DOI: 10.64302/joshusc.v33n2.1361

XÂY DỰNG CHATBOT HỎI ĐÁP VỀ DỮ LIỆU ĐA DẠNG SINH HỌC SỬ DỤNG KỸ THUẬT RETRIEVAL-AUGMENTED GENERATION VÀ GOOGLE GEMINI: NGHIÊN CỨU TRƯỜNG HỢP HỒ AYUN HẠ, TỈNH GIA LAI

Hoàng Đình Trung[1], **Nguyễn Dũng**[2,*], **Nguyễn Ngọc Thuỷ**[2]

[1]Khoa Sinh học, Trường Đại học Khoa học, Đại học Huế

[2]Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

*Email: nguyendung@hueuni.edu.vn

Ngày nhận bài: 22/3/2026; ngày hoàn thành phản biện: 30/3/2026; ngày duyệt đăng: 30/3/2026

TÓM TẮT

Bài báo trình bày nghiên cứu xây dựng hệ thống chatbot thông minh phục vụ tra cứu thông tin đa dạng sinh học ở tỉnh Gia Lai và trong nghiên cứu này lấy hồ Ayun Hạ, tỉnh Gia Lai làm ví dụ, dựa trên kỹ thuật Retrieval-Augmented Generation (RAG) kết hợp mô hình ngôn ngữ lớn Gemini của Google. Dữ liệu thực nghiệm là bộ danh lục 73 loài côn trùng nước thuộc 7 bộ, thu thập tại 9 điểm khảo sát trên hồ Ayun Hạ trong giai đoạn 2020–2022, được số hóa sang định dạng JSON với đầy đủ thông tin phân loại học, đặc điểm sinh thái và phân bố theo điểm. Hệ thống sử dụng mô hình nhúng text-embedding-004 của Google, kho vector FAISS và khung ứng dụng LangChain để xây dựng quy trình truy xuất và sinh câu trả lời. Thực nghiệm cho thấy hệ thống đạt hiệu quả cao, giảm thiểu đáng kể hiện tượng áo giác so với mô hình ngôn ngữ lớn không tích hợp RAG, đồng thời cung cấp câu trả lời có trích dẫn nguồn cụ thể, đáp ứng yêu cầu tra cứu chuyên ngành. Nghiên cứu là bước thử nghiệm thí điểm mở đường cho việc mở rộng hệ thống sang các nhóm sinh vật và hệ sinh thái trên cạn, thủy vực khác tại tỉnh Gia Lai.

Từ khóa: Retrieval-Augmented Generation, Chatbot, Đa dạng sinh học, Google Gemini.

1. MỞ ĐẦU

Dữ liệu đa dạng sinh học ngày càng được tích lũy phong phú qua các công trình điều tra thực địa trên khắp Việt Nam, song việc khai thác các bộ dữ liệu này để phục vụ tra cứu và giáo dục vẫn còn nhiều hạn chế. Các bộ dữ liệu sinh học thường được lưu trữ dưới dạng bảng danh lục hoặc tập tin có cấu trúc phức tạp, chỉ những người có chuyên

môn mới có thể tra cứu hiệu quả. Nghiên cứu của Hoàng Đình Trung (2023) [1] tại hồ Ayun Hạ, tỉnh Gia Lai, đã cung cấp một bộ dữ liệu phân loại học chi tiết về các loài sinh vật tại một trong những hồ nhân tạo lớn nhất Tây Nguyên, tạo nên nguồn tư liệu thực nghiệm có giá trị để thử nghiệm hệ thống tra cứu thông minh.

Trong bối cảnh đó, các mô hình ngôn ngữ lớn (Large Language Models, LLMs) dựa trên kiến trúc Transformer [2] đã mở ra hướng tiếp cận mới đầy triển vọng cho các giao diện hỏi đáp thông minh. Người dùng có thể đặt câu hỏi bằng ngôn ngữ tự nhiên và nhận câu trả lời tức thì thay vì phải tra cứu thủ công qua các bảng danh lục phức tạp. Tuy nhiên, một hạn chế cốt lõi của LLMs là hiện tượng ảo giác (*hallucination)*, tức mô hình tạo ra thông tin sai lệch hoặc không có cơ sở thực tế [3]. Đây là rào cản đặc biệt nghiêm trọng trong lĩnh vực phân loại học, nơi tính chính xác về danh pháp khoa học, vị trí taxon và phân bố địa lý là yêu cầu bắt buộc.

Kỹ thuật Retrieval-Augmented Generation (RAG) được Lewis và cộng sự giới thiệu tại NeurIPS 2020 [4] nhằm khắc phục hạn chế này bằng cách kết hợp khả năng sinh văn bản của LLM với cơ chế truy xuất thông tin trực tiếp từ cơ sở dữ liệu bên ngoài tại thời điểm suy luận. Bài báo này đề xuất và triển khai hệ thống chatbot tra cứu đa dạng sinh học sử dụng kỹ thuật RAG kết hợp với mô hình Gemini của Google [5], với nguồn dữ liệu thực nghiệm là Danh lục thành phần loài từ hồ Ayun Hạ [1] được chuyển đổi sang định dạng JSON có cấu trúc. Các đóng góp chính của nghiên cứu bao gồm: (i) pipeline chuyển đổi dữ liệu phân loại học từ bảng danh lục sang JSON chuẩn hóa phục vụ RAG; (ii) chiến lược thiết kế prompt tối ưu cho domain sinh học; và (iii) đánh giá toàn diện theo khung RAGAS [6] kết hợp nghiên cứu người sử dụng trên bộ câu hỏi kiểm thử chuyên ngành.

