Xây dựng Trợ lý Hỏi Đáp Sử dụng Retrieval Augmented Generation (RAG)

Man Ngo, 2024

Tóm tắt

Dự án này nhằm phát triển một chatbot trả lời các câu hỏi của khách hàng dựa trên cơ sở dữ liệu từ FAQ của website VCB. Tôi đã áp dụng các kỹ thuật tiên tiến nhằm tăng khả năng của RAG như:

- SELF-REFINE【2】giúp các mô hình ngôn ngữ lớn (LLMs) trả lời có trích dẫn trong từng đoạn nhỏ và phù hợp.
- Multi-hop reasoning【1】giải quyết các câu hỏi phức tạp đòi hỏi tổng hợp dẫn chứng từ nhiều truy vấn.

I. Tại sao chọn RAG cho chatbot ngân hàng?

Đặc thù của bài toán:

1. Trả lời câu hỏi trong lĩnh vực tài chính ngân hàng phải chính xác và có trích dẫn, hạn chế tối đa hiện tượng "hallucination".
2. Thông tin thường xuyên thay đổi, đòi hỏi mô hình dễ dàng cập nhật thông tin mới.

Các hướng đi khả thi:

1. Fine-tune với dữ liệu nhân tạo bằng phương pháp Evol-Instruct【3】:
   - Phát triển bộ dữ liệu gốc từ FAQ của VCB thành bộ dữ liệu lớn hơn và đa dạng hơn.
   - Sử dụng bộ dữ liệu này để fine-tune mô hình LLAMA 3.

   Khuyết điểm:
   - Khó trích dẫn và kiểm soát hallucination.
   - Khó cập nhật thông tin mới mà không phải thực hiện lại các bước trên.

2. RAG: gồm 2 bước:
   - Retrieval: Xác định các phần liên quan trong cơ sở dữ liệu FAQ và tài liệu hướng dẫn của ngân hàng, so sánh câu hỏi của người dùng với tài liệu được lập chỉ mục để tìm ra đoạn văn bản chứa câu trả lời.
   - Generation: Từ các văn bản được truy vấn, sử dụng mô hình ngôn ngữ để tạo câu trả lời tự nhiên cho khách hàng.

   Ưu điểm:
   - Tiện lợi cho việc trích dẫn và kiểm soát hallucination.
   - Dễ dàng cập nhật thông tin mới.

   Khuyết điểm:

   Tốc độ sẽ chậm hơn mô hình được finetune theo hướng 1.

Dựa trên phân tích này cũng như thời gian cho phép làm project, tôi quyết định sử dụng RAG để giải quyết vấn đề.

II. Các kỹ thuật cải tiến RAG được sử dụng:

1. Multi-hop reasoning:

- Giải quyết các câu hỏi phức tạp bằng cách liên kết nhiều tài liệu đến từ nhiều truy vấn.
- Tôi sử dụng kỹ thuật trong bài báo [1] để bổ sung thêm các câu hỏi nhằm đào sâu thêm vấn đề nhằm truy vấn đủ tài liệu liên quan cho việc trả lời câu hỏi.

2. SELF-REFINE:

- RAG tự động liệt kê các tài liệu liên quan đến câu hỏi, tổng hợp lại để đưa ra câu trả lời.
- Tuy nhiên, có hai vấn đề chính:
  1. Một số tài liệu không được sử dụng trong câu trả lời.
  2. Không biết phần nào trong câu trả lời liên kết với tài liệu nào, dễ dẫn đến hallucination.

Giải pháp:

- Áp dụng kỹ thuật SELF-REFINE【2】:

  Bằng cách tạo các hàm đánh giá các tiêu chí của câu trả lời như:

  - Ở mỗi 1-2 câu trong câu trả lời có trích dẫn hay không vd: text … [1]. Text …[2]
  - Các đoạn có trích dẫn có thật sự liên quan đến tài liệu được trích dẫn hay không
  - Câu trả lời cuối cùng có liên quan đến câu trả lời hay không (Phần này tôi đã chuẩn bị code, do chưa test kỹ nên đã ẩn). Hay gặp ở các câu hói tiếp tục các câu hỏi trước đó, hoặc các câu không phải câu hỏi.

