UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

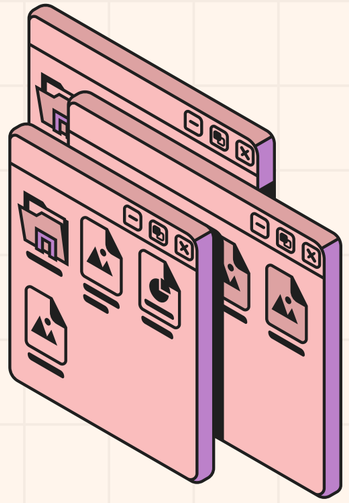
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

ARTIFICIAL INTELLIGENCE PROJECT

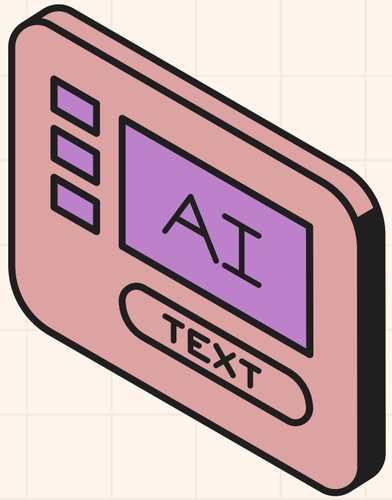
EXPLORING DYNAMIC RAG FRAMEWORK DRAGIN IN COMPARISON WITH TRADITIONAL RAG APPROACH

PETTINARI MARTINO
CREATI DAVIDE

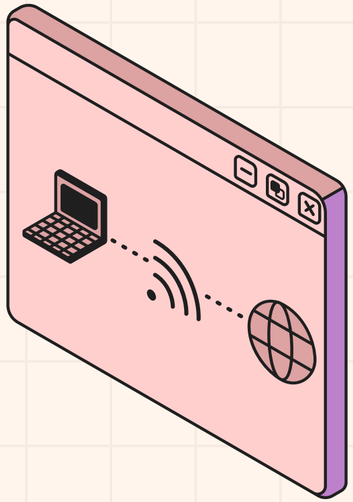
INTRODUCTION



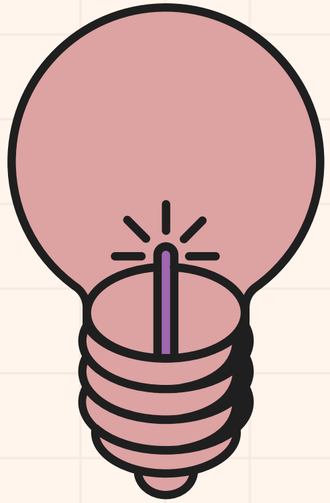
DYNAMIC RAG



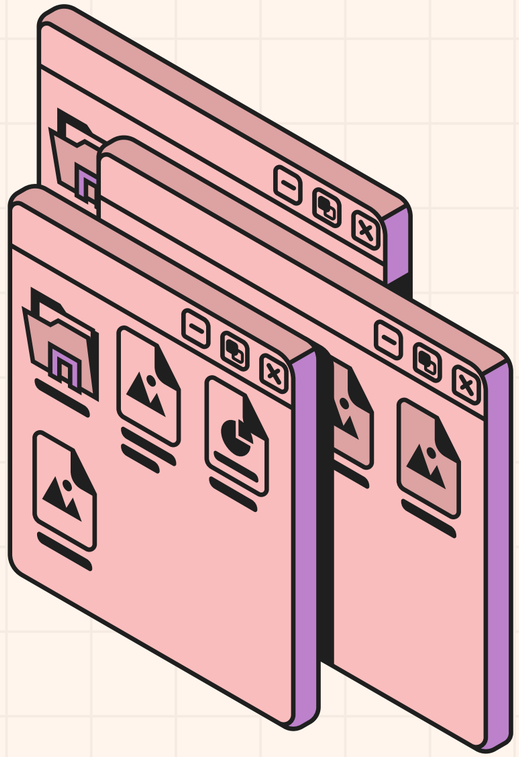
EXPERIMENT



CONCLUSION



INTRODUCTION



STATE OF ART

- **Context**: Highlight the **rapid advancements** in **Artificial Intelligence** and the **growing need** for models to access to **domain-specific information**.
- **Challenge**: Discuss the **limitations** of traditional **Large Language Models** that rely solely on **pre-existing knowledge**, which can lead to **potential inaccuracies** and **outdated responses**.
- **Solution**: Introduce **Retrieval-Augmented Generation** as a technique that enhances LLMs by integrating **external information retrieval** during the text generation process.

