

Title: Modular Memory-augmented Transformer (MoMeT)

Author: Mwaura Vincent Muchai

Introduction

Abstract

Transformer models have emerged as the most transformative and influential architecture in Machine Learning (ML) to date, redefining state-of-the-art performance across fields such as Natural Language Processing (NLP), Computer Vision (CV), and audio processing. This success can be attributed to its novel attention mechanism, which enables models to capture long-range dependencies by dynamically weighting interactions between tokens (smallest unit of input), thereby producing context-sensitive representations that scale efficiently with parallel computation.

However, despite the impressive performance, Transformer models still face numerous challenges:

- In order to achieve state-of-the-art performance, the models require to be scaled up to billions of parameters, trained on a vast dataset, and have a lot of compute for optimal performance. This makes them **computationally and financially infeasible** for a vast majority of people and organizations.
- They are known to suffer from a phenomenon called **hallucinations**. This is where fully trained models generate nonsensical or factually incorrect content frequently enough to be of concern, making them unreliable for most tasks. This can be categorized into three types of hallucinations: input-conflicts where generated content deviates from the source input provided by the user, context-conflicts where generated content conflicts with previously generated information from the model itself, and fact-conflicts where the generated content that is factually incorrect to established world knowledge. This is mostly brought about by the nature of the model to produce coherent and plausible text with no mechanism to ground itself on truthfulness. In addition, this effect could be brought about by low-quality training data and poor generalization of the model during training.
- They are prone to **catastrophic forgetting**. This is where trained Transformer models struggle to learn new tasks and update their knowledge when training on new datasets. This is because the Transformer model implicitly stores information in their parameters (Parametric memory), and therefore any update to them could introduce errors or erase the information already stored. This limits the model's ability to progressively learn without forgetting any previously learned information.

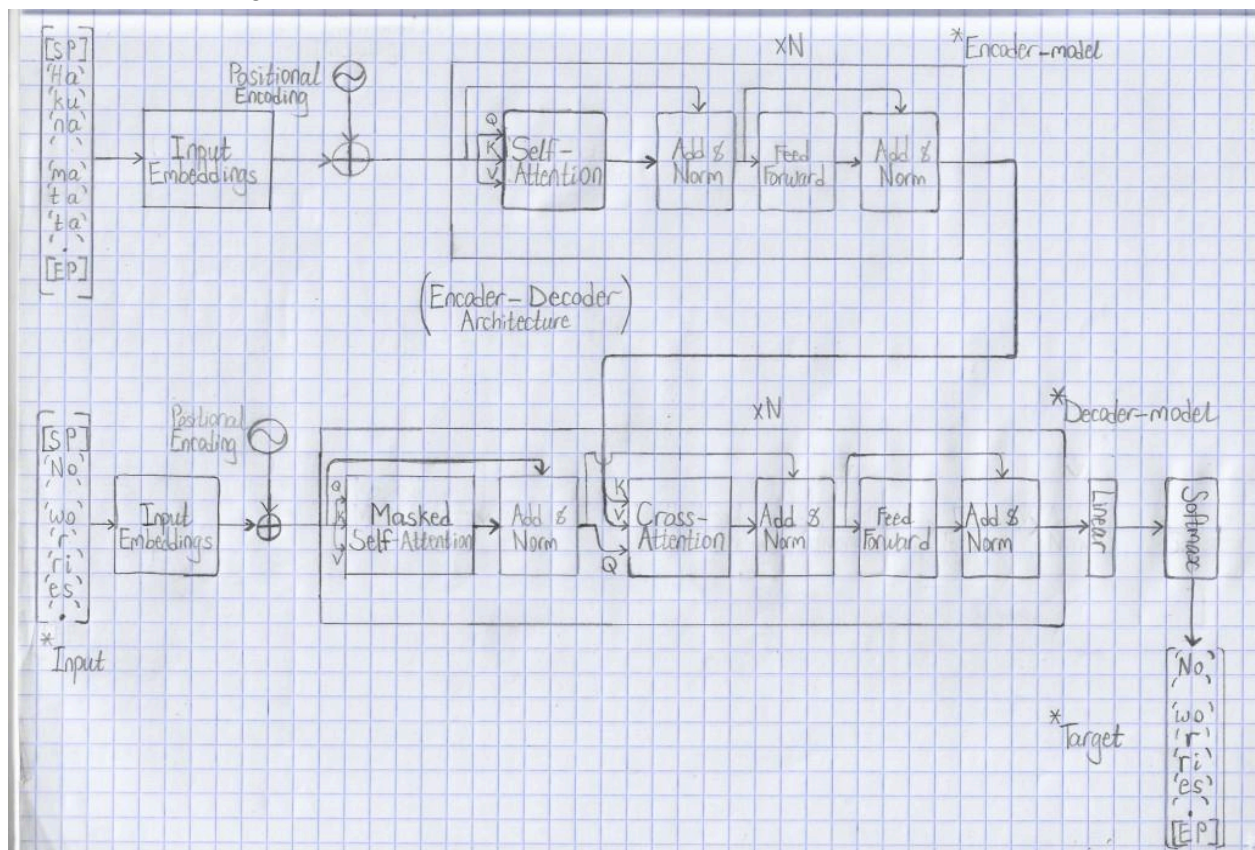
The aim of this project is to try and address the aforementioned challenges by implementing a novel modular system that integrates multiple Transformer models alongside an Information Retrieval (IR) system to enhance knowledge access and contextual reasoning. This is inspired by advances in Retrieval Augmented Generation (RAG) frameworks and Memory-Augmented Transformers, both of which extend the Transformer architecture by introducing memory-like components and knowledge integration mechanisms, albeit with varying degrees of success. The Transformer models will be trained on a partially synthetically generated dataset.

Background

Transformer models

Transformers are a neural network model architecture that rely solely on the attention mechanism, dispensing with recurrence (used by RNNs) and convolutions (used by CNNs) operations entirely. The models generalize well to Natural Language Processing (NLP), Computer Vision (CV), and audio processing. It has become the de facto model architecture for many AI tasks due to its unrivalled performance, ability for parallelization and efficiency during training.

Overview of the original Transformer model architecture and a visual depiction of how it works:



The attention mechanism works by assigning differing weights to tokens passed as input to the model. It does so by mapping the vector representations of said tokens into query (**Q**), key (**K**), and value (**V**) vectors, which are all combined using the formula depicted below:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{\text{Dim}_K}}\right)V$$

There are three types of attention mechanisms:

- **Self-Attention:** Here each token can attend to all other tokens within the same input sequence, enabling contextualized representations.

		0	1	2	3	4	5	6	7	8	9
	QK^T	[SP]	'Ha'	'ku'	'na'	' '	'ma'	'ta'	'ta'	'.'	[EP]
0	[SP]	Q_0K_0	Q_0K_1	Q_0K_2	Q_0K_3	Q_0K_4	Q_0K_5	Q_0K_6	Q_0K_7	Q_0K_8	Q_0K_9
1	'Ha'	Q_1K_0	Q_1K_1	Q_1K_2	Q_1K_3	Q_1K_4	Q_1K_5	Q_1K_6	Q_1K_7	Q_1K_8	Q_1K_9
2	'ku'	Q_2K_0	Q_2K_1	Q_2K_2	Q_2K_3	Q_2K_4	Q_2K_5	Q_2K_6	Q_2K_7	Q_2K_8	Q_2K_9
3	'na'	Q_3K_0	Q_3K_1	Q_3K_2	Q_3K_3	Q_3K_4	Q_3K_5	Q_3K_6	Q_3K_7	Q_3K_8	Q_3K_9
4	' '	Q_4K_0	Q_4K_1	Q_4K_2	Q_4K_3	Q_4K_4	Q_4K_5	Q_4K_6	Q_4K_7	Q_4K_8	Q_4K_9
5	'ma'	Q_5K_0	Q_5K_1	Q_5K_2	Q_5K_3	Q_5K_4	Q_5K_5	Q_5K_6	Q_5K_7	Q_5K_8	Q_5K_9
6	'ta'	Q_6K_0	Q_6K_1	Q_6K_2	Q_6K_3	Q_6K_4	Q_6K_5	Q_6K_6	Q_6K_7	Q_6K_8	Q_6K_9
7	'ta'	Q_7K_0	Q_7K_1	Q_7K_2	Q_7K_3	Q_7K_4	Q_7K_5	Q_7K_6	Q_7K_7	Q_7K_8	Q_7K_9
8	'.'	Q_8K_0	Q_8K_1	Q_8K_2	Q_8K_3	Q_8K_4	Q_8K_5	Q_8K_6	Q_8K_7	Q_8K_8	Q_8K_9
9	[EP]	Q_9K_0	Q_9K_1	Q_9K_2	Q_9K_3	Q_9K_4	Q_9K_5	Q_9K_6	Q_9K_7	Q_9K_8	Q_9K_9

- **Masked Self-Attention:** Here attention is restricted by using -infinite value on specific tokens (upper triangle, shifted once to the right) to ensure that each token can only attend to its own and other preceding tokens in a sequence. This prevents leftward tokens access to future information, tokens to their right. The constraint is typically enforced by applying a masking operation to the QK^T matrix to prevent leftward information flow during computation.

		0	1	2	3	4	5	6	7
	QK^T	[SP]	'No'	' '	'wo'	'r'	'ri'	'es'	'.'
0	[SP]	Q_0K_0	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
1	'No'	Q_1K_0	Q_1K_1	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
2	' '	Q_2K_0	Q_2K_1	Q_2K_2	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
3	'wo'	Q_3K_0	Q_3K_1	Q_3K_2	Q_3K_3	$-\infty$	$-\infty$	$-\infty$	$-\infty$
4	'r'	Q_4K_0	Q_4K_1	Q_4K_2	Q_4K_3	Q_4K_4	$-\infty$	$-\infty$	$-\infty$
5	'ri'	Q_5K_0	Q_5K_1	Q_5K_2	Q_5K_3	Q_5K_4	Q_5K_5	$-\infty$	$-\infty$
6	'es'	Q_6K_0	Q_6K_1	Q_6K_2	Q_6K_3	Q_6K_4	Q_6K_5	Q_6K_6	$-\infty$
7	'.'	Q_7K_0	Q_7K_1	Q_7K_2	Q_7K_3	Q_7K_4	Q_7K_5	Q_7K_6	Q_7K_7

* $\text{Softmax}(-\infty) = 0$

- **Cross-Attention:** Here the queries (**Q**) are derived from the Decoder's input sequences, while the keys (**K**) and values (**V**) originate from the Encoder's output sequence. This allows the Decoder to condition its representations on the Encoder's encoded context, enabling it to generate outputs that are informed by the source input.

