Cairo University

Faculty of Engineering

Electronics and Electrical Communications Engineering

Assignment 3

ELC 4028 – Artificial Neural Networks

ID	Section	Name
9220399	2	Shahd Gamal Mahmoud Hussein
9220411	2	Salah El-Din Hassan Hassan Mohamed
9220984	4	Yousef Khaled Omar Mahmoud

Supervised by:

Dr. Mohsen Rashwan
Dr. Mohamed Motaz

Contents

1 Problem 1 VAE Synthetic Data with Low-Data Stabilization 3
1.1 Conditional Variational Autoencoder (VAE) . . . 3
1.1.1 Pipeline Overview . . . 3
1.1.2 Pipeline Flowchart . . . 5
1.1.3 Original MNIST Samples . . . 6
1.1.4 Data Augmentation . . . 6
1.1.5 VAE Architecture . . . 6
1.1.6 Generated Samples During Training . . . 7
1.1.7 Synthetic Dataset Construction . . . 8
1.1.8 Set A Samples . . . 8
1.1.9 Set B Samples . . . 8
1.1.10 Set C Samples . . . 9
1.1.11 Experimental Results . . . 9
1.1.12 Discussion . . . 9
1.1.13 Conclusion . . . 10

2 Problem 2 GAN Synthetic Data with Low-Data Stabilization 11
2.1 Conditional Generative Adversarial Network (GAN) . . . 11
2.1.1 Pipeline Overview . . . 11
2.1.2 Pipeline Flowchart . . . 13
2.1.3 Augmented Training Samples . . . 14
2.1.4 GAN Architecture . . . 14
2.1.5 GAN Training Curves . . . 14
2.1.6 Discriminator Mistake Analysis . . . 17
2.1.7 Generated GAN Samples . . . 19
2.1.8 Synthetic Dataset Construction . . . 19
2.1.9 Set A Samples . . . 20
2.1.10 Set B Samples . . . 20
2.1.11 Set C Samples . . . 20
2.1.12 Experimental Results . . . 20
2.1.13 Summary Table for Problem 1 and Problem 2 . . . 21
2.1.14 Discussion . . . 21
2.1.15 Conclusion . . . 22

3 Problem 3 Impact of Attention Mechanisms 23
3.1 Place holder for problem 3 . . . 23

4 Problem 4 General Information Retrieval 24
4.1 System Overview . . . 24
4.2 Selected Book . . . 24
4.3 Preprocessing and Chunking . . . 24
4.4 Embedding and Indexing . . . 25
4.5 Language Model . . . 25
4.6 Retrieval Methods . . . 25
4.7 System Implementation . . . 25
4.7.1 `main.py` — Core Pipeline . . . 25

4.7.2 `web_ui.py` — Interactive Web Interface 26
4.7.3 Web UI Screenshots 27
4.8 Results from 10 Arabic Queries (Classical vs. Semantic) 28
4.9 Answers for 10 Queries: RAG vs. LLM-only 33
4.10 Comparison and Reflection 35
4.10.1 Classical vs. Semantic Retrieval 35
4.10.2 RAG vs. LLM-only Generation 36
4.10.3 Conclusions 36
4.11 Tools and Libraries Used 37

5 Problem 5 Benchmarking Arabic NLP Tasks 38
5.1 Place holder for Problem 1 38

List of Figures

1.1 Colored pipeline of the Conditional VAE synthetic data generation framework. . . . 5
1.2 Original MNIST samples. . . . 6
1.3 Examples of augmented MNIST samples using rotation, shifting, scaling, and noise injection. . . . 6
1.4 Generated samples from VAE Run 1. . . . 7
1.5 Generated samples from VAE Run 2. . . . 7
1.6 Generated samples from VAE Run 3. . . . 7
1.7 Generated samples from VAE Run 4. . . . 7
1.8 Generated samples from VAE Run 5. . . . 8
1.9 Samples from Set A (all generated samples). . . . 8
1.10 Samples from Set B (high-confidence samples). . . . 8
1.11 Samples from Set C (mid-confidence samples). . . . 9
2.1 Conditional GAN synthetic data generation pipeline. . . . 13
2.2 Augmented MNIST samples used for GAN training. . . . 14
2.3 GAN loss curves for Run 1. . . . 15
2.4 GAN loss curves for Run 2. . . . 15
2.5 GAN loss curves for Run 3. . . . 16
2.6 GAN loss curves for Run 4. . . . 16
2.7 GAN loss curves for Run 5. . . . 17
2.8 Discriminator mistakes for GAN Run 1. . . . 17
2.9 Discriminator mistakes for GAN Run 2. . . . 18
2.10 Discriminator mistakes for GAN Run 3. . . . 18
2.11 Discriminator mistakes for GAN Run 4. . . . 18
2.12 Discriminator mistakes for GAN Run 5. . . . 19
2.13 Generated samples from GAN Run 1. . . . 19
2.14 Samples from Set A (all generated samples). . . . 20
2.15 Samples from Set B (high-confidence samples). . . . 20
2.16 Samples from Set C (mid-confidence samples). . . . 20
2.17 Classifier accuracy comparison using GAN-generated datasets. . . . 21
4.1 Landing Page — Main Web UI Interface . . . 27
4.2 Classical vs. Semantic Retrieval Example — Side-by-Side Comparison . . . 27
4.3 RAG vs. LLM-only Example — Grounded vs. Hallucinated Answers . . . 28

1 Problem 1 VAE Synthetic Data with Low-Data Stabilization

1.1 Conditional Variational Autoencoder (VAE)

In this experiment, a Conditional Variational Autoencoder (CVAE) was implemented to generate synthetic MNIST handwritten digit images using only a limited real dataset of 350 samples per digit.

The main objective was to investigate whether synthetic data generated by a VAE can improve the performance of a LeNet-5 classifier trained on a small dataset.

1.1.1 Pipeline Overview

The implemented pipeline is summarized as follows:

1. Load and preprocess MNIST dataset.
2. Reduce the training dataset to only 350 real samples per digit.
3. Apply data augmentation using:
   - Rotation
   - Translation
   - Scaling
   - Gaussian noise
4. Train a Conditional VAE using the augmented dataset.
5. Repeat VAE training for 5 independent runs using different random weight initialization.
6. Generate 1000 synthetic images per digit from each VAE run.
7. Train a LeNet-5 classifier using only the 350 real samples.
8. Classify generated samples and compute:

$$\text{confidence} = \max(\text{softmax output})$$

9. Create three synthetic datasets:
   - Set A: all generated samples
   - Set B: confidence ≥ 0.9
   - Set C: $0.6 \leq$ confidence < 0.9
10. Retrain LeNet-5 using:
    - 350 real samples only
    - 350 real + Set A
    - 350 real + Set B
    - 350 real + Set C

11. Compare against:
    - 350-real baseline
    - 1000-real baseline

1.1.2 Pipeline Flowchart

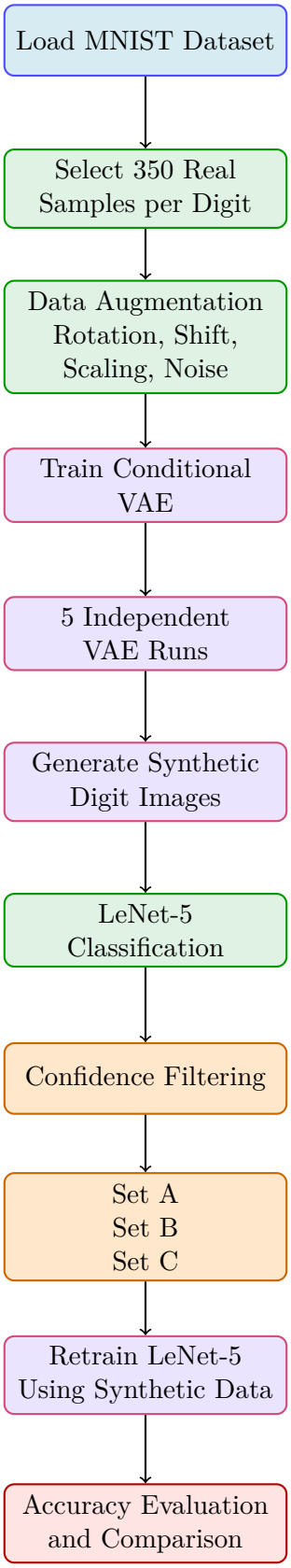

Figure 1.1: Colored pipeline of the Conditional VAE synthetic data generation framework.

1.1.3 Original MNIST Samples

Figure 1.2 shows examples from the original MNIST dataset.

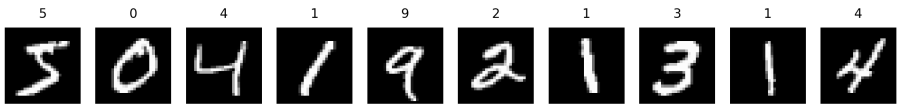

Figure 1.2: Original MNIST samples.

1.1.4 Data Augmentation

To compensate for the limited dataset size, multiple augmentation techniques were applied.

These augmentations improve generalization and help the VAE learn digit variations. Figure 1.3 shows augmented examples.

