

Configuration Manual

MSc Research Project
MSc in Data Analytics

ADDANKI MAHESH KUMAR
Student ID: x22240047

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Addanki Mahesh Kumar
Student ID: x22240047
Programme: MSc in Data Analytics **Year:** 2024-2025
Module: Research Project
Supervisor: Abdul Qayum
Submission Due Date: 15 Sept 2025
Project Title: Multi-LLM Ensemble Architecture for Cross-Modal Document Understanding: A RAG-based Approach

Word Count: 170

Page Count: 07

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Addanki Mahesh Kumar

Date: 15 Sept 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

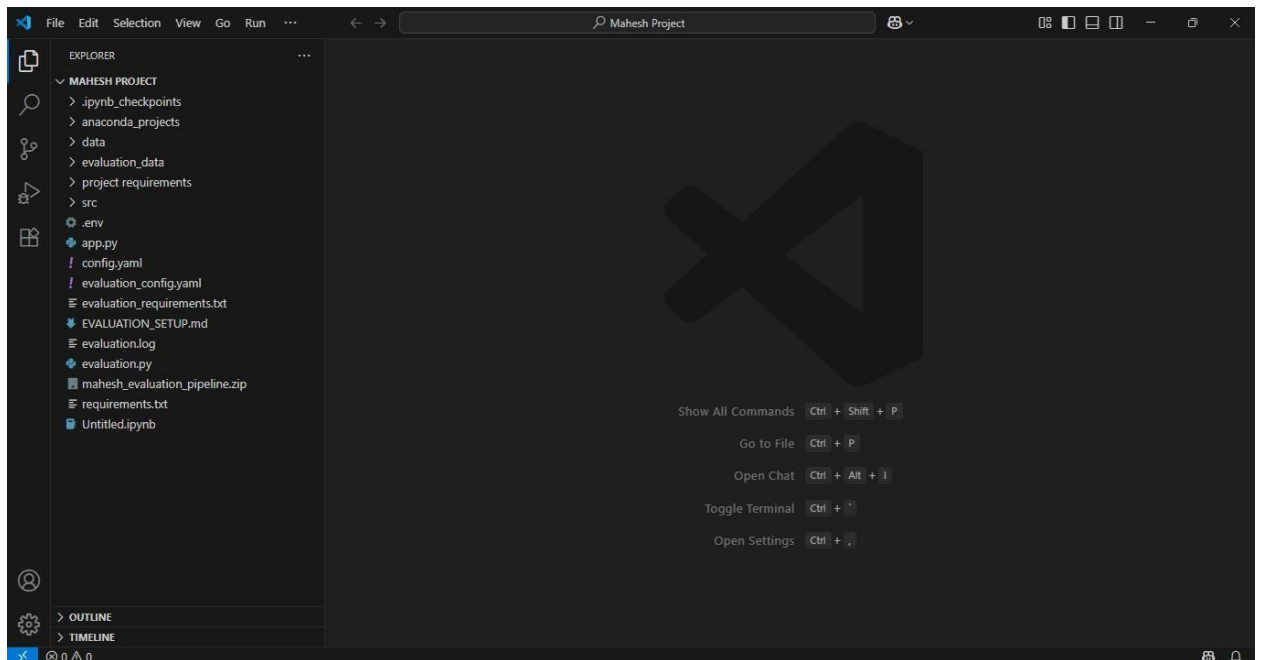
Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