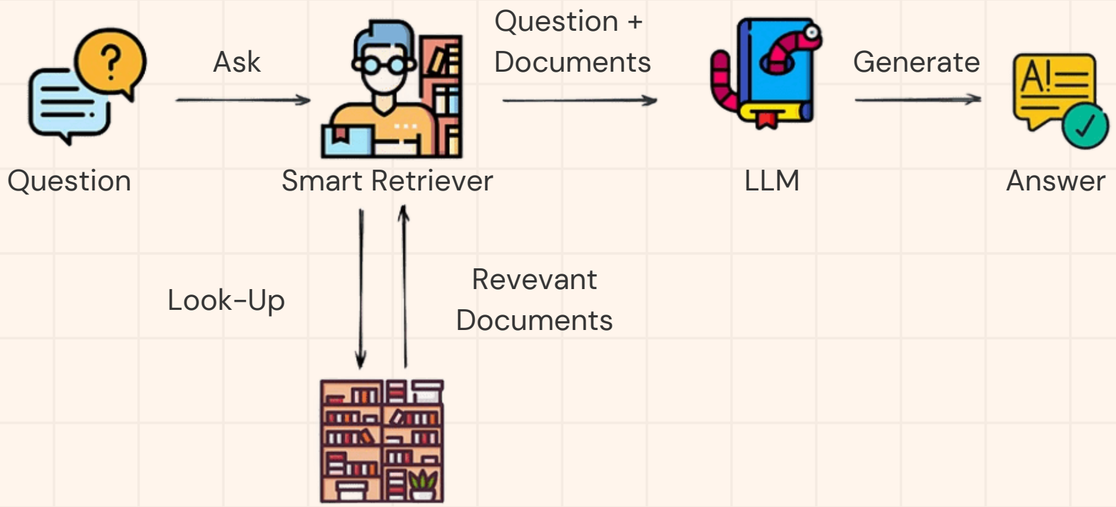
RAG

DYNAMIC RAG

Limitations:

- **Static Retrieval Timing**:
  Retrieval occurs at a fixed point,
  potentially missing context-specific needs.