		0	1	2	3	4	5	6	7	8	9
	QK^T	[SP]	'Ha'	'ku'	'na'	' '	'ma'	'ta'	'ta'	'.'	[EP]
0	[SP]	Q_0K_0	Q_0K_1	Q_0K_2	Q_0K_3	Q_0K_4	Q_0K_5	Q_0K_6	Q_0K_7	Q_0K_8	Q_0K_9
1	'No'	Q_1K_0	Q_1K_1	Q_1K_2	Q_1K_3	Q_1K_4	Q_1K_5	Q_1K_6	Q_1K_7	Q_1K_8	Q_1K_9
2	' '	Q_2K_0	Q_2K_1	Q_2K_2	Q_2K_3	Q_2K_4	Q_2K_5	Q_2K_6	Q_2K_7	Q_2K_8	Q_2K_9
3	'wo'	Q_3K_0	Q_3K_1	Q_3K_2	Q_3K_3	Q_3K_4	Q_3K_5	Q_3K_6	Q_3K_7	Q_3K_8	Q_3K_9
4	'r'	Q_4K_0	Q_4K_1	Q_4K_2	Q_4K_3	Q_4K_4	Q_4K_5	Q_4K_6	Q_4K_7	Q_4K_8	Q_4K_9
5	'ri'	Q_5K_0	Q_5K_1	Q_5K_2	Q_5K_3	Q_5K_4	Q_5K_5	Q_5K_6	Q_5K_7	Q_5K_8	Q_5K_9
6	'es'	Q_6K_0	Q_6K_1	Q_6K_2	Q_6K_3	Q_6K_4	Q_6K_5	Q_6K_6	Q_6K_7	Q_6K_8	Q_6K_9
7	'.'	Q_7K_0	Q_7K_1	Q_7K_2	Q_7K_3	Q_7K_4	Q_7K_5	Q_7K_6	Q_7K_7	Q_7K_8	Q_7K_9
8	[EP]	Q_8K_0	Q_8K_1	Q_8K_2	Q_8K_3	Q_8K_4	Q_8K_5	Q_8K_6	Q_8K_7	Q_8K_8	Q_8K_9

Additionally the Transformer model architecture can be implemented in various configurations:

- **Encoder only model:** Only the Encoder module of the model is used, typically with the Self-Attention mechanism.
- **Decoder only model:** Only the Decoder module of the model is used, typically with the Masked Self-Attention mechanism.
- **Encoder-Decoder model:** Both the Encoder and Decoder modules are used, which jointly utilizes the Masked Self-Attention, and Cross-Attention mechanism. They are usually known as a sequence-to-sequence model.

In order to train a Transformer model, the following stages need to be performed:

1. Data Preparation

- Data collection:** Data can be assembled from a diverse range of sources:
 - Literary and Creative writings** - novels, short stories, poems, plays, fables, myths, and folklore.
 - Educational and Academic materials** - textbooks, research papers, lecture notes, dissertations, scientific articles, and encyclopedias.
 - Technical and Professional Content** - software documentation, source code, API references, patents, engineering manuals, medical guidelines, and legal contracts.
 - News and Journalism:** newspaper articles, magazines, editorials, investigative reports, and press releases.

- v. **Business and Workplace Documents** - emails, memoranda, reports, business plans, meeting transcripts, and financial statements.
 - vi. **Government and Legal Records** - laws, regulations, court rulings, constitutions, treaties, policy briefs, and legislative transcripts.
 - vii. **Everyday and Informal Communication** - social media posts, blogs, forums, product reviews, Q&A threads, and chat logs.
 - viii. **Cultural and Historical Texts** - religious scriptures, speeches, historical archives, and autobiographies.
 - ix. **Instructional and How-to Guides** - tutorials, guides, cookbooks, repair manuals, and user guides.
- b. **Data cleaning:** The collected data are then filtered to remove noise, handle outliers, correct imbalances, and other preprocessing operations. After which deduplication is performed which entails removing duplicates found in the collected data. This is done to reduce redundancy and improve data quality.
2. **Tokenization:** This stage converts the raw text from the collected data into smaller, machine-readable units called tokens. The choice of granularity of these tokens can impact model efficiency and vocabulary sizes. Common approaches include:
- a. **Character-based tokenization:** Here text is broken down into individual characters such as letters, numbers, or other symbols. The major advantage of this approach is there are no or very few unknown or Out-Of-Vocabulary (OOV) words during training and it is simple to implement. The major drawback with this approach is that it produces very long input sequences, which can be computationally expensive and infeasible to process due to the high memory demands of the attention mechanism and limited memory capacity of hardware.
 - b. **Word-based tokenization:** Here text is broken down into words, typically using spaces and punctuation marks as delimiters. The main advantage of this approach is that it is usually simpler to implement and produces shorter input sequences. However the main drawback is that this approach requires extremely large vocabularies to capture all possible words from the text, which can lead to frequent OOV issues.
 - c. **Subword-based tokenization:** Here text is broken down into subwords which are usually the most frequent pairings of adjacent characters. This approach strikes a balance between Character-based tokenizations (very long input sequences but little to no OOV issues) and Word-based tokenizations (very large vocabulary sizes with frequent OOV issues but shorter input sequences). Common algorithms include:
 - i. **BytePairEncoding (BPE):** This is a frequency-based algorithm that iteratively merges the most common adjacent character pairs in a text until a target vocabulary size is reached. At inference, words are decomposed into subword units if unseen, while common words remain whole.
 - ii. **WordPieceEncoding:** This is a probabilistic subword tokenization method that iteratively merges common character sequences, balancing vocabulary size with coverage. At inference, text is segmented into the

longest matching subwords, ensuring out-of-vocabulary (OOV) terms can still be represented.

- iii. **SentencePieceEncoding:** This is a data-driven tokenizer that treats text as a raw sequence of characters, without assuming whitespace as word boundaries. It learns subword vocabularies using algorithms such as BPE or Unigram Language Models, offering robustness across languages and domains.
- 3. **Model Pre-training:** In this stage, the model is trained on the massive data collected using self-supervised objectives e.g. next-token prediction. Leveraging the tokenized data the model learns to capture broad linguistic and world knowledge within its parameters, resulting in foundation models that can serve as the basis for downstream adaptation.
- 4. **Fine-tuning and Instruction Tuning:**
 - a. **Fine-tuning:** The foundation models can be further trained on labeled, task-specific data to improve performance in specialized applications.
 - b. **Instruction Tuning:** The foundation models can be further aligned to follow human-like (natural language) instructions by training on curated instruction-response pairs, increasing their utility and alignment with user intent.

Memory-Augmented Transformer

Despite the state-of-the-art performances in text-generation, language understanding, and efficiency in training, Transformer models still face challenges when it comes to long-range context retention, continual learning, and knowledge integration. To address these challenges, numerous works have turned to neuroscience-inspired dynamic memory mechanisms.

Memory plays a fundamental role in intelligence, enabling learning, reasoning, and adaptability in biological systems. In humans, memory enables the retention, retrieval, and manipulation of information across multiple time scales, supporting complex behaviour such as decision-making and problem-solving. It operates as an interconnected, multi-layer network that comprises three interacting subsystems:

- **Sensory memory:** This is a brief, high capacity buffer that actively holds raw sensory input for milliseconds to seconds before fading or being passed to the working memory.
- **Working memory:** This is a short-term limited capacity system that actively holds and manipulates information needed for reasoning, problem-solving and goal-directed tasks.
- **Long-term memory:** This is the brain's durable storehouse that retains knowledge and experiences for extended periods of times, even lifetimes, organized into episodic (personalized experienced events) and semantic (abstract knowledge, facts, concepts, and meanings) systems with consolidation processes (mechanisms that stabilizes new memories over time) that ensure stability and adaptability with time.

A critical process the brain employs to hold on to memory and avoid forgetting, especially in the Long-term memory, is memory consolidation. This process involves the reactivation and reorganization of memory traces (unit of memory storage), integrating them into existing knowledge networks. Through consolidation, memories become more stable and less susceptible to decay or interference, thereby reducing the likelihood of forgetting.

The Transformer model can be seen as loosely mimicking some of these memory subsystems such as the token embeddings and positional encodings being the raw sensory inputs to the sensory memory. These help in stabilizing the input for downstream layers. The attention mechanism is analogous to the working memory in that it actively maintains and manipulates information required for reasoning, problem-solving, and goal-directed behaviour. Lastly the parameters of the models can be viewed as being analogous to the long-term memory subsystem; they can retain information obtained during training and then retrieve and implicitly integrate them in the generation process (inference).

Despite their numerous successes Transformer models still face several challenges relating to memory:

- Difficulty in storing long-tail and time-sensitive information. Information that constitutes a small fraction of the training dataset is difficult to store and recall on the models efficiently, contributing to less accurate models as there will be some gaps in the knowledge of the models. In addition, due to the static nature of model parameters, they are usually containing out-of-date information.

- Storing information implicitly in the model's parameters (parametric memory) is extremely inefficient as it usually necessitates scaling the model to computationally and financially unsustainable levels to be reliable on a wide range of tasks as well as needing a vast amount of data to be trained on.
- When updating the model's parameters with new information or training on a new task or dataset, there's a likelihood of introducing errors and erasing previously learned information (catastrophic forgetting). This usually leads to a degradation in model performance impacting the reliability of the models.

To address these challenges several studies have explored the use of explicit storage memory in the form of external memory modules to the Transformer models to enable scalable information storage and retrieval all with varying degrees of success. These modules are typically designed to be tightly integrated and differentiable with the model, thereby supporting end-to-end optimizations and enabling more sophisticated memory management strategies.

Some of these implementations include:

- **Memformer**: This is a Transformer model that implements an external dynamic memory for encoding and retrieving past information through timesteps so as to achieve linear time complexity and constant memory space complexity for processing long sequences of input. It uses a slot attention mechanism to write to the memory, where each memory slot attends to the input sequence and to itself to generate an updated memory. Further a forgetting mechanism is implemented to clean up irrelevant or redundant information from the memory.
- **Neural Attention Memory (NAM)**: This implementation reimagines the attention mechanism as a memory architecture for neural networks. It involves writing to a memory matrix using key (**K**) - value (**V**) pairs and reading from it with a query (**Q**) vector. It utilized three differentiable operations for storage and recall, which are read, write and erase.

