**AIRESA: Artificial Intelligence Real Estate Sales Assistant - Hệ Thống AI Hỗ Trợ Bất Động Sản Việt Nam - Roadmap Tổng Quát**

**Ngày lập roadmap**: 28/09/2025

**Mô tả dự án**: PropertAI là một chatbot AI hỗ trợ sales bất động sản (BDS) tại Việt Nam, sử dụng RAG (Retrieval-Augmented Generation) và finetuning model để tư vấn dự án, xử lý objection, thu thập lead, và lên lịch xem nhà. Roadmap thích nghi từ mô hình AI Legal, tập trung vào data BDS biến động (giá, vị trí, tiện ích từ nguồn như batdongsan.com.vn). Scale: 2 người (AI Engineer lead, Fullstack Dev support). Thời gian: 3 tháng (12 tuần, từ 28/09/2025 đến ~28/12/2025). Budget ước tính: 500-1000 USD (GPU cloud, API). Mục tiêu MVP: Accuracy >85% trong matching dự án, tích hợp multichannel (web, Zalo OA, FB Messenger).

Roadmap dựa trên các flowchart cung cấp (tổng quan, chuẩn bị data, xây dựng RAG, training model) và flowchart luồng chạy RAG end-to-end, được tinh chỉnh cho BDS. Nội dung bổ sung từ nguồn (e.g., kịch bản sales từ Omicall, lợi ích AI trong BDS từ FPT.AI).

Flowchart Luồng Chạy (End-to-End RAG cho Chatbot BDS)

Mô tả tổng quát: Flowchart phức tạp mô tả quy trình RAG end-to-end cho hệ thống trả lời câu hỏi sales BDS. Bắt đầu từ "User Question" (e.g., "Tìm căn hộ Q1 dưới 2 tỷ gần trường học?") và kết thúc bằng "Answer" với cơ chế kiểm tra, tối ưu hóa. Flowchart chia thành các phần: Input, Retrieval, Generation, Verification, và Ops (vận hành). Thiết kế để giảm hallucination, đảm bảo data realtime (giá BDS thay đổi hàng ngày), và tích hợp lead capture.

**Cấu trúc chi tiết:**

* **Input:** "User Question" → "QA/FastAPI /ask" (giao diện API nhận query, hỗ trợ multichannel như Zalo/FB) → "Query Understanding" (hiểu query: dịch sang tiếng Việt chuẩn hóa, xử lý timeframe như "dự án mới 2025", extract entity như ngân sách/vị trí qua Underthesea).
* **Retrieval:** 
  + "Retrieval" → "Hybrid Retrieval" (kết hợp BM25 cho tìm kiếm từ khóa như "căn hộ Q1" và Embeddings cho tìm kiếm ngữ nghĩa như "gần trường học, tiện ích gym").
  + Nguồn dữ liệu: "OpenSearch Lexical Index" (chỉ mục từ vựng cho giá/vị trí) và "Vector Store" (cơ sở dữ liệu vector như Qdrant hoặc Pinecone lưu listings BDS chunked).
  + "Top K Candidates" (lấy top K kết quả, e.g., top 20 dự án khớp) → "Cross-encoder Rerank" (rerank bằng cross-encoder để cải thiện độ chính xác, ưu tiên dự án hot từ kịch bản Vpage) → "Diversity Control" (kiểm soát đa dạng để tránh lặp dự án tương tự) → "Effectiveness Filter as of date" (lọc hiệu lực theo ngày, e.g., check availability dự án realtime qua API batdongsan).
  + "Context Builder" (xây dựng context: nhóm theo ý định khách như "mua đầu tư" hoặc "mua ở", thêm metadata như giá, tiện ích, hình ảnh links).
* **Generation**: "Generator (fine tune LLM) RAG Prompt" (sử dụng LLM finetune như Vietcuna/PhoGPT với prompt RAG để generate answer, e.g., "Dựa trên nhu cầu, gợi ý Vinhomes Central Park với giá 3 tỷ, tiện ích...").
* **Verification**: "Verification" (kiểm tra) → Kiểm tra "Citations/effectivity check" với nguồn "PostgresQL Documents + Units + Versions" (DB relational lưu tài liệu BDS, đơn vị dự án, phiên bản cập nhật giá). Nếu "OK" → "Answer + Citations + Links + Confidence + As of" (trả lời với trích dẫn dự án, liên kết hình ảnh/bản đồ, độ tin cậy, ngày cập nhật). Nếu "Mismatch" → "Auto-correct / Add caveat" (tự sửa hoặc thêm lưu ý như "Giá có thể thay đổi, vui lòng kiểm tra realtime").
* **Ops** (Vận hành):
  + "Scheduler Nightly refreshed" (lập lịch cập nhật hàng đêm từ scrape/API BDS để giữ data tươi mới).
  + "Evaluation" (đánh giá: Recall, Citation accuracy, Q – Quality, + BDS metrics như lead quality từ FPT.AI).
  + "Logs" (nhật ký query/sales interaction).
  + "Monitoring/Audit Prompts + Contexts + Versions" (giám sát prompt, context, phiên bản data để debug).
  + "Ingesting Index + Ingest" (ingest data mới vào index, hỗ trợ realtime update từ nguồn như Omicall scripts).