  Các hàm này đều được xây dựng cơ bản với chain-of-thought (CoT) prompt. Các hàm này được dùng như tiêu chí để SELF-REFINE gợi ý việc thay đổi câu trả lời sao cho thỏa mãn các tiêu chí trên, việc này được lặp lại cho tới khi thõa mãn hoàn toàn các tiêu chí.

III. Chi tiết kỹ thuật:

Tôi gửi kèm là code RAG, app bằng Streamlit, kèm docker.

Quy trình liên quan data:

1. Thu thập dữ liệu:
   - Crawling FAQ từ website ngân hàng.
   - Tài liệu PDF khó parse nên tôi chưa sử dụng.

2. Query vào vector database:
   - Tạo hai vector space:

- Vector space 1: embedding vector của [câu hỏi], dùng câu hỏi làm key và value là câu hỏi + câu trả lời.
- Vector space 2: embedding vector của [câu hỏi + câu trả lời], vừa là key vừa là value.

- Lý do: Hai tập key này khác nhau nên không thể sử dụng chung khi query cho một câu hỏi.
- Như vậy 1 câu hỏi sẽ kéo ra 2 tập hợp tài liệu từ, 2 tập này được hợp lại (lọc trùng) rồi trả về.

Quy trình tổng quát:

- Câu hỏi được đưa vào module Multi-hop reasoning (2-hop) để có hai câu hỏi đào sâu thêm, tương ứng sẽ là ba lần query vào (2) vector db, các tài liệu sẽ được lọc trùng.
- Các tài liệu được đưa vào module SELF-REFINE cùng với câu hỏi, để đưa ra câu trả lời có dẫn chứng và được kiểm tra sự phù hợp của dẫn chứng.
- Nếu module SELF-REFINE sau 3 lần lặp vẫn chưa thỏa mãn các tiêu chí thì chuyển sang mô hình RAG đơn giản hơn (chỉ có MultiHop RAG module) nhưng có sử dụng thông tin cuộc trò chuyện ở 3 turn gần nhất (chat memory 3 turns).

Các công cụ được sử dụng:

- Mô hình embedding: text-embedding-ada-002 của OpenAI.
- Mô hình ngôn ngữ lớn: GPT-3.5-turbo của OpenAI.
- Vector database: ChromaDb.
- Framework mô hình ngôn ngữ: DsPY.

IV. Đánh giá

Do giới hạn thời gian, nên tôi chỉ có thể kiểm tra thủ công. Dưới đây là các đề xuất

- **Tập dữ liệu để đánh giá**
  - Thu thập truy vấn thực tế.
  - Kết hợp Synthetic data: Dùng phương pháp Evol-Instruct【3】để tạo thêm các query khó và đa dạng hơn
  - Câu trả lời được tạo ra từ các model mạnh như gpt-4-o và người kiểm tra lại.

- **Đánh giá riêng cho mô hình embedding (retrieve step)**
  - Tạo các cặp (query, doc) gần nhau (positive) và xa nhau (negative).

  Các cặp này được tạo ra bằng: sử dụng các mô hình lớn như gpt-4-o để tạo synthetic data + mining các cặp negative, parapharse lại bộ FAQ, data do người có kiến thức chuyên ngành tạo ra.
  - Đánh giá độ chính xác khi retrievetop 5, top 10

- **Đánh giá chung:**
  - Độ chính xác: Tỷ lệ phần trăm các câu hỏi được trả lời với phản hồi hoàn toàn đúng và phù hợp với ý định của người dùng
  - Độ chính xác của các trích dẫn.

V. Các cải tiến nên làm tiếp:

- **Tốc độ:**
  - Do đang sử dụng SELF-REFINE nên mô hình sẽ bao gồm nhiều bước lặp dẫn đến kéo dài thời gian trả lời, có khi lên đến 60s.
  - Nên có model quyết định có nên tiếp tục lặp hay không.
  - Caching câu query/answer theo semantic vector.