Retrieval Augmented Generation (RAG) framework

Retrieval Augmented Generation (RAG) frameworks function by incorporating information or knowledge from external data sources, to serve as supplementary reference or instructions for the input query or the generated output. This aids the Transformer model in its generation tasks. Unlike the Memory-Augmented Transformers, the Retrieval Augmented Generation (RAG) does not require the data source (Non-parametric memory) to be tightly integrated into the model specifically during the training process.

The main aim of the framework is to reduce the effects of hallucinations. It does so by first invoking a component called a **retriever** to search and extract relevant documents in the external data source that correspond to the input query. The retrieved documents are then combined with the input query as contextually relevant, time-critical, and domain-specific information to enhance the model's generative capability whilst reducing the effect of hallucination.

The framework consists of three major processes:

1. Retrieval

In RAG frameworks, the retriever component functions as the information provider where given an input query from the Transformer model, it retrieves relevant knowledge by estimating similarity between the input query and documents from external sources.

Retrieval methods can be categorized into two types: **sparse** and **dense** retrieval based on how information is encoded. **Sparse** retrieval relies on word-based representations and is commonly used in text retrieval, where documents are selected according to the presence or frequency of specific terms from the input query. Classic algorithms used for this include TF-IDF and BM25. **Dense** retrieval, on the other hand, embeds both the input query and documents into continuous vector space, where relevancy is determined by similarity metrics such as cosine similarity between the input query and documents vector embeddings, enabling more effective retrieval beyond surface-level word overlap.

Retrieval granularity refers to the level of indexing at which a corpus is segmented for retrieval, such as documents, passages, tokens, or entities. The choice of retrieval granularity has a substantial impact on both effectiveness and efficiency. Coarser granularities, such as full-document retrieval, reduce storage overhead and speed up search but may introduce noise by returning large amounts of irrelevant content. Finer granularities, such as token or entity-level retrieval, provide more precise knowledge access but increase database size and computational cost during search.

2. Augmentation

Augmentation describes the technical process that integrates the retrieved information from the retriever component into the generation process. There are three main types of integration:

- **Input-Layer Integration:** A common strategy for integrating retrieved information is to concatenate the retrieved information with the original input query and jointly process

them through the Transformer model. While this approach has proven effective, it is inherently constrained by the context window of the Transformer model.

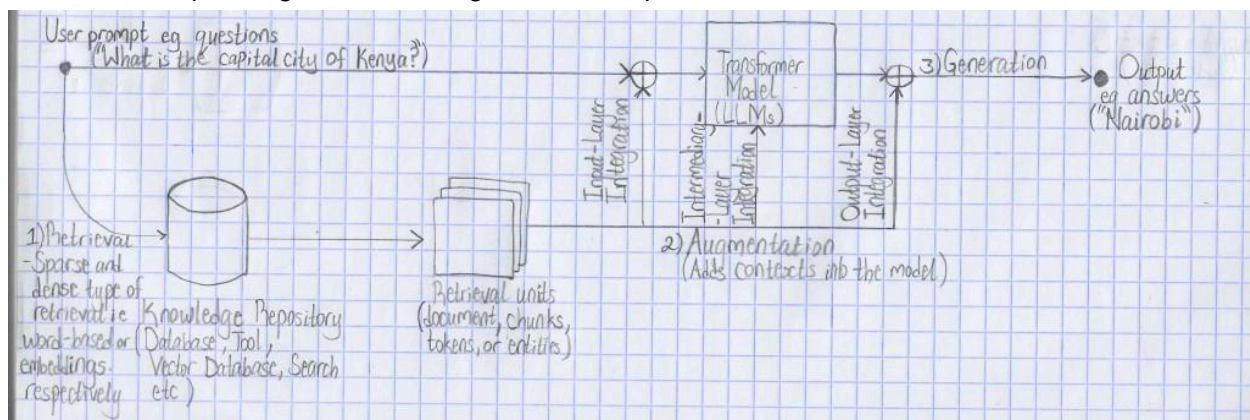
- **Output-Layer Integration:** Another kind of integration strategy is post-hoc, or output-layer integration, in which the retrieved information is combined with the model's generated output rather than the input, effectively merging retrieval and generation results at the final stage.
- **Intermediate-Layer Integration:** This integration strategy involves retrieved information being injected into the Transformer model at select layers. This will allow the external knowledge to influence the hidden representations directly, offering a balance between the contextual richness of input-level integration and the efficiency of output-level integration.

3. Generation

In RAG framework, generators can be broadly classified into two categories:

- **Parameter-Accessible Generators (White-box):** These are generative models whose internal parameters are directly accessible for inspection, and modification. Such accessibility allows the models to be trained or adapted to different retrieval and augmentation strategies, thereby improving the quality and effectiveness of generation.
- **Parameter-Inaccessible Generators (Black-box):** These are generative models whose internal parameters are hidden from the users and are typically accessible only through interfaces or APIs. Interaction is thus limited to providing input queries (prompts) and receiving outputs (responses). In this setting, retrieval and augmentation processes aim to improve performance by enriching the prompt with supplementary knowledge, guidance, or illustrative examples (in-context learning).

Below is a simple diagram illustrating the various processes in a RAG framework:



Information Retrieval (IR) system in the context of Search Engines

An **Information Retrieval (IR)** system is a structured software framework that structures, analyses, organizes, stores, searches, and retrieves information typically in the form of documents. It can handle multiple types of information including text, images, audio, and video. A search engine is usually considered the practical application of information retrieval techniques to large-scale text collection

Users interact with this system through an interface, where they express their information needs as a query. This query is then processed into index terms to be matched against a database. The retrieved documents, and other contents are sorted (ranked) by relevancy and returned to the user as a list, usually this is a list of links to websites in the case of search engines.

The Information Retrieval system generally comprises two core functions:

1. **Index processes:** These processes build the structures that enable efficient search. It involves:
 - a. **Text acquisition:** Identifying and ingesting documents or content to be searched for, usually done by using crawlers in cases of web content. A knowledge repository, often a database (RDBMS), manages the corpus and associated metadata such as document type, document length, etc.
 - b. **Text transformation:** Converting documents into indexable terms or features through the process of tokenization, parsing, and structural recognition such as titles, figures, links, and headings. Additional processes may include stop-word removal, stemming, link analysis (web pages), information extraction (**NER: Named Entity Recognition**), and classification (Identification of class-related metadata for documents or parts of documents). Both documents and queries must undergo consistent transformation to allow effective comparison. The set of all indexed terms forms the index vocabulary.
 - c. **Index creation:** Constructing data structures, such as inverted indices to map terms to the documents in which they appear. This step incorporates document statistics such as counts of index term, number of tokens in documents etc., index term weighting (relative importance of words in documents), and distribution strategies to optimize search performance.
2. **Query processes:** These processes leverage the index structures, created by the index processes, to retrieve relevant content to a user query. It involves:
 - a. **User Interaction:** Managing the interface through which users can submit queries and view search results. The queries are transformed into index terms, which are used by the later processes to get ranked search results for the user. Query refinement techniques such as query suggestions, query expansion, and spell checking can be employed to improve the search performance.
 - b. **Ranking:** The core retrieval operation, which scores documents against the query using a retrieval model such as **TF-IDF**, **BM25**, or neural ranking methods. The results are then ordered by relevancy before being returned to the user. The efficiency of this step depends on the underlying index structures and retrieval model.

- c. **Evaluation:** Assesses both the effectiveness and efficiency of the system. It relies on user interaction data such as log data, click-through rates, or dwell time to measure performance. The insights gained are then used to tune indexing, ranking, and overall system design to improve retrieval outcomes. This is mostly done offline. Other performance metrics that can be measured are:
- i. **Response time:** Delay between user submitting a query and receiving the result list.
 - ii. **Query throughput:** Number of queries that can be processed in a given time.
 - iii. **Indexing speed:** Rate at which text documents can be transformed into indexes for searching.
 - iv. **Coverage:** How much of the existing information has been indexed and stored in the system.
 - v. **Recency or freshness:** “Age” of the stored information.
 - vi. **Index Update Latency:** How fast new data can be incorporated into the indexes.

Retrieval models

This determines how documents are scored and ranked in response to a user query. They determine what is considered relevant by specifying how queries and documents are represented and compared. Some major categories are:

- **Boolean Retrieval models:** Represent queries and documents using boolean logic (AND, OR, NOT) e.g “Machine” AND “Learning” returns documents with both “Machine”, and “Learning” terms.
- **Vector space models:** Documents can be ranked by computing the distance between vector representations of documents and queries.
- **Probabilistic models:** Uses probability theory to estimate the probability that a document is relevant to a query.
- **Neural and Deep learning-base models:** Uses embeddings and neural networks to capture semantic similarity beyond exact word matching.

Social Search

Integrates social signals, such as user interactions, social networks, ratings, likes, shares, tags, or recommendations, into the retrieval and ranking process of a search engine. Unlike traditional systems that rely primarily on textual relevance, social search leverages collective user behaviour and social context to enhance the relevance and personalization of results.

User Tags and Manual Indexing

Many social platforms enable users to assign tags to content, a practice often referred to as collaborative tagging, or folksonomy. The tags serve as additional indexable terms for the IR system, serving as a form of manual indexing in which the content of an object is represented through explicitly assigned descriptors. They can be used for documents, images, videos and audio. In certain areas, the tags could be manually generated by experts, who choose keywords, categories, and other descriptors from a controlled vocabulary (fixed ontology) so as

to ensure the descriptors are standardized.

However, manual indexing is infeasible for large-scale document collections or massive digital media content. To address this, IR systems employ automatic indexing techniques to assign identifiers (terms, phrases, features) to documents during index construction. While automatic methods are more exhaustive and consistent, their quality and accuracy are often lower compared to human-assigned tags.

Categories of tags

Tags can be categorized into several types:

- **Content-based tags:** Tags describing the content of an item e.g. “car”, “woman”, and “sky”.
- **Context-based tags:** Tags that describe the context of an item e.g. “Nairobi City” for a photo of a street there or “KICC” for a document describing the engineering work done on that building.
- **Attribute tags:** Tags that describe implicit attributes of the item e.g. “building” for a document describing the engineering work done on KICC building, “black and white” for a type of movie.
- **Subjective tags:** Tags that subjectively describe an item e.g. “pretty”, “amazing”.
- **Organizational tags:** Tags that help organize items e.g. “todo”, “readme”, “draft”.