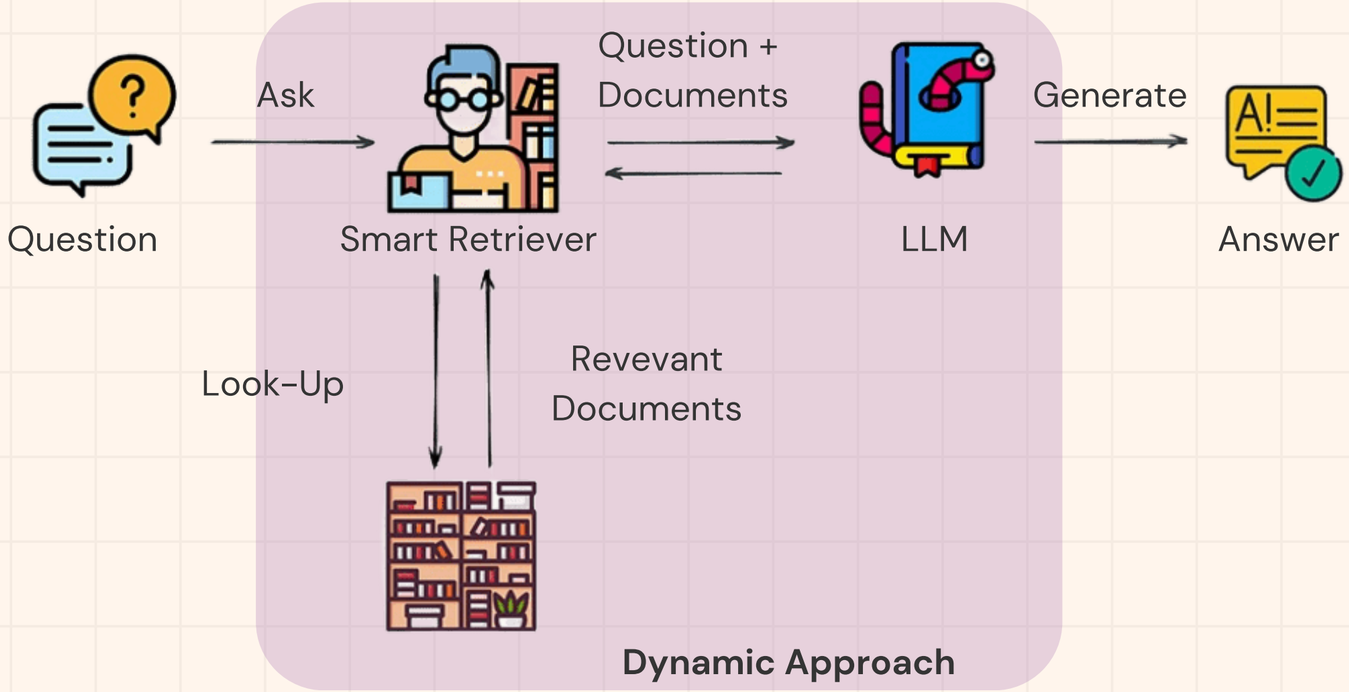
- **Generic Query Formulation**:
  Queries may not fully all information
  required from the model.



- **Potential for Irrelevant Data**:
  Fixed retrieval strategies can introduce information not aligned with the context.

DRAGIN FRAMEWORK

DRAGIN: Present the concept of **Dynamic RAG**, which performs multiple retrievals during the LLM's generation process, involving two steps: identifying the **optimal moment to retrieve** and crafting an **appropriate query**.



PORJECT SCOPE

- **Objective:** This project aims to evaluate the effectiveness of the **DRAGIN framework** by comparing its performance against traditional **RAG methods**.
- **Methodology:** This section outlines the approach, including the customization of the DRAGIN framework and the metrics used for evaluation.
- **Expected Outcome:** The anticipated benefits include improved accuracy and relevance in generated content, demonstrating the superiority of **dynamic retrieval strategies**.

TECNOLOGY

Programming Languages and Frameworks:

- **Python:** languages to implement DRAGIN.
- **PyTorch:** machine learning frameworks .

PyTorch

Integration with LLMs:

- **Hugging Face**: to integrate with different LLMs.

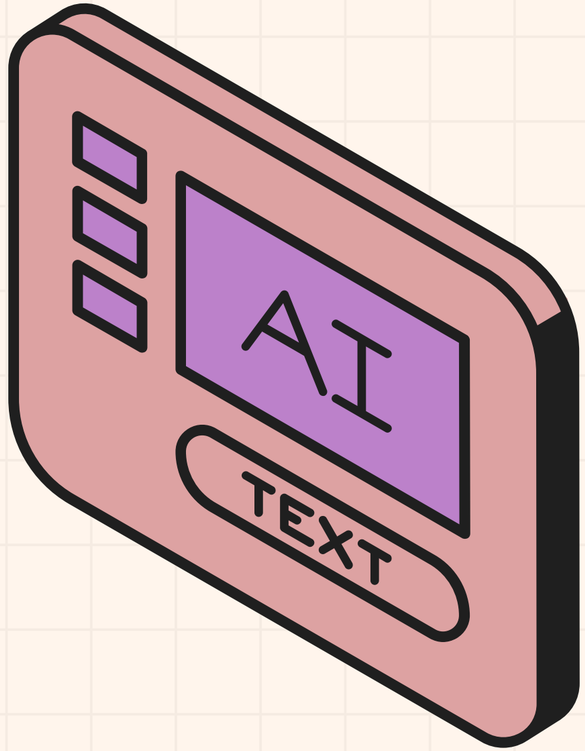
Hugging Face

Retrieval Systems:

