yếu là **n-gram models** (bigram, trigram) để tính xác suất chuỗi từ, hỗ trợ giải mã ASR và gợi ý từ tiếp theo…
- Có thể tích hợp thêm các module NLU như **Rule-Based / Slot-Filling NLU** mà đánh dấu các keyword, regex để "map" keyword/chủ đề thành các intents và slots -> để định danh intent và thực thi lệnh

=> Chính vì dựa trên keyword để trả lời cho các câu truy vấn của người dùng khiến cho các câu trả lời bị thiếu ngữ cảnh, không đúng ý của người dùng

Để việc hiểu biết và nắm bắt ngữ cảnh câu truy vấn tốt hơn, bài báo đề xuất dùng các mô hình ngôn ngữ lớn (Large Language Model - LLM) - mô hình rất là vượt trội trong lĩnh vực nắm bắt ngữ cảnh trong NLP.
Việc mà tích hợp model LLM vốn mạnh bên việc xử lý ngôn ngữ, văn bản sang lĩnh vực giọng nói, âm thanh sẽ cần những điều chỉnh cần thiết, phù hợp với bài toán được đặt ra.

2. Thực nghiệm

Bài báo sẽ thực nghiệm VA tích hợp LLM, lấy ví dụ cụ thể là tích hợp Chat GPT vào Alexa với 3 tình huống được giả định như sau:

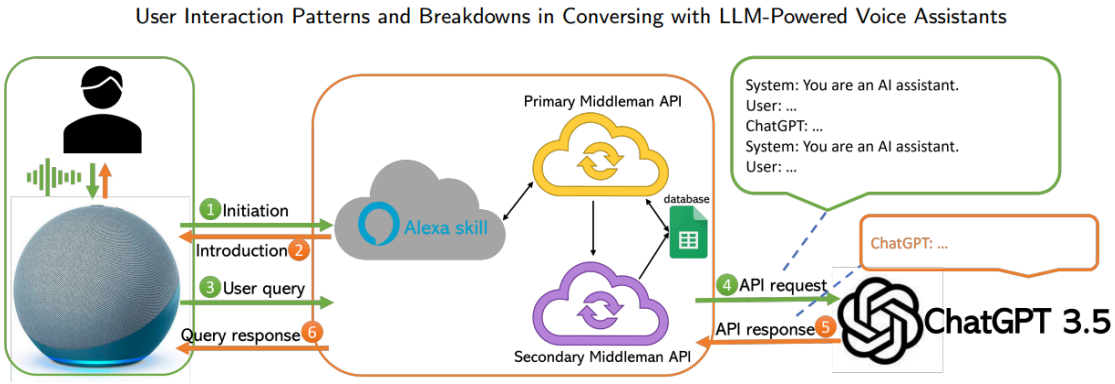
1. medical self-diagnosis - chẩn đoán bệnh
2. creative trip planning - lập lịch
3. discussion with an opinionated AI - thảo luận chuyên sâu với user

Họ đã rút ra được kết luận đó là:

- Tạo ra các mẫu trả lời đa dạng, sát chủ định câu hỏi của user hơn so với VA truyền thống.

- Giảm các lỗi và breakdowns thông qua khả năng hiểu ngữ cảnh và tự khôi phục hội thoại.

3. Kiến trúc VAS tích hợp LLM



Bài báo dùng hệ thống tích hợp ChatGPT vào Alexa bao gồm các thành phần sau:

1. **Thiết bị Echo Dot + Alexa Skill**

    - **Người dùng** : Phát lệnh giọng nói hoặc tiếng ho (đối với kịch bản y tế).
    - **Echo Dot**: Ghi âm, chuyển âm thanh thành văn bản nhờ ASR, phát lại tiếng nói của VA.

2. **Primary Middleman API (đám mây vàng)**

    - Nhận query ngay lập tức, không chờ phản hồi từ ChatGPT.
    - Ghi nhận request của Alexa Skill rồi ngay lập tức trả về “đã tiếp nhận” để không vượt quá giới hạn timeout (8 s).