Terminology

Term	Definition
IR (Information Retrieval)	The field of computer science focused on finding relevant information in large collections of unstructured data (typically text).
AI (Artificial Intelligence)	The broad field of computer science focused on building systems that can perform tasks requiring human-like intelligence, such as reasoning, perception, learning, and language understanding. Usually used as an umbrella term for intelligent systems.
ML (Machine Learning)	A subfield of AI focused on algorithms and models that learn patterns from data rather than relying on explicitly programmed rules. It provides the foundation for modern algorithms like Transformer models.
NLP (Natural Language Processing)	A subfield of AI and ML focused on enabling computers to process and understand human language. It is a core discipline behind chatbots, translation, summarization, and text classification systems.
CV (Computer Vision)	A subfield of AI and ML focused on enabling machines to interpret and analyse visual information from images or video. Its core tasks include object detection, segmentation, and tracking.
RNN (Recurrent Neural Network)	A type of machine learning model designed for sequential data. It processes input step by step while maintaining hidden states that capture information across time.
CNN (Convolutional Neural Network)	A type of machine learning models that apply convolutional operations to capture local patterns in data, most commonly images but can also be applied to NLP, and audio processing tasks.
Corpus	A large and structured collection of text documents used for training, evaluation, or retrieval in NLP and IR systems.
Tokens	The smallest unit of text a machine learning model processes, which may be a word, subword, or character depending on the algorithm used to split them.
Entities	Real-world objects, concepts, or items that can be identified and classified within text e.g. people, organizations, locations, dates, products etc.
Prompt	The input text (instructions, queries, or examples) provided to a language model to guide its output. It can include system messages, user instructions, and context documents.
Inference	The process of using an already trained machine learning model to generate predictions or outputs from new input data. This is distinct from the training process.

Frontier model(s)	The most advanced, large-scale machine learning models that are at the cutting edge of research and deployment, typically trained with massive compute and data resources.
Foundation model(s)	Large-scale machine learning models trained on broad and diverse datasets that can be adapted (via fine-tuning, prompting, or RAG) to many downstream tasks.
BM25 (Best Match 25)	A ranking function used in information retrieval to estimate the relevance of documents to a given query, based on term frequency and inverse document frequency.
TF-IDF (Term Frequency–Inverse Document Frequency)	A statistical weighting method in information retrieval that scores how important a term is within a document relative to a large corpus, by combining term frequency (TF), which is the number of times a term appears in a document, with inverse document frequency (IDF), which is the a measure of how rare or informative a term is across a collection of documents.
Cosine similarity	A metric that measures the similarity between two vectors by calculating the cosine of the angle between them. It ranges from -1 (opposite) to 1 (identical), with 0 meaning orthogonal (no similarity).
In-context learning	The ability of a machine learning model to adapt its behaviour to new tasks by conditioning on examples or instructions provided in the input prompt, without updating the model's parameters.
Continual Learning	Also known as Incremental Learning is the ability of a model to progressively learn new tasks, one at a time, without forgetting the previously learned ones.
Parametric memory	This refers to long-term memory implicitly stored within a model's parameters. It is acquired during training, where they're embedded in the model's parameters and accessed through feedforward computation at inference. It serves as a form of instant, long-term, and persistent memory.
Non-parametric memory	This refers to long-term memory that is stored externally to the machine learning models. It can be stored in various media such as databases, file systems, and computer memory.
Retrieval model	Formal representation of the process of matching a query and a document, file or other media.
RDBMS (Relational Database Management System)	A software system for managing structured data organized into tables with predefined schemas and relationships. Supports SQL for querying, updating, and managing data. Examples: PostgreSQL, MySQL, etc.

Stop-words	Common words in a language like “the”, “of”, “to”, and “for” that help form sentence structure but offer little semantic meaning on their own.
Stemming	Groups words derived from a common stem i.e. “fish”, “fishes”, and “fishing” can be grouped with one designated word “fish”.
Crawlers	Automated programs that systematically browse the web (or other data sources) to collect and index context.
NER (Named-Entity Recognition)	An NLP task that identifies and classifies entities in text into predefined categories such as people, organizations, locations, dates, or products. Often used as a preprocessing step for information extraction, search, etc
Differentiable	A mathematical property of a function that allows its derivatives to be computed. In machine learning, it means the model's parameters can be optimized. This is what makes neural networks trainable.
Subword	A text unit smaller than a full word but larger than a single character. Produced by tokenization methods such as BytePairEncoding (BPE) or WordPieceEncoding.
Out-of-vocabulary (OOV)	A condition where a token (word, subword, or character) is not present in the model's predefined vocabulary.
Unigram Language Model	A probabilistic model that treats each token as independent and estimates the likelihood of a document or query by multiplying the probabilities of its individual words. Used historically in IR as a ranking function.

Out of Scope (Non-Goals)

- **Frontier model performance:** The focus of the project will be on evaluating the proposed model's capabilities and not producing a frontier model.
- **Dynamic knowledge repository:** The knowledge repository will be static with no mechanisms for updating, adding, or removing information from it.
- **External knowledge access:** All the data will be localized to the knowledge repository, with no access to external data sources.
- **General-purpose reasoning across arbitrary domains:** The scope of the model's reasoning capabilities will be constrained to the information present in the knowledge repository.
- **End-to-end training on raw or unlabeled data:** The project will not rely on large-scale unstructured datasets arbitrarily obtained from sources like the internet, like most frontier models, but instead it will operate on a carefully curated and structured dataset.
- **Input multimodality:** The project will only utilize text data.
- **Benchmark priority:** The project will not prioritize defeating any established state-of-the-art frontier models on any benchmark.
- **Agentic system:** The project will not pursue the development of any autonomous agentic system designed to interact with users and environments.
- **Reinforcement Learning alignment:** The project will not implement any reinforcement learning algorithms to improve the model's performance such as Reinforcement Learning with Human Feedback (RLHF).
- **Prompt engineering approaches:** The project will not utilize any of the prompt engineering approaches such as chain-of-thought (CoT), Tree-of-Thought (ToT), etc. to improve the generation process.

Assumptions

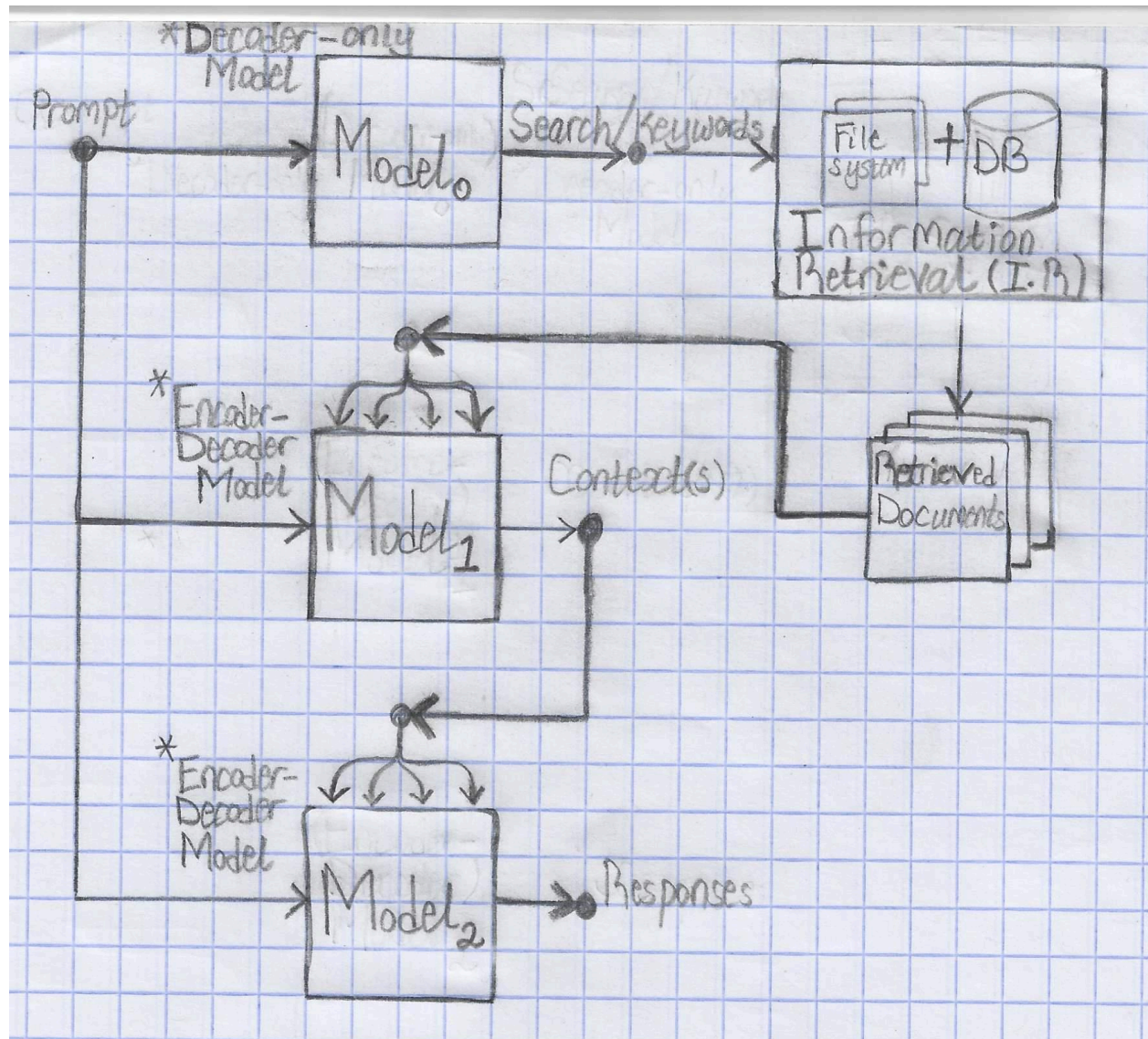
The following assumptions have been made about the project:

- **Model parameter efficiency:** Using non-parametric memory should reduce the model's dependency on parameter scaling (scaling laws), as most knowledge is maintained in external storage rather than embedded in the model's parameters. This should allow for more efficient utilization of the model's parameters for NLP tasks, possibly reducing the training time and computational requirements.
- **Training / Inference cost:** Integrating non-parametric memory into Transformer models should lead to smaller, more performant models which are usually more cost efficient for training and inference.
- **Task decomposition:** By utilizing a modular system that decomposes a complex task into multiple simpler subtasks each with a corresponding subsystem i.e. multiple Transformer models, the overall system should be more scalable, maintainable, and capable as opposed to singular large systems.
- **Dataset creation:** The project requires a well-structured dataset for training. This will be obtained through a combination of automated acquisition and synthetic generation. This process will involve the utilization of crawlers to gather raw data, pre-trained open-source models to generate or refine synthetic samples, and human annotation to ensure quality and relevance specifically with the document tags. The key assumption is this process can be scaled efficiently to generate the needed data for training.
- **Latency:** The modular design of the system will inevitably introduce delays in the information flow between components. To ensure practicality and scalability, these delays must remain within an acceptable limit. In particular, the Information Retrieval system should maintain response time below 100ms, while inter-component communication during inference should achieve comparable performance, preserving overall system efficiency.
- **Generalization:** The model(s) in the overall system should generalize properly to unseen dataset i.e. data held out during training. The unseen data should produce coherent and accurate results during testing.
- **External storage scaling:** Expanding the data storage capacity in the Information Retrieval (IR) system should be cost efficient and scalable.
- **Modularity:** Each component in the overall system should be designed for independent training, updating, and deployment, allowing multiple models to be improved without affecting each other. To preserve flexibility and scalability, the modules must remain loosely coupled, minimizing interdependencies while ensuring smooth integration within the overall architecture.
- **Enhanced reasoning capability:** By offloading knowledge storage to non-parametric memory, the model's capability can be primarily allocated to higher-order language understanding and reasoning, rather than dedicating capacity to memorization of facts.
- **Scaling law:** The performance of the model is positively correlated with the scale of model parameters, size of training dataset, and amount of computation used. The performance of a smaller model should be indicative of the capability of the architecture and if scaled up it should improve in a predictable manner.