2. NGHIÊN CỨU LIÊN QUAN

Nghiên cứu này đặt nền tảng trên sự hội tụ của nhiều nhánh công nghệ, từ kiến trúc mô hình ngôn ngữ lớn, kỹ thuật truy xuất thông tin dày đặc, đến cơ sở hạ tầng lưu trữ vector và các hệ thống dữ liệu sinh học mở. Nền tảng lý thuyết bắt đầu từ kiến trúc Transformer do Vaswani và cộng sự đề xuất năm 2017 [2], trong đó cơ chế tự chú ý đa đầu cho phép mô hình lập mô hình các phụ thuộc tầm xa trong chuỗi văn bản mà không cần cấu trúc hồi quy. Đây là bước ngoặt mang tính cách mạng, tạo tiền đề cho toàn bộ thế hệ LLMs hiện đại. Tiếp nối đó, Devlin và cộng sự [7] giới thiệu BERT, mô hình pre-training hai chiều trên kiến trúc Transformer, đặt nền móng cho các mô hình nhúng văn bản chất lượng cao vốn là thành phần không thể thiếu trong quy trình RAG.

Kỹ thuật Retrieval-Augmented Generation (RAG) được Lewis và cộng sự hệ thống hóa tại NeurIPS 2020 [4], xác lập mô hình lai kết hợp giữa bộ nhớ tham số của mô hình sinh văn bản và bộ nhớ phi tham số tồn tại dưới dạng chỉ mục vec-tơ dày đặc. Gao

và cộng sự [3] đã phân tích sự tiến hóa của RAG qua ba thế hệ: *Naive RAG* với quy trình đơn giản nhất, *Advanced RAG* bổ sung kỹ thuật tối ưu hóa trước và sau khi truy xuất, và *Modular RAG* với kiến trúc linh hoạt cho phép thay thế từng thành phần độc lập. Nghiên cứu của Ma và cộng sự [8] về kỹ thuật viết lại truy vấn cho thấy việc diễn giải lại câu hỏi trước khi truy xuất có thể cải thiện đáng kể chất lượng ngữ cảnh tìm được, đặc biệt với các câu hỏi mơ hồ hoặc sử dụng từ đồng nghĩa không có trong kho dữ liệu.

Thành phần truy xuất dày đặc được làm sáng tỏ qua công trình DPR (Dense Passage Retrieval) của Karpukhin và cộng sự [9], trong đó cả câu truy vấn lẫn đoạn văn bản đều được mã hóa thành vec-tơ dày đặc bởi cùng mô hình BERT, và độ tương đồng được đo bằng tích vô hướng. Kết quả thực nghiệm cho thấy phương pháp truy vấn dày đặc vượt trội so với BM25 [10], với mức cải thiện dao động từ 9 đến 19 điểm phần trăm trên nhiều bộ câu hỏi mở. Để lưu trữ và tìm kiếm hiệu quả trên tập vec-tơ quy mô lớn, thư viện FAISS (Facebook AI Similarity Search) do Johnson, Douze và Jégou phát triển tại Meta AI [11] trở thành lựa chọn hàng đầu nhờ các thuật toán tiên tiến như Inverted File Index (IVF) kết hợp Product Quantization (PQ), cho phép xử lý hàng tỷ vec-tơ với độ trễ mili-giây và tiêu thụ bộ nhớ được nén đáng kể.

Về phía mô hình sinh văn bản, Gemini 1.5 Pro của Google [5] đại diện cho thế hệ LLM đa phương thức quy mô lớn với cửa sổ ngữ cảnh lên đến 1 triệu token và khả năng suy luận vượt trội trên nhiều bài kiểm tra chuyên môn học thuật. Mô hình nhúng text-embedding-004 của Google [12] tạo vector nhúng 768 chiều, được tối ưu hóa đặc biệt cho các tác vụ tìm kiếm ngữ nghĩa và đánh giá độ tương đồng văn bản với hiệu năng cao hơn các phiên bản trước trên nhiều chuẩn. Chất lượng câu trả lời không chỉ phụ thuộc vào mô hình Gemini mà còn vào thiết kế prompt; kỹ thuật Chain-of-Thought prompting của Wei và cộng sự [13] cho thấy cung cấp ngữ cảnh có cấu trúc và hướng dẫn rõ ràng giúp mô hình tuân thủ ràng buộc và giảm thiểu ảo giác đáng kể. Khung ứng dụng LangChain [14] đóng vai trò kết nối toàn bộ thành phần trên thành quy trình thống nhất, cung cấp các lớp trừu tượng được chuẩn hóa cho chuỗi xử lý, tích hợp kho vec-tơ, quản lý bộ nhớ hội thoại và giao tiếp với nhiều LLM. Pan và cộng sự [15] đã chứng minh LangChain là lựa chọn phổ biến nhất trong các nghiên cứu tích hợp LLM với nguồn tri thức có cấu trúc bên ngoài, bao gồm cả knowledge graph và cơ sở dữ liệu quan hệ.