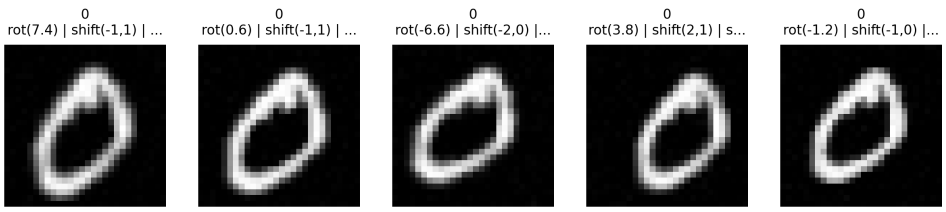

Figure 1.3: Examples of augmented MNIST samples using rotation, shifting, scaling, and noise injection.

1.1.5 VAE Architecture

The implemented Conditional VAE consists of:

- Encoder:
  - Convolutional layers
  - Latent mean and variance estimation
- Decoder:
  - Dense projection
  - Reshape layer
  - Transposed convolution layers
  - Sigmoid output layer

The latent dimension was selected as:

$$z \in \mathbb{R}^8$$

The loss function combines:

- Binary Cross Entropy reconstruction loss
- KL-divergence regularization

1.1.6 Generated Samples During Training

Five independent VAE runs were trained using different random initialization.
Figures 1.4 – 1.8 show generated samples from different runs.

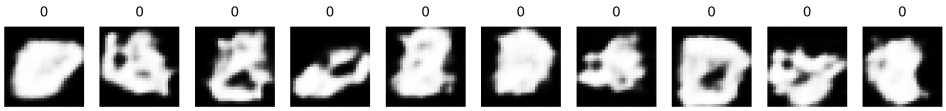

Figure 1.4: Generated samples from VAE Run 1.

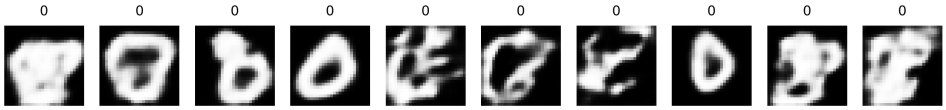

Figure 1.5: Generated samples from VAE Run 2.

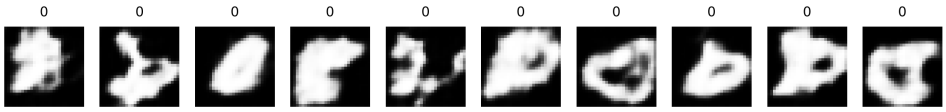

Figure 1.6: Generated samples from VAE Run 3.

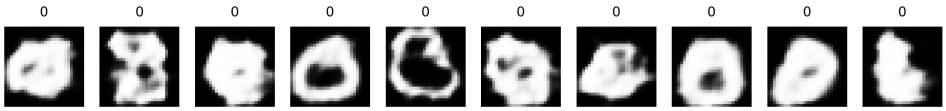

Figure 1.7: Generated samples from VAE Run 4.

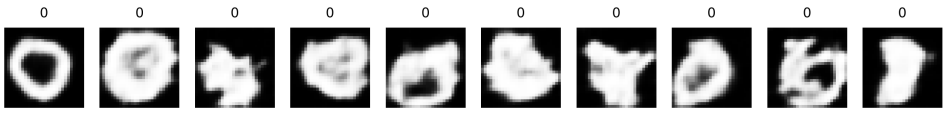

Figure 1.8: Generated samples from VAE Run 5.

The generated images demonstrate that the Conditional VAE successfully learned the structural characteristics of handwritten digits while maintaining variation across samples.

1.1.7 Synthetic Dataset Construction

Generated samples were filtered using the trained LeNet-5 classifier confidence.

Three datasets were created:

- **Set A**: all generated samples
- **Set B**: high-confidence samples
- **Set C**: mid-confidence samples

1.1.8 Set A Samples

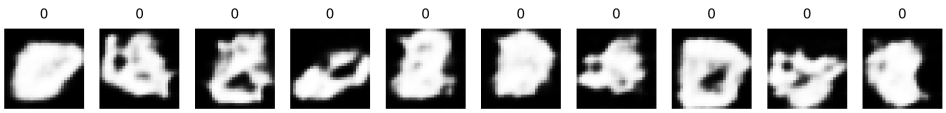

Figure 1.9: Samples from Set A (all generated samples).

Set A contains all generated images without filtering. Some samples are realistic while others contain distortions and noisy structures.

1.1.9 Set B Samples

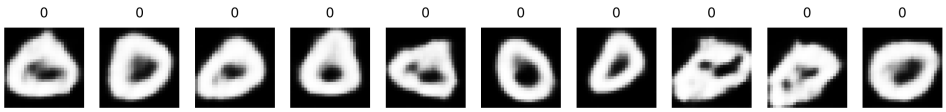

Figure 1.10: Samples from Set B (high-confidence samples).

Set B contains cleaner and more realistic samples because only high-confidence predictions were accepted.

1.1.10 Set C Samples

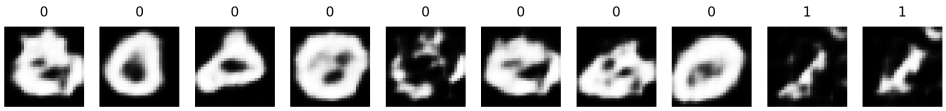

Figure 1.11: Samples from Set C (mid-confidence samples).

Set C contains moderately confident samples that preserve diversity while removing highly corrupted outputs.

1.1.11 Experimental Results

The final LeNet-5 classification accuracies are shown in Table 1.1.

Table 1.1: Comparison of classifier accuracy using different synthetic datasets.

Experiment	**Accuracy**
350 Real Baseline	96.36%
Set A	93.31%
Set B	95.25%
Set C	95.29%
1000 Real Baseline	98.17%

1.1.12 Discussion

The experimental results show that:

- Using only 350 real samples achieved a strong baseline accuracy of 96.36%.
- Adding all generated samples (Set A) reduced performance due to the inclusion of low-quality and distorted generated images.
- Confidence filtering significantly improved performance.
- Set B and Set C achieved accuracies close to the real-only baseline.
- High-confidence filtering successfully removed unrealistic samples while preserving useful digit variations.
- The 1000-real baseline still achieved the best performance because real samples preserve the true MNIST distribution more accurately than synthetic samples.

1.1.13 Conclusion

A Conditional Variational Autoencoder was successfully implemented for MNIST handwritten digit generation.

The generated samples demonstrated that the VAE learned meaningful latent representations despite training on a very limited dataset.

Experimental evaluation showed that synthetic data quality strongly affects classifier performance. Confidence-based filtering improved the usefulness of generated samples and produced classification accuracy close to the real-only baseline.

The results indicate that VAE-generated data can partially compensate for limited datasets, although synthetic samples still do not fully replace real training data.

2 Problem 2 GAN Synthetic Data with Low-Data Stabilization

2.1 Conditional Generative Adversarial Network (GAN)

In this experiment, a Conditional Generative Adversarial Network (CGAN) was implemented to generate synthetic MNIST handwritten digit images using a limited dataset containing only 350 real samples per digit.

The main objective was to investigate whether GAN-generated synthetic data can improve classifier performance and reduce the dependence on large real datasets.

In addition, the experiment compares GAN-generated data against traditional augmentation techniques used in Assignment 2.

2.1.1 Pipeline Overview

The implemented GAN pipeline is summarized as follows:

1. Load and preprocess the MNIST dataset.
2. Reduce the dataset to 350 real samples per digit.
3. Apply augmentation using:
   - Rotation
   - Translation
   - Scaling
   - Gaussian noise
4. Train a Conditional GAN using the augmented dataset.
5. Repeat GAN training for 5 independent runs using different random initialization.
6. Generate synthetic digit samples from each GAN model.
7. Train a LeNet-5 classifier using the reduced real dataset.
8. Classify generated images and compute:

$$\text{confidence} = \max(\text{softmax output})$$

9. Construct three synthetic datasets:
   - Set A: all generated samples
   - Set B: confidence ≥ 0.9
   - Set C: $0.6 \leq$ confidence < 0.9
10. Retrain LeNet-5 using:
    - 350 real samples only
    - 350 real + Set A
    - 350 real + Set B

- 350 real + Set C

11. Compare against:

    - 350-real baseline
    - 1000-real baseline
    - Augmentation-only results from Assignment 2

2.1.2 Pipeline Flowchart

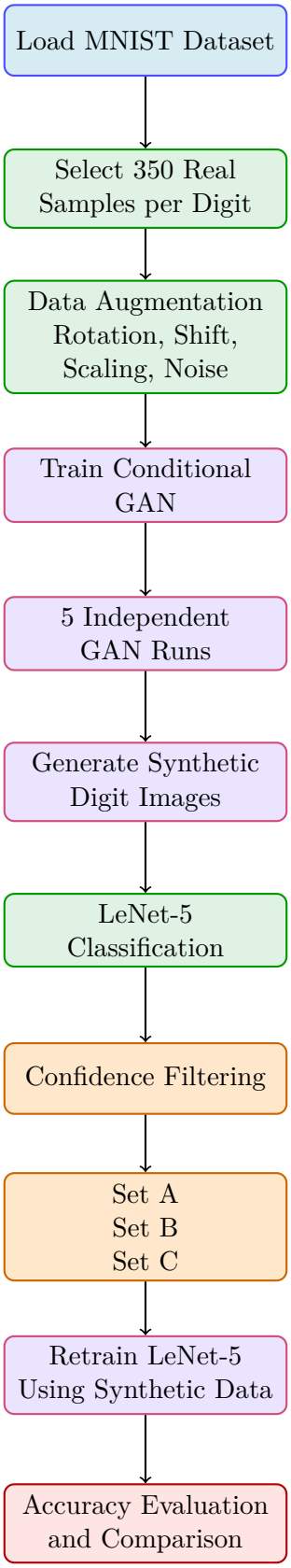

Figure 2.1: Conditional GAN synthetic data generation pipeline.

