

TÀI LIỆU ĐẶC TẢ YÊU CẦU PHẦN MỀM (SOFTWARE REQUIREMENTS SPECIFICATION - SRS)

Dự án: Magic English – Ứng dụng Học tiếng Anh "Tất-cả-trong-một"

1. GIỚI THIỆU

1.1 Mục đích

Tài liệu Đặc tả Yêu cầu Phần mềm (SRS) này được xây dựng nhằm mục đích cung cấp một bản mô tả toàn diện, chi tiết và chính xác về các yêu cầu kỹ thuật cũng như nghiệp vụ cho dự án Magic English. Tài liệu này đóng vai trò là "nguồn sự thật duy nhất" (single source of truth) để định hướng cho toàn bộ vòng đời phát triển phần mềm (SDLC), từ giai đoạn thiết kế kiến trúc, lập trình, kiểm thử (QA/QC) đến giai đoạn triển khai và bảo trì.

Hệ thống Magic English được định nghĩa là một giải pháp ứng dụng học tập tích hợp "All-in-One", được thiết kế để giải quyết triệt để sự phân mảnh trong công cụ học tập của người tự học tiếng Anh. Mục đích cốt lõi của tài liệu này là làm rõ cách thức hệ thống sử dụng công nghệ Trí tuệ nhân tạo (AI) thông qua Ollama Cloud API để tự động hóa quy trình học tập, cung cấp phản hồi tức thì và duy trì động lực người dùng.

Tài liệu này không chỉ liệt kê các tính năng mà còn đi sâu phân tích hành vi của hệ thống trong các điều kiện biên, các kịch bản lỗi (exception handling), và các tiêu chuẩn phi chức năng về hiệu năng và bảo mật nhằm đảm bảo sản phẩm cuối cùng đạt chất lượng cao nhất.

1.2 Phạm vi sản phẩm

Magic English là một ứng dụng đa nền tảng (hỗ trợ Web và Mobile) phục vụ cho nhu cầu cá nhân, phi thương mại. Hệ thống được xây dựng xoay quanh ba trụ cột chức năng chính, tương ứng với ba vấn đề lớn nhất của người tự học:

1. Hệ thống Quản lý Từ vựng Thông minh:

- Giải quyết vấn đề nhập liệu thủ công tốn thời gian và dữ liệu rời rạc.
- Sử dụng AI để tự động trích xuất thông tin ngôn ngữ học (IPA, ngữ nghĩa, loại từ, ví dụ ngữ cảnh) từ một từ khóa đơn giản.
- Tổ chức lưu trữ có cấu trúc để phục vụ việc ôn tập và tra cứu.

2. Hệ thống Phân tích & Sửa lỗi:

- Giải quyết sự thiếu hụt phản hồi tức thì trong kỹ năng Viết.
- Cung cấp khả năng chấm điểm định lượng và định tính theo thời gian thực.
- Đề xuất cải thiện phong cách viết để văn bản tự nhiên hơn, vượt ra ngoài việc

chỉ sửa lỗi ngữ pháp cơ bản.

3. Hệ thống Theo dõi & Tạo động lực:

- Giải quyết tình trạng mất động lực và thiếu kỷ luật.
- Áp dụng cơ chế Gamification (Trò chơi hóa) thông qua việc theo dõi chuỗi ngày học liên tục và trực quan hóa dữ liệu tiến bộ cá nhân.

Phạm vi kỹ thuật của dự án bao gồm việc tích hợp sâu với Ollama Cloud API để tận dụng các mô hình ngôn ngữ lớn (LLMs) mã nguồn mở (như Llama 3, Mistral, hoặc GPT-OSS) nhằm xử lý các tác vụ ngôn ngữ phức tạp ngay trên nền tảng của người dùng hoặc thông qua cloud relay.