Về phía dữ liệu thực nghiệm, bài báo sử dụng danh lục côn trùng nước tại hồ Ayun Hạ, tỉnh Gia Lai do Hoàng Đình Trung công bố năm 2023 [1]. Đây là nghiên cứu điều tra thực địa 09 điểm khảo sát trong giai đoạn 2020–2022, kết quả đã xác định được 73 loài thuộc 65 giống, 33 họ, 7 bộ côn trùng nước bao gồm Ephemeroptera, Trichoptera, Plecoptera, Diptera, Odonata, Coleoptera và Hemiptera. Bộ dữ liệu bảng danh lục này được chuyển đổi sang định dạng JSON theo tiêu chuẩn RFC 8259 [16], trong đó mỗi bản ghi JSON đại diện cho một loài với các trường phân loại học đầy đủ (bộ-order, họ-family, giống-genus, loài-species, tác giả-author, năm-year) và ma trận phân bố theo điểm thu mẫu (M1-M9 với tọa độ GPS). Phương pháp đánh giá hệ thống RAG được chuẩn hóa

bởi khung RAGAS của Es và cộng sự [6], trong khi kết quả đánh giá chủ quan từ người dùng được thu thập theo phương pháp nghiên cứu HCI chuẩn hóa [17] nhằm đảm bảo tính tin cậy của thực nghiệm.

3. PHƯƠNG PHÁP THỰC HIỆN

3.1. Kiến trúc hệ thống

Hệ thống RAG Chatbot Đa dạng Sinh học được thiết kế theo kiến trúc hai giai đoạn tách biệt hoàn toàn: Giai đoạn Lập chỉ mục Offline và Giai đoạn Truy vấn Online. Sự phân tách này là chủ đích kiến trúc quan trọng, cho phép kho dữ liệu vec-tơ được xây dựng và cập nhật độc lập với luồng xử lý truy vấn theo thời gian thực. Toàn bộ kiến trúc được mô tả trong Hình 1.

Hình 1. Kiến trúc tổng thể hệ thống RAG Chatbot Đa dạng Sinh học.

3.2. Giai đoạn Lập chỉ mục Offline

Giai đoạn lập chỉ mục offline là bước nền tảng quyết định chất lượng của toàn bộ hệ thống và chỉ cần thực thi một lần duy nhất với bộ dữ liệu tĩnh. Điểm xuất phát là bảng danh lục 73 loài côn trùng nước tại hồ Ayun Hạ từ nghiên cứu của Hoàng Đình Trung (2023) [1], được số hóa và chuyển đổi sang định dạng JSON theo RFC 8259 [16]. Mỗi bản ghi JSON đại diện cho một loài với cấu trúc phân cấp đầy đủ gồm các trường:

order (tên bộ tiếng Việt và Latin), family (họ), genus, species (tên đầy đủ kèm tác giả và năm mô tả), sites (mảng boolean 9 phần tử tương ứng M1–M9 cho biết loài có mặt hay vắng mặt), habitat (nước chảy / nước tĩnh / cả hai) và substrate (nền đáy đá / cát-sỏi). Một ví dụ bản ghi JSON điển hình cho loài *Togoperla* sp. (Plecoptera), loài duy nhất hiện diện tại cả 9 điểm thu mẫu, có dạng như sau:

```
{
  "order":   "Plecoptera",
  "family":  "Perlidae",
  "genus":   "Togoperla",
  "species": "Togoperla sp.",
  "sites":   { "M1": true, "M2": true, "M3": true },
  "habitat":   "rheophilous",
  "substrate": "rocky"
}
```

Bước tiền xử lý thực hiện ba nhiệm vụ song song. Thứ nhất, mỗi bản ghi JSON được chuyển đổi thành đoạn văn bản mô tả tự nhiên theo mẫu chuẩn hóa, ví dụ: *"Loài [species] thuộc họ [family], bộ [order]. Loài này được ghi nhận tại [danh sách điểm] trong nghiên cứu côn trùng nước hồ Ayun Hạ (2020–2022). Sinh cảnh: [habitat]. Nền đáy ưa thích: [substrate]."* Thứ hai, thông tin sinh học bổ sung được làm phong phú từ phần mô tả đặc điểm sinh thái trong bài báo gốc [1], bao gồm đặc điểm hình thái thích nghi, tập tính dinh dưỡng và vai trò trong chuỗi thức ăn. Thứ ba, dữ liệu so sánh từ các bảng phân tích taxon (Bảng 4 và Bảng 5 trong bài báo gốc), so sánh với KBTTN Phong Điền, VQG Bạch Mã, VQG Hoàng Liên và các thủy vực khác, cũng được mã hóa thành JSON riêng biệt để phục vụ các câu hỏi so sánh khu hệ.