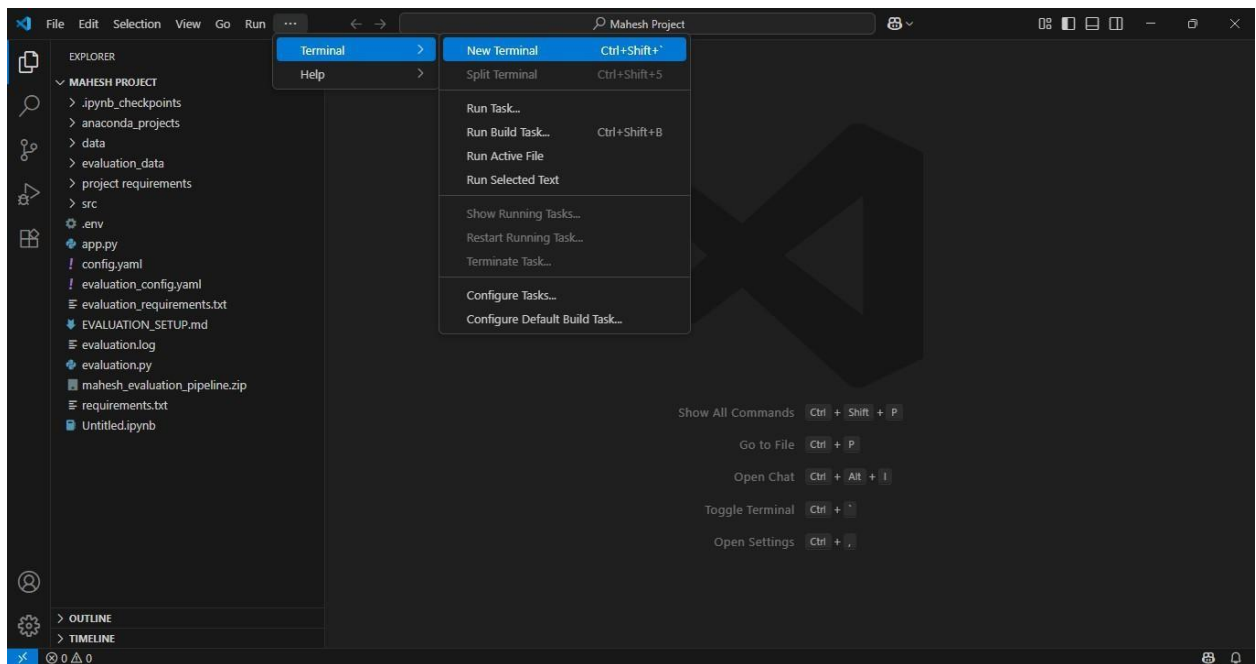
Addanki Mahesh Kumar

Student ID:x22240047

1. Unzip the code and open in VS code

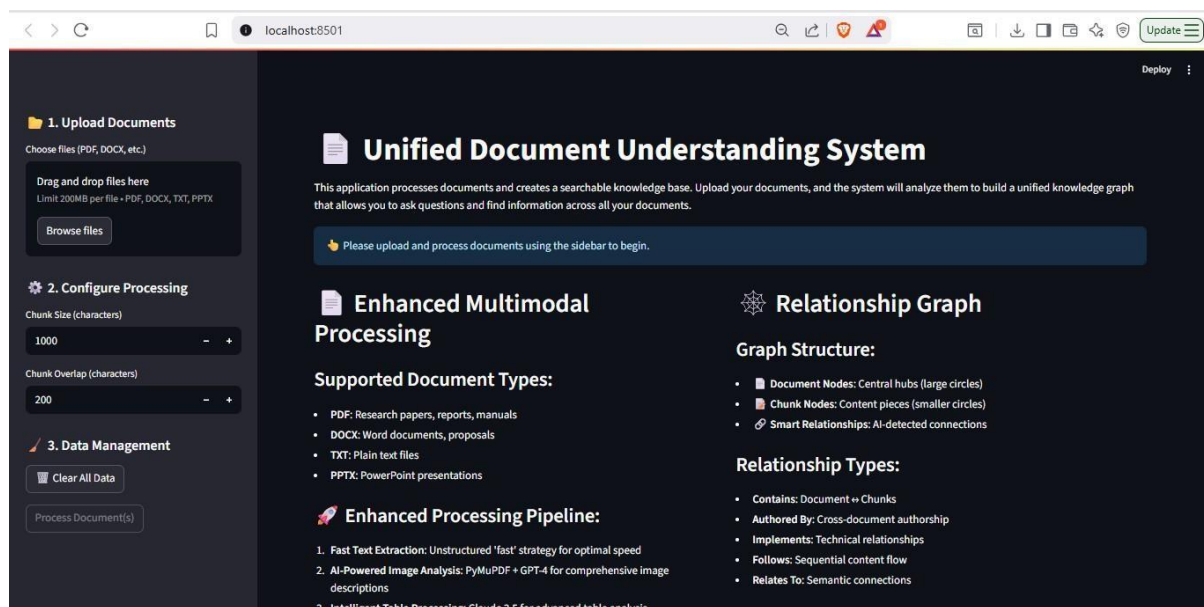


2. Open new terminal as follows