1.3 Đối tượng độc giả (Intended Audience)

Tài liệu này được biên soạn với cấu trúc chuyên sâu, phục vụ các nhóm đối tượng sau:

- Kiến trúc sư hệ thống (System Architects): Để thiết kế cấu trúc tích hợp API, mô hình dữ liệu cục bộ và chiến lược xử lý lỗi (retry logic, circuit breaker) khi làm việc với AI Service bên thứ ba.
- Nhóm phát triển (Developers - Frontend/Backend): Để nắm bắt chính xác logic nghiệp vụ, cấu trúc JSON Request/Response cần thiết khi giao tiếp với Ollama API, và các quy tắc giao diện người dùng (UI rules).
- Đội ngũ đảm bảo chất lượng (QA/Testers): Để xây dựng bộ Test Case chi tiết bao gồm cả kiểm thử chức năng và kiểm thử biên đối với các phản hồi không xác định của AI.
- Stakeholders (Giảng viên/Người dùng): Để xác nhận sự phù hợp giữa giải pháp công nghệ và nhu cầu thực tế của người học.

1.4 Định nghĩa và Thuật ngữ

Thuật ngữ / Viết tắt	Định nghĩa chi tiết
SRS	Software Requirements Specification - Đặc tả yêu cầu phần mềm.
LLM	Large Language Model - Mô hình ngôn ngữ lớn, công nghệ lõi đằng sau các tính năng thông minh của ứng dụng.
Ollama Cloud API	Dịch vụ cung cấp khả năng truy cập các mô hình LLM mã nguồn mở qua giao thức HTTP RESTful.

IPA	International Phonetic Alphabet - Bảng ký hiệu ngữ âm quốc tế, dùng để hiển thị cách phát âm chuẩn xác.
CEFR	Common European Framework of Reference for Languages - Khung tham chiếu trình độ ngôn ngữ chung của Châu Âu (A1-C2).
Streak	Chuỗi ngày liên tục mà người dùng thực hiện ít nhất một hoạt động học tập có ý nghĩa.
JSON Mode	Một chế độ hoạt động của Ollama API ép buộc mô hình trả về dữ liệu dưới định dạng JSON có cấu trúc thay vì văn bản tự do.
Hallucination	Hiện tượng AI tạo ra thông tin sai lệch hoặc không có thực nhưng được trình bày một cách thuyết phục.

2. MÔ TẢ TỔNG QUAN

2.1 Viễn cảnh sản phẩm

Magic English được định vị là một ứng dụng "Client-centric" nhưng phụ thuộc nặng vào dịch vụ "AI-as-a-Service". Sản phẩm không hoạt động độc lập mà là một phần của hệ sinh thái tích hợp.

- **Mô hình Kiến trúc:** Ứng dụng hoạt động như một giao diện người dùng thông minh. Logic xử lý ngôn ngữ tự nhiên không nằm trong mã nguồn ứng dụng mà được ủy quyền cho Ollama Cloud API. Điều này giúp ứng dụng nhẹ, nhưng tạo ra sự phụ thuộc mạng và tính sẵn sàng của dịch vụ bên thứ ba.
- **Quản lý Dữ liệu:** Khác với các hệ thống tập trung, Magic English ưu tiên lưu trữ dữ liệu cá nhân cục bộ hoặc trên cloud riêng của người dùng để đảm bảo quyền riêng tư và tốc độ truy xuất nhanh cho tính năng "Sổ tay từ vựng".

2.2 Đặc trưng người dùng

Dựa trên tài liệu đề tài, hệ thống tập trung phục vụ một chân dung người dùng cụ thể: Người Tự học Tiếng Anh có Kỷ luật.

Việc phân tích sâu đặc trưng của nhóm này là tiền đề cho các yêu cầu chức năng:

- **Hành vi học tập:** Họ học theo kiểu "Micro-learning" (học nhỏ giọt), tranh thủ thời gian ngắn trong ngày. -> Yêu cầu hệ thống phải khởi động cực nhanh và thao tác tối giản.
- **Tâm lý:** Họ rất sợ sự phiền toái. Việc phải mở 3 tab khác nhau để tra từ, nghe phát âm và chép vào Excel là lý do chính khiến họ bỏ cuộc. -> Yêu cầu tính năng "Một chạm" của Magic Vocab.
- **Động lực:** Họ cần nhìn thấy kết quả hữu hình để duy trì thói quen. -> Yêu cầu hệ thống Stats & Streaks phải trực quan, cập nhật tức thì.
- **Năng lực công nghệ:** Họ có kỹ năng sử dụng công nghệ cơ bản nhưng không phải chuyên gia, do đó việc cấu hình API hay Model AI cần được đơn giản hóa tối đa.