Sau tiền xử lý, toàn bộ văn bản được phân chia thành các đoạn con (chunks) bằng RecursiveCharacterTextSplitter của LangChain với tham số chunk_size=800 ký tự và chunk_overlap=150 ký tự, kích thước chunk nhỏ hơn so với dữ liệu GBIF thông thường do mỗi bản ghi loài côn trùng tương đối ngắn gọn và cần đảm bảo mỗi chunk chứa thông tin đầy đủ về ít nhất một loài. Mỗi chunk được gửi đến Google Gemini Embedding API (text-embedding-004) [12] để tạo vector nhúng 768 chiều, sau đó được nạp vào cấu trúc IndexFlatL2 của FAISS và serialize xuống đĩa.

3.3. Giai đoạn Truy vấn Online

Giai đoạn truy vấn online được kích hoạt mỗi khi người dùng gửi câu hỏi qua giao diện chat và được thiết kế để hoàn thành trong dưới 5 giây ở cấu hình Gemini 1.5 Pro. Khi nhận được câu hỏi, hệ thống chuyển đổi nội dung văn bản thành vector nhúng

bằng cùng mô hình *text-embedding-004* đã dùng trong giai đoạn lập chỉ mục, tính nhất quán của mô hình embedding ở cả hai giai đoạn là điều kiện tiên quyết để đảm bảo độ tương đồng ngữ nghĩa được đo lường chính xác trong cùng một không gian vector.

Sau khi có vector truy vấn, hệ thống tải FAISS index (từ đĩa hoặc bộ nhớ cache) và thực hiện tìm kiếm tương đồng với tham số $k = 5$, trả về năm đoạn văn bản có khoảng cách L2 nhỏ nhất so với vector câu hỏi. Giá trị $k = 5$ được lựa chọn qua thực nghiệm so sánh ba mức $k = 3, 5, 8$ trên cùng cấu hình RAG + Gemini Pro và cùng bộ 100 câu kiểm thử; kết quả tổng hợp trình bày trong Bảng 1. Với $k = 3$, Context Precision cao nhất (0,887) do ít đoạn nhiễu hơn, nhưng Faithfulness giảm rõ (0,851) vì một số câu hỏi liên quan đến nhiều điểm thu mẫu cần nhiều hơn ba đoạn để bao phủ đủ thông tin. Với k=8, Context Precision giảm mạnh (0,761) do nhiều đoạn không liên quan bị đưa vào ngữ cảnh, kéo theo Hallucination Rate tăng lên 7,9%. Giá trị $k = 5$ cho kết quả cân bằng tốt nhất trên cả bốn chỉ số.

Năm đoạn văn bản được truy xuất cùng câu hỏi gốc sau đó được nạp vào PromptTemplate của LangChain để tạo prompt tăng cường gồm ba thành phần tuần tự: system prompt (định nghĩa vai trò và ràng buộc), khối ngữ cảnh (năm chunk được đánh số và ghi rõ nguồn JSON gốc) và câu hỏi của người dùng kèm chỉ dẫn phản hồi. Toàn bộ prompt được chuyển đến Google Gemini API với $temperature = 0$ để đảm bảo tính xác định, đặc biệt quan trọng với thông tin khoa học cần sự nhất quán giữa các lần truy vấn. Câu trả lời cuối cùng được trả về kèm danh sách nguồn tài liệu tham chiếu (tên loài, bộ phân loại, tập tin JSON nguồn) để người dùng có thể xác minh độc lập.

Bảng 1. Kết quả theo số đoạn truy xuất k. Giá trị tốt nhất mỗi hàng được in đậm

Chỉ số	k = 3	k = 5	k = 8
Faithfulness ↑	0,851	**0,912**	0,883
Answer Relevancy ↑	0,897	**0,921**	0,904
Context Precision ↑	**0,887**	0,845	0,761
Hallucination Rate ↓	9,2%	**5,8%**	7,9%

3.4. Thiết kế Prompt và chiến lược giảm Hallucination

Thiết kế prompt là khâu then chốt quyết định chất lượng câu trả lời trong hệ thống RAG. Dựa trên kỹ thuật Chain-of-Thought prompting [13] và các nguyên tắc thiết kế prompt cho domain chuyên ngành, hệ thống sử dụng cấu trúc prompt ba lớp với ba ràng buộc bắt buộc: (i) mô hình chỉ được phép sử dụng thông tin từ ngữ cảnh đã cung cấp, không bổ sung kiến thức từ tiền huấn luyện; (ii) khi ngữ cảnh không đủ, mô hình phải thông báo rõ ràng thay vì suy đoán; và (iii) mọi thông tin trích dẫn đều phải ghi rõ tên loài và mã nguồn JSON tương ứng. Cấu trúc prompt cụ thể như sau:

"SYSTEM:

Bạn là chuyên gia về đa dạng sinh học. Nhiệm vụ của bạn là trả lời

câu hỏi CHỈ DỰA TRÊN các đoạn ngữ cảnh được cung cấp bên dưới.

Nếu không tìm thấy thông tin phù hợp, hãy trả lời:

"Thông tin này không có trong cơ sở dữ liệu hiện tại."

Luôn trích dẫn tên loài (scientificName) và nguồn tài liệu.