Công nghệ ngầm định:

* API: FastAPI cho endpoint /ask, hỗ trợ asynchronous để xử lý query nhanh.
* Retrieval: Hybrid (BM25 + Embeddings), Cross-encoder (từ Sentence Transformers như "cross-encoder/ms-marco-MiniLM-L-6-v2" finetune tiếng Việt), Vector Store (Qdrant/Pinecone), OpenSearch (Elasticsearch-based cho lexical search).
* DB: PostgresQL cho metadata (dự án, phiên bản giá, availability).
* LLM: Fine-tuned LLM (Vietcuna, PhoGPT, hoặc Llama finetune cho domain BDS/sales).
* Khác: Scheduler (Cron jobs cho refresh), Monitoring (Prometheus hoặc ELK Stack cho logs/audit), Redis caching cho query phổ biến.

Ý nghĩa cho BDS: Flowchart nhấn mạnh tính chính xác và cập nhật (tránh tư vấn dự án hết hàng), phù hợp với thị trường VN biến động (theo VnExpress: AI định hình BDS bằng dự báo giá).

**Phân Tích Flowchart Hình Ảnh**

Dựa trên 3 hình cung cấp (tổng quan, chuẩn bị data, xây dựng RAG, training model – từ AI Legal, thích nghi cho BDS).

* **Flowchart Tổng Quan (FLOWCHART AI LEGAL)**: Overview tuyến tính 4 giai đoạn. Bắt đầu từ data nguồn (bộ luật → listings BDS), qua chuẩn bị data (①), xây dựng RAG (②), training model (③), đến xây dựng chatbot (④). Thêm eval lặp (e.g., "Đánh giá model cơ sở"). Tinh chỉnh cho BDS: Thay "Data đầy đủ các bộ luật" bằng "Data listings dự án BDS từ scrape".
* **Flowchart Chuẩn Bị Data (CHUẨN BỊ DATA)**: Phân nhánh từ nguồn (listings BDS, case sales, QA) đến xử lý (scrape BeautifulSoup/Selenium, clean/chunk). Thêm personalization (QA mang tính cá nhân như "Dựa ngân sách 2 tỷ"). Sử dụng datasets sẵn (duane1024/newyork\_real\_estate adapt VN). Phân ngày cho sales expert thiết kế.
* **Flowchart Xây Dựng RAG (XÂY DỰNG RAG)**: Từ eval embedding (benchmark thiết kế) đến build pipeline (retrieval/re-ranking), chọn vector DB (Qdrant), embed chunks, test performance. Thêm hybrid search cho query BDS chính xác (e.g., "Dự án Q1").
* **Flowchart Training Model (TRAINING MODEL)**: Lặp eval base/finetuned/hyperparams, train model, re-eval. Tinh chỉnh: RLHF từ feedback sales thật để handle objection.

