TEAM 3

Production-Style RAG System for Neural Network PDFs

MOTIVATION & PROBLEM

- STUDENTS READ LONG NEURAL NETWORK PDFS (PAPERS, NOTES, SLIDES).
- Manual search (scrolling, Ctrl+F) is slow and error-prone.
- LLMs alone may give generic or syllabus-misaligned answers.
- Goal: Answer course questions grounded in the actual PDFs.

What is Retrieval-Augmented Generation (RAG)?

- Pipeline: Retrieval + Generation.
- Retriever finds top-k relevant chunks from a document corpus.
- LLM generates answers using retrieved chunks as context.
- Benefits: fewer hallucinations, course alignment, traceability.

Data & Ingestion Pipeline

- FOLDER /CONTENT/MATERIALS CONTAINS MULTIPLE PDFS:
  - 1810.04805v2.pdf (BERT), 1706.03762v7.pdf (Transformer)
    - ResNets.pdf, Dropouts-1.pdf, Back-Propagation-1.pdf, RNN,LSTM,GRU.pdf
- Use PdfReader to extract text page by page.
- Normalize whitespace with ' '.join(full_text.split()).
- Store each document as {id, filename, text} in documents list.

Chunking Strategies & Statistics

- FIXED-SIZE WORD CHUNKS: 200 WORDS PER CHUNK.
  - Produced 270 fixed-size chunks across all PDFs.
- Sentence-based chunks: 3 sentences per chunk.
  - Produced 3,700 sentence-based chunks.
  - Trade-off: uniform length vs. better semantic coherence.

Embeddings & Retrieval

- Open-source embeddings: SentenceTransformer all-MiniLM-L6-v2.
- Fixed-size chunks: embeddings_open_fixed shape (270, 384).
- Sentence-based chunks: embeddings_open_sentence shape (3700, 384).
- Closed-source embeddings: OpenAI text-embedding-3-small (optional).
- Custom cosine_similarity + retrieve_top_k for top-k chunk retrieval.

Example Query 1: 'Explain what a dropout layer does'

- Retriever uses open-source embeddings over fixed-size chunks.
- Top result: Dropouts-1.pdf, chunk 2 (score ≈ 0.541).
- Chunk explains dropout as a powerful regularization technique:
  - Randomly “drops” neurons during training.
- Other hits include dropout citations and related deep network sections.

Example Query 2: 'Where is the dropout part?'

- Top-5 chunks (open-source, fixed-size):

• 1706.03762v7.pdf (chunk 27), RNN,LSTM,GRU.pdf (chunk 21),
• Dropouts-1.pdf (chunk 2), 1706.03762v7.pdf (chunk 17), ResNets.pdf (chunk 4).

- Similarity scores from ≈ 0.312 down to ≈ 0.233.
- Shows that retriever surfaces both dropout-focused and related contexts.

RAG vs Baseline Answers

- With RAG (context from retrieved chunks):

• Answer focuses on dropout as NN regularisation.

• Explains random deactivation of neurons and improved generalisation.

- Without RAG (no context):

• Gives multiple meanings of 'dropout' (education, health, NN).

• Less aligned with the specific neural network course setting.

Discussion & Future Work

- RAG improves specificity and course alignment for NN questions.
- Fixed-size vs sentence-based chunks: speed vs conceptual coherence.
- Open-source vs OpenAI embeddings: cost vs slight semantic gains.
- Future work:

• Metadata-aware retrieval, hybrid search, UI for students,
• Labelled evaluation set for quantitative metrics.

Thank you!

RAJ SINGH
SRISHTI SRINIVASAN
VRINDA THAKUR