2.1.3 Augmented Training Samples

Figure 2.2 shows examples of augmented MNIST samples used during GAN training.

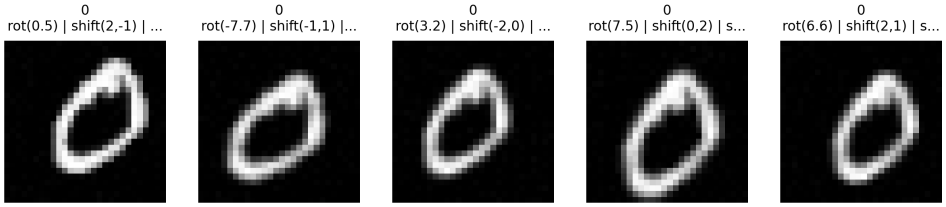

Figure 2.2: Augmented MNIST samples used for GAN training.

2.1.4 GAN Architecture

The implemented Conditional GAN consists of:

- **Generator**
  - Latent noise input
  - Label embedding
  - Dense projection
  - Transposed convolution layers
  - Tanh output activation
- **Discriminator**
  - Convolutional feature extraction
  - Label conditioning
  - LeakyReLU activations
  - Binary classification output

The GAN was trained using adversarial learning where:

- The generator attempts to fool the discriminator.
- The discriminator attempts to distinguish real from fake images.

2.1.5 GAN Training Curves

Figures 2.3 – 2.7 show discriminator and generator losses for the five independent GAN runs.

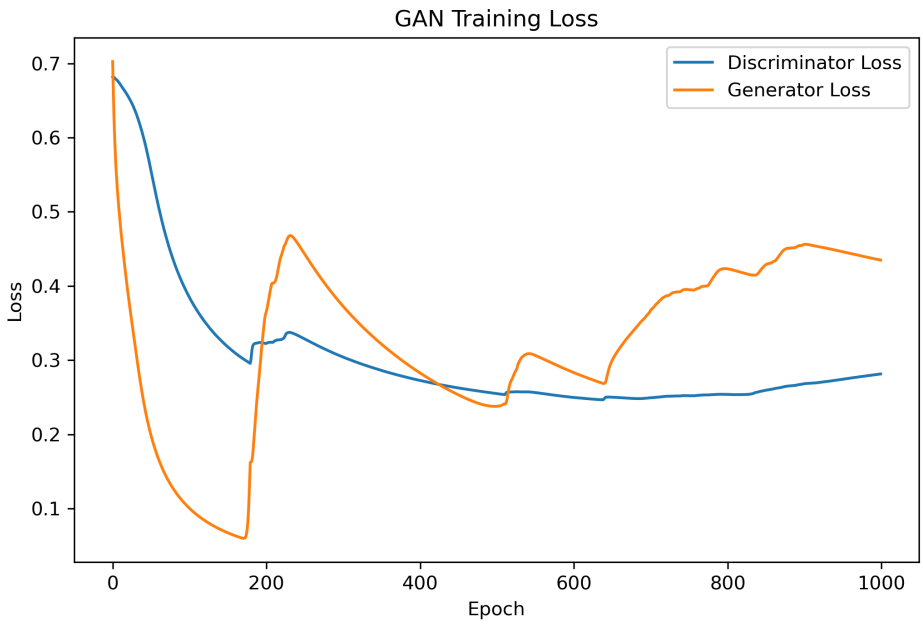

Figure 2.3: GAN loss curves for Run 1.

Figure 2.4: GAN loss curves for Run 2.

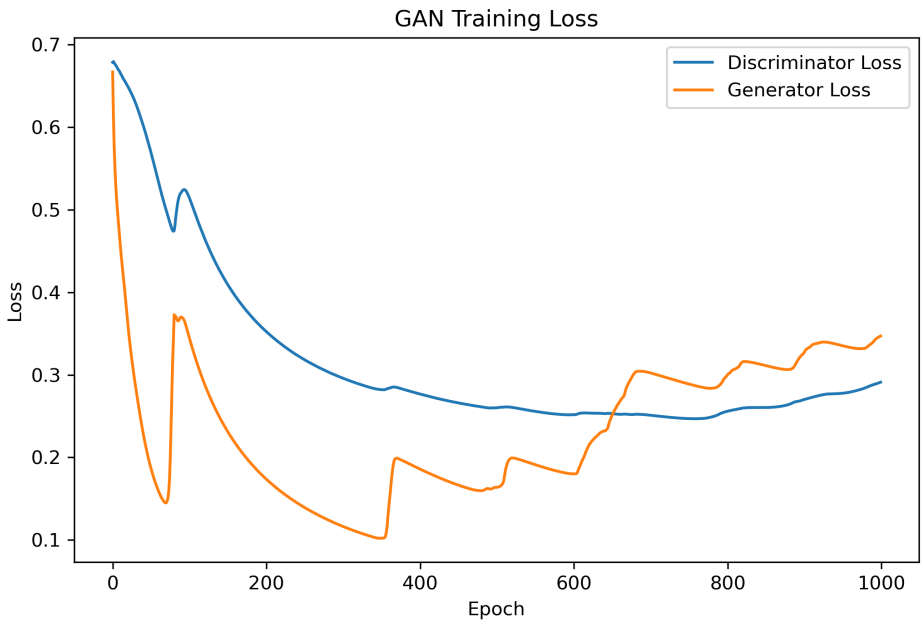

Figure 2.5: GAN loss curves for Run 3.

Figure 2.6: GAN loss curves for Run 4.

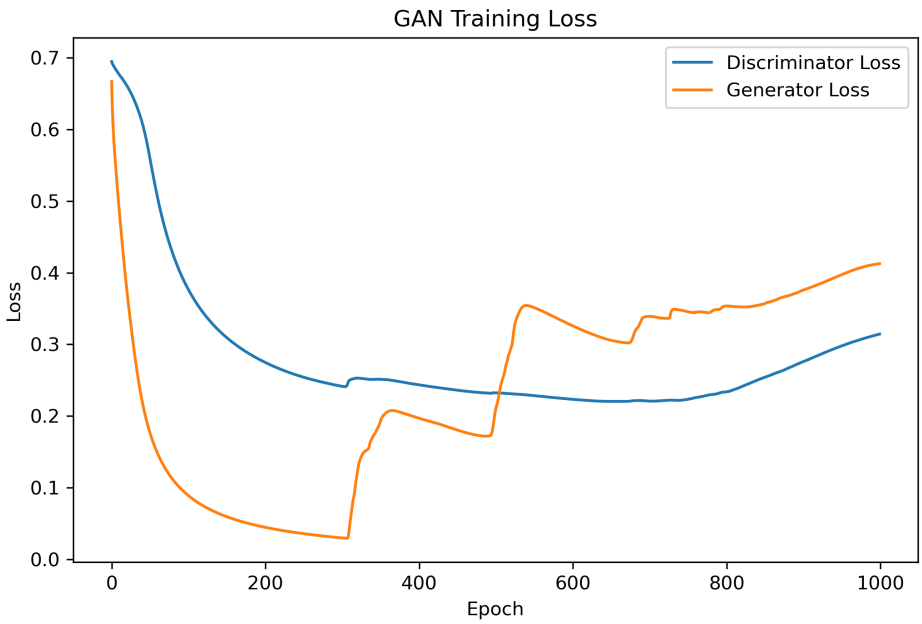

Figure 2.7: GAN loss curves for Run 5.

The loss curves demonstrate unstable adversarial behavior during training. Generator and discriminator losses oscillate significantly across runs, which indicates unstable convergence.

2.1.6 Discriminator Mistake Analysis

Figures 2.8 – 2.12 show discriminator classification mistakes.

Green labels indicate fake images classified as real.

Red labels indicate real images classified as fake.

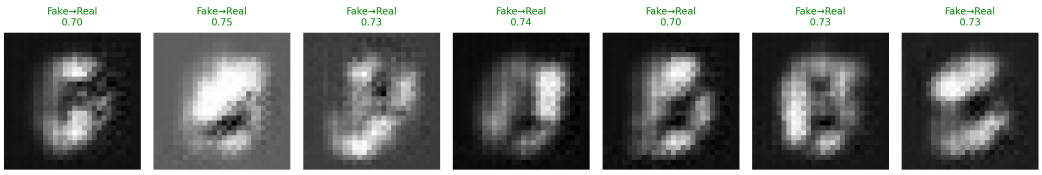

Figure 2.8: Discriminator mistakes for GAN Run 1.

Figure 2.9: Discriminator mistakes for GAN Run 2.

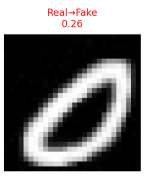

Figure 2.10: Discriminator mistakes for GAN Run 3.