Proposed Design

The project aims to develop a system of modular components arranged sequentially, consisting of multiple Transformer models alongside an Information Retrieval subsystem. Each module is responsible for a narrowly defined task, adhering strictly to the separation-of-concerns principles and maintaining minimal interdependencies.

System Architecture



The system consists of three Transformer models, one Decoder-only Transformer model and two Encoder-Decoder Transformer models, along with an Information Retrieval (IR) subsystem. These modules are all interconnected in a sequential pipeline, where the output of each component serves as the input to the next.

The overall system workflow can be outlined as follows:

1. A person provides a **prompt** into the system in the form of a question, or an instruction.
2. The **prompt** is initially passed to **Model₀**, a Decoder-only Transformer, whose role is to generate **Search terms** or **Keywords** (Tags) relating to the prompt. These Tags serve as cues for the retrieval process in the succeeding component to help in filtering and ranking stored data.
3. The **Information Retrieval (IR)** subsystem will then take the tags provided by the previous component, **Model₀**, and search its internal database using basic algorithms like **Boolean retrieval**, to filter and retrieve them as documents. Then using a scoring algorithm similar to **TF-IDF**, it will rank them by order of relevancy.
4. The retrieved documents are then passed to the next component, **Model₁**, an Encoder-Decoder Transformer model. In this stage, the Encoder transforms the retrieved documents into a compatible representation, which is subsequently integrated with the initial user prompt in the Decoder. **Model₁** initially evaluates the relevance of each retrieved document to the user prompt, filtering out those deemed irrelevant. The relevant documents are then used to generate summaries, illustrative examples, or refined step-by-step instructions, which will serve as enriched contextual information (**contexts**) for the succeeding component: **Model₂**.
5. This component, **Model₂**, is an Encoder-Decoder Transformer model in which the Encoder processes the **contexts** from **Model₁** into compatible representations, and the Decoder integrates them using an Intermediate-Layer Integration strategy with the original user prompt. The role of the model is to generate a coherent response to the user prompt in the form of an answer to a question or any other appropriate text.

The overall system will first be **pre-trained** with a well-structured dataset to strengthen their respective language understanding and reasoning capability. The dataset will initially consist of two parts:

- **Content:** The full source text.
- **Context:** A concise summary of the Content text strictly following a well-defined standard.

The models will be trained as follows:

1. **Model₀**: This model will be trained exclusively on the Content.
2. **Model₁**: This model will be trained on the Context augmented by the Content.
3. **Model₂**: This model will be inversely trained on the Content augmented by the Context.

The trained models will then be **fine-tuned** on synthetically generated datasets organized into four parts:

- **Content:** The full source text.
- **Context:** A concise summary of the Content, highlighting relevant information about the prompts.
- **Tags:** A list of descriptive keywords associated with the prompt.
- **Prompt:** Synthesized user input usually in the form of questions or instructions.

The models will be fine-tuned as follows:

1. **Model₀**: This model will be trained on Prompts combined with Tags.

2. **Model₁**: This model will be trained on Prompts combined with Context, and augmented by the Content.
3. **Model₂**: This model will inversely be trained on Prompts combined with Content, and augmented by the Context.

The Content and their respective Tags will need to be stored internally in the Information Retrieval (IR) subsystem for retrieval during the inference process. This subsystem will need to store numerous documents from a wide variety of domains for the overall system to adequately function in a real-world scenario. The documents will be stored in a file system, while their associated tags will be maintained in a relational database (RDBMS).

In addition the system's component will need to utilize specialized tokens to signal special operations to be performed and to clearly delineate the boundaries of information segments, such as indicating the start and end of specific content. These tokens will be used to guide the flow of information within the components, and signal for actions that need to be performed on the data. They include:

- **[null]**: Null token; Standard token used for the absence of context in **Model₂**.
- **[Pad]**: Padding token; Standard placeholder token to ensure inputs to a model are of the same context size, it is usually the max context window each model can support.
- **[SP] / [EP]**: Start of Prompt / End of Prompt tokens; denotes the start and end of user prompt(s) used by all models.
- **[ST] / [ET]**: Start of Tags / End of Tags tokens; denotes the start and end of tags generated by **Model₀**.
- **[SR] / [ER]**: Start of Response / End of Response tokens; denotes the start and end of **Model₂** output response.
- **[RLV] / [NRLV]**: Relevant / Not Relevant tokens; specifies whether a retrieved document from the **Information Retrieval** subsystem is relevant to the user prompt (determined by **Model₁**).
- **[SDocument] / [EDocument]**: Start of Document / End of Document tokens; denotes the start and end of retrieved document content used by **Model₁**.
- **[SContext] / [EContext]**: Start of Context / End of Context; denotes the start and end of context(s) generated by **Model₁**.

The core innovation of this system lies in the utilization of an automated form of in-context learning in **Model₂**. The user doesn't need to supplement the prompt with additional contextual information, but rather the system stores information internally on numerous domains which can be retrieved, filtered, and processed into a contextually rich representation that guides the generation process of the Transformer model.

A key distinction of this approach, compared to most Retrieval-Augmented Generation (RAG) implementations, lies in its training process: contextual information is explicitly incorporated into the models (Intermediate Layer Integration) to encourage them to prioritize language understanding and reasoning over memorization. In addition, instead of relying on a single general-purpose model to perform multiple tasks, the system will employ a sequential pipeline of

modules, comprising multiple Transformer models and an Information Retrieval (IR) subsystem, each specialized for a narrowly-defined task.

Dataset

Training Data

The training data collected from numerous sources will need to have synthetically generated summaries and user-like prompts for training. This can be done using open-source multimodal frontier models, such as **gpt-oss-20b**.

To ensure the summaries are generated consistently and usable by the model, the following guidelines should be applied (**Work In Progress**):

- Breaking down content:
 - Extract all key facts and information from content into concise summaries.
 - Do not copy sentences verbatim unless unavoidable; always condense.
 - Context must remain traceable to the original content.
 - Use plain, simple English. Don't mimic the source's content writing style.
- Context Labels and Categories:
 - Context Format: **<Context Label>::<Summary>**
 - Summaries can be subdivided into categories for conciseness and clarity using the format: **<Categories>:<Summaries>**. If the summary is purely descriptive rather than factual, the **<Category>** label may be omitted.
 - **<Context Labels>** and **<Categories>** can be inferred from the content.
 - Use **Title Case** naming convention for labels (e.g **Content Type, Heading**).
- Integrity of Content:
 - Don't add, interpret, or speculate beyond the source content.
 - Preserve the names of places, people, organizations, abbreviations and specific domain terminologies.
 - Keep the order and structure of information (e.g. **Content, Headings, SubHeadings**).
- Formatting rules:
 - Separate contexts with ;;
 - Separate disjointed summaries belonging to the same **<Context Labels>** with ;
 - Supplementary information can be added in parentheses ().
 - If ;; is unavoidable in a summary, escape with \;;.
- Standard **<Context Labels>** (Non-exhaustive):
 - General:
 - **Content Type**: The type of the content e.g. story, poem, mathematical problem, play, etc.
 - **Writing Style**: The style of writing used in the source content e.g. expository, descriptive, narrative, dramatic, poetic, technical, journalistic etc.
 - Literature:
 - **Heading**: The main title at the beginning of an article, essay, or section.
 - **SubHeading**: The secondary heading under a **Heading**, or other **SubHeadings**.
 - **Dialogue**: The content of the dialogues in plays.
 - **From**: The sender of an email.

- **To:** The recipient of an email.
- **Subject:** The subject of an email.
- **Content:** A distinct section of a piece of writing, usually dealing with a single theme and indicated by a new line, indentation, or numbering.
- **Stanza:** A group of lines separated by spacing, functioning like a paragraph within a poem.
- **Act:** The major divisions of a play, often marking shifts in time, location, or dramatic arc.
- **Scene:** The subdivisions of an act, usually tied to a particular location or group of characters in a play.
- Code (Language Agnostic)
 - **Imports:** Name of the external code or libraries being imported by the code.
 - **Operation:** Describes the functionality or purpose of each line of code (in simple english).
- Mathematics / Technical:
 - **Problem Statement:** The task or question posed in text.
 - **Solution:** The resolved outcome or result in the text.
 - **Solution Step:** A unit of reasoning or calculation in the process in the text.

The processed data will then need to be organized into a structured format, similarly to benchmark datasets like the **Stanford Question Answering Dataset (SQuAD)**.

