TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC QUY NHƠN

Xây dựng hệ thống Chatbot tư vấn tuyển sinh hoạt động trên thiết bị biên sử dụng nền tảng Rasa: Thiết kế, Triển khai và Đánh giá

Phan Đức Thiện[1,*], Nguyễn Thị Thu Thủy[1], Trần Văn Hạnh[1], Trần Thị Yến[1]

[1] Trường Đại học Sư phạm Kỹ thuật Nam Định, Ninh Bình, Việt Nam

Ngày nhận bài: dd/mm/yyyy

TÓM TẮT

Trong bối cảnh chuyển đổi số giáo dục, việc ứng dụng trí tuệ nhân tạo vào công tác tư vấn tuyển sinh là xu hướng cấp thiết nhằm giảm tải áp lực cho nhân sự hành chính và đáp ứng nhu cầu tương tác tức thời của người học thế hệ mới. Bài báo này đề xuất thiết kế và triển khai hệ thống trợ lý ảo (Chatbot) tương tác bằng giọng nói tiếng Việt, hoạt động độc lập trên thiết bị nhúng Raspberry Pi 4. Nghiên cứu sử dụng khung làm việc mã nguồn mở Rasa (gồm Rasa NLU và Rasa Core) kết hợp với mô hình nhận dạng giọng nói Whisper để xây dựng giải pháp tính toán biên (Edge Computing) không phụ thuộc vào dịch vụ đám mây. Kết quả thực nghiệm tại Trường Đại học Sư phạm Kỹ thuật Nam Định cho thấy hệ thống vận hành ổn định trong môi trường offline, đạt độ chính xác nhận diện ý định trung bình trên 92% đối với các nhóm câu hỏi về mã ngành, học phí và quy chế đào tạo. Giải pháp đề xuất không chỉ đảm bảo an toàn dữ liệu mà còn tối ưu hóa chi phí đầu tư phần cứng (dưới 5 triệu đồng), chứng minh tính khả thi cao khi ứng dụng tại các cơ sở giáo dục đại học.

Từ khóa: *Rasa Framework, Chatbot, Raspberry Pi, Tư vấn tuyển sinh, Tính toán biên.*

[*] *Tác giả liên hệ chính*
Email: phanducthien@nute.edu.vn

QUY NHON UNIVERSITY
JOURNAL OF SCIENCE

Development of an Edge-based Admissions Counseling Chatbot System using Rasa Framework: Design, Implementation, and Evaluation

Duc-Thien Phan[1,*], Thi-Thu-Thuy Nguyen[1], Van-Hanh Tran[1], Thi-Yen Tran[1]

[1] *Nam Dinh University of Technology Education, Ninh Binh, Viet Nam.*

Received: dd/mm/yyyy

ABSTRACT

In the context of digital transformation in education, applying artificial intelligence to admissions counseling is an urgent trend to reduce the workload for administrative staff and meet the immediate interaction needs of new-generation learners. This paper proposes the design and implementation of a Vietnamese voice-activated virtual assistant (Chatbot) system operating independently on the Raspberry Pi 4 embedded device. The study utilizes the open-source Rasa framework (comprising Rasa NLU and Rasa Core) combined with the Whisper speech recognition model to build an Edge Computing solution that does not rely on cloud services. Experimental results at Nam Dinh University of Technology Education show that the system operates stably in an offline environment, achieving an average intent recognition accuracy of over 92% for inquiries regarding major codes, tuition fees, and training regulations. The proposed solution not only ensures data privacy but also optimizes hardware investment costs (under 5 million VND), demonstrating high feasibility for application in higher education institutions.

Keyworks: *Rasa Framework, Chatbot, Raspberry Pi, Admissions Counseling, Edge Computing.*

1. MỞ ĐẦU

Trong bối cảnh Cuộc Cách mạng công nghiệp 4.0 diễn ra mạnh mẽ, quá trình chuyển đổi số đang trở thành yêu cầu tất yếu đối với các cơ sở giáo dục đại học. Việc ứng dụng công nghệ số không chỉ góp phần tối ưu hoá hoạt động quản lý mà còn nâng cao chất lượng dịch vụ hỗ trợ người học. Tại Trường Đại học Sư phạm Kỹ thuật Nam Định, công tác tư vấn tuyển sinh giữ vai trò then chốt trong việc định hướng và thu hút thí sinh tiềm năng. Tuy nhiên, các phương thức tư vấn truyền thống như điện thoại, email hay trao đổi thủ công qua mạng xã hội đang bộc lộ nhiều hạn chế: quá tải nhân sự vào các giai đoạn cao điểm, thời gian phản hồi kéo dài, thiếu tính đồng bộ thông tin và khó duy trì chất lượng dịch vụ ở mức ổn định [1].