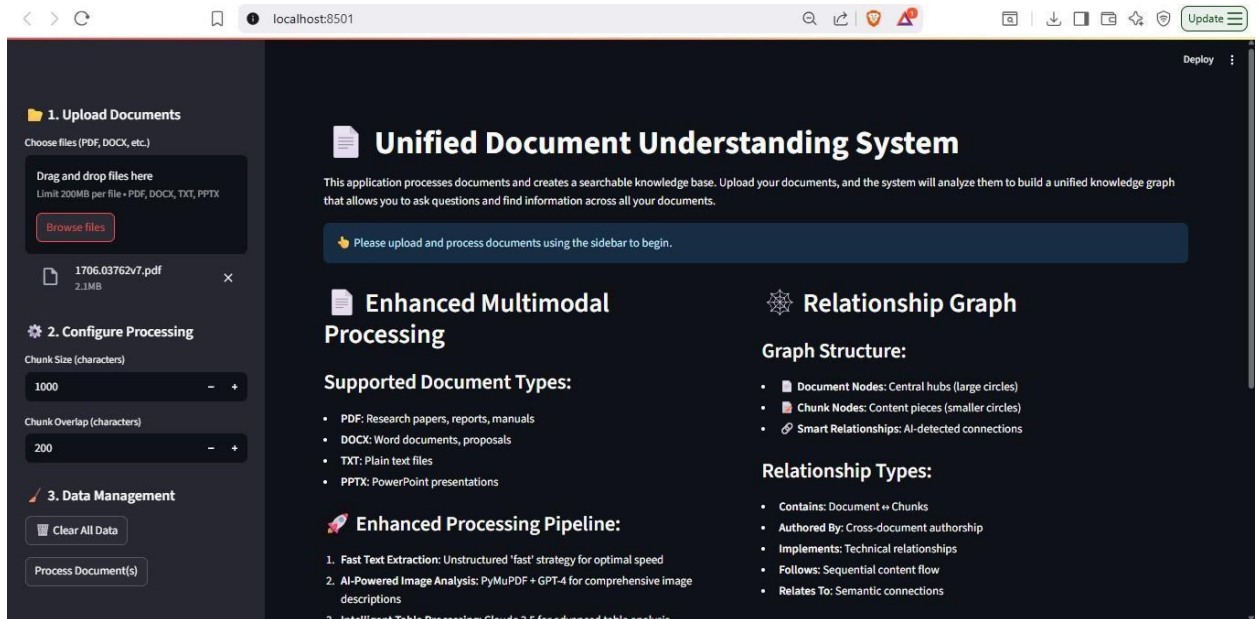


3. Run the following commands in the new terminal
 - a. `conda activate --name my_project python=3.10 -y`
 - b. `pip install -r requirements.txt`
 - c. `streamlit run app.py`