- **Elastic Search**: search engines that store and provide the external information used during generation.

elasticsearch

DYNAMIC RAG



DRAGIN

DYNAMIC RAG

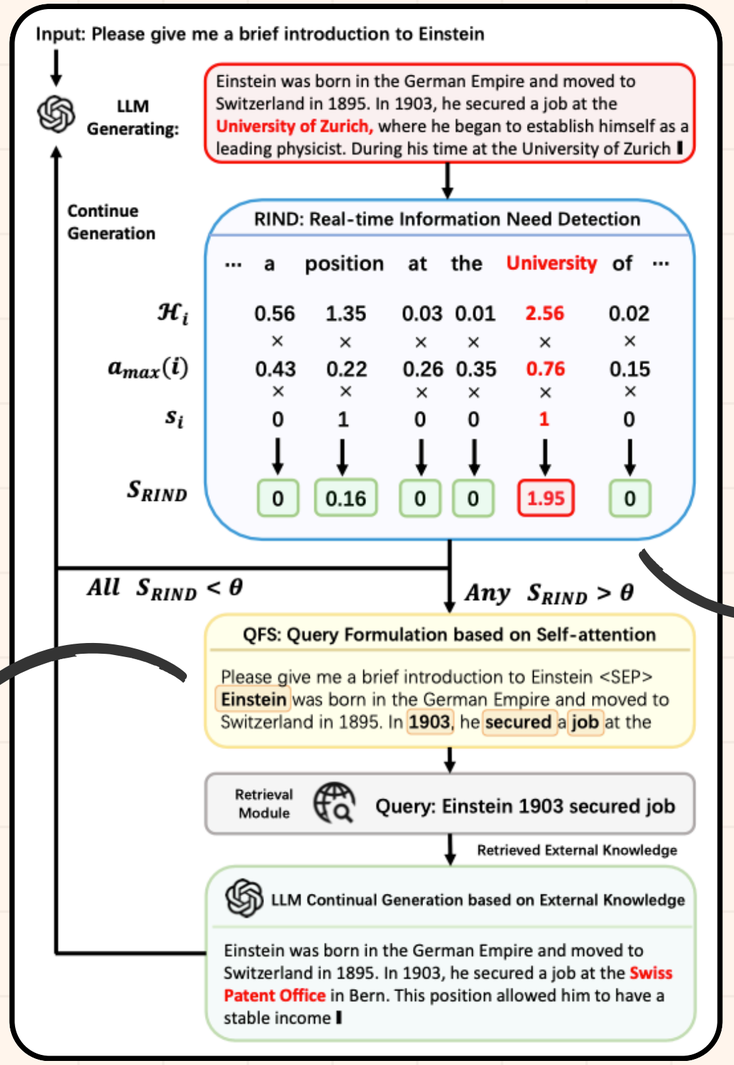
Main components:

- **Real-time Information Needs Detection** (RIND):
  - Evaluate uncertainty about generated content.
  - Identify the precise moments during text generation when retrievals are necessary.
- **Query Formulation based on Self-attention** (QFS):
  - Decide what information retrieve.
  - Formulate effective queries that address the model's information needs.

Workflow Integration: RIND and QFS work in sequence to detect the need for new data and extract information.