3. **Secondary Middleman API (đám mây tím)**

    - Thực sự chịu trách nhiệm gọi tới ChatGPT API, kèm theo toàn bộ lịch sử hội thoại.
    - Khi ChatGPT phản hồi, viết kết quả vào cơ sở dữ liệu chung (Google Sheets).

4. **ChatGPT 3.5-turbo**
    - Xử lý yêu cầu kèm lịch sử cuộc trò chuyện, sinh văn bản trả lời.
5. **Cơ sở dữ liệu chung (Google Sheets)**
    - Luồng hai chiều giữa Primary và Secondary để trao đổi response từ ChatGPT.

Luồng hoạt động (đánh số 1→6)

1. **Initiation**: Người dùng khởi động cuộc trò chuyện bằngl (“Alexa, let’s chat” hoặc ho) → Alexa Skill nhận lệnh.
2. **Introduction**: Alexa Skill trả lời mở màn (“You are an AI assistant…”) theo prompt kịch bản.
3. **User query**: Người dùng hỏi nội dung chính → Alexa Skill thu văn bản.
4. **API request**: Primary Middleman gửi yêu cầu kèm lịch sử cuộc trò chuyện đến Secondary, Secondary gọi ChatGPT.
5. **API response**: ChatGPT sinh câu trả lời, Secondary Middleman nhận và lưu vào Google Sheets.
6. **Query response**: Primary Middleman lấy câu trả lời từ Sheets, Alexa Skill phát lại qua loa Echo Dot.

Sau khi hoàn tất một lượt hỏi–đáp, chu trình từ bước 3→4→5→6 lặp lại cho mỗi truy vấn tiếp theo, duy trì multi-turn conversation.

Kiến trúc ba module:

1. Echo Dot (ghi âm & transcription)
2. Alexa skill + dual middleman APIs (Primary để phản hồi tức thì, Secondary liên kết trực tiếp với ChatGPT và lưu lịch sử trên Google Sheets)
3. ChatGPT API (gpt-3.5-turbo)

Xử lý latency:

- 2 s → VA sử dụng filler như (“I’m on it”) trám vào thời gian chờ answer của user
- 6 s → VA sử dụng các câu chuyện phiếm/small talk như (“How’s your day going?”)
- Sau đó quay lại kết quả khi ready