Bên cạnh đó, đối tượng tuyển sinh chính hiện nay là thế hệ Gen Z - nhóm người học có đặc trưng nổi bật về việc tiếp nhận thông tin nhanh, ưa thích các tương tác tức thời, cá nhân hoá và thường xuyên sử dụng các nền tảng số. Gen Z kỳ vọng dịch vụ tư vấn hoạt động 24/7, phản hồi chính xác và nhất quán, điều mà các mô hình tư vấn truyền thống khó đáp ứng hiệu quả [2,6]. Điều này đặt ra yêu cầu cấp thiết đối với các cơ sở đào tạo trong việc xây dựng các hệ thống hỗ trợ thông minh, linh hoạt và tự động hoá cao.

Hiện nay, nhiều nền tảng Chatbot dựa trên trí tuệ nhân tạo (AI) đã được phát triển và đưa vào ứng dụng, phổ biến nhất là các giải pháp dựa trên Google Dialogflow, Microsoft Bot Frame-

[*] *Corresponding author*
Email: phanducthien@nute.edu.vn

work hay OpenAI API. Mặc dù mang lại khả năng xử lý ngôn ngữ tự nhiên tốt và có độ chính xác cao, các giải pháp này tiềm ẩn nhiều hạn chế: chi phí duy trì lớn, phụ thuộc hoàn toàn vào kết nối Internet ổn định, khó kiểm soát dữ liệu người dùng và gia tăng rủi ro rò rỉ thông tin nhạy cảm [3]. Đây là những yếu tố đặc biệt quan trọng trong lĩnh vực giáo dục, nơi yêu cầu về bảo mật và quyền riêng tư được đặt lên hàng đầu.

Để khắc phục những tồn tại trên, hướng tiếp cận sử dụng mã nguồn mở **Rasa** kết hợp với phần cứng giá rẻ **Raspberry Pi** theo mô hình tính toán biên (Edge Computing) đang ngày càng nhận được sự quan tâm. Mô hình này cho phép xử lý dữ liệu ngay tại thiết bị cục bộ thay vì phụ thuộc vào dịch vụ đám mây, từ đó giảm độ trễ, đảm bảo tính riêng tư, tiết kiệm chi phí và duy trì khả năng hoạt động ngay cả khi kết nối Internet không ổn định. Đây là một hướng tiếp cận phù hợp với điều kiện triển khai thực tế của các cơ sở giáo dục tại Việt Nam.

Bài báo này trình bày quy trình xây dựng, huấn luyện và triển khai trợ lý ảo tư vấn tuyển sinh - **NUTE-Bot**. Điểm mới của nghiên cứu là (i) tối ưu hoá mô hình ngôn ngữ tiếng Việt để vận hành hiệu quả trên thiết bị có tài nguyên hạn chế như Raspberry Pi 4; (ii) xây dựng kịch bản hội thoại mang tính sư phạm, chuyển từ văn phong hành chính sang văn phong thân thiện, đồng hành; và (iii) thử nghiệm mô hình trong môi trường thực tế để đánh giá hiệu năng, khả năng mở rộng cũng như mức độ chấp nhận của người dùng. Kết quả nghiên cứu góp phần cung cấp một giải pháp khả thi, chi phí thấp nhưng hiệu quả, hướng tới mục tiêu chuyển đổi số bền vững trong công tác tư vấn tuyển sinh tại các cơ sở giáo dục đại học.

2. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP

Việc xây dựng một hệ thống trợ lý ảo tư vấn tuyển sinh yêu cầu sự kết hợp giữa các mô hình xử lý ngôn ngữ tự nhiên, các cơ chế quản lý hội thoại thông minh và phương pháp triển khai tối ưu trên thiết bị biên (Edge Device). Hình 1 minh họa tổng quan cấu trúc kết nối phần cứng cùng luồng xử lý dữ liệu trên Raspberry Pi trong nghiên cứu này.