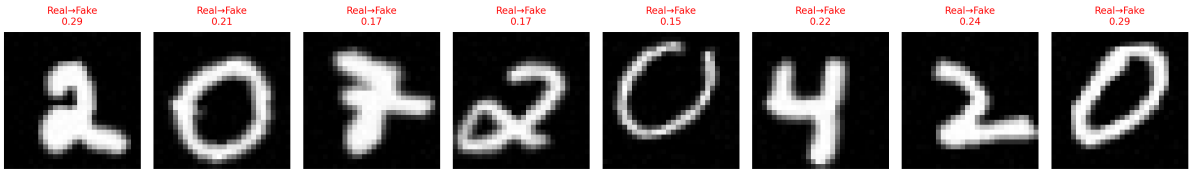

Figure 2.11: Discriminator mistakes for GAN Run 4.

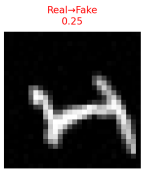

Figure 2.12: Discriminator mistakes for GAN Run 5.

The discriminator mistakes show that some generated samples successfully fooled the discriminator despite containing visible distortions and unrealistic structures.

2.1.7 Generated GAN Samples

Figure 2.13 shows generated samples from one GAN run.

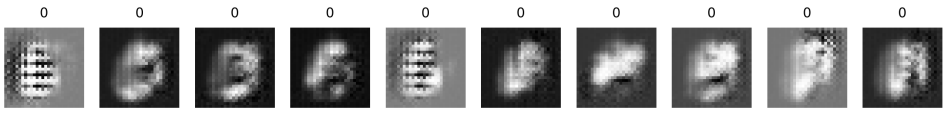

Figure 2.13: Generated samples from GAN Run 1.

The generated digits demonstrate that the GAN learned general digit structure. However, many samples remain blurry or contain artifacts.

Some generated samples appear realistic, while others collapse into noisy or distorted patterns.

2.1.8 Synthetic Dataset Construction

Generated samples were filtered using LeNet-5 classifier confidence.

Three datasets were constructed:

- **Set A**: all generated samples
- **Set B**: confidence ≥ 0.9
- **Set C**: $0.6 \leq$ confidence < 0.9

2.1.9 Set A Samples

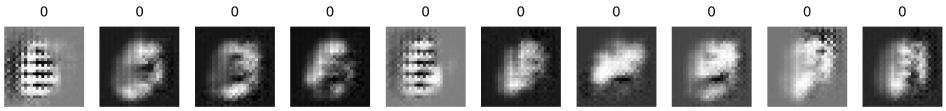

Figure 2.14: Samples from Set A (all generated samples).

Set A includes all generated images. Many samples contain severe artifacts and unstable structures.

2.1.10 Set B Samples

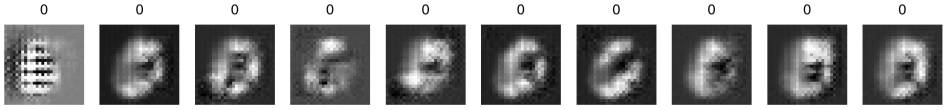

Figure 2.15: Samples from Set B (high-confidence samples).

Set B contains higher-quality samples selected using confidence filtering.

2.1.11 Set C Samples

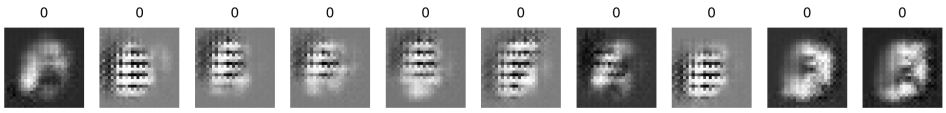

Figure 2.16: Samples from Set C (mid-confidence samples).

Set C preserves moderately confident samples while removing highly corrupted outputs.

2.1.12 Experimental Results

Figure 2.17 summarizes classifier performance using GAN-generated synthetic datasets.

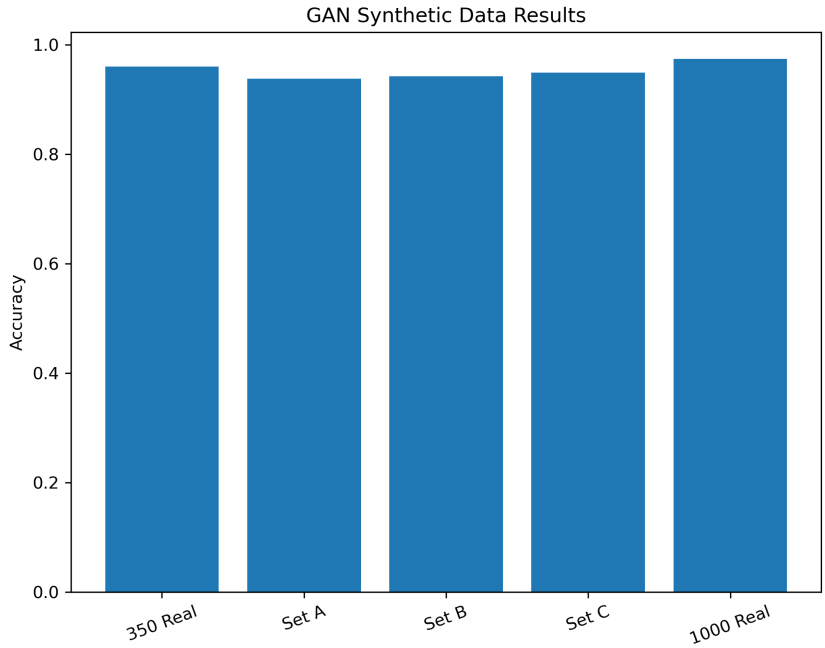

Figure 2.17: Classifier accuracy comparison using GAN-generated datasets.

2.1.13 Summary Table for Problem 1 and Problem 2

Table 2.1 summarizes the results of both VAE and GAN approaches.

Table 2.1: Comparison of VAE and GAN synthetic data strategies.

Model Used	No. of Examples	Confidence Level	Accuracy
VAE	All generated data	Set A	93.31%
VAE	CL > 0.9	Set B	95.25%
VAE	$0.6 <$ CL < 0.9	Set C	95.29%
GAN	All generated data	Set A	93.8%
GAN	CL > 0.9	Set B	94.4%
GAN	$0.6 <$ CL < 0.9	Set C	94.8%
350 Real Baseline	Real only	—	96.36%
1000 Real Baseline	Real only	—	98.17%

2.1.14 Discussion

The experimental results demonstrate several important observations:

- GAN training was significantly more unstable than VAE training.
- The adversarial nature of GAN optimization caused oscillations between generator and discriminator performance.
- Generated GAN samples varied greatly in quality across runs.
- Some generated digits appeared realistic, while others were blurry or highly distorted.

- Confidence filtering improved the usefulness of generated samples by removing low-quality outputs.
- Set B and Set C consistently performed better than Set A because filtering removed unrealistic samples.
- Despite improvements, GAN-generated data still did not outperform the real-only baseline.
- Compared to Assignment 2 augmentation, traditional augmentation remained more stable and reliable.
- The experiments show that synthetic data can partially compensate for limited datasets, but fully replacing real data remains difficult.

2.1.15 Conclusion

A Conditional GAN was successfully implemented for MNIST handwritten digit generation using a highly limited dataset.

The generated samples demonstrated that the GAN learned meaningful digit structures; however, training instability caused large quality variations across runs.

Visual inspection confirmed that some generated samples were realistic while others suffered from mode collapse, blur, and severe artifacts.

Experimental evaluation showed that confidence-based filtering improved classifier performance by selecting more reliable synthetic samples.

Overall, GAN-generated data partially reduced the need for real data, but the results indicate that synthetic GAN samples still cannot fully replace real training samples.

Compared to VAE generation and traditional augmentation, GAN training was less stable and more sensitive to initialization and adversarial dynamics.

3 Problem 3 Impact of Attention Mechanisms

3.1 Place holder for problem 3

test

4 Problem 4 General Information Retrieval

4.1 System Overview

This problem focuses on Arabic Information Retrieval (IR) and Question Answering (QA) using classical keyword retrieval, dense semantic retrieval, and Retrieval-Augmented Generation (RAG). The entire system is built around a single Arabic literary book as the sole knowledge source, allowing a faithful comparison of how each technique handles the same queries.

4.2 Selected Book

The selected book is الأيام (*Al-Ayyam*, "The Days") by طه حسين (Taha Hussein), one of the most celebrated figures in modern Arabic literature. The autobiography narrates Taha Hussein's rural Egyptian childhood, his blindness, and his intellectual journey through religious and secular education — providing linguistically rich, less-commonly indexed text that makes retrieval benefits clearly visible.

- **Book Title:** الأيام (*Al-Ayyam*)
- **Author:** طه حسين (Taha Hussein)
- **Language:** Classical / Modern Standard Arabic
- **Source:** Public-domain Arabic text

4.3 Preprocessing and Chunking

The raw text underwent the following normalization steps:

- Removal of unnecessary whitespace, line breaks, and page numbers
- Removal of Arabic diacritics (tashkeel) via Unicode range
- Character normalization:
  - أ / إ / آ → ا
  - ى → ي
  - ة → ه

The normalized text was split into overlapping chunks:

- **Chunk size:** 4 sentences **Overlap:** 2 sentences (50 % stride)
- Chunks with fewer than 100 Arabic characters were discarded
- Copyright / metadata patterns were filtered out

This yielded approximately **304 chunks** and ∼1 200 sentences from the full text.