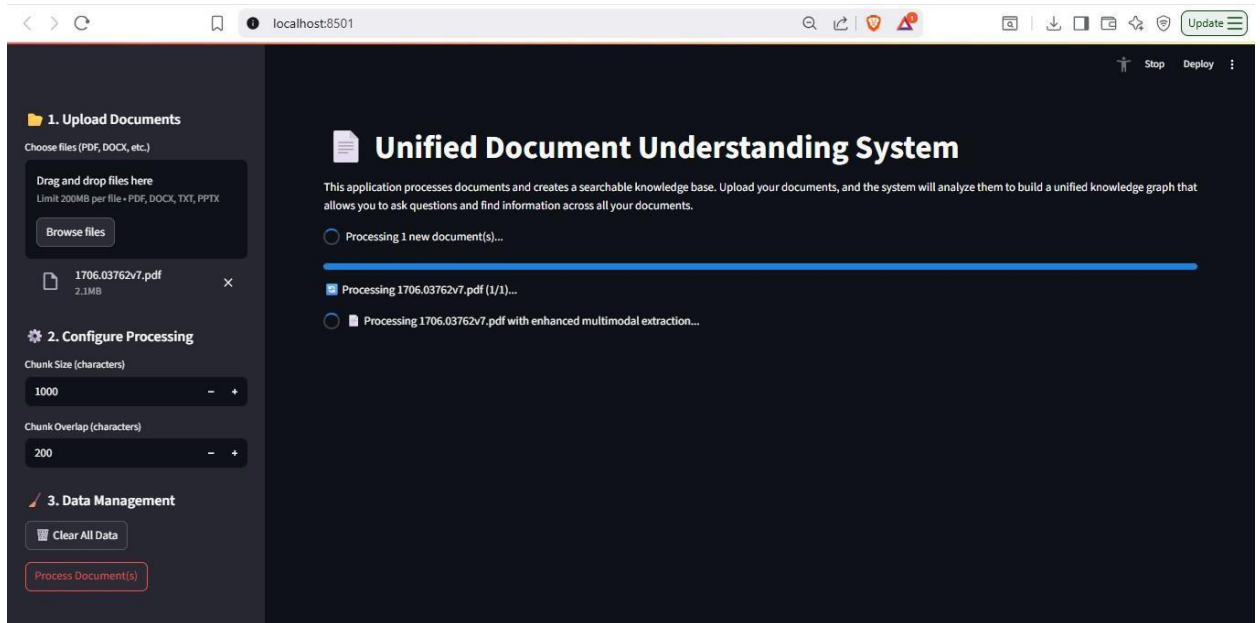
4. Then you can see an UI like this

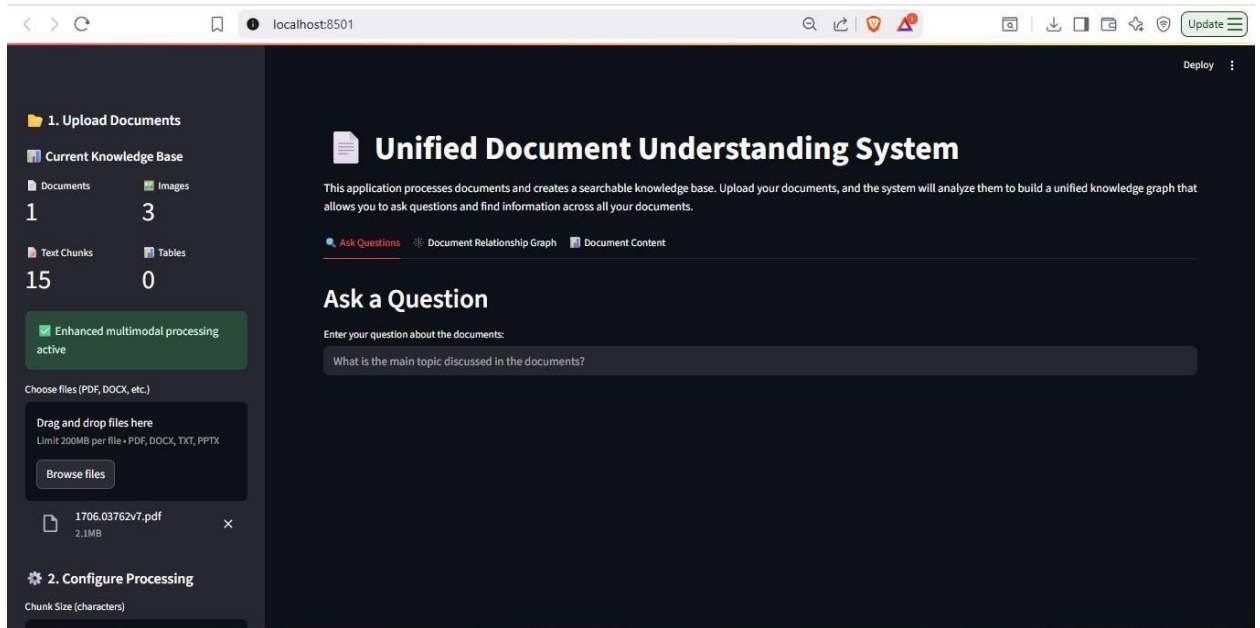


5. Click on Browse files and upload document as follows

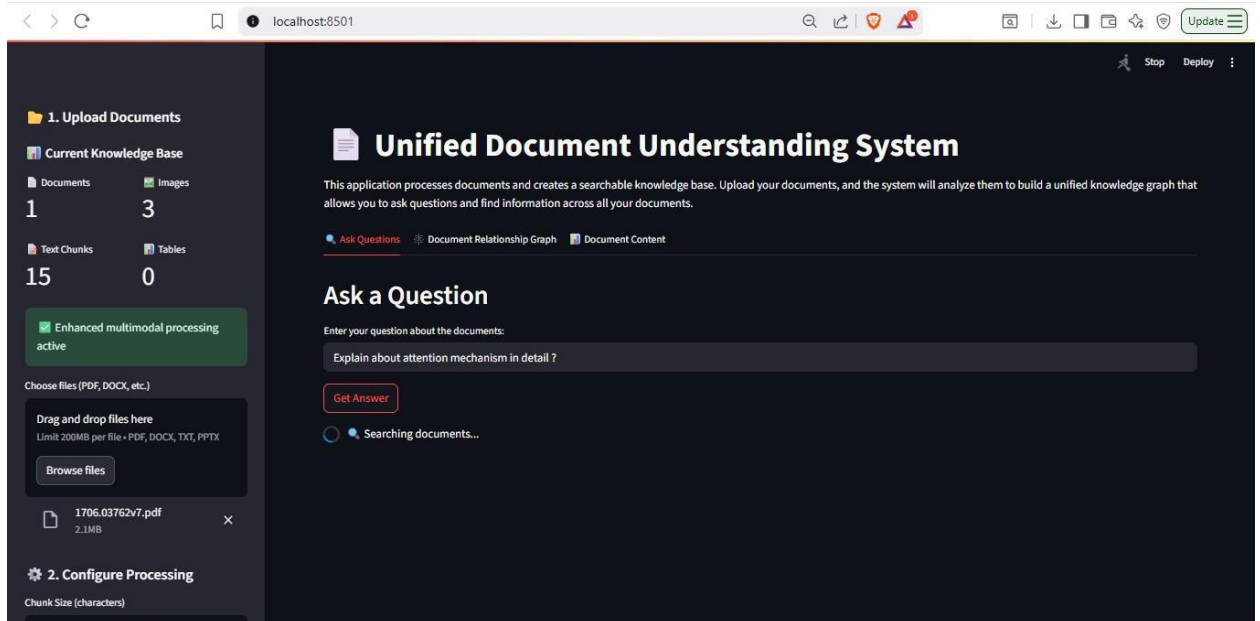


6. Click on process documents, then you can see the below screen





7. Enter question and click on get answer button



1. Upload Documents

Current Knowledge Base

Documents

1

Images

3

Text Chunks

15

Tables

0

Enhanced multimodal processing

active

Choose files (PDF, DOCX, etc.)

Drag and drop files here

Limit 200MB per file • PDF, DOCX, TXT, PPTX

Browse files

1706.03762v7.pdf

2.1MB

2. Configure Processing

Chunk Size (characters)

Ask a Question

Enter your question about the documents:

Explain about attention mechanism in detail ?

Get Answer

Answer

The attention mechanism is a fundamental concept in modern neural network architectures, particularly in the field of natural language processing (NLP). It allows models to focus on specific parts of the input sequence when producing each part of the output sequence, thereby improving the model's ability to handle long-range dependencies and complex relationships within the data. Here's a detailed explanation of the attention mechanism, particularly as it is used in the Transformer model:

1. Basic Concept of Attention

The attention mechanism can be thought of as a way to dynamically weight the importance of different parts of the input data when making predictions. It involves three main components:

- **Queries (Q):** These are the vectors that represent the current focus of the model, typically derived from the current state of the decoder in sequence-to-sequence models.
- **Keys (K):** These vectors represent the elements of the input sequence that the model can attend to.
