Đánh giá độ chính xác Chatbot RAG (Retrieval + Agentic) cho Document Search

Nội dung dạng slide-by-slide để copy/paste vào PowerPoint.

Thông tin điền thêm:

- Người thực hiện: {Tên}
- Ngày: {dd/mm/yyyy}
- Phạm vi: {Nội bộ / PoC / Production}

Slide 1. Bài toán và mục tiêu

- Chatbot dùng để tìm kiếm tài liệu và trả lời trong hội thoại (multi-turn).
- Mục tiêu
  - Tìm đúng tài liệu trong kho (1000 files).
  - Chống nhiễu (noise/distractor files).
  - Trả lời đúng và bám nguồn (grounded, ít hallucination).
- Định nghĩa Accuracy
  - Retrieval đúng (có file/đoạn đúng).
  - Agent dùng đúng nguồn đã retrieve.
  - Answer đúng và có trích dẫn minh chứng.

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 2. Kiến trúc hệ thống cần test

- Corpus (1000 files) -> Chunking -> Embedding -> Vector Index
- Retriever (top-k) + (optional) Reranker
- Agentic loop: nhiều bước (query rewrite / iterative retrieval / refine)
- Generator (LLM) tạo câu trả lời + trích dẫn

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 3. Nguyên tắc đánh giá (tách tầng)

- Tầng 1: Retrieval
  - Kéo đúng tài liệu/đoạn giữa 'đống nhiễu'.
- Tầng 2: Agent
  - Gọi tool đúng? Có dùng đúng context đã retrieve?
- Tầng 3: Answer
  - Trả lời đúng? Có bịa ngoài context? Có cite đúng?

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 4. Thiết kế dữ liệu test (1000 files + nhiễu)

- Chia nhóm
  - Relevant pool: file có chứa đáp án cho bộ câu hỏi.
  - Noise pool: file cùng domain/từ khóa gần giống nhưng không chứa đáp án.
  - Optional: adversarial noise (cụm từ dễ gây hiểu nhầm).
- Split gợi ý
  - Dev (tuning): {x}% queries
  - Test (report): {y}% queries
- Schema tối thiểu mỗi query
  - question, ground_truth_answer (nếu có), relevant_doc_ids, (optional) relevant_spans

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 5. Lấy file để build corpus: nguồn nội bộ

- Nguồn file nội bộ
  - {ShareDrive / Confluence / Google Drive / Git repo / S3 / local folder...}
- Chuẩn hóa format ingest
  - .txt / .md / .html / .jsonl / .pdf (pdf nên extract text trước)
- Quy ước metadata
  - doc_id, source, title, url, created_at

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 6. Đưa corpus lên Hugging Face Hub

- Tạo Dataset repository trên Hugging Face (UI: New Dataset).
- Sắp xếp file theo cấu trúc repo (splits / data files) để load_dataset() đọc tự động.
- Format phổ biến: txt/csv/jsonl/parquet/zip đều có thể host và load.
- Lý do
  - Versioning (commit/tag), chia split rõ (train/dev/test).
  - Dễ chia sẻ trong team, tái lập kết quả benchmark.

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 7. Cách download/load dataset từ Hugging Face

- Trên trang dataset: tab 'Files and versions' -> tải file hoặc lấy tên repo {namespace}/{dataset_name}.
- Load bằng datasets.load_dataset() từ Hub.
- Tip
  - Giữ mapping doc_id <-> file/path <-> URL để truy vết nguồn khi audit.

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 8. Nguồn datasets trên HF để test Retrieval (IR)

- Nhóm IR benchmark (query + corpus + qrels)
  - BEIR (nhiều dataset IR đa dạng) - BeIR/*
  - MIRACL (multilingual retrieval) - miracl/miracl
  - MS MARCO (passage ranking) - msmarco_* (corpus/queries)
- Dùng để đo
  - Recall@k, MRR, nDCG cho document/passage retrieval.
  - Khả năng chịu nhiễu (distractors).

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 9. Nguồn datasets cho Conversational Search/QA

- Conversational benchmarks
  - TREC CAsT (conversational search)
  - QReCC (question rewriting + retrieval + reading)
  - TopiOCQA (open-domain conversational QA)
- Phù hợp vì
  - Có lịch sử hội thoại + nhu cầu rewrite/resolve context.