**Giai Đoạn 1: Chuẩn Bị Data (Tuần 1-3)**

* Tập trung thu thập/ xử lý data BDS (listings, giá, vị trí, kịch bản sales từ A Tiến/Omicall). Clean/normalize/chunk. Augment QA (PhoGPT paraphrase). Split 80/10/10. Công nghệ: Crawling (BeautifulSoup/Selenium), Processing (Pandas/Underthesea), Datasets (Hugging Face, e.g., duane1024/newyork\_real\_estate adapt).
* **Tuần 1: Thu Thập Data Gốc (AI Lead, Dev Support)** 
  + **Tasks:** 
    - **Nghiên cứu nguồn: Scrape batdongsan.com.vn, homedy.com, alonhadat.com.vn (sử dụng BeautifulSoup/Selenium cho listings dự án, giá realtime). Thu thập ~5k records (dự án, vị trí, tiện ích).**
    - **Thiết kế QA data: Tạo 500 QA pairs (e.g., "Tư vấn dự án Vinhomes Ocean Park?") dựa trên sales scripts từ Omicall/Vpage (từ search results).**
    - **Phân công: AI: Script scrape + initial clean (remove duplicates via pandas). Dev: Setup repo Git, env (Docker, Python 3.12).**
  + **Milestones: Dataset raw ~2GB JSON. Tools: Datasets lib (Hugging Face).**
  + **Deliverables: Repo Git với data raw, script scrape (test trên 100 records). Thời gian: 40h/người.**
* **Tuần 2: Làm Sạch & Thiết Kế Data (AI Lead)** 
  + **Tasks:** 
    - **Clean data: Normalize (UTF-8, standardize giá/vị trí), anonymize (loại info cá nhân), augment QA (paraphrase bằng PhoGPT để ~2k pairs).**
    - **Thiết kế metadata: Thêm tags (loại BDS: căn hộ/đất nền, phân khúc: cao cấp/trung cấp). Sử dụng Underthesea cho entity extraction (e.g., "Q1 TP.HCM").**
    - **Phân công: AI: Full clean + QA design. Dev: Build data pipeline (Airflow đơn giản cho ETL).**
  + **Milestones: Dataset sạch, split 80/10/10 (train/val/test).**
  + **Deliverables: JSON dataset validated (fact-check bởi sales expert), pipeline script. Thời gian: 35h/người.**
* **Tuần 3: Chunking & Initial Eval (AI Lead, Dev Review)** 
  + **Tasks:** 
    - **Chunk data: Chia listings thành segments (256 tokens, overlap 32) bằng TokenChunker (tương tự Import\_Data.py).**
    - **Initial eval: Kiểm tra quality (duplication <5%, coverage lĩnh vực BDS >90%).**
    - **Phân công: AI: Chunk + eval metrics (TF-IDF similarity). Dev: Integrate vào DB temp (SQLite).**
  + **Milestones: Dataset chunked sẵn cho embedding (~10k chunks).**
  + **Deliverables: Report eval (PDF), chunked data folder. End giai đoạn: Data foundation ready.**

**Giai Đoạn 2: Xây Dựng RAG (Tuần 4-6)**

**Thích nghi từ Legal: Build retrieval cho matching property (hybrid search: keyword "căn hộ Q1" + semantic "gần trường học"). Eval trên benchmark sales (precision@5 cho top matches).**

* **Tuần 4: Eval Embedding & Setup RAG Core (AI Lead)** 
  + **Tasks:** 
    - **Eval models: Test Sentence Transformers (e.g., paraphrase-multilingual-MiniLM cho VN BDS) trên custom benchmark (200 queries: "Tìm dự án giá <2 tỷ").**
    - **Setup RAG: Xây pipeline cơ bản (LangChain: retriever + reranker cross-encoder).**
    - **Phân công: AI: Model eval + prototype RAG. Dev: Setup Qdrant collection (COSINE distance).**
  + **Milestones: Top embedding model selected (accuracy >80%).**
  + **Deliverables: RAG prototype script, benchmark report.**
* **Tuần 5: Hybrid Retrieval & Re-ranking (AI Lead)** 
  + **Tasks:** 
    - **Implement hybrid: BM25 (keyword cho giá/vị trí) + embeddings (ngữ nghĩa nhu cầu khách).**
    - **Re-rank: Cross-encoder rerank top-20 candidates, add diversity (tránh duplicate dự án).**
    - **Phân công: AI: Core logic. Dev: API stub (FastAPI endpoint /retrieve).**
  + **Milestones: Retrieval latency <1s, recall@10 >85%.**
  + **Deliverables: Integrated RAG module, test queries dataset.**
* **Tuần 6: Embed & Index Data + Test (AI Lead, Dev Integrate)** 
  + **Tasks:** 
    - **Embed chunks: Sử dụng model selected, index vào Qdrant (upsert ~10k points).**
    - **Test pipeline: Eval end-to-end (MRR, Hit Rate) trên sales scenarios (e.g., lead qualification).**
    - **Phân công: AI: Embed/index. Dev: Connect to BE prototype.**
  + **Milestones: Full RAG operational, data indexed.**
  + **Deliverables: Qdrant dashboard, eval report. End giai đoạn: Retrieval ready for generation.**