- **Fine-tuning embedding model:** finetune để mô hình thích nghi với ngôn ngữ/thuật ngữ riêng của ngành ngân hàng và ngôn ngữ nói tự nhiên của người dùng (bao gồm cả lỗi chính tả).

- **Finetune Instruction prompt and few-shot-example of LLM:** ta sẽ dựa vào tập test và áp dụng các kỹ thuật teleprompters [4]. Đây cũng chính là lý do tôi chọn LM framework là DsPY thay vì các framework khác như Langchain hay Llama-Index.

- **Function calling:** đưa chức năng gọi hàm nhằm tăng khả năng tính toán của llm, ví dụ: tính toán với ngày/tháng/thứ, tính tiền lời theo lãi suất,...

- **Cải thiện Knowledge Base:**
  - Chuẩn bị sẵn cho tình huống không có FAQ, chỉ có Document hướng dẫn chung, tình huống này ta có thể dùng các phương pháp như Self-instruct [5].
  - Cơ chế cho nhân viên sửa những trường hợp sai, hoặc manually đưa thêm thông tin.

Tài liệu tham khảo:

[1] Khattab, Omar, Christopher Potts, and Matei Zaharia. "Baleen: Robust multi-hop reasoning at scale via condensed retrieval." Advances in Neural Information Processing Systems 34 (2021): 27670-27682.

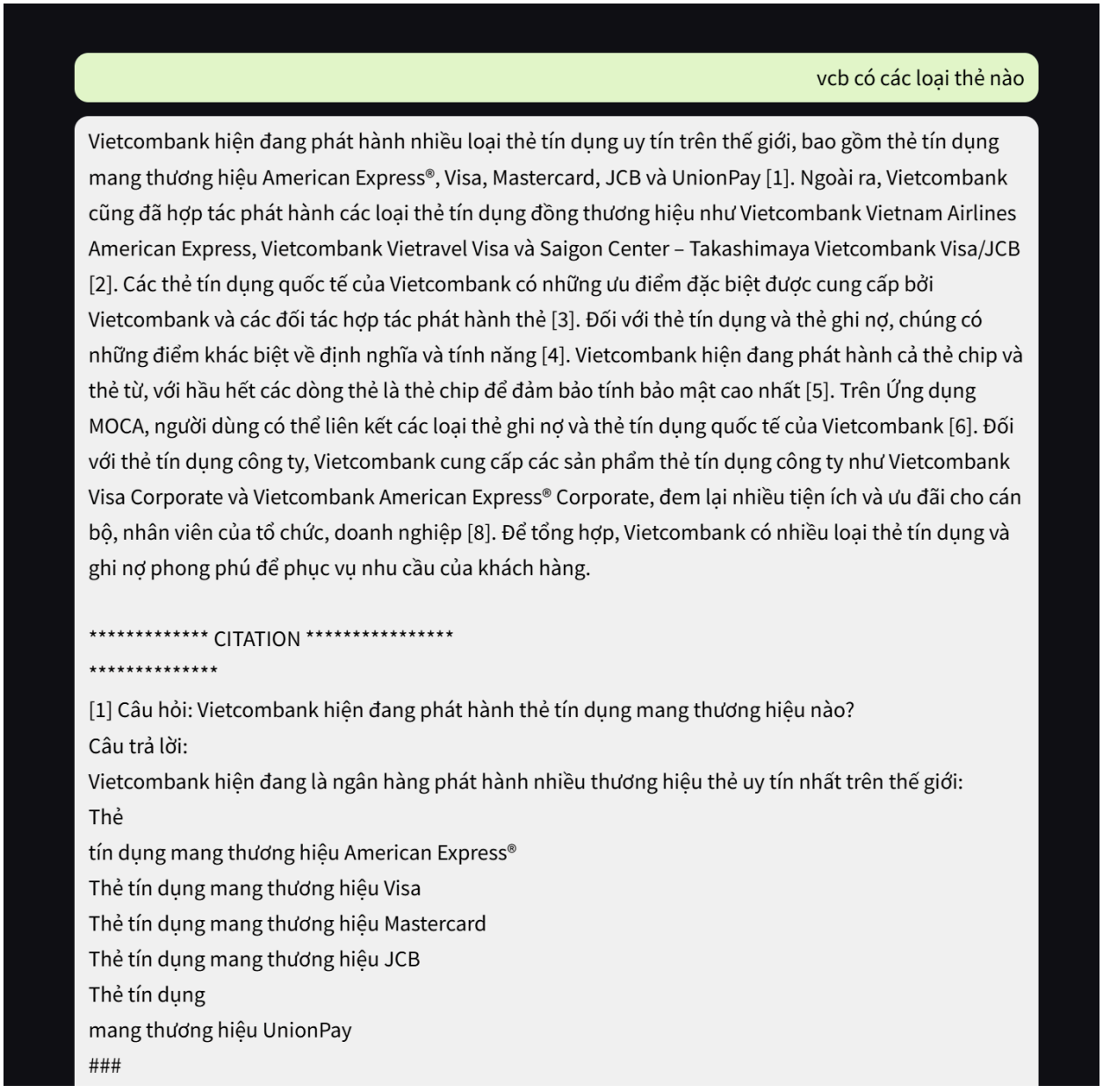
[2] Madaan, Aman, et al. "Self-refine: Iterative refinement with self-feedback." Advances in Neural Information Processing Systems 36 (2024).