WORKFLOW

DYNAMIC RAG



RIND evaluation

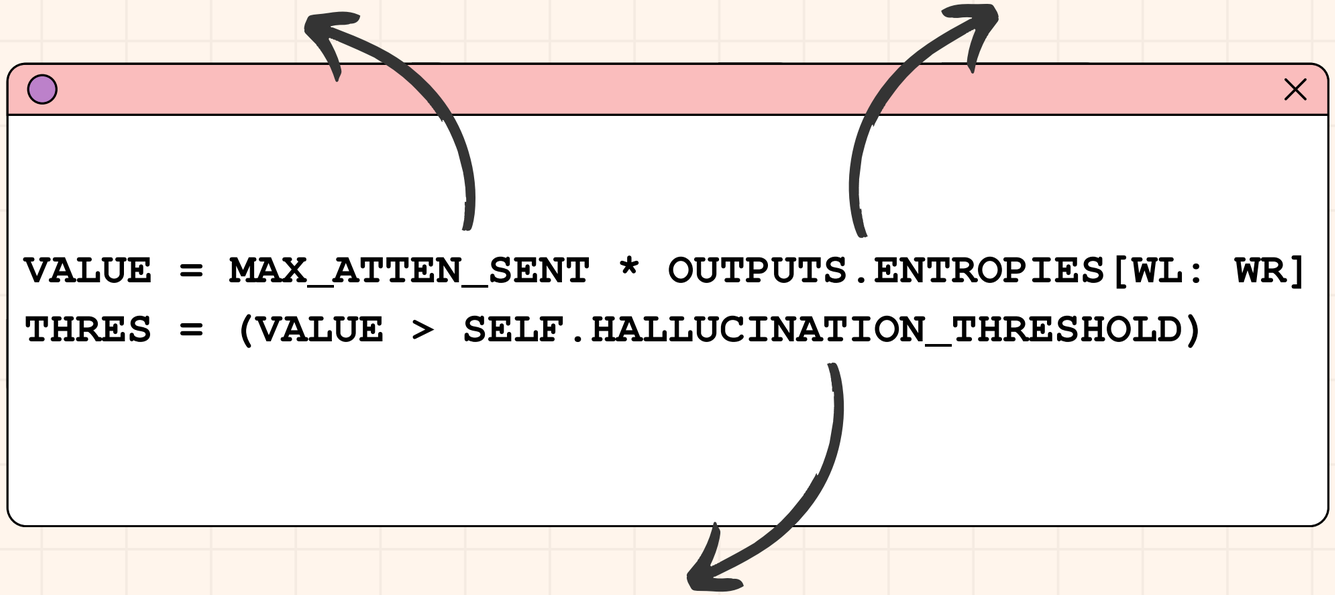
QFS query generation

HALLUCINATION (RIND)

DYNAMIC RAG

Maximum **attention scores** for the words in the sentence.

Entropy values corresponding to the words in the sentence.



Predefined value that determines if a **token** is considered **hallucinated**.

RETRIEVE (QFS) | 1

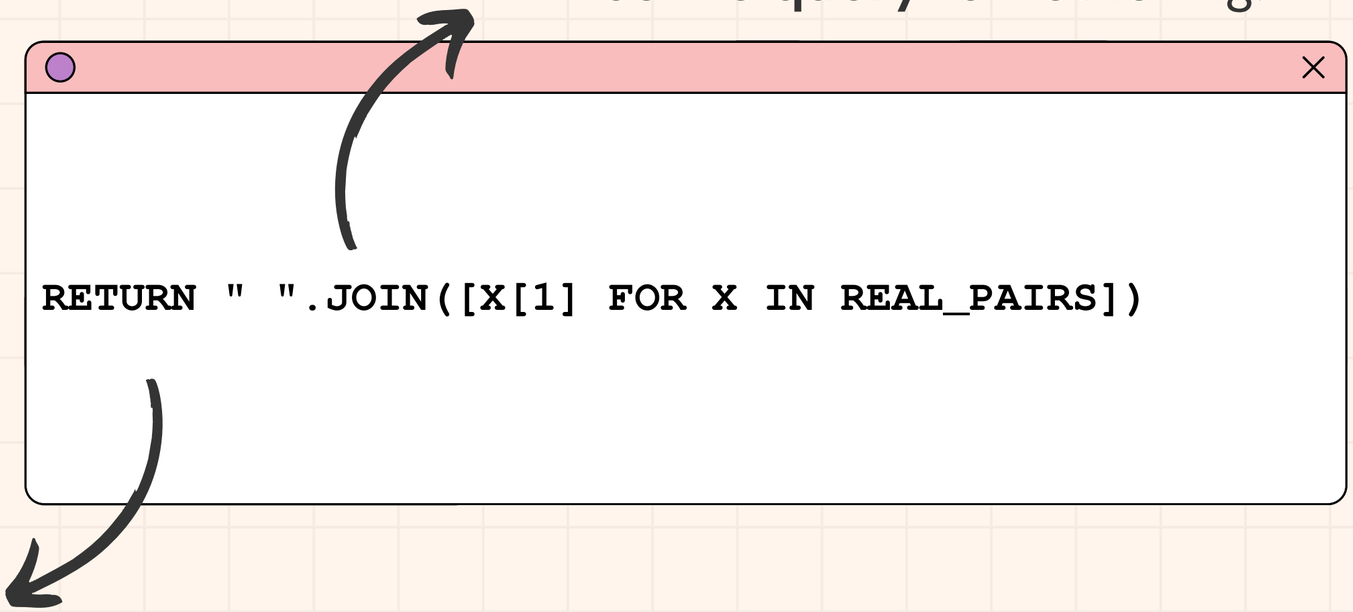
Retrieves the **number of words** in a specific block of text to retrieve information.

```
QUES_ST = OUTPUTS.BLOCKS[0].LEN_WORDS +
  OUTPUTS.BLOCKS[1].LEN_WORDS
TEXT_EN = TEXT_ST + OUTPUTS.BLOCKS[4].LEN_WORDS +
  CHECK_INFO.CURR_ST
```