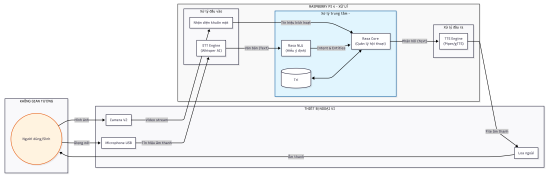

Figure 1. Sơ đồ kết nối phần cứng và luồng xử lý dữ liệu trên Raspberry Pi

2.1. Xây dựng mô hình Rasa NLU và Core

Trong nghiên cứu này, Rasa Framework được lựa chọn làm nền tảng vì tính mở, khả năng tùy chỉnh sâu và khả năng triển khai cục bộ mà không phụ thuộc vào dịch vụ đám mây. Thành phần NLU (Natural Language Understanding) sử dụng kiến trúc DIET (Dual Intent and Entity Transformer) [4,5], vốn nổi bật bởi khả năng học đa nhiệm và hiệu quả tính toán phù hợp với môi trường triển khai hạn chế tài nguyên.

Pipeline xử lý dữ liệu đầu vào của Rasa được mô tả như sau:

$$\text{Pipeline} = [\text{Tokenizer} \rightarrow \text{Featurizer} \rightarrow \text{Classifier}] \quad (1)$$

Trong nghiên cứu này, tệp *config.yml* được thiết kế và tối ưu nhằm đạt hiệu năng tốt nhất khi huấn luyện và suy luận mô hình:

- **Tokenizer:** Sử dụng *WhitespaceTokenizer* kết hợp tiền xử lý tách từ tiếng Việt. Điều này đảm bảo giảm thiểu hiện tượng nhập nhằng từ vựng trong văn bản tiếng Việt, đặc biệt khi xử lý các cụm danh từ dài.
- **Featurizer:** Kết hợp *CountVectorsFeaturizer* (biểu diễn văn bản bằng mô hình Bag-of-Words) và *LexicalSyntacticFeaturizer* để mô hình hóa đặc trưng ngữ nghĩa và cú pháp.
- **Classifier:** Thành phần chính là *DIETClassifier* với 100 epochs, được lựa chọn nhằm cân bằng giữa độ chính xác và thời gian huấn luyện. Kiến trúc DIET cho phép mô hình học đồng thời ý định (Intent) và thực thể (Entity), giúp hệ thống hiểu ngữ cảnh tốt hơn.

Bên cạnh đó, cơ sở tri thức được số hoá từ các văn bản chính thức như quy chế đào tạo, thông báo tuyển sinh, đề án tuyển sinh và các quyết định liên quan (ví dụ: Quyết định 414/QĐ-ĐHSPKTNĐ). Tập dữ liệu huấn luyện (*nlu.yml*) bao gồm hơn 25 nhóm ý định (intents) tiêu biểu: *hoi_ma_truong, hoi_hoc_phi, hoi_ky_tuc_xa, hoi_nganh_cntt*, ... Mỗi ý định được xây dựng với tập câu mẫu đa dạng nhằm tăng khả năng tổng quát hóa của mô hình trong thực tế.

Để duy trì độ chính xác lâu dài, hệ thống được tích hợp giao diện giám sát hội thoại (Hình 2), cho phép người quản trị xem lại hội thoại, gán nhãn lại các câu mà mô hình hiểu sai và bổ sung vào tập huấn luyện. Đây là cơ chế học liên tục (Continual Learning) giúp mô hình ngày càng hoàn thiện và thích ứng với nhu cầu thực tiễn.

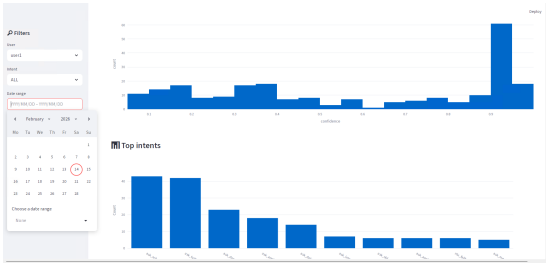

Figure 2. Giao diện giám sát hội thoại hỗ trợ tinh chỉnh tri thức và huấn luyện lại (Retrain)

2.2. Quản lý hội thoại và tối ưu hóa trên thiết bị biên

Trong cấu trúc Rasa, quản lý hội thoại được thực hiện bởi Rasa Core. Tuy nhiên, trên các thiết bị biên như Raspberry Pi, việc sử dụng cơ chế *Stories* phức tạp sẽ tiêu tốn nhiều tài nguyên tính toán do yêu cầu mô hình dự đoán chuỗi hành động tiếp theo dựa trên lịch sử hội thoại. Điều này làm tăng thời gian phản hồi và có thể gây quá tải CPU.