The following is a hypothetical example of data structured for **Pre-training**:

```
{
  "Content": "Kenya, officially the Republic of Kenya, is a country located in East Africa. With an estimated population of more than 52.4 million as of mid-2024, Kenya is the 27th-most-populous country in the world and the 7th most populous in Africa. Kenya's capital and largest city is Nairobi. The second-largest and oldest city is Mombasa, a major port city located on Mombasa Island. Other major cities within the country include Kisumu, Nakuru and Eldoret. Going clockwise, Kenya is bordered by South Sudan to the northwest (though much of that border includes the disputed Ilemi Triangle), Ethiopia to the north, Somalia to the east, the Indian Ocean to the southeast, Tanzania to the southwest, and Lake Victoria and Uganda to the west.",
  "Context": "Content Type::Encyclopedic (Geography);;Writing Style::Expository;;Content::Kenya (Officially the Republic of Kenya): East African country;Population::~~52.4 million as of mid-2024 (27th in the world, 7th in Africa);Nairobi: Capital city (largest city);Mombasa: 2nd largest and oldest city (major port city located on Mombasa Island);Major cities: Kisumu, Nakuru, and Eldoret;Bordering (Clockwise): South Sudan (NorthWest, though much of the border includes the disputed Ilemi Triangle),Ethiopia(North),Somalia(East),Indian Ocean(SouthEast), Lake Victoria and Uganda(West).;;",
}
```


The following is a hypothetical example of data structured for **Fine-tuning**:

```
{  
    "Content": "Kenya, officially the Republic of Kenya, is a country located in East Africa. With an estimated population of more than 52.4 million as of mid-2024, Kenya is the 27th-most-populous country in the world and the 7th most populous in Africa. Kenya's capital and largest city is Nairobi. The second-largest and oldest city is Mombasa, a major port city located on Mombasa Island. Other major cities within the country include Kisumu, Nakuru and Eldoret. Going clockwise, Kenya is bordered by South Sudan to the northwest (though much of that border includes the disputed Ilemi Triangle), Ethiopia to the north, Somalia to the east, the Indian Ocean to the southeast, Tanzania to the southwest, and Lake Victoria and Uganda to the west.",  
    "Tags": ["Kenya", "Capital City"]  
    "Context": "[RLV]Content Type::Encyclopedic (Geography);;Nairobi: Capital city (largest city);;",  
    "Prompt": "[SP]What is the capital city of Kenya?[EP][SR]The capital of Kenya is Nairobi.[ER]"  
}
```

References

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, June 12; revised August 2, 2023). *Attention Is All You Need* (arXiv:1706.03762) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021, June 8; revised June 15). *A survey of Transformers* (arXiv:2106.04554) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2106.04554>
- He, Z., Lin, W., Zheng, H., Zhang, F., Jones, M. W., Aitchison, L., Xu, X., Liu, M., Kristensson, P. O., & Shen, J. (2024, November 1; revised January 12, 2025). *Human-inspired perspectives: A survey on AI long-term memory* (arXiv:2411.00489) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2411.00489>
- Omidi, P., Huang, X., Laborieux, A., Nikpour, B., Shi, T., & Eshaghi, A. (2025, August 14; revised August 16). *Memory-Augmented Transformers: A Systematic Review from Neuroscience Principles to Enhanced Model Architectures* (arXiv:2508.10824v2) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2508.10824>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020, January 23). *Scaling laws for neural language models* (arXiv:2001.08361) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2001.08361>
- Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search Engines: Information Retrieval in Practice*. Addison-Wesley.
- Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024, May 10). *A survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models* (arXiv:2405.06211) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2405.06211>
- Khosla, S., Zhu, Z., & He, Y. (2023, December 11; revised December 13). *Survey on Memory-Augmented Neural Networks: Cognitive Insights to AI Applications* (arXiv:2312.06141v2) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2312.06141>
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024, February 9; revised March 23, 2025). *Large language models: A survey* (arXiv:2402.06196v3) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2402.06196>
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023, September 3; revised September 24). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models* (arXiv:2309.01219v2) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2309.01219>

Aleixo, E. L., Colonna, J. G., Cristo, M., & Fernandes, E. (2023, December 16). *Catastrophic forgetting in deep learning: A comprehensive taxonomy* (arXiv:2312.10549) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2312.10549>

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024, May 31; revised February 7, 2025). *The rising costs of training frontier AI models* (arXiv:2405.21015v2) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2405.21015>

Appendices

Appendix A: Examples of various Pre-training data by categories

EXAMPLE 1: Literary and Creative writings

Frankenstein; Novel

Content:

Letter 1

To Mrs. Saville, England.

St. Petersburg, Dec. 11th, 17—.

You will rejoice to hear that no disaster has accompanied the commencement of an enterprise which you have regarded with such evil forebodings. I arrived here yesterday, and my first task is to assure my dear sister of my welfare and increasing confidence in the success of my undertaking.

I am already far north of London, and as I walk in the streets of Petersburg, I feel a cold northern breeze play upon my cheeks, which braces my nerves and fills me with delight. Do you understand this feeling? This breeze, which has travelled from the regions towards which I am advancing, gives me a foretaste of those icy climes. Inspirited by this wind of promise, my daydreams become more fervent and vivid. I try in vain to be persuaded that the pole is the seat of frost and desolation;; it ever presents itself to my imagination as the region of beauty and delight. There, Margaret, the sun is for ever visible, its broad disk just skirting the horizon and diffusing a perpetual splendour. There—for with your leave, my sister, I will put some trust in preceding navigators—there snow and frost are banished;; and, sailing over a calm sea, we may be wafted to a land surpassing in wonders and in beauty every region hitherto discovered on the habitable globe. Its productions and features may be without example, as the phenomena of the heavenly bodies undoubtedly are in those undiscovered solitudes. What may not be expected in a country of eternal light? I may there discover the wondrous power which attracts the needle and may regulate a thousand celestial observations that require only this voyage to render their seeming eccentricities consistent for ever. I shall satiate my ardent curiosity with the sight of a part of the world never before visited, and may tread a land never before imprinted by the foot of man. These are my enticements, and they are sufficient to conquer all fear of danger or death and to induce me to commence this laborious voyage with the joy a child feels when he embarks in a little boat, with his holiday mates, on an expedition of discovery up his native river. But supposing all these conjectures to be false, you cannot contest the inestimable benefit which I shall confer on all mankind, to the last generation, by discovering a passage near the pole to those countries, to reach which at present so many months are requisite;; or by ascertaining the secret of the magnet, which, if at all possible, can only be effected by an undertaking such as mine.

Context:

Content Type::Letter;;Writing Style::Epistolary / Narrative;;Heading::"Letter 1";;SubHeading::"To Mrs. Saville, England.";;SubHeading::"St. Petersburg, Dec. 11th, 17—.";;Content::Writer assures sister;No disaster at start;Arrived yesterday;Well-being confirmed;Confidence in success;;Content::Far north of London;In Petersburg;Cold breeze delights;Foretaste of icy regions;Pole imagined as beauty not desolation;Sun ever visible;Calm seas;Undiscovered land of wonders;Hopes to find magnetic power;Curiosity drives voyage;Joy compared to child's river expedition;Even if false, journey benefits mankind;Possible pole passage;Secret of magnet;;

EXAMPLE 2: Educational and Academic materials

Wikipedia Article (Kenya)

Content:

Kenya, officially the Republic of Kenya, is a country located in East Africa. With an estimated population of more than 52.4 million as of mid-2024, Kenya is the 27th-most-populous country in the world and the 7th most populous in Africa. Kenya's capital and largest city is Nairobi. The second-largest and oldest city is Mombasa, a major port city located on Mombasa Island. Other major cities within the country include Kisumu, Nakuru and Eldoret. Going clockwise, Kenya is bordered by South Sudan to the northwest (though much of that border includes the disputed Ilemi Triangle), Ethiopia to the north, Somalia to the east, the Indian Ocean to the southeast, Tanzania to the southwest, and Lake Victoria and Uganda to the west.

Context:

Content Type::Encyclopedic (Geography);;Writing Style::Expository;;Content::Kenya (Officially the Republic of Kenya): East African country;Population:~52.4 million as of mid-2024 (27th in the world, 7th in Africa);Nairobi: Capital city (largest city);Mombasa: 2nd largest and oldest city (major port city located on Mombasa Island);Major cities: Kisumu, Nakuru, and Eldoret;Bordering (Clockwise): South Sudan (NorthWest, though much of the border includes the disputed Ilemi Triangle),Ethiopia(North),Somalia(East),Indian Ocean(SouthEast), Lake Victoria and Uganda(West).;;

EXAMPLE 3: Technical and Professional Content

Python Code

Content:

Solve the quadratic equation $ax^2 + bx + c = 0$

```
# import complex math module
import cmath
```

```
a = 1
b = 5
c = 6
```

```
# calculate the discriminant
d = (b**2) - (4*a*c)
```

```
# find two solutions
sol1 = (-b-cmath.sqrt(d))/(2*a)
sol2 = (-b+cmath.sqrt(d))/(2*a)
```

```
print('The solution are {0} and {1}'.format(sol1,sol2))
```

Context:

Content Type::Python Code;;Writing Style::Technical / Coding;;Problem Statement::Solve quadratic equation $ax^2 + bx + c = 0$;;Imports::cmath;;Operation::Define coefficients $a=1$; $b=5$; $c=6$;;Operation::Compute discriminant $d = (b^2) - (4ac)$;;Operation::Compute root $sol1 = \frac{-b - \sqrt{d}}{2a}$;;Operation::Compute root $sol2 = \frac{-b + \sqrt{d}}{2a}$;;Operation::Print solutions $sol1$ and $sol2$;;

EXAMPLE 4: News and Journalism

Newspaper Article

Content:

Alzheimer's disease: Causes, symptoms, diagnosis, and treatment

Wednesday, October 12, 2022

By Denis Mutua

What you need to know:

- Alzheimer's disease is a deteriorative condition of the brain.
- It results in progressive loss of cognitive and memory function, and is most associated with ageing.

Alzheimer's disease is an irreversible, progressive brain disease characterised by the destruction and loss of function of brain cells. The destruction and loss of function of the brain cells leads to dementia i.e., the loss of cognitive function - memory, reasoning, and thinking.

Based on this relationship, Alzheimer's disease is described both as a cause and a type of dementia. As the condition progresses, the loss of cognitive function hinders the person's ability to carry out the simplest tasks of daily living. Advanced Alzheimer's disease also affects the patients' social well-being as they forget the people in their lives and their relationship.

What causes Alzheimer's?

It has been over a century since Alois Alzheimer observed anomalies in the brains of patients with dementia. Since then, scientists have quite a clear understanding of the biochemical changes that happen in the brain of patients with Alzheimer's disease.

Despite the remarkable progress in understanding the disease, it is still not clear what causes the biochemical changes that lead to the deterioration of the brain and subsequent dementia.

However, some risk factors have been identified as possible contributors to the development of the disease.