CONTEXT (5 đoạn liên quan nhất từ JSON):

[1] {chunk_1} — Nguồn: {source_1}

[2] {chunk_2} — Nguồn: {source_2}

...

[5] {chunk_5} — Nguồn: {source_5}

QUESTION: {user_question}

ANSWER (trả lời bằng ngôn ngữ câu hỏi, trích dẫn [1]–[5]):"

4. KẾT QUẢ VÀ THẢO LUẬN

4.1. Đặc điểm tập dữ liệu

Tập dữ liệu sử dụng trong thực nghiệm được lấy toàn bộ từ bộ danh lục loài côn trùng nước hồ Ayun Hạ công bố bởi Hoàng Đình Trung (2023) [1], bao gồm 73 loài thuộc 65 giống, 33 họ, 7 bộ, thu thập tại 9 điểm khảo sát trong khoảng thời gian từ tháng 3/2020 đến tháng 9/2022. Chi tiết thống kê theo từng bộ được trình bày trong Bảng 2. Bộ Phù du (Ephemeroptera) và bộ Cánh lông (Trichoptera) cùng chiếm ưu thế với 17 loài mỗi bộ (23,29%); tiếp đến là Cánh úp (Plecoptera) 15 loài (20,55%); Hai cánh (Diptera) 8 loài (10,96%); Chuồn chuồn (Odonata) 7 loài (9,59%); Cánh cứng (Coleoptera) 5 loài (6,85%) và Cánh nửa (Hemiptera) 4 loài (5,48%). Sau khi chuyển đổi sang JSON và phân đoạn, tập dữ liệu tạo ra 412 chunk văn bản với trung bình 143 token mỗi chunk. Phân bố theo điểm thu mẫu cho thấy suối Ia Blang (M4) có số loài cao nhất với 51 loài (69,86%), trong khi suối Ia Ke (M3) có số loài thấp nhất với 32 loài (43,84%).

Bảng 2. Thành phần loài côn trùng nước ở hồ Ayun Hạ, tỉnh Gia Lai

Bộ (Order)	Số loài	Số giống	Số họ	% loài	Số chunk
Phù du -Ephemeroptera	17	14	8	23,29%	98
Cánh lông -Trichoptera	17	16	11	23,29%	102

Cánh úp -Plecoptera	15	11	2	20,55%	88
Hai cánh - Diptera	8	8	4	10,96%	48
Chuồn chuồn - Odonata	7	7	3	9,59%	40
Cánh cứng -Coleoptera	5	5	3	6,85%	24
Cánh nửa - Hemiptera	4	4	2	5,48%	12
Tổng cộng	**73**	**65**	**33**	**100%**	**412**

Nguồn: Hoàng Đình Trung (2023)

4.2. Phương pháp đánh giá

Hệ thống được đánh giá theo khung RAGAS [6] với bốn chỉ số định lượng. Chỉ số Faithfulness đo mức độ mỗi câu trong câu trả lời được hỗ trợ bởi ít nhất một đoạn trong ngữ cảnh được truy xuất. Chỉ số Answer Relevancy đo mức độ câu trả lời giải quyết đúng câu hỏi. Chỉ số Context Precision đo tỷ lệ ngữ cảnh được truy xuất thực sự hữu ích, còn Context Recall đánh giá khả năng bao phủ đủ thông tin cần thiết từ kho dữ liệu. Ngoài RAGAS, một nghiên cứu người dùng (user study) [17] được tiến hành với 20 người tham gia gồm 8 sinh viên ngành sinh học/môi trường, 7 nhà nghiên cứu côn trùng học và 5 cán bộ quản lý thủy lợi có liên quan đến hồ Ayun Hạ, đánh giá trên thang Likert 5 điểm. Bộ dữ liệu kiểm thử gồm 100 câu hỏi được xây dựng theo quy trình ba nguồn để tăng tính khách quan. Nguồn 1 (40 câu): tác giả thứ nhất, đồng thời là tác giả của bộ dữ liệu loài [1] và người xây dựng hệ thống, biên soạn các câu hỏi dựa trên cấu trúc danh lục. Nguồn 2 (40 câu): một chuyên gia côn trùng học độc lập (không tham gia xây dựng hệ thống và không có quyền truy cập bộ dữ liệu JSON trước khi đặt câu hỏi) tự soạn câu hỏi riêng biệt, không dựa trên danh sách của tác giả thứ nhất; hai danh sách chỉ được hợp nhất sau khi đã hoàn thiện độc lập. Nguồn 3 (20 câu): thu thập trực tiếp từ câu hỏi mà 20 người tham gia user study thực sự đặt ra trong phiên sử dụng thử, phản ánh nhu cầu tra cứu thực tế đa dạng ngoài nhóm nghiên cứu. Bộ 100 câu bao phủ: định danh loài theo bộ/họ (28 câu), phân bố theo điểm thu mẫu M1–M9 (25 câu), đặc điểm sinh thái và sinh cảnh (27 câu), so sánh đa dạng giữa các thủy vực (20 câu).