Represents the **starting position** of **hallucination detection.**.

RETRIEVE (QFS) | 2

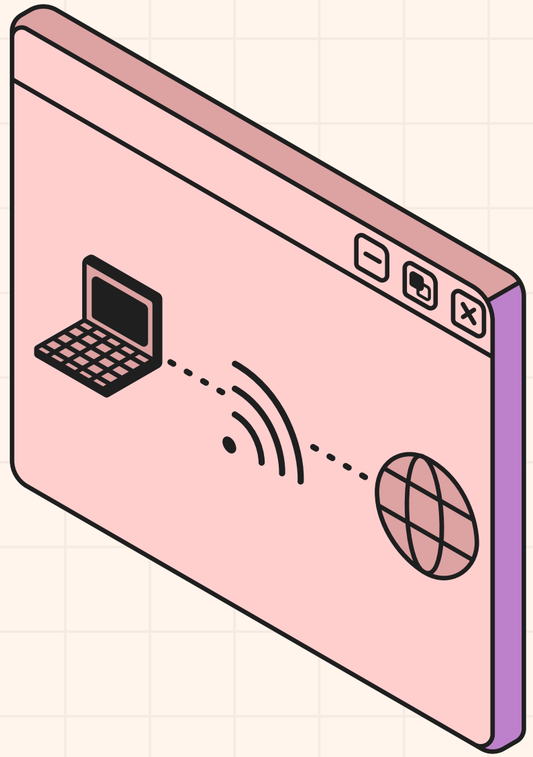
Join words into a string that serves as the **query** for retrieving.



```
RETURN " ".JOIN([X[1] FOR X IN REAL_PAIRS])
```

Return the final query by taking the selected meaningful words.

EXPERIMENT



EXPERIMENTAL SETUP

EXPERIMENT

Goal: Evaluate the performance of the **LLM** across three configurations:

- Without Retrieval-Augmented Generation (Non-RAG)
- Single Round RAG
- DRAGIN (Dynamic RAG)

Model Details: Utilized the **Llama-3.2-1B-Instruct** parameter model, recognized for its balance between performance and computational efficiency.

Evaluation Metrics: Assessed models based on **accuracy**, **relevance of generated content**, and **computational efficiency**.

Data: WikiMultiHopQA designed to evaluate reasoning. Composed with multiple **Wikipedia articles**. The dataset **challenges models** to **retrieve and integrate information** from various sources to answer complex queries effectively.

RAG COMPARISON

EXPERIMENT

- **Non-RAG Approach**: The model generates responses based solely on pre-existing knowledge without external data retrieval.
- **Single Round RAG Approach**: Incorporates a static retrieval mechanism, fetching external information once before response generation.
- **Dynamic RAG with DRAGIN**: Utilizes the DRAGIN framework to perform multiple, context-aware retrievals during the generation process, optimizing both the timing and content of information retrieval.