- **Values (V):** These are the actual data points that are being weighted and combined to produce the output.

The attention mechanism computes a weighted sum of the values, where the weights are determined by a compatibility function that measures how well each key matches the query.

2. Scaled Dot-Product Attention

In the Transformer model, the attention mechanism is implemented as "Scaled Dot-Product Attention." The process involves:

- Calculating the dot product of the query with all keys to obtain a set of scores.

Deploy

8. Explore multiple section of the app

1. Upload Documents

Current Knowledge Base

Documents

1

Images

3

Text Chunks

15

Tables

0

Enhanced multimodal processing

active

Choose files (PDF, DOCX, etc.)

Drag and drop files here

Limit 200MB per file • PDF, DOCX, TXT, PPTX

Browse files

1706.03762v7.pdf

2.1MB

2. Configure Processing

Chunk Size (characters)

1000

Chunk Overlay (characters)

200

3. Data Management

Clear All Data

Process Document(s)

This application processes documents and creates a searchable knowledge base. Upload your documents, and the system will analyze them to build a unified knowledge graph that allows you to ask questions and find information across all your documents.

Ask Questions

Document Relationship Graph

Document Content

Document Relationship Graph

Unified knowledge graph showing documents, their content (text, images, tables), and relationships

Documents

1

Text Chunks

15

Images

3

Tables

0

Graph Type

Document-Centric

Document-Centric Knowledge Graph

Documents appear as central nodes with their content elements connected around them. Related documents are linked based on content similarity.

