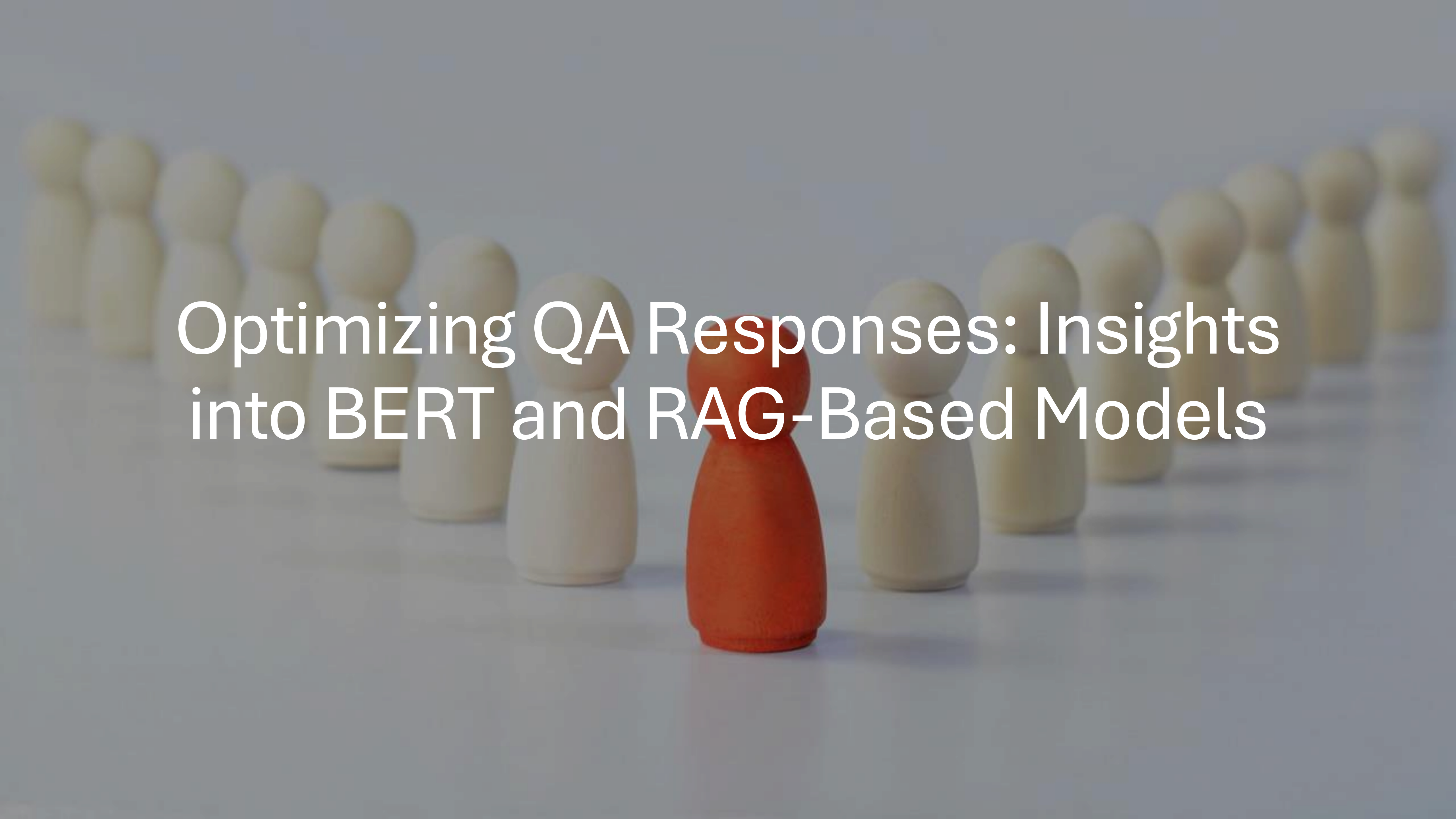
Optimizing QA Responses: Insights into BERT and RAG-Based Models



Background and Objectives

- **Background:**
- Started with a basic BERT model (`deepset/bert-base-cased-squad2`) for QA tasks.
- Document handling included DOCX, PDF, and Excel formats.
- Limited in handling complex queries and large datasets.
- **Objective:**
- Enhance QA system with a sophisticated Retrieval-Augmented Generation (RAG) model.
- Improve handling of complex queries by combining document retrieval with generative language models.