[3] Xu, Can, et al. "Wizardlm: Empowering large language models to follow complex instructions." arXiv preprint arXiv:2304.12244 (2023).

[4] Khattab, Omar, et al. "DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines." *The Twelfth International Conference on Learning Representations*. 2023.

[5] Wang, Yizhong, et al. "Self-instruct: Aligning language models with self-generated instructions." *arXiv preprint arXiv:2212.10560* (2022).

Một vài ví dụ:



[2] Câu hỏi: Vietcombank đã phát hành thẻ tín dụng đồng thương hiệu nào?
Câu trả lời:
Hiện Vietcombank đã hợp tác phát hành thẻ tín dụng đồng thương hiệu sau:
Vietcombank Vietnam Airlines American Express: là thẻ đồng thương hiệu do Vietcombank hợp tác với VietnamAirlines. Trong đó, thẻ Vietcombank Vietnam Airlines Platinum American Express (hạng bạch kim) mang lại cho chủ thẻ nhiều ưu đãi hấp dẫn ngoài tính năng cơ bản của thẻ tín dụng đồng thương hiệu.
Vietcombank Vietravel Visa: là thẻ đồng thương hiệu do Vietcombank hợp tác với Vietravel. Chủ thẻ được hưởng những ưu đãi hấp dẫn từ Vietcombank và Vietravel.
Thẻ Saigon Center – Takashimaya Vietcombank Visa/JCB: là thẻ đồng thương hiệu do Vietcombank hợp tác với Takashimaya, chủ thẻ được tham gia chương trình tích lũy điểm thưởng để quy đổi thành thẻ trả trước sử dụng tại Trung tâm thương mại Saigon Centre với nhiều ưu đãi hấp dẫn.
###
****
****
[3] Câu hỏi: Thẻ tín dụng quốc tế nói chung của Vietcombank và thẻ tín dụng quốc tế Đồng thương hiệu Vietcombank có khác biệt gì?
Câu trả lời:
Thẻ tín dụng quốc tế Đồng thương hiệu Vietcombank ngoài các tính năng của sản phẩm thẻ tín dụng mang thương hiệu của TCTQT còn được áp dụng thêm những ưu đãi đặc biệt được cung cấp bởi Vietcombank và các đối tác hợp tác phát hành thẻ cùng Vietcombank.