2.3 Môi trường vận hành

- **Thiết bị:** Hoạt động trên trình duyệt web hiện đại (Chrome, Edge, Firefox, Safari) và thiết bị di động (Android/iOS thông qua PWA hoặc Native shell).
- **Kết nối:** Yêu cầu kết nối Internet liên tục để giao tiếp với AI API.
- **API Service:** Hệ thống giao tiếp với Ollama API qua các endpoint RESTful chuẩn như /api/generate và /api/chat.

2.4 Giả định và Ràng buộc

Ràng buộc:

- **Phụ thuộc bên thứ ba:** Tính năng cốt lõi (FR1, FR2) sẽ lệ thuộc hoàn toàn nếu Ollama Cloud API gặp sự cố hoặc thay đổi đột ngột cấu trúc phản hồi.
- **Giới hạn của AI:** Chất lượng của việc chấm điểm và tra từ phụ thuộc vào năng lực của Model được chọn (ví dụ: Model 7B tham số sẽ kém hơn Model 70B tham số). Hệ thống phần mềm không thể cam kết độ chính xác 100% về mặt nội dung ngôn ngữ học.
- **Chi phí và Tài nguyên:** Nếu sử dụng các model lớn trên Cloud, có thể phát sinh chi phí hoặc giới hạn số lượng request.

Giả định:

- Người dùng sở hữu thiết bị có khả năng kết nối internet ổn định.
- Người dùng có quyền truy cập hợp lệ (API Key) vào dịch vụ Ollama Cloud hoặc có một instance Ollama chạy cục bộ có thể truy cập qua mạng.
- Độ trễ mạng đến máy chủ AI nằm trong mức chấp nhận được (< 2 giây) để đảm bảo

trải nghiệm người dùng không bị gián đoạn.

3. CÁC CHỨC NĂNG CHÍNH VÀ CHI TIẾT DÒNG SỰ KIỆN

Phần này mô tả chi tiết các yêu cầu chức năng, được chia nhỏ theo từng module. Mỗi chức năng (Feature) được trình bày kèm theo dòng sự kiện (Use Case Scenarios) chi tiết để làm rõ logic xử lý.

3.1 MODULE "MAGIC VOCAB"

Đây là phân hệ cốt lõi giúp giải quyết vấn đề quản lý từ vựng rời rạc.

Bảng tóm tắt các chức năng Module Magic Vocab

ID	Tên chức năng	Mô tả ngắn	Tác nhân
FR1.1	Nhập từ mới	Giao diện nhập liệu đơn giản cho từ vựng.	Người học
FR1.2	Làm giàu dữ liệu AI	Gọi API để lấy IPA, nghĩa, loại từ, ví dụ, CEFR.	Hệ thống
FR1.3	Lưu trữ cá nhân	Lưu kết quả vào CSDL người dùng.	Hệ thống
FR1.4	Hiển thị sổ tay	Xem, tìm kiếm và lọc danh sách từ vựng.	Người học

Chi tiết Dòng sự kiện: FR1.1 & FR1.2 - Thêm từ và Tự động làm giàu dữ liệu

Mô tả: Người dùng chỉ cần nhập từ tiếng Anh, hệ thống tự động sử dụng AI để điền toàn bộ thông tin chi tiết còn lại.

Tiền điều kiện:

- Người dùng đã đăng nhập.
- Hệ thống đã kết nối thành công với Ollama API.