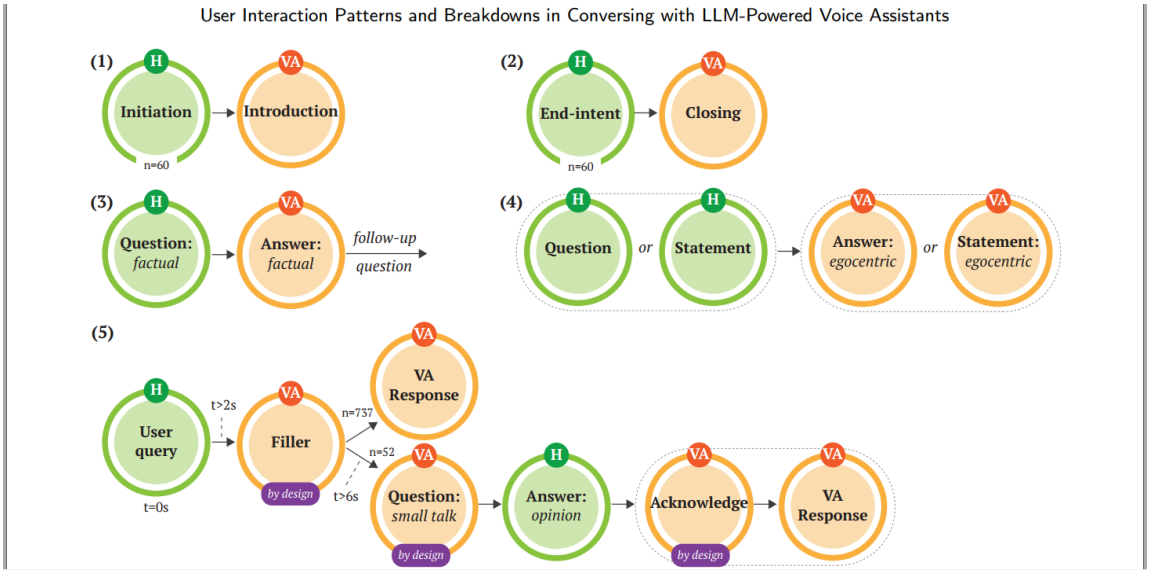
4. Kịch bản thử nghiệm

- Người tham gia: 20 người tham gia (10 nam, 10 nữ; độ tuổi 19–57; kinh nghiệm sử dụng VA trung bình) thực hiện luân phiên 3 kịch bản

1. **Medical: Self-diagnosis**
    - **Mục tiêu**: Mô phỏng quá trình thăm khám bệnh với triệu chứng (cough, fever…), hỏi thuốc OTC, cách chữa trị/phòng bệnh, dấu hiệu nhập viện.
    - **Prompt**: ChatGPT hành xử như “AI bác sĩ giọng nói”, hỏi từng bước, tránh lặp cảnh báo, giới hạn độ dài answer < 100 từ
2. **Creative planning: Plan a day**
    - **Kịch bản**: Người dùng hỏi VA lập kế hoạch như là đi chơi cho mình với những giới hạn nhất định như “phải gần hotel ở, phù hợp với sở thích, thời tiết…”l.
    - **Prompt**: ChatGPT nhớ vị trí, gợi ý tổng quát + chi tiết theo follow-up, < 100 từ .
3. **Discussion with AI: Opposing stance**
    - **Chủ đề**: “Đại học nên có lực lượng cảnh sát riêng không?”
    - **Prompt**: ChatGPT hỏi ý kiến user trước, giữ vững lập trường quan điểm đã được cài đặt từ trước, sau khi user chọn phe sẽ phản biện/đồng tình, không thay đổi quan điểm, < 100 từ

5. Nhận xét

- Dữ liệu tương tác từ người tham gia có 969 lượt; mỗi lượt bao gồm một cặp query user - answer VA
- Đặc trưng chung trong giao tiếp với VA ở cả 3 kịch bản gồm 5 đặc trưng chính:



1. **Initiation → Introduction**

   – **Pattern**: User khởi đầu 1 cuộc trò chuyện với VA bằng lệnh hoặc bằng tín hiệu mà VA bắt được(VD: tiếng ho trong kịch bản y tế) → VA giới thiệu (“I just heard you cough… Maybe I can help?”).
   – **Tần suất**: n = 60 .

2. **End-intent → Closing**

   – User đưa ra ý định kết thúc (“That’s all.”) → VA nói lời tạm biệt (“Goodbye!”). n = 60 .

3. **Factual question → Factual answer**

   – Hỏi các câu đòi hỏi các tri thức thực tế/hiển nhiên “factual” → VA đáp lại bằng câu trả lời mang tính chất thực tế “factual”, dẫn đến các câu hỏi liên quan từ user “follow-up”.

4. **Perspective of speech → Answer/Statement: egocentric**

   – VA luôn dùng ngôi “you” dù cho user dùng ngôi thứ 3 hay thứ nhất. .