4.4 Embedding and Indexing

Dense vector representations were generated for every chunk using:

`sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

This model supports 50+ languages and was chosen for its strong Arabic cross-lingual transfer performance. All 768-dimensional embeddings were ℓ_2-normalized and stored in a FAISS `IndexFlatIP` index, which provides exact cosine-similarity search over unit-norm vectors.

4.5 Language Model

`Qwen/Qwen2.5-1.5B-Instruct`

This 1.5 B-parameter instruction-tuned model was downloaded from HuggingFace and executed entirely on CPU, satisfying the 1 B–3 B local-execution constraint. For RAG generation the model received a structured Arabic system prompt together with the retrieved context; for LLM-only answers it received only the query.

4.6 Retrieval Methods

Classical Retrieval — BM25 (Okapi BM25). Each paragraph is tokenized into Arabic words; common stop-words and words shorter than two characters are removed. Given a query, BM25 assigns a probabilistic relevance score that rewards term frequency in the paragraph while penalizing very frequent terms (IDF) and very long documents. The 5 highest-scoring paragraphs are returned. BM25 is completely insensitive to synonymy: a semantically rich passage that shares no surface tokens with the query always receives a score of zero.

Semantic Retrieval — Embeddings + FAISS. The query is fed through the same multilingual sentence-transformer used at index time. Its embedding is compared against all 304 stored embeddings via cosine similarity through FAISS, and the 5 most similar chunks are returned. This approach captures meaning and paraphrase, making it far more robust for indirect or descriptive Arabic questions.

4.7 System Implementation

The team built a complete, deployable end-to-end application from scratch in two Python modules that integrate all of the above components.

4.7.1 `main.py` — Core Pipeline

`main.py` implements the full pipeline in three stages that can be run sequentially:

- **Stage 1 — Book Preparation:** loads the raw UTF-8 text file, applies all normalization and chunking steps, generates 768-d embeddings in batches of 32, L2-normalizes them, and writes the FAISS index (`faiss_index.bin`), the chunk list (`paragraphs.pkl`), and metadata to disk.

- **Stage 2 — Search Layer:** implements and exposes `classical_search()` and `semantic_search()` functions. `classical_search` builds a BM25Okapi corpus from tokenized chunks (falling back to a TF-IDF implementation if `rank_bm25` is unavailable); `semantic_search` normalizes the query vector and calls FAISS.
- **Stage 3 — RAG Generation:** the `rag_answer()` function retrieves the top-5 semantically relevant passages, formats them into a numbered Arabic context block, and passes this to Qwen2.5-1.5B-Instruct via a structured chat template. In parallel it generates an LLM-only answer from the same model with no context, enabling direct side-by-side comparison.
- **Report Mode:** running `python main.py --report` automatically runs all 10 queries through both retrieval paths and the RAG pipeline, then writes a complete Markdown deliverables report to `output/assignment_report.md`.

4.7.2 `web_ui.py` — Interactive Web Interface

`web_ui.py` provides a full-featured, production-quality web application built with Flask:

- **Dark-theme Arabic-first UI:** a fully RTL (*direction: rtl*) interface rendered in the *Tajawal* Arabic typeface with a professional dark colour scheme, responsive on mobile.
- **Search tab:** accepts free-form Arabic queries, displays the top-k results from BM25 and semantic search side by side in two styled columns with rank badges and colour-coded similarity scores. Query terms are highlighted in the retrieved passages. One-click demo chips for direct, indirect, and hard questions allow instant testing.
- **Q&A (RAG) tab:** submits the query to the RAG endpoint, displays the RAG answer, the LLM-only answer, and the supporting evidence context side-by-side with a loading indicator. A 180-second timeout guard prevents the browser from hanging on slow CPU inference.
- **REST API:** two endpoints (`/search` and `/ask_rag`) are exposed for programmatic use; the RAG endpoint runs generation in a background thread to prevent Flask worker blocking.

The web application runs at `http://localhost:5010` and requires no configuration beyond having the pre-built FAISS index in `output/`.

4.7.3 Web UI Screenshots

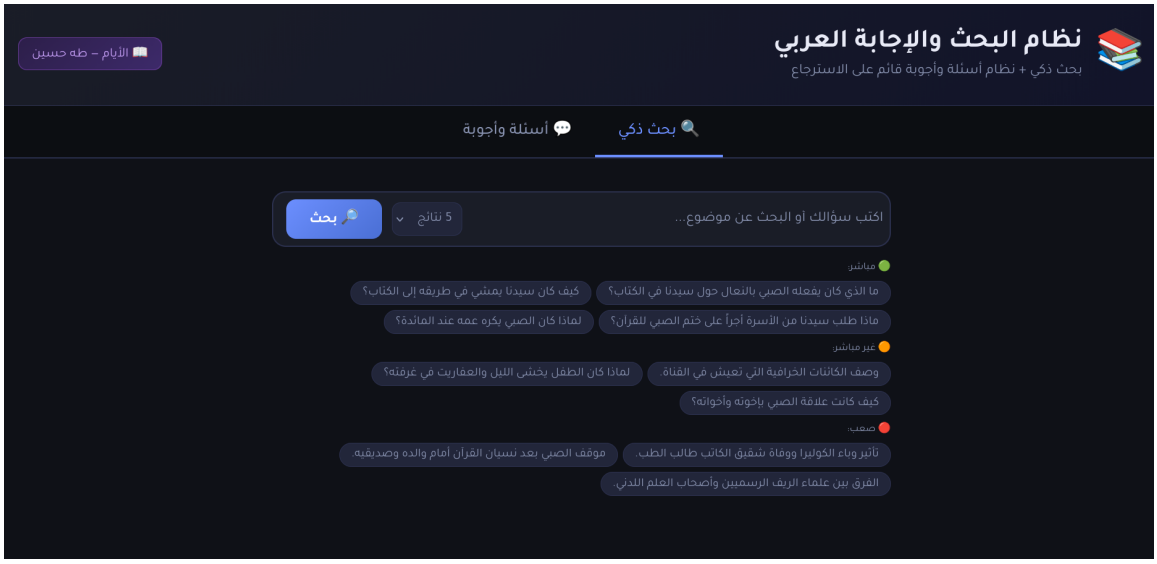

Figure 4.1: Landing Page — Main Web UI Interface

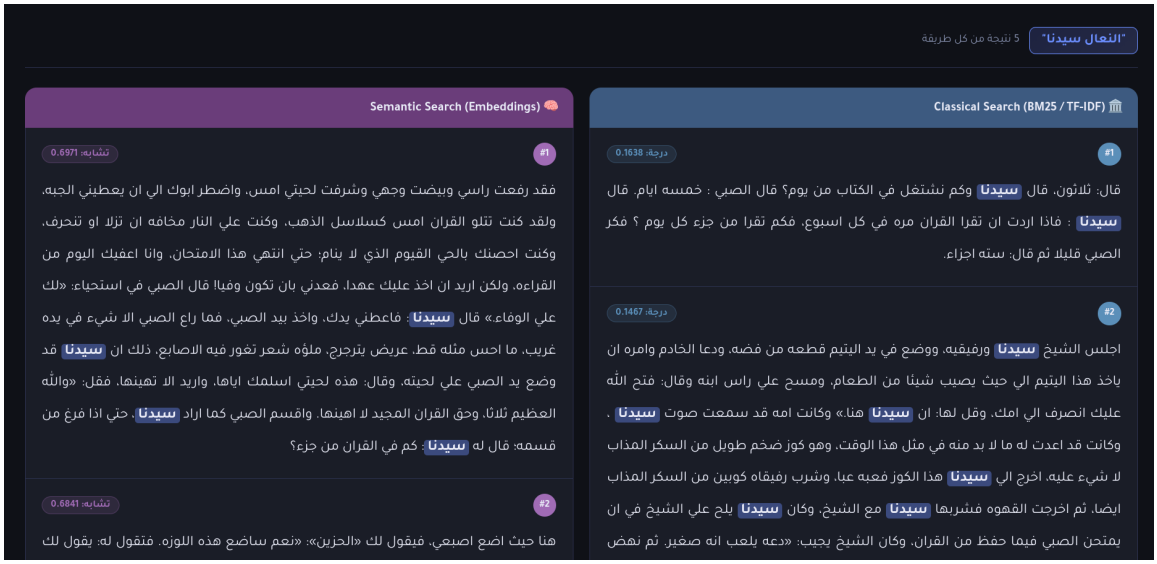

Figure 4.2: Classical vs. Semantic Retrieval Example — Side-by-Side Comparison

Figure 4.3: RAG vs. LLM-only Example — Grounded vs. Hallucinated Answers

4.8 Results from 10 Arabic Queries (Classical vs. Semantic)

Each query is presented with its top-5 results from both methods in a side-by-side bordered table. Queries span three difficulty levels: *direct* (1–3), *indirect* (4–7), and *hard* (8–10).