Dòng sự kiện chính:

1. Khởi tạo: Tại màn hình chính, người dùng nhấn nút "Thêm từ mới" (+).
2. Hiển thị: Hệ thống hiển thị Modal nhập liệu với trỏ chuột focus vào ô "English Word".
3. Nhập liệu: Người dùng nhập từ (ví dụ: "Benevolent") và nhấn Enter hoặc nút "Magic

Analyze".

4. Khóa giao diện: Hệ thống chuyển trạng thái nút bấm sang "Loading", vô hiệu hóa ô nhập liệu để ngăn gửi trùng lặp.
5. Xây dựng Prompt: Hệ thống tạo một bản tin (prompt) gửi tới Ollama API (POST /api/generate hoặc /api/chat).
 - Kỹ thuật: Prompt được thiết kế kèm theo yêu cầu format: "json" để đảm bảo tính cấu trúc.
 - Nội dung Prompt: "Define the word 'Benevolent'. Return JSON with keys: ipa, vietnamese_meaning, part_of_speech, example_sentence, cefr_level."
6. Giao tiếp API: Hệ thống gửi request tới Ollama Cloud.
7. Nhận phản hồi: Hệ thống nhận về JSON Object từ API (HTTP 200 OK).
8. Phân tích (Parsing): Hệ thống đọc dữ liệu JSON và điền tự động vào các trường tương ứng trên Form nhập liệu:
 - IPA: /bə'nev.əl.ənt/
 - Meaning: Nhân từ, rộng lượng
 - Type: Adjective
 - Example: "He was a benevolent old man and wouldn't hurt a fly."
 - CEFR: C1
9. Xác nhận: Người dùng xem lại thông tin. (Người dùng có thể sửa đổi nếu muốn).
10. Lưu: Người dùng nhấn "Lưu vào Sổ tay".
11. Ghi dữ liệu: Hệ thống lưu bản ghi vào CSDL.
12. Cập nhật: Hệ thống cập nhật bộ đếm từ vựng và gửi sự kiện cập nhật Streak (FR3).
13. Kết thúc: Đóng Modal và hiển thị thông báo "Đã thêm từ thành công".

Dòng sự kiện phụ:

- A1: Lỗi kết nối API
 - Tại bước 6: Request bị timeout hoặc nhận mã lỗi 503 (Service Unavailable).
 - Hệ thống: Hiển thị thông báo: "Kết nối AI gặp sự cố. Đang thử lại..." và thực hiện cơ chế Retry (thử lại 1-2 lần).
 - Nếu vẫn thất bại: Hiển thị: "Không thể kết nối AI. Vui lòng nhập thông tin thủ công" và mở các ô nhập liệu trống để người dùng tự điền.
- A2: Từ vựng không hợp lệ / AI không hiểu

- Tại bước 7: AI trả về nội dung cho biết từ không tồn tại hoặc JSON rỗng.
- Hệ thống: Hiển thị cảnh báo: "Không tìm thấy định nghĩa cho từ này. Hãy kiểm tra chính tả."
- A3: Từ đã tồn tại.
- Tại bước 3: Hệ thống kiểm tra CSDL cục bộ ngay khi nhập.
 - Hệ thống: Hiển thị cảnh báo "Từ này đã có trong sổ tay".
 - Người dùng: Có thể chọn "Xem từ đã lưu" hoặc "Vẫn thêm (nghĩa khác)".

Chi tiết Dòng sự kiện: FR1.4 - Hiển thị Sổ tay Từ vựng (Vocabulary Notebook)

Mô tả: Giao diện hiển thị danh sách từ đã học, hỗ trợ tìm kiếm và lọc thông minh.

Dòng sự kiện chính:

1. Truy cập: Người dùng vào mục "Magic Vocab".
2. Truy vấn: Hệ thống tải danh sách từ từ CSDL, sắp xếp theo "Mới nhất".
3. Hiển thị: Danh sách hiển thị dưới dạng thẻ. Mỗi thẻ có:
 - Từ vựng (In đậm, size lớn).
 - Phiên âm IPA.
 - Nghĩa tiếng Việt.
 - Badge màu sắc cho Loại từ (Noun=Blue, Verb=Red...).
4. Xem chi tiết: Người dùng nhấn vào thẻ để xem câu ví dụ và cấp độ CEFR (hiệu ứng Accordion hoặc Modal).