###
****
****
[4] Câu hỏi: Thẻ tín dụng và thẻ ghi nợ của Vietcombank có điểm gì khác biệt?
Câu trả lời: 1. Định nghĩa
Thẻ
tín
dụng: Là thẻ được sử dụng để chi tiêu trước, trả tiền sau trong một hạn mức nhất định do ngân hàng cấp, với thời gian miễn lãi từ 45 đến 55 ngày tùy theo từng sản phẩm thẻ.

thẻ nào có hạn miễn lãi trên 50 ngày

Các thẻ tín dụng của Vietcombank có hạn miễn lãi trên 50 ngày bao gồm thẻ TDQT Vietcombank Visa Signature/ Vietcombank Mastercard World và thẻ tín dụng công ty Vietcombank Visa Corporate, Vietcombank American Express Corporate. Theo thông tin từ Vietcombank, thẻ TDQT Vietcombank Visa Signature/ Vietcombank Mastercard World có thời gian miễn lãi tối đa là 55 ngày, trong khi thẻ tín dụng công ty Vietcombank Visa Corporate, Vietcombank American Express Corporate có thời gian miễn lãi tối đa là 57 ngày [2]. Do đó, các thẻ này đều có hạn miễn lãi trên 50 ngày.

********** CITATION **********
****
[2] Câu hỏi: Ngân hàng quy định thời gian miễn lãi của các thẻ tín dụng như thế nào?
Câu trả lời: Đối với thẻ TDQT Vietcombank Visa Signature/ Vietcombank Mastercard World: thời gian miễn lãi tối đa 55 ngày.
Đối với thẻ TDQT cá nhân mang thương hiệu American Express: thời gian miễn lãi tối đa 50 ngày.
Đối với thẻ TDQT cá nhân mang thương hiệu Visa/Mastercard/JCB/UnionPay khác: thời gian miễn lãi tối đa 45 ngày.

Đối với thẻ tín dụng công ty Vietcombank Visa Corporate, Vietcombank American Express Corporate: thời gian miễn lãi tối đa 57 ngày.
###
****

quên mã pin

Khi quên mã PIN thẻ, có nhiều cách để khắc phục tình huống này. Trong trường hợp thẻ chưa bị khóa, khách hàng có thể tự chủ động cấp lại mật khẩu trên ứng dụng Digibank của Vietcombank [1]. Nếu thẻ đã bị khóa do nhập sai mã PIN quá 3 lần hoặc khách hàng không đăng ký sử dụng ứng dụng Digibank, quý khách cần mang chứng minh thư tới điểm giao dịch Vietcombank gần nhất để được hỗ trợ [2]. Để tránh tình huống khó khăn khi quên mã PIN, quý khách nên ghi nhớ và bảo mật mã PIN của mình [2]. Nếu có bất kỳ vấn đề nào khác liên quan đến thẻ, khách hàng có thể liên hệ tổng đài hoặc tới điểm giao dịch của Vietcombank để được giải quyết [3].

********** CITATION **********
****
[1] Câu hỏi: Tôi bị quên mã PIN thẻ của mình, tôi phải làm thế nào để được cấp lại PIN?
Câu trả lời: 1. Trường hợp thẻ của khách hàng chưa bị khóa:
Khách hàng có thể chủ động cấp lại mật khẩu thẻ trên ứng dụng Digibank → Dịch vụ thẻ → Dịch vụ thẻ khác → Tạo mới/Đổi mã PIN và lựa chọn thông tin.
2. Trường hợp thẻ của khách hàng đã bị khóa do nhập sai mã Pin 3 lần hoặc khách hàng không đăng ký sử dụng ứng dụng Digibank:
Quý khách vui lòng mang chứng minh thư tới điểm giao dịch VCB gần nhất để được phục vụ.
###
****
****
[2] Câu hỏi: Tôi phải làm gì khi bị quên mã PIN?
Câu trả lời: Quý khách nên ghi nhớ và bảo mật mã PIN của mình.
1. Trường hợp Quý khách quên mã Pin, Quý khách có thể cấp lại mã PIN trên ứng dụng VCB Digibank theo hướng dẫn
tại đây
.
2. Trường hợp khách hàng không đăng ký sử dụng ứng dụng Digibank: Quý khách vui lòng tới Điểm giao dịch của VCB gần nhất để được trợ giúp.
###