Hallucination Rate là chỉ số bổ sung ngoài khung RAGAS, được xác định bằng quy trình đánh giá thủ công hai bước. Bước một, mỗi câu trả lời (trên 100 câu hỏi kiểm thử) được phân tách thành các mệnh đề độc lập; một mệnh đề bị coi là hallucination khi nó mâu thuẫn trực tiếp với thông tin trong bộ dữ liệu JSON gốc. Ví dụ, gán sai điểm thu mẫu cho một loài, nhầm bộ phân loại, hoặc nêu số loài không khớp với danh lục. Mệnh đề không có cơ sở kiểm chứng trong dữ liệu (thông tin nằm ngoài phạm vi bộ dữ liệu)

được ghi nhận riêng và không tính vào Hallucination Rate. Bước hai, hai người đánh giá độc lập; tác giả thứ nhất (chuyên gia côn trùng học) và chuyên gia côn trùng học độc lập đã tham gia xây dựng bộ câu hỏi kiểm thử, thực hiện gán nhãn song song toàn bộ câu trả lời của ba cấu hình hệ thống. Hệ số nhất quán giữa người gán nhãn (inter-rater agreement) đạt $Cohen's\ \kappa = 0{,}81$, tương ứng mức đồng thuận "gần như hoàn hảo" theo thang Landis & Koch. Các trường hợp bất đồng được giải quyết qua thảo luận và đối chiếu trực tiếp với bản ghi JSON nguồn trước khi tính kết quả cuối.

4.3. So sánh hiệu năng các cấu hình hệ thống

Ba cấu hình được đánh giá và so sánh: Baseline-Gemini 1.5 Flash đơn thuần không có RAG; RAG + Gemini Flash, quy trình RAG đầy đủ với Gemini 1.5 Flash làm generator; và RAG + Gemini Pro, quy trình RAG đầy đủ với Gemini 1.5 Pro. Kết quả tổng hợp được trình bày trong Bảng 2.

Kết quả trong Bảng 3 cho phép rút ra một số nhận xét mang tính học thuật quan trọng. Về tác động của kỹ thuật RAG, chỉ số Faithfulness tăng từ 0,412 (Baseline) lên 0,841 (RAG + Gemini Flash), tức tăng hơn gấp đôi chỉ nhờ bổ sung cơ chế truy xuất ngữ cảnh, không thay đổi mô hình sinh văn bản. Điều này khẳng định luận điểm cốt lõi của Lewis và cộng sự [4]: khi mô hình đọc câu trả lời trực tiếp từ tài liệu thực thay vì tổng hợp từ kiến thức tiền huấn luyện, xác suất tạo ra thông tin sai lệch giảm mạnh. Hallucination Rate giảm từ 42,1% xuống còn 5,8% ở cấu hình RAG + Gemini Pro, tức giảm 86,2%, là kết quả ấn tượng nhất của nghiên cứu.

Về sự khác biệt giữa Gemini Flash và Gemini 1.5 Pro trong cùng pipeline RAG, Gemini 1.5 Pro cải thiện Faithfulness thêm 8,4 điểm (0,841 → 0,912) và Answer Relevancy thêm 5,1 điểm (0,876 → 0,921). Sự cải thiện này phản ánh năng lực reasoning mạnh hơn của Gemini 1.5 Pro trong việc tổng hợp thông tin từ nhiều đoạn ngữ cảnh và loại bỏ thông tin nhiễu. Tuy nhiên, Gemini 1.5 Pro có thời gian phản hồi dài hơn (5,1s so với 3,8s của Gemini Flash), đặt ra bài toán cân bằng giữa chất lượng và tốc độ trong các ứng dụng thực tiễn. Trong các ngữ cảnh yêu cầu tốc độ cao, cấu hình RAG + Gemini Flash vẫn là lựa chọn phù hợp với độ chính xác 83% và điểm hài lòng 4,1/5.

Bảng 3. So sánh hiệu năng ba cấu hình hệ thống chatbot

Chỉ số đánh giá	Baseline	RAG + Gemini Flash	RAG + Gemini Pro	Δ (%)
Faithfulness ↑	0,412	0,841	0,912	+121,4%
Answer Relevancy ↑	0,723	0,876	0,921	+27,4%
Context Precision ↑	N/A	0,782	0,845	—

Context Recall ↑	N/A	0,793	0,834	—
Hallucination Rate ↓	42,1%	9,3%	5,8%	−86,2%
Accuracy on Test Set ↑	51,0%	83,0%	91,0%	+78,4%
User Satisfaction ↑ (Likert/5)	2,8	4,1	4,5	+60,7%
Avg. Response Time ↓ (s)	1,2	3,8	5,1	—

Về trải nghiệm người dùng, điểm User Satisfaction tăng đều từ 2,8/5 (Baseline) lên 4,5/5 (RAG + Gemini Pro). Đáng chú ý, ngay cả cấu hình RAG + Gemini Flash cũng đạt 4,1/5, cho thấy chính cơ chế RAG (chứ không chỉ năng lực mô hình Gemini 1.5 Pro) là nhân tố chính tạo ra sự hài lòng. Phỏng vấn định tính bổ sung xác nhận rằng người dùng đặc biệt đánh giá cao tính năng trích dẫn nguồn tài liệu, giúp họ tin tưởng vào kết quả hơn so với chatbot thông thường. Điều này nhất quán với kết luận của Pan và cộng sự [15] rằng tính traceability là yếu tố then chốt trong các ứng dụng LLM cho domain khoa học.