Dòng sự kiện phụ:

- A1: Tìm kiếm: Người dùng nhập vào thanh tìm kiếm -> Hệ thống lọc danh sách theo thời gian thực (Real-time filter).
- A2: Lọc nâng cao: Người dùng chọn Filter -> Chọn "CEFR Level: C1" -> Hệ thống chỉ hiện các từ mức C1.

3.2 MODULE "GRAMMAR & STYLE CHECKER"

Phân hệ này giải quyết nhu cầu phản hồi tức thì về kỹ năng viết, đóng vai trò như một gia sư AI.

Bảng tóm tắt các chức năng Module Grammar Checker

ID	Tên chức năng	Mô tả ngắn	Tác nhân
FR2.1	Nhập văn bản	Soạn thảo hoặc dán đoạn văn tiếng Anh.	Người học

FR2.2	Phân tích AI	Gửi văn bản tới AI để tìm lỗi và đánh giá.	Hệ thống
FR2.3	Hiển thị kết quả	Hiển thị điểm số, bôi đỏ lỗi, gợi ý sửa.	Hệ thống

Chi tiết Dòng sự kiện: FR2.1 - FR2.3 - Kiểm tra và Chấm điểm

Mô tả: Quy trình gửi văn bản đến AI và nhận về phản hồi chi tiết về ngữ pháp và phong cách.

Tiền điều kiện: Người dùng có đoạn văn bản cần kiểm tra.

Dòng sự kiện chính:

1. Nhập liệu: Người dùng nhập đoạn văn vào vùng soạn thảo lớn.
2. Kích hoạt: Người dùng nhấn nút "Chấm điểm ngay".
3. Tiền xử lý: Hệ thống kiểm tra độ dài (giới hạn ví dụ 2000 ký tự) để đảm bảo hiệu suất API.
4. Gửi Request: Hệ thống gửi prompt tới Ollama API.
 - Prompt Logic: Yêu cầu AI đóng vai trò giám khảo IELTS/TOEFL. Output yêu cầu JSON gồm: score (0-10), mistakes (array chứa vị trí lỗi, loại lỗi, gợi ý), improved_version (đoạn văn viết lại).
5. Phản hồi: Hệ thống nhận JSON kết quả.
6. Trực quan hóa:
 - Điểm số: Hiển thị vòng tròn điểm số (Ví dụ: 8.5) với màu sắc tương ứng (Xanh lá > 8.0, Vàng > 5.0, Đỏ < 5.0).
 - Highlight Lỗi: Hệ thống quét đoạn văn gốc, bôi đỏ các từ/cụm từ có trong danh sách mistakes trả về.
 - Gợi ý: Khi người dùng di chuột (hover) vào lỗi, hiển thị Tooltip: "Lỗi: Sai thì - > Gợi ý: 'went' thay vì 'go'".
7. Sửa lỗi: Người dùng nhấn vào gợi ý -> Hệ thống thay thế từ sai bằng từ đúng trong văn bản gốc.
8. Tham khảo Style: Hệ thống hiển thị box "Đề xuất viết lại" chứa nội dung improved_version để người dùng học cách diễn đạt tự nhiên hơn.

Dòng sự kiện phụ:

- A1: Văn bản quá ngắn: Nếu người dùng nhập < 5 từ, hệ thống báo lỗi: "Văn bản quá ngắn để phân tích".
- A2: Lỗi phân tích JSON: Nếu AI trả về định dạng không đúng JSON (do model kém chất lượng), hệ thống hiển thị văn bản thô của AI và cảnh báo: "Không thể trích xuất chi tiết lỗi, hiển thị nhận xét thô từ AI".

3.3 MODULE "STATS & STREAKS"

Phân hệ này sử dụng tâm lý học hành vi để duy trì động lực người học.