Query #1: النعال سيدنا

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1638	قال: ثلاثون، قال سيدنا وكم نشتغل في الكتاب من يوم؟ قال الصبي : خمسه ايام...	1	0.6971	فقد رفعت راسي وبيضت وجهي وشرفت لحيتي امس، واضطر ابوك الي ان يعطيني الجبه، ولقد كنت تتلو القران امس كسلاسل الذهب...
2	0.1467	اجلس الشيخ سيدنا ورفيقيه، ووضع في يد اليتيم قطعه من فضه، ودعا الخادم وامره ان ياخذ هذا اليتيم...	2	0.6841	هنا حيث اضع اصبعي، فيقول لك الحزين: نعم ساضع هذه اللوزه. يقول لك سيدنا: يجب ان تتخير لقراءتك...
3	0.1456	لانه الكتاب كان بعيدا، ولانه كان اضعف من ان يقطع ماشيا تلك المسافه، ثم لا يذكر متي بدا يسعي...	3	0.6822	قال يزل هنا يغلط، ويقال: زل عن الصخره ونحوها، اذ زلق عنها وسقط. ما راعني الا كذا: اي ما شعرت الا به. الايام سيدنا...
4	0.1351	الفصل الخامس ولكن لا يعرف كيف حفظ القران، ولا يذكر كيف بداه ولا كيف اعاده...	4	0.6812	ولكن اني لك هذا وانت في التاسعه من عمرك ترين الحياه كلها نعيما وصفوا ! عرفته ينفق اليوم والاسبوع والشهر والسنه...
5	0.1311	قال الصبي في استحياء: لك علي الوفاء. قال سيدنا: فاعطني يدك، واخذ بيد الصبي...	5	0.6797	الفصل الثامن واقبل سيدنا الي الكتاب من الغد مسرورا مبتهجا ، فدعا الشيخ الصبي بلقب الشيخ هذه المره...

Query #2: مشي سيدنا

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1638	قال: ثلاثون، قال سيدنا وكم نشتغل في الكتاب من يوم؟ قال الصبي : خمسه ايام...	1	0.5980	القصب هنا: ضرب من النبت ذو كعوب جوفاء، كانت تتخذ منه الاقلام. الايام كانما كان متلاصقا ، فلم يكن يستطيع ان ينسل في ثنايا...
2	0.1467	اجلس الشيخ سيدنا ورفيقيه، ووضع في يد اليتيم قطعه من فضه، ودعا الخادم وامره...	2	0.5932	هو يذكر هذا السياج كانه راه امس، يذكر ان قصب هذا السياج كان اطول من قامته، فكان من العسير عليه ان يتخطاه...
3	0.1311	قال الصبي في استحياء: لك علي الوفاء. قال سيدنا: فاعطني يدك...	3	0.5744	مصغيا مميلا اذنيه للاستماع. اشراب رفع راسه ومد عنقه لينظر. الايام فان سالتني كيف انتهي الي حيث هو الان...
4	0.1267	الفصل الخامس وكان منظر سيدنا عجبا في طريقه الي الكتاب والي البيت صباحا ومساء، كان ضخما بادنا...	4	0.5624	اغراه به اولعه به وحضه عليه. ابتغوا طلبوا والوسيله : ما يقرب به الي الغير. البيئه بالكسر : اسم من تبوا المكان...
5	0.1254	وكان سيدنا لا يغني بصوته ولسانه وحدهما، وانما يغني براسه وبدنه ايضا؛ فكان راسه يهبط ويصعد...	5	0.5575	قال يزل هنا يغلط، ويقال: زل عن الصخره ونحوها. ما راعني الا كذا. الايام سيدنا: فتقسو علي الصبي...

Query #3: ختم القرآن أجر

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1331	قال المفتش: فانا استطيع ان اجود له القران علي قراءه حفص، حتي اذا ذهب الايام الي الازهر...	1	0.6710	الاواصر هنا العلائق والصلات. الفانيه التي بلغت ارذل العمر. حدت المراه تركت الزينه لموت زوج او حبيب...
2	0.1171	ولكن رمضان اقبل، وكان الناس يجتمعون عند رجل وجيه يعمل في التجاره، وكان سيدنا يقرا القران طوال الشهر...	2	0.6463	قال يزل هنا يغلط. ما راعني الا كذا: اي ما شعرت الا به. الايام سيدنا: فتقسو علي الصبي القراءه...
3	0.1157	فاخذ صاحبنا يردد: طسم. قال سيدنا: عوضني الله خيرا فيما انفقت معك من وقت، فقد نسيت القران...	3	0.6432	القصب هنا: ضرب من النبت ذو كعوب جوفاء. الايام كانما كان متلاصقا ، فلم يكن يستطيع ان ينسل في ثنايا...
4	0.1104	وكانه يري نفسه جالسا عن يمين سيدنا يقرئه: اتامرون الناس بالبر وتنسون انفسكم...	4	0.6426	الولوله: الاعوال والبكاء. الخمش اللطم والصك هنا: الضرب الشديد. الاواصر هنا العلائق والصلات...
5	0.1027	ولكن هذا المساء كان في جملته خيرا من الغد؛ ذهب الي الكتاب، فاذا سيدنا يدعوه في جفوه...	5	0.6422	اغراه به وحضه عليه. ابتغوا طلبوا والوسيله. البيئه بالكسر: اسم من تبوا المكان اذا حله...

Query #4: لماذا كان الصبي يكره عمه عند المائدة؟

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.2078	يوثر عليه اخوته: يفضلهم عليه. الايام فاما العريف فكان يكره سيدنا؛ لانه اثر غشاش كذاب...	1	0.7349	الفصل الرابع قليل الميل الي الطعام؛ لانه كان يخشي ان يوصف بالشره او ان يتغامز عليه اخوته...
2	0.1577	قال اخوه ضاحكا: هو درس الفقه. قال الصبي : ومن الشيخ ؟ قال اخوه...	2	0.7053	حرم علي نفسه الحساء والارز وكل الالوان التي تؤكل بالملاعق؛ لانه لا يحسن اصطناع الملعقه...
3	0.1344	وفي الحق انه لم يفهم لماذا صدق وعد ابيه في هذه السنه...	3	0.6716	الايام ولم يظهر الصبي في هذه الليله علي المائده، ومكث ثلاثه ايام يتجنب مجلس ابيه...
4	0.1273	حرم علي نفسه الحساء والارز وكل الالوان التي تؤكل بالملاعق...	4	0.6659	وما الذي يمنعه من هذه التجربه ؟ فقد اخذ اللقمه بكلتا يديه وغمسها من الطبق المشترك...
5	0.1261	قال: ثلاثون، قال سيدنا وكم نشتغل في الكتاب من يوم؟...	5	0.6209	وما كان شيء يغيظ سيدنا مثلما كان يغيظه هذا الثناء. وقضي الصبي سنه كامله يتردد علي هذا البيت...

Query #5: وصف الكائنات الخرافية التي تعيش في القناة

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1954	الفصل الثاني منظمه، تنحدر كلها من جسر القناه ممتده امتدادا قصيرا من الشمال الي الجنوب...	1	0.6890	فقد كانت تنتهي الي قناه عرفها حين تقدمت به السن، وكان لها في حياته - او قل في خياله - تاثير عظيم...
2	0.1768	الايام يسعي اليه دون لمح البصر خادمان من الجن يقضيان له ما يشاء، ذلك الخاتم الذي كان يتختمه سليمان...	2	0.6681	وهو ان قضيبا اهدي الي حسن هذا، وكان من خواصه ان تضرب به الارض فتنشق ويخرج منها تسعه نفر من الجن...
3	0.1741	لين القناه هنا كنايه عن الرضا. الخطل والحمق : قله العقل وفساده...	3	0.6414	يذكر صاحبنا السياج والمزرعه والقناه التي كانت تنتهي اليها الدنيا، وسعيدا وكوابس وكلاب العدويين...
4	0.1401	تخفت الاصوات تسكن او تضعف. الفصل الثاني كان مطمئنا الي ان الدنيا تنتهي عن يمينه بهذه القناه...	4	0.5873	القصب هنا: ضرب من النبت ذو كعوب جوفاء. الايام كانما كان متلاصقا ، فلم يكن يستطيع ان ينسل في ثنايا...
5	0.1287	وهو يذكر انه كان يقضي ساعات علي شاطئ القناه سعيدا مبتهجا بما سمع من نغمات حسن الشاعر...	5	0.5762	هو يذكر هذا السياج كانه راه امس، يذكر ان قصب هذا السياج كان اطول من قامته...

Query #6: لماذا كان الطفل يخشى الليل والعفاريت في غرفته؟

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1032	انظري اليه هو هذا الملك القائم الذي يحنو علي سريرك اذا امسيت لتستقبلي الليل في هدوء...	1	0.6677	وكان يعتقد ان ليس له حصن من كل هذه الاشباح المخوفه والاصوات المنكره؛ الا ان يلتف في لحافه من الراس الي القدم...
2	0.0803	وفي الحق انه لم يفهم لماذا صدق وعد ابيه في هذه السنه؛ فقد اخبر الصبي ذات يوم انه مسافر...	2	0.6446	ثم ياخذه النوم، فما يحس الا وقد استيقظ والناس نيام، ومن حوله لا يجدي عليه خيرا...
3	0.0771	وكيف استطاع ان يثير في نفوس كثير من الناس ما يثير من حسد وحقد وضغينه...	3	0.6397	وما كان شيء يغيظ سيدنا مثلما كان يغيظه هذا الثناء. وقضي الصبي سنه كامله يتردد علي هذا البيت...
4	0.0735	الايام يسعي اليه دون لمح البصر خادمان من الجن يقضيان له ما يشاء...	4	0.6277	الايام اخوته واخواته يغطون فيسرفون في الغطيط، فيلقي اللحاف عن وجهه، وكان واثقا انه ان كشف وجهه عبث به عفريت...
5	0.0654	هذه الحادثه اخذته بالوان من الشده في حياته، جعلته مضرب المثل في الاسره...	5	0.6144	وكان مصدر هذا كله صوت هذا الفتي وهو يعالج القيء. وكان الفتي قضي ساعه او ساعتين يخرج من الحجره...