Để khắc phục, nghiên cứu lựa chọn cơ chế *Rules*, vốn đơn giản, trực quan và phù hợp với các tác vụ tư vấn dạng truy vấn-trả lời. Quy tắc (rule) được khai báo trong tệp *rules.yml* theo định dạng:

```
- rule: Tra loi ve hoc phi
  steps:
  - intent: hoi_hoc_phi
  - action: utter_hoc_phi
```

Việc sử dụng Rules giúp đảm bảo tốc độ phản hồi ổn định, giảm đáng kể lượng tài nguyên cần thiết để suy luận hành động. Đây là yếu tố quan trọng khi triển khai trên Raspberry Pi, nơi CPU và bộ nhớ RAM đều có giới hạn.

Quá trình huấn luyện mô hình (Train) được thực hiện trên máy tính cá nhân có cấu hình cao nhằm tối ưu thời gian đào tạo và tránh gây quá tải thiết bị biên. Sau khi huấn luyện hoàn tất, mô hình được đóng gói dưới dạng tệp nén `.tar.gz` và chuyển sang Raspberry Pi để triển khai (Deploy). Cách tiếp cận này không chỉ đảm bảo tính ổn định khi vận hành mà còn kéo dài tuổi thọ phần cứng nhờ giảm nhiệt lượng sinh ra trong quá trình tính toán nặng.

Ngoài ra, mô hình còn được tối ưu bằng cách vô hiệu hóa các thành phần không cần thiết trong Rasa (như telemetry, tracker persistence nâng cao), giúp giảm mức sử dụng CPU và bộ nhớ. Nhờ đó hệ thống có thể hoạt động liên tục 24/7, đáp ứng yêu cầu tư vấn tuyển sinh trong thời gian cao điểm.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Đánh giá hiệu năng mô hình NLU

Mô hình NLU sau khi huấn luyện được kiểm thử trên tập dữ liệu chiếm 20% tổng số mẫu, đảm bảo tính khách quan trong đánh giá. Bộ chỉ số gồm Precision, Recall và F1-Score được lựa chọn vì khả năng phản ánh toàn diện mức độ hiểu đúng ý định người dùng. Các kết quả trên một số nhóm ý định phổ biến được tổng hợp trong Bảng 1.

Table 1. Kết quả đánh giá mô hình NLU trên các Intent chính

Intent (Ý định)	Precision	Recall	F1-Score
hoi_ma_truong	1.00	1.00	1.00
hoi_hoc_phi	0.94	0.92	0.93
hoi_nganh_cntt	0.89	0.91	0.90
chao_hoi	0.98	0.97	0.97
hoi_ktx	0.92	0.88	0.90
Trung bình	**0.94**	**0.93**	**0.94**

Kết quả cho thấy mô hình đạt hiệu năng tổng thể tốt (F1 trung bình 0.94). Các ý định có cấu trúc ngắn gọn, ngữ nghĩa rõ ràng như *hoi_ma_truong* đạt độ chính xác tuyệt đối. Đối với các ý định có từ khóa tương đồng (ví dụ: sự gần nhau giữa "Công nghệ thông tin" và "Công nghệ kỹ thuật máy tính"), độ chính xác có giảm nhẹ do hiện tượng giao thoa ngữ nghĩa. Tuy nhiên, hệ thống fallback hoạt động hiệu quả, giúp người dùng được định hướng hỏi lại theo cấu trúc rõ ràng hơn.

Đáng chú ý, trong tập kiểm thử xuất hiện nhiều mẫu có cách hành văn đa dạng, bao gồm cả các câu hỏi viết không dấu. Điều này cho thấy mô hình có khả năng tổng quát hóa tốt, đáp ứng các tình huống sử dụng thực tế tại môi trường tư vấn tuyển sinh.

3.2. Triển khai thực tế và hiệu năng phần cứng

Hệ thống sau khi hoàn thiện phần mềm đã được đóng gói và triển khai thực nghiệm tại khu vực tư vấn tuyển sinh của Trường Đại học Sư phạm Kỹ thuật Nam Định. Thiết bị Raspberry Pi 4 (4GB RAM) vận hành độc lập với nguồn điện 5V-3A, không yêu cầu kết nối Internet trong quá trình tư vấn (Hình 3). Kết nối mạng chỉ được kích hoạt khi cập nhật tri thức hoặc nâng cấp phần mềm.