5. **Wait patterns (fillers & small talk)**

   – **Short wait** (2–6 s): Nếu việc truy xuất thông tin mất hơn 2 giây, VA sẽ cung cấp các câu lệnh filler trám vào thời gian chờ như “I’m looking it up”, chiếm 76% lượt.
   – **Long wait** (> 6 s): VA sẽ bắt đầu 1 cuộc trò chuyện phiếm/small talk như hỏi về món ăn yêu thích của người dùng rồi sau khi nhận được response cho câu hỏi ban đầu thì VA sẽ kết thúc small talk và trả lời câu hỏi ban đầu, chiếm 5.37% lượt

5.1. Kịch bản 1: Chẩn đoán y tế

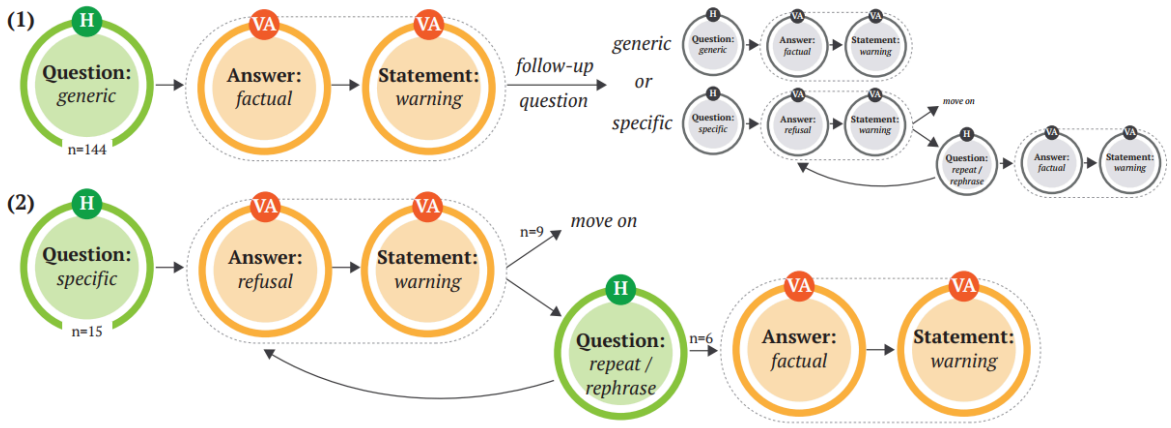


Figure 6: User interaction patterns in medical self-diagnosis: Participants’ frequent *generic* questions were answered in a *factual* style, often including a cautionary *warning* (1). For *specific* questions, which were less frequent, the VA demonstrated a reluctance to answer (*refusal*) and instead issued a warning (2). However, when participants reformulated *(repeat/rephrase)* the question, the VA responded in a factual manner (2).

Question: generic → Answer: factual + statement: warning

– Hỏi chung chung/generic như (“unusual symptom…”) → VA sẽ trả lời các thông tin liên quan kèm theo đó tham khảo thêm thông tin từ chuyên gia (n = 123). .

Question: specific → Answer: refusal + statement: warning

– Hỏi cụ thể/specific như (thành phần thuốc, hãng thuốc,..) → hướng dẫn rõ + tham khảo thêm từ chuyên gia (n = 97).

5.2. Kịch bản 2: Lập kế hoạch

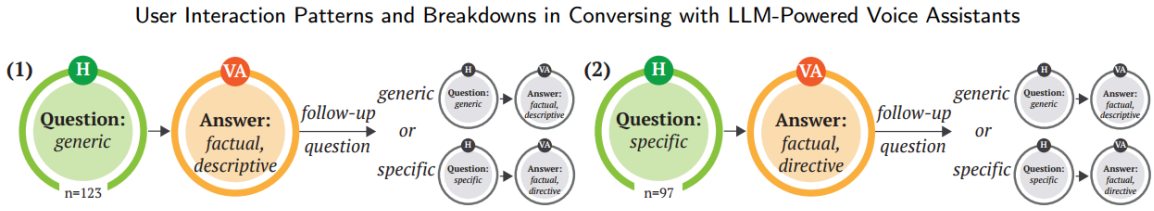


Figure 7: User interaction patterns in creative trip planning: Participant's *generic* questions were answered by VA in a *descriptive* style (1) while *specific* ones in a *directive* style (2).