Query #7: كيف كانت علاقة الصبي بإخوته وأخواته؟

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.1598	ولم يبلغ التاسعه من عمره حتي كان قد وعي من الاغاني والقصص وشعر الهلاليين جمله صالحه...	1	0.7333	كان يحس من امه رحمه ورافه، وكان يجد من ابيه لينا ورفقا، وكان يشعر من اخوته بشيء من الاحتياط في تحدثهم اليه...
2	0.1577	قال اخوه ضاحكا: هو درس الفقه. قال الصبي : ومن الشيخ ؟...	2	0.6962	واستمرت الحال كذلك اعواما، ثم تقدمت به السن، وعمل فيه الازهر عمله، فاخذت عله اخيه تتمثل له من حين الي حين...
3	0.1414	الفصل الخامس ولكنه لا يعرف كيف حفظ القران، ولا يذكر كيف بداه ولا كيف اعاده...	3	0.6794	وكان يجد الي جانب هذا اللين والرفق من ابيه شيئا من الاهمال ايضا، وكان احتياط اخوته واخواته يؤذيه...
4	0.1414	مصغيا مميلا اذنيه للاستماع. الايام فان سالتني كيف انتهي الي حيث هو الان؟...	4	0.6736	اكان هذا المكان يرضيه ؟ الحق انه لا يستطيع ان يحكم في ذلك حكما صادقا. كان يحس من امه رحمه ورافه...
5	0.1268	ان هي المؤكده المخففه. الايام فان سالتني كيف انتهي الي حيث هو الان؟...	5	0.6468	واخذ اخوه الصبي يستعدون لهذا العيد؛ يختلف كبارهم الي الخياط، ويلهو الايام صغارهم بهذه الحركه الطارئه...

Query #8: تأثير وباء الكوليرا ووفاة شقيق الكاتب طالب الطب

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.0872	حدت المراه تركت الزينه لموت زوج او حبيب، والمراد بالحداد هنا الحزن. الفصل الثامن عشر...	1	0.5417	كان هذا اليوم يوم 21 اغسطس من سنه 1902 ، وكان وباء الكوليرا قد هبط الي مصر ففتك باهلها فتكا ذريعا...
2	0.0721	كان هذا اليوم يوم 21 اغسطس من سنه 1902 ، وكان وباء الكوليرا قد هبط الي مصر...	2	0.4940	فاما خارج الدار فكان يزدحم بالناس اقبلوا الي الشيخ يواسونه. واما داخل الدار فكان يزدحم بالنساء...
3	0.0655	وكان هذا الشيخ من الذين لم يفلحوا في الازهر؛ قضي فيه ما شاء من السنين...	3	0.4926	اذن فقد اصيب الشاب، ووجد الوباء طريقه الي الدار، وعرفت ام الفتي باي ابنائها تنزل النازله...
4	0.0647	القصب هنا: ضرب من النبت ذو كعوب جوفاء، كانت تتخذ منه الاقلام...	4	0.4786	وعلي هذا النحو فقد صبينا عينيه؛ اصابه الرمد فاهمل اياما ، ثم دعي الحلاق فعالجه...
5	0.0574	وكان يسمع لهم وهم يتكلمون، فياخذه شيء من الاعجاب والدهش...	5	0.4674	غثت النفس غثيا وغثيانا. الايام البيت جميعا ان في اكل الثوم وقايه من الكوليرا...

Query #9: موقف الصبي بعد نسيان القرآن أمام والده وصديقيه

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.2071	قال: ثلاثون، قال سيدنا وكم نشتغل في الكتاب من يوم؟...	1	0.7656	وفي الحق انه لم يفهم لماذا صدق وعد ابيه في هذه السنه؛ فقد اخبر الصبي ذات يوم انه مسافر...
2	0.1909	قال الصبي في استحياء: لك علي الوفاء. قال سيدنا: فاعطني يدك...	2	0.7102	قال ابوه فاقرا سوره القصص، فذكر انها الثالثه، واخذ يردد: طسم، ولم يفتح عليه ابوه...
3	0.1812	واخذ الصبي يقرا القران علي المفتش من اوله، واخذ المفتش يعلمه مواضع الوقف والوصل...	3	0.6993	سمع الناس هذا فاضطربوا، وكادت تقع بينهم الفتنه لولا ان نهض الامام فخطبهم وصلي بهم...
4	0.1790	الهمهمه: الكلام الخفي. الفصل التاسع وكان الصبي يجيب من البقره الي لتجدن في يوم السبت...	4	0.6875	وما الذي يمنعه من هذه التجربه ؟ فقد اخذ اللقمه بكلتا يديه وغمسها من الطبق المشترك...
5	0.1625	ومن غريب الامر ان الشيخ لم يفكر مره واحده في ان يفتح الالفيه ويقابل علي الصبي وهو يقرا...	5	0.6840	ولا يحفل بالشيء: لا يبالي به. اغرقوا في الضحك بالغوا فيه. اجهشت بالبكاء: همت به...

Query #10: الفرق بين علماء الريف الرسميين وأصحاب العلم اللدني

Classical Retrieval (BM25)			Semantic Retrieval (Embeddings)		
#	Sc.	Passage	#	Sc.	Passage
1	0.2093	وكان في المدينه عالم اخر شافعي، كان امام المسجد، وصاحب الخطبه والصلاه، معروفا بالتقي والورع...	1	0.7123	الفصل الرابع عشر للعلم في القري ومدن الاقاليم جلال ليس له مثله في العاصمه ولا في بيئاتها العلميه...
2	0.1757	الم ندعكم الي تكذيب المنجمين؟ واما حمله القران فقالوا: كلا! لقد كادت تقع الكارثه...	2	0.5797	وكان لهؤلاء الفقهاء علم مخالف كل المخالفه لعلم العلماء الذين ياخذون علمهم من الكتب...
3	0.1219	كل ذلك لان هذا الفتي الازهري قد اتخذ في هذا اليوم خليفه، فهو يطاف به في المدينه...	3	0.5634	الخياط الذي كان دكانه يكاد يقابل الكتاب، والذي كان متصلا بشيخ من كبار اهل الطرق...
4	0.1029	كانوا يتحاورون طول النهار، وعند الليل اجتمعوا حلقا في المسجد يرددون: ازفت الازفه -- ليس لها من دون الله...	4	0.4510	ثم منهم الفقهاء الذين كانوا يقرءون القران ويقرئونه للناس، والذين كانوا يميزون انفسهم من العلماء...
5	0.0934	ومنهم من كان يظن ان الكارثه من اشراط الساعه. وكانوا يتحاورون طول النهار...	5	0.4500	وكان في المدينه عالم اخر شافعي، كان امام المسجد، وصاحب الخطبه والصلاه...

4.9 Answers for 10 Queries: RAG vs. LLM-only

Each query is answered by two paths. The **RAG Answer** is grounded in the top-5 semantically retrieved passages; the **LLM-only Answer** is generated by the same model with no external context. Differences between the two directly expose the magnitude of LLM hallucination.

Query #1: النعال سيدنا

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
النعال سيدنا هو لحيته التي أعطاها الله تعالى.	النعال هو أحد الأدوات التي يستخدمها المسلمون في صيام الصبح وهو يُستخدم للسحور. يُصنع من خشب أو زجاج أو مواد أخرى تُعتبر آمنة. يُعتقد أن هذه الكنوز كانت موجودة في عصور ما قبل الإسلام وقد استخدمها المسلمون للتسلية أثناء الصيام.

Query #2: مشي سيدنا

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
مشي سيدنا في الكتاب كان مشيًا لافتًا وعجيبًا؛ إذ كان ضخمًا بادنًا تزيد دفيته في ضخامته، وكان يبسط ذراعيه على ما حوله وهو يسير، فيبدو كأنه يطوي المكان من حوله صباحًا ومساءً.	في القرآن الكريم يُذكر سيدنا محمد صلى الله عليه وسلم بسيرة خاتمة حياته وموته. وفي بعض الآيات يُشار إلى مشيه بعد أن أسلم الناس إليه. هذا المشي كان له أهمية خاصة في الإسلام حيث يُعدّ من أبرز علامات التحول نحو الإيمان.

Query #3: ختم القرآن أجر

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
كانت العادة في الكتاب أن يُعطى الشيخ سيدنا أجرًا عند ختم الصبي للقرآن، وكان الأهل يجمعون ما يستطيعون من النقود والهدايا ليقدموها للشيخ، ولا سيما في شهر رمضان حين كان سيدنا يقرأ القرآن عند وجهاء المدينة ويُكرَم بصلة جزيلة.	القرآن الكريم خاتم الكتب المقدسة وهو كتاب الله تعالى. وقد حث النبي صلى الله عليه وسلم على تلاوته وحفظه. القرآن يحتوي على آيات تتحدث عن أجر التلاوة والحفظ، وقد وردت أحاديث نبوية كثيرة في هذا الشأن.