Bảng tóm tắt các chức năng Module Stats

ID	Tên chức năng	Mô tả ngắn	Tác nhân
FR3.1	Theo dõi tự động	Ghi log hành động học tập (Tracking).	Hệ thống (Ngầm)
FR3.2	Tính toán Streak	Logic tính chuỗi ngày liên tục.	Hệ thống (Ngầm)
FR3.3	Thống kê tổng quan	Hiển thị tổng số từ, huy hiệu thành tích.	Người học
FR3.4	Biểu đồ hóa	Vẽ biểu đồ tròn/cột về phân bố từ vựng.	Hệ thống

Chi tiết Dòng sự kiện: FR3.1 & FR3.2 - Tính toán Streak

Mô tả: Logic nền tảng để xác định người dùng có giữ được chuỗi học tập hay không.

Dòng sự kiện chính:

1. Lắng nghe: Hệ thống lắng nghe sự kiện ACTION_COMPLETED (khi người dùng thêm từ mới hoặc chấm điểm bài viết).
2. Xác định thời gian: Lấy thời gian hiện tại của hệ thống.
3. So sánh:
 - Lấy Last_Study_Date từ CSDL.
 - Nếu Current_Date == Last_Study_Date: Không làm gì (đã tính streak cho hôm nay).
 - Nếu Current_Date == Last_Study_Date + 1 ngày: Tăng biến Streak_Count lên 1. Cập nhật Last_Study_Date thành hôm nay.
 - Nếu Current_Date > Last_Study_Date + 1 ngày: (Người dùng bỏ học hôm

qua) -> Reset Streak_Count về 1.

4. Thông báo: Nếu Streak tăng, hiển thị thông báo chúc mừng: "Bạn đã đạt chuỗi 5 ngày! Tiếp tục phát huy!".

Dòng sự kiện phụ:

- A1: Múi giờ thay đổi: Nếu người dùng di chuyển sang múi giờ khác, hệ thống giữ nguyên mốc thời gian của lần học cuối cùng để đảm bảo tính công bằng (hoặc cho phép một khoảng ân hạn - grace period).

1. Chi tiết Dòng sự kiện: FR3.4 - Biểu đồ hóa dữ liệu

Mô tả: Hiển thị biểu đồ phân tích chất lượng học tập.

Dòng sự kiện chính:

2. Tổng hợp: Hệ thống thực hiện truy vấn tổng hợp trên CSDL từ vựng:
 - Đếm số lượng từ theo part_of_speech (Danh từ, Động từ...).
 - Đếm số lượng từ theo cefr_level (A1, A2...).
3. Vẽ biểu đồ:
 - Pie Chart: Hiển thị tỷ lệ loại từ. Giúp người học nhận ra sự mất cân đối (ví dụ: học quá nhiều danh từ, ít động từ).
 - Column Chart: Hiển thị tháp trình độ. Giúp người học thấy sự tiến bộ từ cơ bản lên nâng cao.
4. Tương tác: Người dùng nhấn vào một phần của biểu đồ (ví dụ: cột C1) -> Hệ thống chuyển sang màn hình Sổ tay và tự động lọc các từ thuộc nhóm C1 đó.

3.4 CẤU HÌNH HỆ THỐNG (SYSTEM CONFIGURATION)

FR4: Tùy chọn Model AI

Mô tả: Cho phép người dùng chọn "bộ não" cho ứng dụng.

Dòng sự kiện chính:

1. Vào cài đặt: Người dùng mở Settings -> AI Config.
2. Quét Model: Hệ thống gọi API GET /api/tags tới Ollama để liệt kê các model đã tải về (llama3, mistral, gemma...).
3. Chọn: Người dùng chọn model mong muốn từ danh sách.
4. Lưu: Hệ thống lưu tên model vào biến cấu hình toàn cục. Các lệnh gọi API sau đó sẽ sử dụng tham số model: "tên_model_đã_chọn".