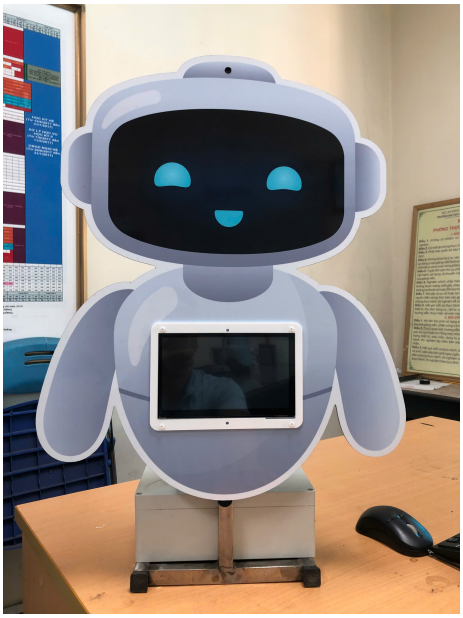

Figure 3. Hình ảnh thực tế hệ thống Robot

Để đánh giá hiệu năng trên thiết bị biên có tài nguyên hạn chế, nhóm nghiên cứu tiến hành theo dõi mức tiêu thụ tài nguyên trong hai trạng thái vận hành: *Idle* và *Active*. Các kết quả ghi nhận được phân tích như sau:

- **Mức tiêu thụ RAM:** Hệ thống sử dụng trung bình 1.2-1.5 GB RAM, nằm trong ngưỡng an toàn của Raspberry Pi 4 (4 GB). Điều này cho phép hệ điều hành và các tiến trình nền hoạt động ổn định mà không xảy ra tình trạng tràn bộ nhớ.
- **Tải trọng CPU:** Khi ở trạng thái chờ, mức tải CPU duy trì ở mức thấp. Khi xử lý giọng nói (Whisper nhỏ gọn) và suy luận NLU, CPU tăng lên 70-85% trong thời gian ngắn (1.5-2 giây). Sau đó, tải giảm nhanh, giúp thiết bị vận hành liên tục mà không gặp nguy cơ quá nhiệt.
- **Độ trễ (Latency):** Thời gian phản hồi trung bình từ lúc người dùng dứt lời đến khi hệ thống trả lời khoảng 3-4 giây. Mặc dù cao hơn so với các hệ thống dựa trên cloud API, đây là mức chấp nhận được đối với một mô hình xử lý tại chỗ (on-premise) vận hành hoàn toàn offline, ưu tiên bảo mật dữ liệu.

Kết quả trên khẳng định khả năng ứng

dụng thực tiễn của việc triển khai mô hình Rasa trên thiết bị nhúng giá rẻ. Điều này đặc biệt có ý nghĩa trong bối cảnh các cơ sở đào tạo mong muốn triển khai hệ thống tư vấn tự động nhưng có hạn chế về kinh phí và hạ tầng.

3.3. Đánh giá tính sư phạm và mức độ tương tác

Khía cạnh sư phạm trong giao tiếp là yếu tố quan trọng của hệ thống NUTE-Bot. Thay vì sử dụng ngôn ngữ hành chính khô khan, các câu trả lời được tái thiết kế với phong cách thân thiện, đồng hành và dễ tiếp cận. Ví dụ, thay vì trả lời: *"Theo Điều 5, khoản 2 trong Quy chế tuyển sinh..."*, bot sẽ sử dụng: *"Chào bạn! Về vấn đề học phí, Nhà trường đang áp dụng mức..."*.

Một khảo sát nhỏ với 50 sinh viên tình nguyện cho thấy:

- 85% người tham gia đánh giá giọng điệu trả lời là dễ hiểu và thân thiện.
- 78% cho biết bot giúp họ tiếp cận nhanh với thông tin họ cần.
- 12% phản hồi rằng bot đôi khi chưa hiểu đúng các câu hỏi phức hợp hoặc đa ý.

Mặc dù còn hạn chế trong việc xử lý các câu hỏi dài, đa tầng nghĩa hoặc chứa cảm xúc mạnh, hệ thống bước đầu đã thể hiện tính hiệu quả, hỗ trợ người dùng tìm kiếm thông tin nhanh và giảm tải đáng kể cho đội ngũ tư vấn tuyển sinh.