**Giai Đoạn 3: Training Model (Tuần 7-9)**

**Thích nghi từ Legal: Finetune LLM cho conversational sales (e.g., objection handling: "Giá cao quá?" → Suggest alternatives). Sử dụng QA data cho SFT.**

* **Tuần 7: Eval Base Model & Prep Training (AI Lead)** 
  + **Tasks:** 
    - **Eval base: Test PhoGPT/Vietcuna trên 500 QA (perplexity, BLEU score).**
    - **Prep data: Format instruction-tuning (prompt: "Khách: Tìm căn hộ 3PN. Bot: Dựa trên data, gợi ý Vinhomes...").**
    - **Phân công: AI: Full eval/prep. Dev: Setup GPU env (Colab hoặc Vast.ai).**
  + **Milestones: Base model baseline (<70% accuracy).**
  + **Deliverables: Training dataset formatted, hyperparams grid (Optuna).**
* **Tuần 8: Finetune Model (AI Lead)** 
  + **Tasks:** 
    - **Train: SFT với PEFT (LoRA) trên train set, epochs=3, batch=8 (GPU 1x A100).**
    - **Monitor: Track loss, early stopping nếu val loss tăng.**
    - **Phân công: AI: Run training. Dev: Build monitoring dashboard (TensorBoard).**
  + **Milestones: Finetuned model checkpoint.**
  + **Deliverables: Trained model (Hugging Face repo), loss curves.**
* **Tuần 9: Eval & Iterate Model (AI Lead, Dev Review)** 
  + **Tasks:** 
    - **Eval: So sánh base vs. finetuned (F1>85% cho sales responses, hallucination <5%).**
    - **Iterate: Tune hyperparams (learning rate 1e-5), retrain nếu cần.**
    - **Phân công: AI: Eval/iterate. Dev: Integrate model vào RAG stub.**
  + **Milestones: Model ready (accuracy >90% trên test set).**
  + **Deliverables: Final model, eval report. End giai đoạn: AI core complete.**

**Giai Đoạn 4: Xây Dựng Chatbot (FE, BE, Integration) (Tuần 10-12)**

**Thích nghi từ Legal: Build fullstack với verification (e.g., confirm lead data), multi-turn chat. Deploy MVP.**

* **Tuần 10: BE Integration & API (Dev Lead, AI Support)** 
  + **Tasks:** 
    - **BE: FastAPI endpoints (/chat: query → RAG + generate; /lead: store customer info).**
    - **Integrate: Chain RAG + finetuned LLM (LangChain), add verification (confidence score >0.7).**
    - **Phân công: Dev: API build. AI: Hook model/RAG.**
  + **Milestones: BE API testable (Postman).**
  + **Deliverables: API docs (Swagger), integration tests.**
* **Tuần 11: FE Development & UI/UX (Dev Lead)** 
  + **Tasks:** 
    - **FE: React app với chat interface (multi-turn, buttons cho quick replies e.g., "Xem giá", "Lên lịch").**
    - **Features: Property cards (từ retrieval), Zalo/FB integration (Messenger API).**
    - **Phân công: Dev: Full FE. AI: Provide mock responses.**
  + **Milestones: FE prototype connected to BE.**
  + **Deliverables: Deployable FE (Vercel), UI wireframes.**
* **Tuần 12: Testing, Deploy & Polish (Dev Lead, AI Review)** 
  + **Tasks:** 
    - **End-to-end test: User scenarios (lead gen, sales flow), fix bugs (latency, errors).**
    - **Deploy: Dockerize, host trên AWS/Heroku, setup monitoring (Sentry).**
    - **Polish: Add analytics (lead conversion rate), nightly data refresh (Cron).**
    - **Phân công: Dev: Test/deploy. AI: Performance tuning.**
  + **Milestones: MVP live, 95% uptime.**
  + **Deliverables: Demo video, user guide, final report (ROI projection: tăng lead 30%). End dự án: Launch ready.**

**Tổng Kết: Roadmap linh hoạt, với buffer 10% thời gian cho iterations. Theo dõi bằng Trello/Jira. Dự kiến output: Chatbot tăng hiệu quả sales 2x (dựa trên case Omicall). Nếu cần adjust, review hàng tháng.**

**Nguồn**

[**https://omicall.com/kich-ban-chatbot-bat-dong-san/**](https://omicall.com/kich-ban-chatbot-bat-dong-san/) **(**