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 10. Nguồn datasets cho RAG end-to-end

- Benchmarks
  - KILT tasks (knowledge intensive; đánh giá retrieval + task).
  - RAGBench (benchmark RAG; đo end-to-end).
- Dùng để đo
  - retrieve đúng + trả lời đúng + groundedness.

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 11. Quy trình test tổng thể (end-to-end)

- 1) Chuẩn hóa corpus (1000 files + noise files)
- 2) Chunking (size/overlap/strategy)
- 3) Embed + Index
- 4) Retriever + (optional) Reranker
- 5) Agent policy (max steps, retrieve/step, rewrite...)
- 6) Run evaluation set (multi-turn nếu có)
- 7) Tính metrics (retrieval / agent / answer)
- 8) Ablation (thay 1 tham số/lần) -> báo cáo

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 12. Test thủ công (Human eval) để audit chất lượng hội thoại

- Lấy mẫu 50-100 queries (dễ/khó/nhiễu).
- Hiển thị cho người chấm
  - Câu hỏi + lịch sử chat
  - Top-k passages (kèm nguồn)
  - Agent steps (nếu có)
  - Final answer + citations
- Rubric (0/1 hoặc 1-5)
  - Retrieve đúng doc?
  - Context có quá nhiễu?
  - Agent có dùng đúng doc?
  - Answer đúng?
  - Có hallucination?

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 13. Test tự động (Automated) cho 100% queries

- A) Retrieval metrics
  - Recall@k (k = 3/5/10/20)
  - MRR@k
  - nDCG@k
  - Báo cáo theo noise ratio: 0% / 50% / 80%
- B) Agent metrics
  - Tool call rate / fail rate
  - Avg steps / max steps hit rate
  - Used-docs correctness: used_docs ⊆ retrieved_docs; used_docs ∩ relevant_docs ≠ ∅
- C) Answer metrics
  - QA ngắn: EM / F1
  - QA dài: LLM-as-judge (correctness + faithfulness) + spot-check human

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 14. Bộ thông số cần chạy (Ablation grid)

- Chunking
  - chunk_size: 256 / 512 / 1024 tokens
  - overlap: 0 / 50 / 100
  - strategy: fixed / semantic
- Retrieval
  - embedding model: {bge-m3 / e5 / ...}
  - top_k: 3 / 5 / 10 / 20
  - similarity: cosine / dot
  - reranker: off / on
  - rerank_top_n: 20 / 50
- Agent
  - max_steps: 1 / 3 / 5
  - retrieve_per_step: 3 / 5
  - query_rewrite: off / on
  - tool_budget: giới hạn / unlimited
- Generation
  - temperature: 0.0 / 0.2 / 0.7
  - max_new_tokens: 256 / 512
  - citation_policy: off / on
- Noise
  - noise_ratio: 0% / 50% / 80%

- noise_type: same-domain / keyword-overlap / adversarial

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 15. Template báo cáo kết quả

- Bảng tổng
  - Config | Recall@5 | MRR@10 | nDCG@10 | Faithfulness | Answer Acc | Avg Latency
- Biểu đồ
  - Recall@5 theo noise_ratio
  - Faithfulness theo agent max_steps
- Top failure categories
  - Query rewrite lỗi
  - Chunking làm mất evidence
  - Reranker sai
  - Answer hallucination khi context nhiễu

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.

Slide 16. Checklist nguồn dataset (điền link và tên repo)

- Corpus nội bộ
  - Nguồn: {...}
  - Cách export: {...}
  - Đường dẫn / repo: {...}
- Hugging Face Datasets
  - BEIR: BeIR/*
  - MIRACL: miracl/miracl
  - MS MARCO: msmarco_*
  - TREC CAsT: trec-cast-*
  - QReCC: svakulenk0/qrecc
  - TopiOCQA: McGill-NLP/TopiOCQA
  - KILT: facebook/kilt_tasks
  - RAGBench: rungalileo/ragbench

Gợi ý: Bạn có thể rút gọn còn 4-6 bullet/slide để phù hợp phong cách trình bày.