Context:

Content Type::Health Article;;Writing Style::Expository / Informative;;Heading::Alzheimer's disease: Causes, symptoms, diagnosis, and treatment;;SubHeading::Wednesday, October 12, 2022;;Author::Denis Mutua;;SubHeading::What you need to know;;Content::Alzheimer's disease;Deteriorative brain condition;Progressive cognitive and memory loss;Most associated with ageing;;Content::Irreversible;Destruction and loss of brain cell function;Leads to dementia;Loss of memory, reasoning, thinking;;Content::Alzheimer's as cause and type of dementia;Progression hinders daily tasks;Advanced stage affects social well-being;Patients forget people and relationships;;SubHeading::What causes Alzheimer's?;;Content::Over a century since Alois Alzheimer's observation;Scientists understand biochemical brain changes;;Content::Cause of biochemical changes unclear;Progress of understanding significant;;Content::Risk factors identified;Possible contributors to disease development;;

EXAMPLE 5: Business and Workplace Documents

Email

Content:

From: Steve Jobs

Sent: Thursday, May 26 2005 9:36 AM

To: Bruce Chizen

Subject: Recruiting

Bruce,

Adobe is recruiting from Apple. They have hired one person already and are calling lots more. I have a standing policy with our recruiters that we don't recruit from Adobe. It seems you have a different policy. One of us must change our policy. Please let me know who.

Steve

Context:

Content Type::Email;;Writing Style::Direct / Business;;From::Steve Jobs;;Sent::Thursday, May 26 2005, 9:36 AM;;To::Bruce Chizen;;Subject::Recruiting;;Content::Adobe recruiting from Apple;One Apple employee hired;More being contacted;Apple policy: no recruiting from Adobe;Adobe policy: different;Jobs demands policy change;Requests clarification who must change;;

EXAMPLE 6: Government and Legal Records

Constitution of Kenya

Content:

THE CONSTITUTION OF KENYA CHAPTER ONE—SOVEREIGNTY OF THE PEOPLE AND SUPREMACY OF THIS CONSTITUTION

Sovereignty of the people.

1. (1) All sovereign power belongs to the people of Kenya and shall be exercised only in accordance with this Constitution.

(2) The people may exercise their sovereign power either directly or through their democratically elected representatives.

(3) Sovereign power under this Constitution is delegated to the following State organs, which shall perform their functions in accordance with this Constitution—

- (a) Parliament and the legislative assemblies in the county Governments;
- (b) the national executive and the executive structures in the county governments; and
- (c) the Judiciary and independent tribunals.

(4) The sovereign power of the people is exercised at—

- (a) the national level; and
- (b) the county level.

Supremacy of this Constitution.

2. (1) This Constitution is the supreme law of the Republic and binds all persons and all State organs at both levels of government.

(2) No person may claim or exercise State authority except as authorised under this Constitution.

(3) The validity or legality of this Constitution is not subject to challenge by or before any court or other State organ.

(4) Any law, including customary law, that is inconsistent with this Constitution is void to the extent of the inconsistency, and any act or omission in contravention of this Constitution is invalid.

(5) The general rules of international law shall form part of the law of Kenya.

Context:

Content Type::Legal Document;;Writing Style::Formal / Legislative;;Heading::THE CONSTITUTION OF KENYA;;SubHeading::CHAPTER ONE—SOVEREIGNTY OF THE PEOPLE AND SUPREMACY OF THIS CONSTITUTION;;SubHeading::Sovereignty of the People;;Content::All sovereign power belongs to the people;Exercised only per Constitution;;Content::Power exercised directly or through elected representatives;;Content::Sovereign power delegated to State organs;Organs: Parliament and county assemblies;National executive and county executive;Judiciary and independent tribunals;;Content::Sovereign power exercised at national level;County level;;SubHeading::Supremacy of this Constitution;;Content::Constitution is supreme law;Binds all persons and State organs at both government levels;;Content::No person may claim or exercise State authority except as authorised;;Content::Validity or legality of Constitution not subject to challenge;;Content::Law inconsistent with Constitution is void;Act or omission contravening Constitution is invalid;;Content::General rules of international law form part of law of Kenya;;

EXAMPLE 7: Everyday and Informal Communication

Social Media Posts (Reddit - GreenText)

Content:

Anonymous 02/22/17(Wed)09:11:28 No.79676775

>> 79675930 #

This reminds me of something i read from the Bush War

> Rhodes recover enemy rifles after each engagement

> They keep finding rifles with the rear sight dialed in to the highest setting, practically f*cking indirect fire

> They are extremely confused by this

> Eventually they ask a prisoner about it during interrogation

> Literally none of them know what the sights are for, they think that choosing the bigger number makes the gun shoot harder

Warfare in Africa is truly unbelievable

>>79678011 # >>79678022 # >> 79678742 #

Context:

Content Type::Forum Thread Post;;Writing Style::Informal / Anecdotal;;Post::Anonymous;Date: 02/22/17 (Wed) 09:11:28;Post No: 79676775;;Reference::Replies to 79675930;;Content::Story from Bush War;Rhodesian soldiers recover enemy rifles;Rifles often found with rear sight at highest setting;Confusion among soldiers;Interrogation of prisoner;Enemy fighters unaware of sight purpose;Belief: bigger number = gun shoots harder;Commentary: Warfare in Africa described as unbelievable;;Reference::Replies received:79678011;79678022;79678742;;

EXAMPLE 8: Cultural and Historical Texts

Speech by the Rev. Martin Luther King Jr. at the “March on Washington,” 1963

Content:

I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

Five score years ago a great American in whose symbolic shadow we stand today signed the Emancipation Proclamation. This momentous decree is a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity. But 100 years later the Negro still is not free. One hundred years later the life of the Negro is still badly crippled by the manacles of segregation and the chains of discrimination. One hundred years later the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later the Negro is still languished in the corners of American society and finds himself in exile in his own land. So we've come here today to dramatize a shameful condition.

In a sense we've come to our nation's capital to cash a check. When the architects of our Republic wrote the magnificent words of the Constitution and the Declaration of Independence, they were signing a promissory note to which every American was to fall heir. This note was a promise that all men—yes, black men as well as white men—would be guaranteed the unalienable rights of life, liberty and the pursuit of happiness. . . .

Context:

Content Type::Speech;;Writing Style::Oratory / Persuasive;;Content::Opening;Greatest demonstration for freedom;Historic significance;;Content::Five score years ago;Emancipation Proclamation signed;Beacon of hope for slaves;End of captivity;;Content::100 years later;Negro still not free;Segregation;Discrimination;Poverty;Exile in own land;Shameful condition dramatized;;Content::Nation's capital;Cash a check metaphor;Constitution and Declaration;Promissory note;Guarantee of unalienable rights;Life, liberty, pursuit of happiness;;

EXAMPLE 9: Instructional and How-to Guides

How-To guide on drinking water from a cup.

Content:

To drink water from a cup, grasp the cup, bring it to your lips, tilt it to allow a small sip to flow into your mouth, swallow, and then lower the cup. For a better experience, sit down, drink slowly, and avoid tilting the cup too far back to prevent spills.

Steps to Drink from a Cup

1. **Grasp the cup:** Hold the cup firmly with your dominant hand or both hands.
2. **Bring to lips:** Lift the cup and bring it to your mouth.
3. **Tilt and sip:** Gently tilt the cup to let water flow into your mouth.
4. **Swallow:** Swallow the water, relaxing your throat to help it go down smoothly.
5. **Lower the cup:** Move the cup away from your mouth and lower it.

Tips for a Better Drinking Experience

- **Sit down:** Sitting while drinking can help you avoid issues and fully benefit from the water.
- **Drink slowly:** Take small sips to avoid choking or spluttering.
- **Don't overfill the cup:** A full cup is more likely to spill.
- **Mindfulness:** Pay attention to the sensation of the water as it enters your mouth, which can be a moment for mindfulness.
- **Pinch paper cups:** When drinking from a paper cup during a race, you can pinch the top to create a smaller spout, making it easier to drink without splashing.
- **Relax your throat:** Relaxing your throat muscles can make swallowing easier.

Context:

Content Type::Instructional;;Writing Style::Step-by-Step / Practical;;Heading::Steps to Drink from a Cup;;Step::Grasp cup;Hold firmly with dominant hand or both hands;;Step::Bring to lips;Lift cup to mouth;;Step::Tilt and sip;Gently tilt;Allow water to flow;;Step::Swallow;Relax throat;Ease water down;;Step::Lower cup;Move away from mouth;Return to rest;;Heading::Tips for Better Drinking Experience;;Tip::Sit down;Avoid issues;Enhance benefit;;Tip::Drink slowly;Small sips;Prevent choking;;Tip::Don't overfill;Reduce spills;;Tip::Mindfulness;Notice sensation;Moment of awareness;;Tip::Pinch paper cup;Create spout;Prevent splashing;;Tip::Relax throat;Ease swallowing;;

Appendix B: Hypothetical example of how the system could function (Model₀ + IR Subsystem + Model₁ + Model₂).

NOTE:

- The documents in the file storage are grouped hierarchically using the following structure: **Domain/Topic/Sub-Topic(s)**.
- Comments about any of the component's operations will use the following format: ** This is a comment **

Example 1

Model₀

[SP]

<User>

Is this email a scam:

"Your Apple ID will be Disabled Because of Some Violated Polices

Dear Customer,

For your safety, your Apple ID has been disabled because some information Appears to be missing or invalid. And its against out policy terms of service to Give fake identity in yout apple account. therefore we need to re-verify your account data at this link: 'http://thisisascamlol.com'. If you did not verify your account within 48 hrs yout account will be permanently locker."

</User>

[EP]

[ST]"Organizations/Apple", "IT/Cyber Crime", "Crime/Scam", "IT/Email", "IT/Phishing", "Crime/Identity Theft", "Crime/Investment Scams", "IT/Malware", "Crime/Extortion"**[ET]**

Information Retrieval (IR) Subsystem

The following operations will be performed:

1. The text enclosed between the **[ST]** / **[ET]** tokens is parsed into a list of individual tags, separated by commas.
2. The Information Retrieval subsystem searches its database for these tags and their associated documents and retrieves them.
3. The retrieved documents are ranked according to their relevance, using a weighting score similar to the TF-IDF algorithm.
4. The top K documents are then passed to the next component, where K is the maximum number of documents that can be retrieved.

Model₁

[SP]

<User>

Is this email a scam:

"Your Apple ID will be Disabled Because of Some Violated Policies

Dear Customer,

For your safety, your Apple ID has been disabled because some information Appears to be missing or invalid. And its against out policy terms of service to Give fake identity in yout apple account. therefore we need to re-verify your account data at this link: 'http://thisisascamlol.com'. If you did not verify your account within 48 hrs yout account will be permanently locker."