Evalueted with the same **metrics** and **data**.

METRICS

EXPERIMENT

Exact Match (EM): Measures whether the model's answer is **100% identical** to the correct answer.

F1 Score: Balances **precision** and **recall**, capturing partial correctness when answers aren't exact.

Precision: Evaluates the proportion of retrieved/generated information that is **correct** and **relevant**.

Recall: Measures the extent to which the correct answer was **successfully retrieved**, ensuring that **key details** are not missed.

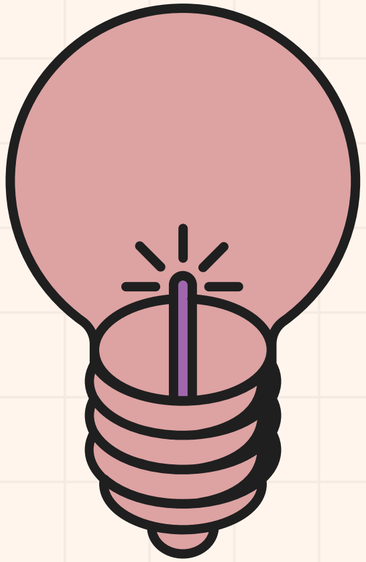
RESULTS

EXPERIMENT

Metric	Non-RAG	Standard RAG	DRAGIN
Exact Match (EM)	0.0	0.3	**0.333**
F1 Score	0.04	0.322	**0.407**
Precision	0.05	0.333	**0.407**
Recall	0.033	0.317	**0.407**

Based on 10 question in the dataset.

CONCLUSION



CONSIDERATIONS

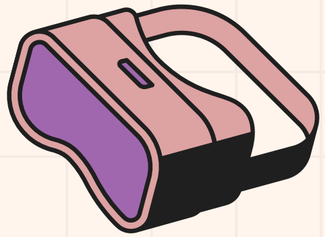
CONCLUSION

Work takeaways:

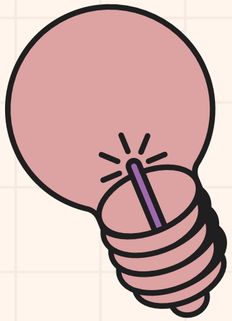
- **Dynamic RAG (DRAGIN):** outperform both standard RAG and non-RAG approaches.
- Higher **F1**, **Precision**, and **Recall** indicate better retrieval and generation accuracy.
- Enhanced **hallucination detection** reduces misinformation in generated responses.

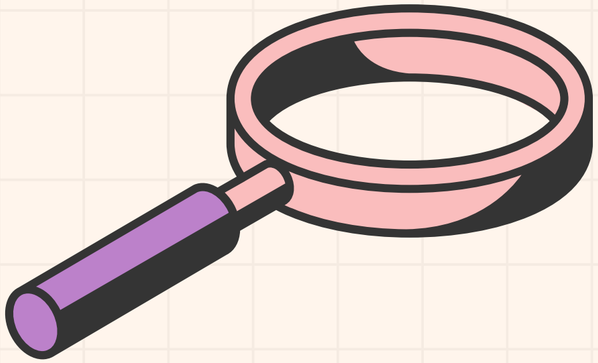
Practical considerations and limitations:

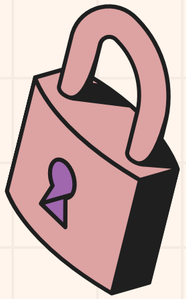
- **Increase computational requirements**: dynamic retrieval adds complexity.
- **Hardware limitations** impact performance, especially for local execution.
- Optimizing **retrieval mechanisms** is crucial for improving efficiency.



THANK YOU







SOURCES

DRAGIN: https://arxiv.org/abs/2403.10081

RAG Survey: https://arxiv.org/abs/2402.19473

WikiMultihopQA: https://arxiv.org/pdf/2011.01060