Document-Centric Knowledge Graph

Documents appear as central nodes with their content elements connected around them. Related documents are linked based on content similarity.

1. Upload Documents

Current Knowledge Base

Documents

Images

1

3

Test Chunks

Tables

15

0

Enhanced multilingual processing
active

Choose files (PDF, DOCX, etc.)

Drag and drop files here
Limit: 50MB per file • PDF, DOCX, TXT, PPTX

Browse Files

1708.03742d7.pdf
3.1 MB

2. Configure Processing

Chunk Size (characters)

1000

Chunk Overlap (characters)

200

3. Data Management

Clear All Data

Process Document(s)

localhost:8501

Unified Document Understanding System

This application processes documents and creates a searchable knowledge base. Upload your documents, and the system will analyze them to build a unified knowledge graph that allows you to ask questions and find information across all your documents.

Ask QuestionsDocument Relationship GraphDocument Content

Document Relationship Graph

Unified knowledge graph showing documents, their content [text, images, tables], and relationships

Documents

Test Chunks

Images

Tables

1

15

3

0

Graph Type

Similarity Network

Similarity-Based Content Network

Content elements are positioned by similarity with connections showing relationships between different parts.

Document Content Network

UnstructuredText (1)

Image (2)

StructuredElement (2)

Nodes prioritized by relevance/similarity

1. Upload Documents

Current Knowledge Base

Documents

3

Images

0

Text Chunks

15

Tables

0

Enhanced multimodal processing active

Choose files (PDF, DOCX, etc.)

Drag and drop files here

Limit: 100MB per file • PDF, DOCX, TXT, PPTX

Browse Files

1706.03762v7.pdf

2.1MB

2. Configure Processing

Chunk Size (characters)

1000

Chunk Overlay (characters)

200

3. Data Management

Clear All Data

Process Document(s)

localhost:8501

Unified Document Understanding System

Ask Questions

Document Relationship Graph

Document Content

This application processes documents and creates a searchable knowledge base. Upload your documents, and the system will analyze them to build a unified knowledge graph that allows you to ask questions and find information across all your documents.

Document Content

Content Overview

Filter by Document:

All Documents

Filter by Content Type:

All Types

Total Items

Documents

Content Types

19

1

3

Content Table

Document Name	Chunk ID	Type	Title	Raw Content
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_0	NarrativeText	Provided! Google Research AI	g v A 2 [1 C . s t] Provided proper attribution is provided, Google hereby grants permission to rep
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_1	NarrativeText	Introduction Recurrent approaches modeling	Introduction Recurrent neural networks, long short-term memory (LSTM) and gated recurrent [7] neu
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_2	NarrativeText	Figure 1: The Transformer model	Figure 1: The Transformer model architecture. The Transformer follows this overall architecture v
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_3	NarrativeText	Scaled Dot Product Attention Multi-Head	Scaled Dot Product Attention Multi-Head Attention Figure 2: [8] Scaled Dot Product Attention (v
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_4	NarrativeText	output values	output values. These are concatenated and once again projected, resulting in the final values, w d d
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_5	NarrativeText	Table Maximum Layer Type	Table 1: Maximum path lengths, per layer complexity and minimum number of sequential operatio
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_6	NarrativeText	models	length is smaller than the representation dimensionality d, which is modd often the case with seq
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_7	NarrativeText	Model BLEU F1N DE EN FR 23	Model BLEU F1N DE EN FR 23.17 Training Cost (11.0M) EN DE EN FR Residual Dropout We apply d
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_8	NarrativeText	Table Variations Transformer Unilted	Table 3: Variations on the Transformer architecture, Unilted values are identical to those of the bas
1706.03762v7.pdf	CHUNK_1706.03762v7.pdf_9	NarrativeText	(2006) [26] Zhu et al	Table 4: The Transformer generates well-to-English consistency parsing (usually are on Section 2)

Report Content as CSV

References

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A survey. arXiv (Cornell University).
<https://doi.org/10.48550/arxiv.2312.10997>

He, H., Ye, M., Zhang, J., Cai, X., Liu, J., Du, B., & Tao, D. (2025, May 19). Reasoning-OCR: Can Large Multimodal Models Solve Complex Logical Reasoning Problems from OCR Cues? arXiv.org.
<https://arxiv.org/abs/2505.12766>