Design Decisions and Approach

- **A. Data Processing and Integration:**
- **Document Processing:**
  - Initial: Used `pandas`, `docx`, `pdfminer` for text extraction.
  - Enhanced: Adopted `PyPDF2`, `pptx`, and `langchain` for advanced handling and embedding.
- **Switch to RAG Model:**
  - Initial: Basic BERT model.
  - RAG Model: Integrated RAG with FAISS for similarity search and document retrieval, used `HuggingFaceEmbeddings` and `Cohere`.

RAG Model Integration

1. Document Embedding and Indexing:

- **Embedding:** `HuggingFaceEmbeddings` with `sentence-transformers/all-MiniLM-L6-v2`.
- **Indexing:** FAISS for fast similarity search.

2. Retriever Setup:

- FAISS-based retriever for document retrieval.

3. Prompt Template:

- Designed prompt to ensure answers based on context.

4. Answer Generation:

- Constructed RAG chain using `Cohere`, `format_docs`, and `generate_answer` function.

API Integration with FastAPI

1. API Design:

- **Endpoints:**
- POST /answer/: Handles QA requests, generates answers, returns result and latency.
- GET /evaluate_latency/: Evaluates latency for sample questions.

2. Error Handling:

- Managed exceptions and provided informative error messages.

3. Latency Measurement:

- Measured latency for QA requests to assess performance.

Technical Details

1. Libraries and Frameworks:

- `HuggingFaceEmbeddings`, `FAISS`, `Cohere`, `langchain`, `FastAPI`.

2. Code Implementation:

- Document processing, RAG chain construction, API endpoints.

Dataset

1. Benchmark Datasets:

- **SQuAD Dataset:**
  - **Description:** Stanford Question Answering Dataset (SQuAD) is a widely used benchmark dataset for evaluating QA systems. It contains questions based on paragraphs from Wikipedia articles.
  - **Purpose:** Provides a standard for assessing model performance on question-answering tasks.

2. Custom QA Dataset:

- **Description:** Created from various document formats including DOCX, PDFs, and Excel files. Contains 48 questions designed to evaluate the model’s ability to handle diverse document types.
- **Purpose:** Tailored to test the model's performance on real-world documents and queries beyond the standard SQuAD dataset.

Performance Evaluation Metrics

1. Initial BERT Model Metrics:

a. **Latency:** Time taken to generate answers using the basic BERT model.

b. **Accuracy:** Proportion of correct answers among all questions.

c. **F1 Score:** Harmonic mean of precision and recall, measuring the model's balance between precision and recall.

2. RAG Model Metrics:

a.Latency: Time taken for the RAG-based model to generate answers. Typically higher due to the complexity of retrieval and generation processes.

b.Similarity Metrics:

Average Cosine Similarity: Measures the cosine of the angle between the embeddings of the generated answer and the reference answers, indicating how similar they are in vector space.

Average Embedding Similarity: Measures the similarity between the embeddings of the generated answers and reference answers, indicating the quality of embeddings used.

Average BERTScore: Measures the relevance and quality of generated answers using contextual embeddings from the BERT model. Higher BERTScore indicates better alignment with reference answers.

Summary and Recommendations

- **1. Performance Insights:**
- Initial BERT model: Faster responses but less accurate for complex queries.
- RAG model: Sophisticated but higher latency and lower similarity scores.
- **2. Recommendations:**
- **Optimization:** Reduce latency in retrieval and generation.
- **Enhancement:** Fine-tune or explore alternative models.
- **Evaluation:** Continuous evaluation with additional metrics.

THANK YOU