RAG Demo Q&A; Battle Card

1. Core Definitions

Embedding: A list of numbers (vector) that captures the *meaning* of text, like a GPS coordinate on a map of meanings.

Indexing: Organizing embeddings in a vector database (like Chroma) for fast semantic search.

Naive RAG: Retrieve docs by embedding the query directly and sending retrieved context to the LLM.

HyDE RAG: Generate a hypothetical answer, embed that, and retrieve better context before answering.

2. Explanations & Analogies

• Embeddings = GPS coordinates for meaning. Similar ideas are close together.

• Indexing = Library catalog pointing to the right shelf instantly.

• Naive RAG = Asking a friend with exact words. HyDE = Friend imagines what you mean and answers smarter.

3. Naive RAG Workflow

1. Ingest: Chunk → Embed → Store in vector DB.

2. Query: Embed query → Retrieve top-k docs.

3. Generate: LLM answers using query + docs.

Key: Frozen LLM, no retraining, context injection only.

4. HyDE Workflow

1. User query → LLM drafts hypothetical answer.

2. Embed hypothetical answer → Retrieve semantically richer docs.

3. Combine query + docs → Final LLM answer.

Key: Improves recall & accuracy, especially for vague queries.

5. Naive RAG vs HyDE

Feature	Naive RAG	HyDE RAG
Embeds query directly	■	■
Embeds hypothetical answer	■	■
Handles vague queries	Weak	Strong
Retrieval accuracy	Good	Better

6. Quick Audience Q&A;

Q: What is an embedding?
A: A numerical representation of text meaning, not raw words.

Q: Why embeddings over keywords?
A: They capture meaning, so even different wording retrieves the right doc.

Q: What is indexing?
A: Organizing embeddings in a vector DB for fast semantic search.

Q: Why Naive RAG?
A: Simple and effective for many QA tasks with static knowledge bases.

Q: Why HyDE?
A: Handles vague queries better, improves recall and accuracy.

Q: Do embeddings store text?
A: No, only numbers — privacy safe.

Q: What metric measures similarity?
A: Cosine similarity (angle between vectors).

Q: How often to re-index?
A: Only when documents are updated or new ones added.

Q: Are embeddings the same across models?
A: No — size and quality differ (MiniLM: 384D, OpenAI Ada: 1536D).