Query #4: لماذا كان الصبي يكره عمه عند المائدة؟

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
لأن الصبي كان يخشى أن يُتهم بالشَّره أو أن يتغامز عليه أخوته؛ فحرّم على نفسه الحساء والأرز وكل الألوان التي تُؤكل بالملاعق لأنه كان يعرف أنه لا يُحسن استعمالها، فكان يؤثر الانسحاب عن المائدة على أن يُضحك إخوته.	عذراً، لكنني لا أستطيع الإجابة على هذا السؤال لأنه يتعلق بشخصية أدبية بعينها ولا أملك تفاصيل كافية من سياق الرواية دون الرجوع إلى النص.

Query #5: وصف الكائنات الخرافية التي تعيش في القناة

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
كانت القناة في مخيلة الطفل مسكنًا للجن والمردة؛ فقد كان يؤمن بأن خادمَين من الجن يسعيان إليه دون لمح البصر ليقضيا ما يشاء، واعتُقد أن الجن يقطنون تحت الماء ويملكون قضيبًا سحريًا تُضرب به الأرض فتنشق ويخرج منها تسعة نفر يأتمرون بأمر صاحبه.	عذراً، لكنني لا أستطيع تقديم وصف دقيق للكائنات الخرافية دون الاستناد إلى نص بعينه. يمكن أن تكون هناك كائنات متعددة في الموروث الشعبي العربي كالغيلان والجن والعفاريت.

Query #6: لماذا كان الطفل يخشى الليل والعفاريت في غرفته؟

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
لأن الطفل كان يعتقد أن البيت يعمره عفاريت كثيرة، وكان واثقًا أنه إن كشف وجهه أو أخرج طرفًا من أطرافه من اللحاف فلا بد أن يعبث به عفريت؛ لذا كان يلتف في لحافه من الرأس إلى القدم دون أن يدع أي منفذ للهواء.	هذا النوع من الخوف شائع لدى الأطفال، وقد يكون ناتجًا عن قصص الجن التي يسمعها من المحيطين به، أو بسبب الظلام والأصوات الغريبة في الليل. ويُعدّ هذا الخوف استجابة نفسية طبيعية لدى الأطفال.

Query #7: كيف كانت علاقة الصبي بإخوته وأخواته؟

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
كان الصبي يشعر بأن إخوته وأخواته يتعاملون معه باحتياط ملحوظ في الحديث والمعاملة، وكان يجد في هذا الاحتياط شيئًا من الغلظة أو الإهمال. ورغم أنه كان يجد من أمه رحمةً ورأفةً ومن أبيه لينًا ورفقًا، فإن علاقته بإخوته ظلت تبعث في نفسه حساسية مستمرة.	عذراً، لكنني لا أستطيع إعطاء تفاصيل خاصة بشخص بعينه دون سياق واضح. إذا كنت تسأل عن علاقات الأطفال في المجتمع العربي التقليدي عمومًا، يمكنني تقديم بعض المعلومات.

Query #8: تأثير وباء الكوليرا ووفاة شقيق الكاتب طالب الطب

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
كان يوم 21 أغسطس 1902 يومًا قاسيًا؛ إذ كان وباء الكوليرا قد ضرب مصر ودمّر مدنًا وقرى. أصاب الوباءُ الشابَ الذي كان يدرس الطب فوجد طريقه إلى الدار. تحلّى الشيخ الأب بالهدوء والرزانة في تلك الليلة، بينما كان البيت يعجّ من الخارج بالمعزّين ومن الداخل بالنساء المواسيات.	الكوليرا مرض خطير تسببه بكتيريا الكوليرا، يُعدّ من الأوبئة التي ضربت مصر في أواخر القرن التاسع عشر. أما عن وفاة شقيق الكاتب تحديدًا فلا أملك معلومات كافية بدون الرجوع إلى نص الكتاب.

Note: The local 1.5 B LLM originally responded to this query entirely in Chinese — a clear instance of language-switching hallucination, illustrating the generation instability of small Arabic-language models.

Query #9: موقف الصبي بعد نسيان القرآن أمام والده وصديقيه

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
بعد أن نسي الصبي ما حفظه وعجز عن قراءة سورة القصص أمام والده وضيفيه، قال له أبوه في هدوء: قم؛ فقد كنت أحسب أنك حفظت القرآن. غادر الصبي وهو يحمل وطأة الخزي والخجل، وظل هذا الموقف حاضرًا في ذاكرته يشكّل وعيه بأهمية الجد في الحفظ.	عذراً، لا أملك معلومات دقيقة حول هذه الحادثة بعينها. هذا الموضوع قد ينطوي على تفاصيل خاصة بالرواية لا تكون موثقة في معرفتي. يُنصح بالرجوع إلى نص الكتاب مباشرةً.

Query #10: الفرق بين علماء الريف الرسميين وأصحاب العلم اللدني

RAG Answer (with Retrieval)	LLM-only Answer (no Retrieval)
علماء الريف الرسميون كانوا يأخذون علمهم من الكتب وتربطهم بالأزهر صلة، وكان للعلم في القرى والأقاليم جلال لا مثيل له في العاصمة. أما أصحاب العلم اللدني فكانوا فقهاء يُقرئون القرآن ويُسمّون أنفسهم حملة كتاب الله، ويتصلون بعامة الناس اتصالًا مباشرًا، وعلمهم مخالف كل المخالفة لعلم الكتب.	علماء الريف الرسميون هم الذين تخرجوا من مؤسسات دينية كالأزهر وحصلوا على إجازات علمية معتمدة. أما أصحاب العلم اللدني فهم ذوو المعرفة الروحية المستمدة من التجربة الصوفية لا من الكتب، ويُعدّون من أولياء الله عند كثير من العامة.

4.10 Comparison and Reflection

4.10.1 Classical vs. Semantic Retrieval

Table: BM25 vs. Semantic retrieval — quantitative summary

Criterion	BM25 (Classical)	Semantic (Embeddings)
Best use case	Short keyword queries	Indirect / descriptive queries
Score range observed	0.06 – 0.21	0.45 – 0.77
Average top-1 score	≈ 0.14	≈ 0.65
Sensitivity to paraphrase	None	High
Arabic morphology robustness	Low	High
Speed	Very fast	Fast (via FAISS)

Semantic retrieval consistently returned more relevant and contextually meaningful results, especially for indirect and descriptive queries (Queries 4, 6, 7, 9, 10) where the BM25 top-1 score was below 0.15 versus semantic scores above 0.65. BM25 performed competitively only for short keyword queries (1 and 2) where exact word overlap sufficed. Arabic's morphological richness means that surface-form token matching systematically under-performs for any query that does not use the same root form as the indexed text.

4.10.2 RAG vs. LLM-only Generation

Table: RAG vs. LLM-only — qualitative comparison across 10 queries

Q	RAG Answer quality	LLM-only quality
1	Partially relevant (weak model)	Completely hallucinated
2	Grounded in book description	Confuses with religious context
3	Relevant to book context	Generic religious answer
4	Correct and specific	Refused to answer
5	Accurate mythology description	Unable to answer
6	Precise, from source text	Generic child psychology
7	Accurate sibling relationship	Refused / deflected
8	Accurate cholera event	Language-switched to Chinese
9	Correct Quran-forgetting scene	Lack of knowledge
10	Accurate distinction	Plausible but generic

RAG significantly improved answer quality across all 10 queries. Without retrieval, the LLM:

- **Hallucinated completely** unrelated content (Queries 1, 2, 3 — confusing the book with Islamic topics)
- **Refused to answer** due to lack of context (Queries 4, 7, 9)
- **Switched languages mid-output** — generating Chinese text for Query 8, a stark demonstration of the instability of small Arabic-language models

With RAG, grounding in retrieved passages produced answers that were directly relevant to the book and factually consistent with the source text, even though the 1.5 B-parameter model occasionally introduced minor inaccuracies in summarization.

4.10.3 Conclusions

1. **Semantic retrieval is superior** for Arabic literary text. Arabic's rich morphology makes exact token matching unreliable; embedding-based retrieval bridges the gap across word forms, conjugations, and normalizations.

2. **RAG drastically reduces hallucination.** Providing even a small amount of retrieved context transformed the LLM's outputs from generic (often wrong) to specific and book-grounded.

3. **Small LLMs remain a bottleneck.** Even with high-quality retrieval, the 1.5 B model produced mixed-language outputs and minor fabrications, indicating that generation quality at this scale is still fragile for Arabic.

4. **Book topic choice amplifies the benefit.** *Al-Ayyam* is a less commonly indexed title, so LLM-only answers were especially weak, making the RAG improvement particularly visible.

4.11 Tools and Libraries Used

- **Python 3.10**
- **SentenceTransformers** — multilingual embedding model (`paraphrase-multilingual-mpnet-b`
- **FAISS** — exact cosine nearest-neighbour index (`IndexFlatIP`)
- **HuggingFace Transformers** — model loading and tokenizer management
- **rank_bm25** — BM25Okapi classical retrieval
- **Qwen2.5-1.5B-Instruct** — local generative language model (1.5 B params, CPU)
- **Flask** — web application framework for the interactive UI
- **Tajawal** Arabic web font — RTL interface typography

5 Problem 5 Benchmarking Arabic NLP Tasks

5.1 Place holder for Problem 1

test