4. YÊU CẦU PHI CHỨC NĂNG

4.1 Hiệu suất - NFR1

- Độ trễ phản hồi:
 - Giao diện người dùng phải phản hồi ngay lập tức ($< 100\text{ms}$) khi người dùng thao tác (nhấn nút, chuyển tab).
 - Các tác vụ AI (Tra từ, Chấm điểm) phải hoàn thành trong vòng 3-5 giây trong điều kiện mạng tiêu chuẩn. Nếu quá 5 giây, phải có thanh tiến trình (progress bar) hoặc thông báo vui nhộn để giữ chân người dùng.
- Khởi động: Ứng dụng (Web/App) phải có thời gian tải ban đầu dưới 1.5 giây.

4.2 Tính Khả dụng - NFR2

- Giao diện: Tuân thủ nguyên tắc thiết kế tối giản. Loại bỏ mọi thành phần thừa thãi.
 - Font chữ dễ đọc, độ tương phản cao.
 - Không gian trắng được sử dụng hợp lý để giảm tải nhận thức.
- Trải nghiệm: Quy trình nhập từ mới phải tối ưu hóa để không vượt quá 3 thao tác (Mở app -> Nhập từ -> Enter).

4.3 Độ tin cậy và Chính xác (Reliability & Accuracy) - NFR3

- Dữ liệu AI: Dữ liệu trích xuất từ AI (IPA, Nghĩa) phải đạt độ chính xác cao. Hệ thống cần sử dụng Prompt Engineering kỹ lưỡng để giảm thiểu ảo giác. Ví dụ: Prompt phải chứa chỉ thị *"Only return standard dictionary definitions"*.
- Tính toàn vẹn dữ liệu: Bộ đếm Streak không được phép sai sót. Dữ liệu này cần được backup định kỳ (nếu có tính năng Cloud Sync).

4.4 Khả năng cấu hình - NFR4

- Lựa chọn Model: Hệ thống không được "hard-code" một model AI cụ thể. Người dùng phải có quyền chuyển đổi giữa các model như GPT-OSS 20B, Llama 3, v.v., tùy thuộc vào cấu hình máy chủ Ollama của họ.
- Endpoint: Cho phép cấu hình URL của Ollama API (mặc định `http://localhost:11434`, nhưng có thể đổi thành IP server trong mạng LAN).

4.5 Bảo mật

- An toàn API: Nếu người dùng sử dụng Ollama Cloud có xác thực, API Key phải được lưu trữ an toàn trên thiết bị, không được truyền đi dưới dạng văn bản thô trong các log không cần thiết.
- Riêng tư: Dữ liệu học tập cá nhân (từ vựng, bài viết) mặc định được lưu trữ cục bộ, đảm bảo quyền riêng tư tuyệt đối cho người học.

5. PHỤ LỤC

5.1 Xử lý Lỗi API

Hệ thống cần xử lý các mã trạng thái HTTP từ Ollama API như sau:

Mã lỗi	Ý nghĩa kỹ thuật	Thông báo hiển thị cho User	Hành động hệ thống
200 OK	Thành công	(Không hiển thị, xử lý dữ liệu)	Parse JSON và hiển thị
400 Bad Request	Request sai định dạng	"Lỗi dữ liệu gửi đi. Vui lòng thử lại."	Log lỗi, không retry
401 Unauthorized	Sai API Key	"Vui lòng kiểm tra API Key trong Cài đặt."	Chuyển hướng trang Settings
404 Not Found	Model chưa tải/Sai endpoint	"Model AI chưa sẵn sàng. Hãy kiểm tra Ollama."	Hướng dẫn pull model
503 Service Unavailable	Ollama quá tải/Đang tắt	"Dịch vụ AI đang bận hoặc chưa bật."	Thử lại sau 3s (Exponential Backoff)

5.2 Mẫu Prompt Kỹ thuật

Để đảm bảo FR1.2 hoạt động ổn định, Prompt gửi đi cần tuân thủ cấu trúc nghiêm ngặt.

Mẫu cho Magic Vocab:

```
{
  "model": "llama3",
  "format": "json",
  "stream": false,
  "messages":
}
```