</User>

[EP]

** Each document from the IR subsystem is processed independently by **model₁**. **

[SDocument:Organizations/Apple/Products.pdf]

[NRLV]

[EDocument]

[SDocument:Organizations/Apple/Policies.pdf]

[RLV]

- Apple policies don't explicitly detail reasons for Apple ID disabling but rather provide guidelines for what to do when it occurs, which often happens due to security concerns, multiple incorrect password attempts, or billing and payment issues, though violations of Terms of Service are also a possibility.
- An Apple ID might be disabled for several reasons, often related to security or billing: Security Reasons, Billing and Payment Issues, Inactivity, and Violation of Terms of Service.

[EDocument]

[SDocument:Organizations/Apple/Website.pdf]

[RLV]

- Apple's Account link: "<https://support.apple.com/en/>".

[EDocument]

[SDocument:IT/Email/Etiquette.pdf]

[RLV]

- Business Email Etiquette: Most emails from businesses tend to always state their user's account name, and have proper spelling and grammar.

[EDocument]

[SDocument:IT/Scam/Phishing.pdf]

[RLV]

- Phishing is a type of cybercrime and social engineering attack where criminals impersonate legitimate individuals or organizations to trick victims into revealing sensitive information, such as passwords, bank details, or personal data.

[EDocument]

...

Model₂

**** Contexts are aggregated and passed to the Encoder model. ****

[SContext]

- Apple policies don't explicitly detail reasons for Apple ID disabling but rather provide guidelines for what to do when it occurs, which often happens due to security concerns, multiple incorrect password attempts, or billing and payment issues, though violations of Terms of Service are also a possibility.
- An Apple ID might be disabled for several reasons, often related to security or billing: Security Reasons, Billing and Payment Issues, Inactivity, and Violation of Terms of Service.
- Apple's Account link: "<https://support.apple.com/en/>".
- Phishing is a type of cybercrime and social engineering attack where criminals impersonate legitimate individuals or organizations to trick victims into revealing sensitive information, such as passwords, bank details, or personal data.
- Business Email Etiquette: Most emails from businesses tend to always state their user's account name, and have proper spelling and grammar.
- ...

[EContext]

**** Prompt and Response is passed to the Decoder model. ****

[SP]

<User>

Is this email a scam:

"Your Apple ID will be Disabled Because of Some Violated Polices

Dear Customer,

For your safety, your Apple ID has been disabled because some information

Appears to be missing or invalid. And its against out policy terms of service to

Give fake identity in yout apple account. therefore we need to re-verify your

account data at this link: 'http://thisisascamlol.com'. If you did not verify your account within 48 hrs yout account will be permanently locker."

</User>

[EP]

[SR]

<Chatbot>

Yes, the email is a scam (a classic phishing attempt). Here are the key red flags:

Grammar and spelling errors: Legitimate Apple emails are carefully proofread. Phrases like

"violated polices" and "against out policy" would never appear in official communication.

Suspicious link: Apple will never ask you to verify your account by clicking on a non-Apple URL.

They only direct you to apple.com.

Generic greeting: Apple usually addresses you by your real name, not "Dear Customer."

</Chatbot>

[ER]

Example 2

Model₀

[SP]

<User>

Translate this phrase to English: "Fisi na Kunguru walikuwa marafiki sana ingawa walikuwa tofauti. Kunguru aliweza kuruka lakini Fisi alitembea tu."

</User>

[EP]

[ST] "Kiswahili-English:Fisi", "Kiswahili-English:Kunguru", "Kiswahili-English:marafiki", "Kiswahili-English:kuruka", ... **[ET]**

Information Retrieval (IR) Subsystem

The following operations will be performed:

1. The text enclosed between the **[ST]** / **[ET]** tokens is parsed into a list of individual tags, separated by commas.
2. The Information Retrieval subsystem searches its database for these tags and their associated documents and retrieves them.
3. The retrieved documents are ranked according to their relevance, using a weighting score similar to the TF-IDF algorithm.
4. The top K documents are then passed to the next component, where K is the maximum number of documents that can be retrieved.

Model₁

[SP]

<User>

Translate this phrase to English: "Fisi na Kunguru walikuwa marafiki sana ingawa walikuwa tofauti. Kunguru aliweza kuruka lakini Fisi alitembea tu."

</User>

[EP]

**** Each document from the IR subsystem is processed independently by model₁. ****

[SDocument:Kiswahili-English/Fisi]

[RLV]

- Fisi - Swahili word meaning hyena.

[EDocument]

[SDocument:Kiswahili-English/Kunguru]

[RLV]

- Kunguru - Swahili word meaning crow.

[EDocument]

...

Model₂

**** Contexts are aggregated and passed to the Encoder model. ****

[SContext]

- Fisi - hyena.
- Kunguru - crow.
- ...

[EContext]

**** Prompts and responses are processed by the Decoder model. ****

[SP]

<User>

Translate this phrase to English: "Fisi na Kunguru walikuwa marafiki sana ingawa walikuwa tofauti. Kunguru aliweza kuruka lakini Fisi alitembea tu."

</User>

[EP]

[SR]

<Chatbot>

Here's the translation into English:

"The Hyena and the Crow were very good friends even though they were different. The Crow could fly but the Hyena only walked."

</Chatbot>

[ER]

Example 3

Model₀

[SP]

<User>

Differentiate:

$f(x) = x^2 \sin(x^3) / \sqrt{x + 1}$

</User>

[EP]

[ST]"Calculus", "Derivatives"**[ET]**

Information Retrieval (IR) Subsystem

The following operations will be performed:

1. The text enclosed between the **[ST]** / **[ET]** tokens is parsed into a list of individual tags, separated by commas.
2. The Information Retrieval subsystem searches its database for these tags and their associated documents and retrieves them.
3. The retrieved documents are ranked according to their relevance, using a weighting score similar to the TF-IDF algorithm.
4. The top K documents are then passed to the next component, where K is the maximum number of documents that can be retrieved.

Model₁

[SP]

<User>

Differentiate:

$$f(x) = x^2 \sin(x^3) / \sqrt{x + 1}$$

</User>

[EP]

**** Each document is processed independently by `model1`. ****

...

[SDocument:Mathematics/Derivatives/Rules]

[RLV]

- Product rule (two factors): $(uv)' = u'v + uv'$.
- Product rule (three factors): $(uvw)' = u'vw + uv'w + uvw'$.
- Chain rule: $d/dx[g(h(x))] = g'(h(x))h'(x)$.
- Power rule: $d/dx(x^n) = nx^{n-1}$.
- Trig derivative: $d/dx(\sin(u)) = \cos(u)u'$.

[EDocument]

[SDocument:Mathematics/Derivatives/Examples]

[RLV]

- Worked out examples:

Differentiate: $g(x) = x \sin(x^2)$.

Steps:

- Let $u(x) = x$, $v(x) = \sin(x^2)$.
- For u : $u'(x) = 1$.
- For v : $d/dx(\sin(x^2)) = \cos(x^2)x^{2'} = \cos(x^2)2x$.
- Product rule: $g'(x) = u'v + uv' = 1 * \sin(x^2) + x * (2x\cos(x^2))$.
- So: $g'(x) = \sin(x^2) + 2x^2 \cos(x^2)$.

[EDocument]

...

Model₂

** Contexts are aggregated and passed to the Encoder model. **

[SContext]

- Product rule (two factors): $(uv)' = u'v + uv'$.
- Product rule (three factors): $(uvw)' = u'vw + uv'w + uvw'$.
- Chain rule: $d/dx[g(h(x))] = g'(h(x))h'(x)$.
- Power rule: $d/dx(x^n) = nx^{n-1}$.
- Trig derivative: $d/dx(\sin(u)) = \cos(u)u'$.
- Worked out examples:

Differentiate: $g(x) = x\sin(x^2)$.

Steps:

- Let $u(x) = x$, $v(x) = \sin(x^2)$.
- For u : $u'(x) = 1$.
- For v : $d/dx(\sin(x^2)) = \cos(x^2)x^{2'} = \cos(x^2)2x$.
- Product rule: $g'(x) = u'v + uv' = 1 * \sin(x^2) + x * (2x\cos(x^2))$.
- So: $g'(x) = \sin(x^2) + 2x^2\cos(x^2)$.
- ...

[EContext]

** Prompts and responses are processed by the Decoder model. **

[SP]

<User>

Differentiate:

$$f(x) = x^2\sin(x^3) / \sqrt{x + 1}$$

</User>

[EP]

[SR]

<Chatbot>

Step-by-step solution:

1. Rewrite as a product of three factors:

$$f(x) = u(x)v(x)w(x) \text{ with } u=x^2, v=\sin(x^3), w=(x+1)^{-0.5}.$$

2. Compute derivatives of each factor:

- a. For u : $u'(x) = 2x$

- b. For v : $v(x) = \sin(x^3)$: $v'(x) = \cos(x^3)3x^2 = 3x^2\cos(x^3)$.

- c. For w : $w(x) = (x + 1)^{-1/2}$: $w'(x) = -1/2(x+1)^{-3/2}$.

3. Apply the product rule for three factors: $f'(x) = u'vw + uv'w + uvw'$.

4. Substitute the pieces:

$$f'(x) = (2x)\sin(x^3)(x + 1)^{1/2} + x^2(3x^2\cos(x^3))(x + 1)^{-1/2} + x^2\sin(x^3)(-1/2(x + 1)^{-3/2}).$$

</Chatbot>

[ER]

Example 4

Model₀

[SP]

<User>

What is the meaning of “shahint”?

</User>

[EP]

[ST]”Dictionary:shahint”, ”Slang:shahint”**[ET]**

Information Retrieval (IR) Subsystem

The following operations will be performed:

1. The text enclosed between the **[ST]** / **[ET]** tokens is parsed into a list of individual tags, separated by commas.
2. The Information Retrieval subsystem searches its database for these tags and their associated documents and retrieves them.
3. The retrieved documents are ranked according to their relevance, using a weighting score similar to the TF-IDF algorithm.
4. The top K documents are then passed to the next component, where K is the maximum number of documents that can be retrieved.

Model₁

** If no document is passed from the IR subsystem, then the model is skipped. **

Model₂

** Contexts are aggregated and passed to the Encoder model. **

[SContext]

[null]

[EContext]

** Prompts and responses are processed by the Decoder model. **

[SP]

<User>

What is the meaning of “shahint”?

</User>

[EP]

[SR]

<Chatbot>

“Shahint” doesn’t ring a bell in English.

</Chatbot>

[ER]