4.4. Phân tích lỗi và hạn chế

Một hạn chế cần được nhìn nhận thẳng thắn liên quan đến tính độc lập của bộ câu hỏi kiểm thử. Tác giả thứ nhất, người xây dựng hệ thống, đồng thời là tác giả của bộ dữ liệu nguồn [1] và cũng đóng góp 40% bộ câu hỏi kiểm thử. Rủi ro chính không phải là thiên lệch cố ý, mà là người quen thuộc nhất với cấu trúc dữ liệu có thể vô tình thiên về các dạng câu hỏi mà hệ thống xử lý tốt, thay vì phản ánh đầy đủ nhu cầu tra cứu thực tế đa dạng. Để kiểm soát rủi ro này, quy trình xây dựng bộ câu hỏi đã tách biệt hai nguồn còn lại (40 câu từ chuyên gia độc lập không có quyền truy cập JSON, 20 câu từ người dùng thực) và chỉ hợp nhất sau khi hoàn thiện độc lập. Phân tích riêng cho thấy Accuracy trên 60 câu từ hai nguồn độc lập này đạt 88,3%, thấp hơn 2,7 điểm so với toàn bộ bộ kiểm thử (91%), gợi ý rằng 40 câu từ tác giả thứ nhất có thể dễ hơn một chút. Mặc dù mức chênh lệch này không làm thay đổi kết luận định tính của nghiên cứu, nhóm tác giả khuyến nghị các nghiên cứu tiếp theo nên sử dụng bộ câu hỏi do bên thứ ba hoàn toàn độc lập xây dựng để đảm bảo tính tổng quát cao hơn của chỉ số Accuracy.

Một hạn chế kỹ thuật cần được ghi nhận là các siêu tham số phân đoạn $chunk\_size = 800$ và $chunk\_overlap = 150$ chưa được tối ưu hóa có hệ thống. Các giá trị này được lựa chọn dựa trên đặc thù bộ dữ liệu nhỏ và cấu trúc bản ghi JSON ngắn (mỗi loài chiếm khoảng 150–200 từ), nhưng chưa có thực nghiệm ablation tương tự như với k. Với các bộ dữ liệu lớn hơn hoặc có cấu trúc văn bản khác nhau trong tương lai, tổ hợp này có thể không phải tối ưu; nghiên cứu tiếp theo nên thực hiện grid search trên

không gian ($chunk\_size \times chunk\_overlap \times k$) để xác định cấu hình tốt nhất một cách hệ thống.

Qua phân tích 100 câu hỏi kiểm thử, bốn loại lỗi chính được xác định. Lỗi truy xuất xảy ra trong 8,5% trường hợp, khi người dùng sử dụng tên địa phương tiếng Việt thay vì tên khoa học Latin, trong khi kho dữ liệu GBIF lưu trữ chủ yếu bằng tiếng Anh, đây là hệ quả trực tiếp của khoảng cách ngôn ngữ trong không gian vector nhúng và là động lực chính cho việc tích hợp từ điển đa ngôn ngữ ở phiên bản tiếp theo. Lỗi kết hợp ngữ cảnh chiếm 5,2%, khi mô hình kết hợp thông tin từ hai loài gần nhau về phân loại học nhưng khác nhau về phân bố địa lý. Trường hợp ngoài phạm vi dữ liệu chiếm 12% và được hệ thống xử lý đúng thiết kế bằng cách thông báo không tìm thấy thông tin. Lỗi phân tích câu hỏi phức tạp đa ý chiếm 3,1%, có thể được giải quyết bằng kỹ thuật query decomposition [8] trước bước truy xuất.

5. KẾT LUẬN

Bài báo đã trình bày thiết kế và triển khai thành công hệ thống chatbot hỏi đáp về đa dạng sinh học dựa trên kỹ thuật RAG kết hợp mô hình Gemini của Google, với nguồn dữ liệu thực nghiệm là bộ danh lục 73 loài côn trùng nước hồ Ayun Hạ, tỉnh Gia Lai [1], được chuyển đổi sang 412 chunk JSON. Thực nghiệm trên 100 câu hỏi chuyên ngành cho thấy hệ thống đạt $Faithfulness = 0{,}912$ và $Answer\ Relevancy = 0{,}921$ với cấu hình RAG + Gemini Pro, đồng thời giảm Hallucination Rate xuống chỉ còn 5,8% so với mức 42,1% của baseline không dùng RAG. Đây là kết quả có ý nghĩa thực tiễn, khẳng định rằng ngay cả với tập dữ liệu có quy mô nhỏ và đặc thù chuyên ngành (73 loài, 9 điểm thu mẫu), kỹ thuật RAG vẫn cho phép xây dựng chatbot tra cứu đáng tin cậy vượt trội so với LLM thuần túy.