4. KẾT LUẬN

Nghiên cứu đã xây dựng thành công hệ thống trợ lý ảo phục vụ tư vấn tuyển sinh tại Trường Đại học Sư phạm Kỹ thuật Nam Định dựa trên nền tảng mã nguồn mở Rasa và phần cứng chi phí thấp Raspberry Pi. Kết quả thực nghiệm cho thấy hệ thống hoạt động ổn định, đáp ứng được nhu cầu tư vấn cơ bản của thí sinh trong bối cảnh số hóa hoạt động giáo dục ngày càng cấp thiết. Các đóng góp chính của bài báo có thể được tóm tắt như sau:

1. **Đề xuất kiến trúc Chatbot hoạt động độc lập (offline):** Hệ thống vận hành hoàn toàn cục bộ trên thiết bị biên, không phụ thuộc vào dịch vụ đám mây, qua đó đảm bảo an toàn dữ liệu cá nhân và giảm chi phí vận hành. Giải pháp phần cứng có tổng chi phí dưới 5 triệu đồng, phù hợp với điều kiện triển khai của các cơ sở đào tạo tại Việt Nam.
2. **Phát triển bộ dữ liệu và cấu hình tối ưu cho tiếng Việt:** Nhóm nghiên cứu đã xây dựng tập dữ liệu đa lĩnh vực liên quan đến tư vấn tuyển sinh, kết hợp với pipeline xử lý ngôn ngữ được tối ưu cho tiếng Việt. Mô hình DIET đạt độ chính xác trung bình 94%, cho thấy Rasa là nền tảng phù hợp để phát triển trợ lý ảo trong môi trường giáo dục.
3. **Khẳng định tính khả thi của Edge AI trong giáo dục:** Việc triển khai mô hình trên Raspberry Pi 4 chứng minh rằng các mô hình AI có thể được "đưa xuống biên" (edge computing) mà vẫn đảm bảo hiệu năng và trải nghiệm người dùng. Đây là hướng tiếp cận đặc biệt hữu ích đối với các đơn vị có ngân sách hạn chế nhưng vẫn cần xây dựng hệ thống thông minh phục vụ người học.

Mặc dù hệ thống đáp ứng tốt các yêu cầu cơ bản, vẫn còn tồn tại những thách thức như độ trễ xử lý giọng nói, hạn chế trong việc hiểu các câu hỏi đa ý hoặc mang yếu tố cảm xúc. Do vậy, trong tương lai nhóm nghiên cứu sẽ tập trung vào những hướng phát triển sau:

- **Tối ưu tốc độ suy luận** thông qua các kỹ thuật lượng tử hoá (quantization) và rút gọn mô hình (model pruning), nhằm giúp hệ thống phản hồi nhanh hơn mà không làm tăng chi phí phần cứng.
- **Mở rộng khả năng tích hợp** với hệ thống quản lý đào tạo, hệ thống tra cứu điểm, đăng ký nguyện vọng và các nền tảng số của nhà trường để nâng cao tính hữu ích.
- **Phát triển khả năng xử lý ngôn ngữ nâng cao**, bao gồm hiểu câu hỏi phức hợp, phân tích cảm xúc cơ bản và cá nhân hoá trải nghiệm người dùng.

- **Đánh giá dài hạn** về mức độ hài lòng của thí sinh khi tương tác với hệ thống trong các mùa tuyển sinh liên tiếp.

Nhìn chung, nghiên cứu mở ra hướng đi tiềm năng trong việc ứng dụng trí tuệ nhân tạo biên vào các hoạt động giáo dục, góp phần thúc đẩy chuyển đổi số bền vững, tiết kiệm và phù hợp với thực tiễn tại các trường đại học Việt Nam.

REFERENCES

1. Trần Thị Hương, Lê Văn An. Xu hướng sử dụng công cụ Chatbot trong học tập của sinh viên đại học: Một nghiên cứu thực nghiệm, *Tạp chí Giáo dục*, **2024**, 24(1), 45-52.

2. J. E. Chukwuere. The future of generative AI chatbots in higher education, *arXiv preprint arXiv:2403.13487*, **2024**.

3. T. Bocklisch, J. Faulkner, N. Pawlowski, A. Nichol. Rasa: Open Source Language Understanding and Dialogue Management, *arXiv preprint arXiv:1712.05181*, **2017**.

4. D. Jurafsky, J. H. Martin. *Speech and Language Processing (3rd ed.)*, Pearson, **2021**.

5. B. Balon, M. Simic. Using Raspberry Pi Computers in Education, *Proceedings of the 30th DAAAM International Symposium*, **2019**, 0197-0204.

6. A. Khumairo, A. Al Halik, L. K. Ningrum, C. Zahra. Analysis of the Impact of Using AI-Based Chatbots in Guidance and Counseling Services in Schools, *International Journal of Education and Literature*, **2025**, 6(1), 1-10.