Question: generic → Answer: factual, descriptive

– Hỏi chung (“unusual experiences…”) → mô tả sinh động (n = 123). .

Question: specific → Answer: factual, directive

– Hỏi specific (directions, hours) → hướng dẫn rõ (n = 97).

5.3. Kịch bản 3: Tranh luận với VA có lập trường quan điểm được thiết lập sẵn

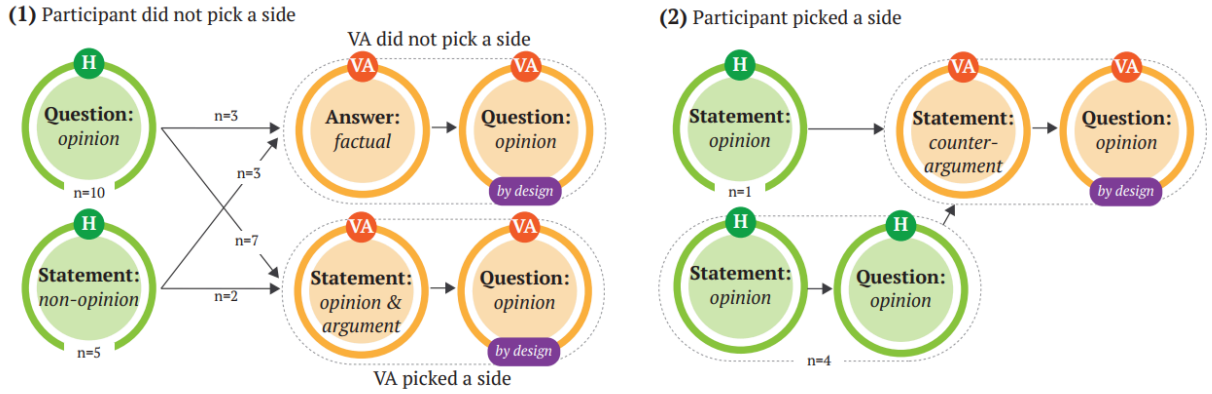


Figure 8: Interaction patterns for the commencement of a discussion with an opinionated AI. Participants either remain neutral (1) or pick a side (2). Each discussion starts only once per participant, totaling 20 patterns.The VA state "Question: opinion" is marked "by design" since we prompt ChatGPT to ask a question at end of each turn (see Appendix A.1).

Discussion commencement

– User neutral với user chọn phe → VA hỏi lại user lập trường quan điểm trước khi tranh luận (20 patterns). .

Discussion progression

– **Q–A** để làm rõ stance (n = 69),
– **Phản biện qua lại** (n = 73),
– **Đồng tính với ý kiến** user + VA củng cố quan điểm (n = 18).