[**https://botpress.com/vi/blog/chatbot-for-real-estate**](https://botpress.com/vi/blog/chatbot-for-real-estate) **(hướng dẫn build)**

[**https://fpt.ai/vi/bai-viet/chatbot-va-nhung-loi-ich-trong-linh-vuc-bat-dong-san/**](https://fpt.ai/vi/bai-viet/chatbot-va-nhung-loi-ich-trong-linh-vuc-bat-dong-san/)**(Tối ưu UX/UI)**

[**https://emandai.net/blog/vi/ung-dung-callbot-bat-dong-san-xu-huong-moi-cho-nganh-bat-dong-san/**](https://emandai.net/blog/vi/ung-dung-callbot-bat-dong-san-xu-huong-moi-cho-nganh-bat-dong-san/)

[**https://baodautu.vn/batdongsan/ai-tra-loi-the-nao-cau-hoi-ve-bat-dong-san-viet-nam-d246931.html**](https://baodautu.vn/batdongsan/ai-tra-loi-the-nao-cau-hoi-ve-bat-dong-san-viet-nam-d246931.html)

[**https://vnexpress.net/ai-gop-phan-dinh-hinh-thi-truong-bat-dong-san-nhu-the-nao-4904607.html**](https://vnexpress.net/ai-gop-phan-dinh-hinh-thi-truong-bat-dong-san-nhu-the-nao-4904607.html)

[**https://vpage.nhanh.vn/blog/kich-ban-chatbot-tu-van-bat-dong-san-cac-du-an-hot-dang-ban-bao-gia-a118.html**](https://vpage.nhanh.vn/blog/kich-ban-chatbot-tu-van-bat-dong-san-cac-du-an-hot-dang-ban-bao-gia-a118.html)

[**https://www.linkedin.com/pulse/%2525E1%2525BB%2525A9ng-d%2525E1%2525BB%2525A5ng-tr%2525C3%2525AD-tu%2525E1%2525BB%252587-nh%2525C3%2525A2n-t%2525E1%2525BA%2525A1o-trong-b%2525E1%2525BA%2525A5t-%2525C4%252591%2525E1%2525BB%252599ng-s%2525E1%2525BA%2525A3n-c%2525C3%2525A1ch-ai-t%2525C3%2525A1i-%2525C4%252591%2525E1%2525BB%25258Bnh-aczlc**](https://www.linkedin.com/pulse/%2525E1%2525BB%2525A9ng-d%2525E1%2525BB%2525A5ng-tr%2525C3%2525AD-tu%2525E1%2525BB%252587-nh%2525C3%2525A2n-t%2525E1%2525BA%2525A1o-trong-b%2525E1%2525BA%2525A5t-%2525C4%252591%2525E1%2525BB%252599ng-s%2525E1%2525BA%2525A3n-c%2525C3%2525A1ch-ai-t%2525C3%2525A1i-%2525C4%252591%2525E1%2525BB%25258Bnh-aczlc)

[**https://hbr.edu.vn/ung-dung-ai-trong-bat-dong-san**](https://hbr.edu.vn/ung-dung-ai-trong-bat-dong-san)

[**https://www.facebook.com/groups/cto.platform/posts/2082397275539637/**](https://www.facebook.com/groups/cto.platform/posts/2082397275539637/)

[**https://botpress.com/vi/blog/how-to-build-your-own-ai-chatbot**](https://botpress.com/vi/blog/how-to-build-your-own-ai-chatbot)

[**https://cmcts.com.vn/vi/ai-chatbot-tu-dong.html**](https://cmcts.com.vn/vi/ai-chatbot-tu-dong.html)

[**https://www.facebook.com/groups/bdtd.vn/posts/1540427299963267/**](https://www.facebook.com/groups/bdtd.vn/posts/1540427299963267/)

[**https://www.facebook.com/groups/market.report/posts/1584855392472416/**](https://www.facebook.com/groups/market.report/posts/1584855392472416/)

[**https://bizfly.vn/techblog/tuyet-chieu-tao-chatbot-hieu-qua-cho-nganh-bat-dong-san.html**](https://bizfly.vn/techblog/tuyet-chieu-tao-chatbot-hieu-qua-cho-nganh-bat-dong-san.html)

[**https://coka.ai/tin-tuc/chatbot-bat-dong-san/**](https://coka.ai/tin-tuc/chatbot-bat-dong-san/)

**https://tinhte.vn/thread/9-top-real-estate-chatbot-2024-tang-doanh-thu-tang-lead-hieu-qua.3838758/**