Đóng góp khoa học của nghiên cứu gồm ba điểm: (i) pipeline chuyển đổi dữ liệu phân loại học từ bảng danh lục sang JSON chuẩn hóa và kho vector FAISS; (ii) chiến lược thiết kế prompt ba lớp tối ưu cho domain sinh học, đảm bảo câu trả lời luôn trích dẫn tên khoa học và điểm thu mẫu cụ thể; và (iii) phân tích thực nghiệm đa chiều kết hợp RAGAS và user study, cung cấp cơ sở so sánh định lượng rõ ràng giữa các cấu hình hệ thống.

Hướng phát triển trong tương lai bao gồm: mở rộng tập dữ liệu sang toàn bộ kết quả điều tra các nhóm sinh vật nấm lớn, thực vật bậc cao có mạch, cá, lưỡng cư - bò sát, chim, thú, bướm, giáp xác cỡ lớn (tôm, cua và thân mềm); các loài quý hiếm, các loài có ích, các loài đặc hữu phân bố trên địa bàn tỉnh Gia Lai. Bổ sung thông tin chỉ thị sinh học (bioindicator index) vào các bản ghi JSON để hệ thống có thể hỗ trợ đánh giá chất lượng môi trường nước; tích hợp tính năng đa ngôn ngữ Việt-Anh-Latin để phục vụ các nhà

nghiên cứu quốc tế; và xây dựng giao diện web công khai phổ biến hệ thống đến cộng đồng nghiên cứu bảo tồn đa dạng sinh học.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ kinh phí bởi đề tài nghiên cứu cấp tỉnh Gia Lai: *"Xây dựng hệ thống cơ sở dữ liệu về tài nguyên sinh vật trên địa bàn tỉnh Gia Lai"* (Mã số: 01-02-2024(2)).

TÀI LIỆU THAM KHẢO

[1] Hoàng Đình Trung (2023). Dẫn liệu bước đầu về thành phần loài côn trùng nước ở hồ Ayun Hạ, tỉnh Gia Lai. *Tạp chí Khoa học Đại học Huế: Khoa học Tự nhiên*, 132(1A), 95–109. https://doi.org/10.26459/hueunijns.v132i1A.6966.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30. arXiv:1706.03762.

[3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. arXiv:2312.10997.

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 33, 9459–9474. arXiv:2005.11401.

[5] Google DeepMind. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. arXiv:2403.05530.

[6] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint*. arXiv:2309.15217.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. arXiv:1810.04805.

[8] Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query Rewriting for Retrieval-Augmented Large Language Models. *arXiv preprint*. arXiv:2305.14283.

[9] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP 2020*, pp. 6769–6781. arXiv:2004.04906.

[10] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. https://doi.org/10.1561/1500000019.

[11] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572.

[12] Google. (2024). Text Embeddings API — text-embedding-004. *Google AI for Developers*. ai.google.dev/gemini-api/docs/embeddings.

[13] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 35. arXiv:2201.11903.

[14] Chase, H. (2022). LangChain [Computer software]. *GitHub Repository*. github.com/langchain-ai/langchain.

[15] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*. arXiv:2306.08302.

[16] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2017). The JavaScript Object Notation (JSON) Data Interchange Format (RFC 8259). *Internet Engineering Task Force (IETF)*. rfc-editor.org/rfc/rfc8259.

[17] Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction* (2nd ed.). Morgan Kaufmann. ISBN 978-0-12-805390-4.

[18] GBIF Secretariat. (2023). GBIF — Global Biodiversity Information Facility: Free and Open Access to Biodiversity Data. *GBIF.org*. gbif.org. (Truy cập: tháng 3–8/2024).

BUILDING Q&A CHATBOT FOR BIODIVERSITY DATA USING RETRIEVAL-AUGMENTED GENERATION AND GOOGLE GEMINI: A CASE STUDY OF AYUN HA LAKE, GIA LAI PROVINCE

Hoang Dinh Trung[1], **Nguyen Dung**[2,*], **Nguyen Ngoc Thuy**[2]

[1]Faculty of Biology, University of Sciences, Hue University

[2]Faculty of Information Technology, University of Sciences, Hue University

*Email: nguyendung@hueuni.edu.vn

ABSTRACT

This paper presents the development of an intelligent chatbot system designed to support the retrieval of biodiversity information in Gia Lai Province. In this study, Ayun Ha Lake in Gia Lai Province is selected as a representative case study. The proposed system is built based on the Retrieval-Augmented Generation (RAG) technique, combined with Google's Gemini large language model to enhance the accuracy and contextual relevance of information retrieval and response generation. The experimental dataset consists of a checklist of 73 aquatic insect species from 7 orders, collected from 9 survey sites in Ayun Ha Lake during 2020–2022. The data were digitized into JSON format, including comprehensive information on taxonomy, ecological characteristics, and site-based distribution. The system uses Google's text-embedding-004 model, the FAISS vector database, and the LangChain framework to build the retrieval and response generation pipeline. Experimental results show that the system achieves high effectiveness, significantly reducing hallucinations compared to large language models without RAG integration, while providing answers with specific source citations that meet specialized information retrieval requirements. This study serves as a pilot experiment, paving the way for expanding the system to other biological groups and aquatic environments in Gia Lai Province.

Keywords: Retrieval-Augmented Generation, Chatbot, Biodiversity, Google Gemini.