6. Sự cố và lỗi xảy ra trong thực nghiệm

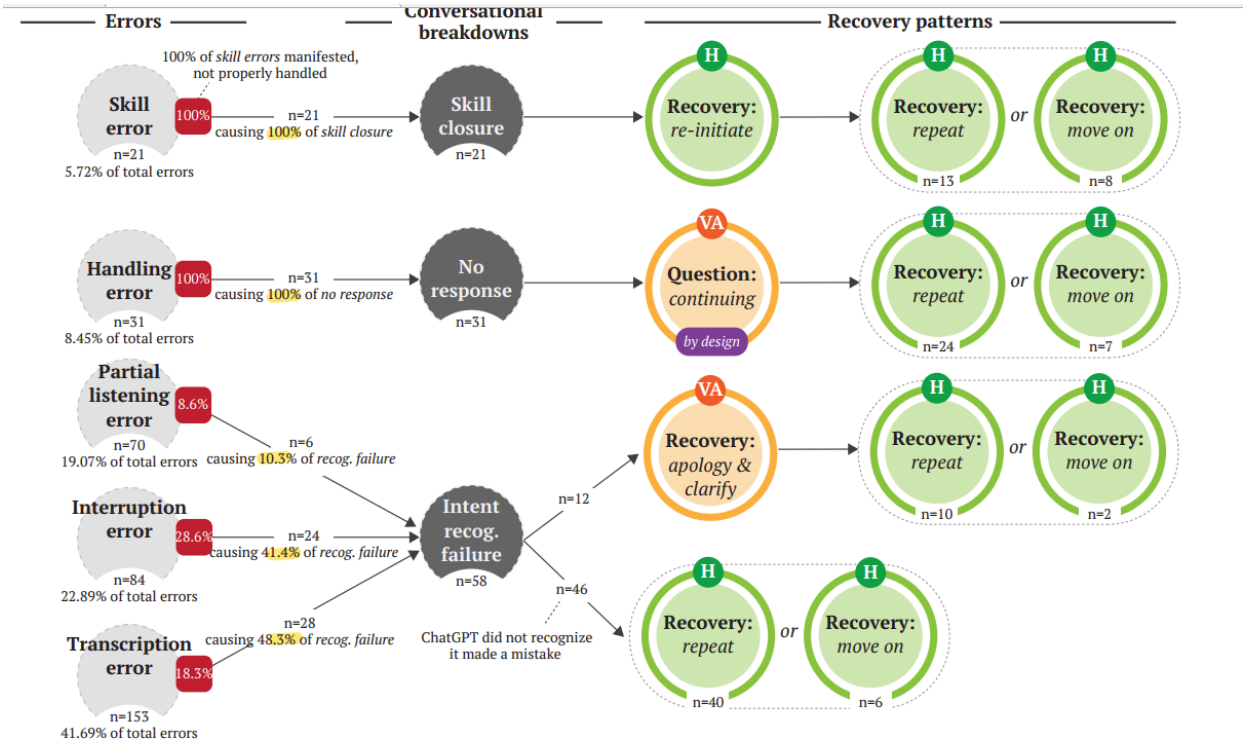
Error Type	Causes and Breakdowns
skill	**Cause:** Issues related to our system implementation, such as API response error. **Breakdown:** Skill closure after Alexa's announcement: "There was a problem with requested skill's response."
listening	**Cause:** User speaking when Alexa is not listening. **Breakdown:** Nothing happens.
handling	**Cause:** Alexa fails to pass transcribed speech to the ChatGPT skill. **Breakdown:** *No VA response.*
partial listening	**Cause:** Alexa only partially captures user speech. **Breakdown:** User intent recognition failure.
interruption	**Cause:** Alexa interrupts or cuts off user. **Breakdown:** User intent recognition failure.
transcription	**Cause:** Alexa transcribes user speech incorrectly. **Breakdown:** User intent recognition failure.
Recovery Strategy	**Definition**
repeat/rephrase	User repeats their query with added details or changed wording.
move on	User overlooks the unanswered or wrongly answered query and proceeds with a new one.
apology and clarify	VA apologizes and asks user to clarify their query before responding.

Table 9
Total number of errors for each scenario. Note: all skill and handling errors are manifested—resulted in breakdown.

Task	Skill	Handling	Partial listening	Manifested	Interruption	Manifested	Transcription	Manifested
Medical	2	12	20	1	22	7	37	6
Trip	16	13	32	4	24	10	62	15
Discussion	3	6	18	1	38	7	54	7

- **Tổng số lỗi**: 37.87% lượt (n = 367).
- **Các lỗi phổ biến**: transcription - biên dịch từ voice sang text (41.7%), interruption, partial listening…
- **Manifested**: lỗi từ VA chỉ ~30% (n = 110) dẫn đến breakdowns

Chiến lược VA dùng để khôi phục lại cuộc trò chuyện và sửa lỗi sự cố/breakdown



Skill Closure → Recovery

– Skill error → VA shutdown (n = 21); user re-initiate + repeat query .

No VA response → Recovery

– Handling error → VA tiếp “Any other questions?”; user repeat hoặc move on (n = 31) .

Intent recognition failure → Recovery

– Partial listening/interruption/transcription → VA apology+clarify (n = 12), user repeat (n = 40) hoặc move on (n = 6)

Trải nghiệm của user về giao tiếp với VA:

- VA cho các reponse dài dòng và lặp lại "Verbose & repetitive": người dùng than phiền nhưng vẫn chấp nhận các warning do cảm thấy cũng phù hợp trong việc cung cấp các thông tin cần kiểm chứng.
- Cảm nhận riêng với từng kịch bản:
    - Medical → "cautious"
    - Creative → "information provider"
    - Discussion → "opinionated but not aggressive"
- Dễ phục hồi cuộc trò chuyện nhờ lưu lại lịch sử hội thoại, nhưng đôi khi phải lặp lại thông tin đã có từ trước để VA "nắm" lại.

7. Vấn đề còn tồn tại - Hướng giải quyết tiếp theo

1. **Repetitiveness of content**
    - **Vấn đề**: VA lặp cảnh báo và filler quá nhiều.
    - **Giải pháp**: Prompt engineering để giảm lặp cảnh báo, đa dạng hóa phrasing. .
2. **Oversharing: Density of information**
    - **Vấn đề**: Phản hồi dài dòng, gây quá tải cho người nghe.
    - **Giải pháp**: Áp dụng **hierarchical responses**: cho response ngắn gọn, bao quát được ý đồ của user rồi đi vào chi tiết thêm khi user yêu cầu; thêm pauses/fillers tự nhiên. .
3. **Potential discrepancies in users' mental models**
    - **Vấn đề**: Sau breakdown, user nghĩ VA đã quên hết thông tin từ trước đó nên phải hỏi lại từ các bước đầu.
    - **Giải pháp**: VA nên gợi ý có thể tiếp tục cuộc trò chuyện dang dở ("Welcome back! Pick up where we left off?") để điều chỉnh kỳ vọng. .

Capabilities of LLM-Powered VAs: Potential and Design Guidelines

1. **Conversational resilience: The role of LLMs in overcoming VA disruptions**
    - **Tiềm năng**: LLM giảm 81.1% lỗi không xác định được ý đồ của người tham gia nhờ giữ context, tự xin lỗi+làm rõ lại ý đồ 20.7%; tăng tính bền bỉ và nắm bắt ngữ

cảnh/lịch sử trò chuyện tốt hơn. .

2. **Balancing proactive recovery and contextual comprehension**
    - **Thách thức**: Quá nhiều clarification làm gián đoạn, quá ít thì bỏ lỡ breakdown.
    - **Giải pháp**: Tune prompt + model parameters để cân bằng – chỉ hỏi khi thực sự cần .
3. **Retaining conversational history for post-error recovery**
    - **Tiềm năng**: VA lưu lịch sử khi skill đóng, giúp resume mạch hội thoại sau lỗi.
    - **Giải pháp**: Thiết kế để mọi lỗi không mất hết ngữ cảnh

Hạn chế của thực nghiệm:

- Đây chỉ là lỗi và nhận định trên từ VA là Alexa chưa phải là các VA khác ví dụ như Siri của Apple.
- Được thực hiện trong phòng thí nghiệm, có kịch bản từ trước chưa thực hiện trên thực tế với đông người tham gia
- Chưa thử mixed-initiative VA-initiated prompts

Hướng tương lai: Khai thác proactive mixed-initiative, tích hợp sâu hơn với speech-to-speech LLMs

8. Kết luận

LLM (ChatGPT) khi tích hợp vào VA cho thấy:

- Giữ ngữ cảnh tốt, đa mẫu tương tác
- Giảm breakdowns, tự phục hồi
- Khả năng cung cấp đa dạng phong cách trả lời theo nhiệm vụ

Hướng triển khai tiếp theo: Áp dụng các giải pháp nêu trên để phát triển VA voice cao cấp hơn. .