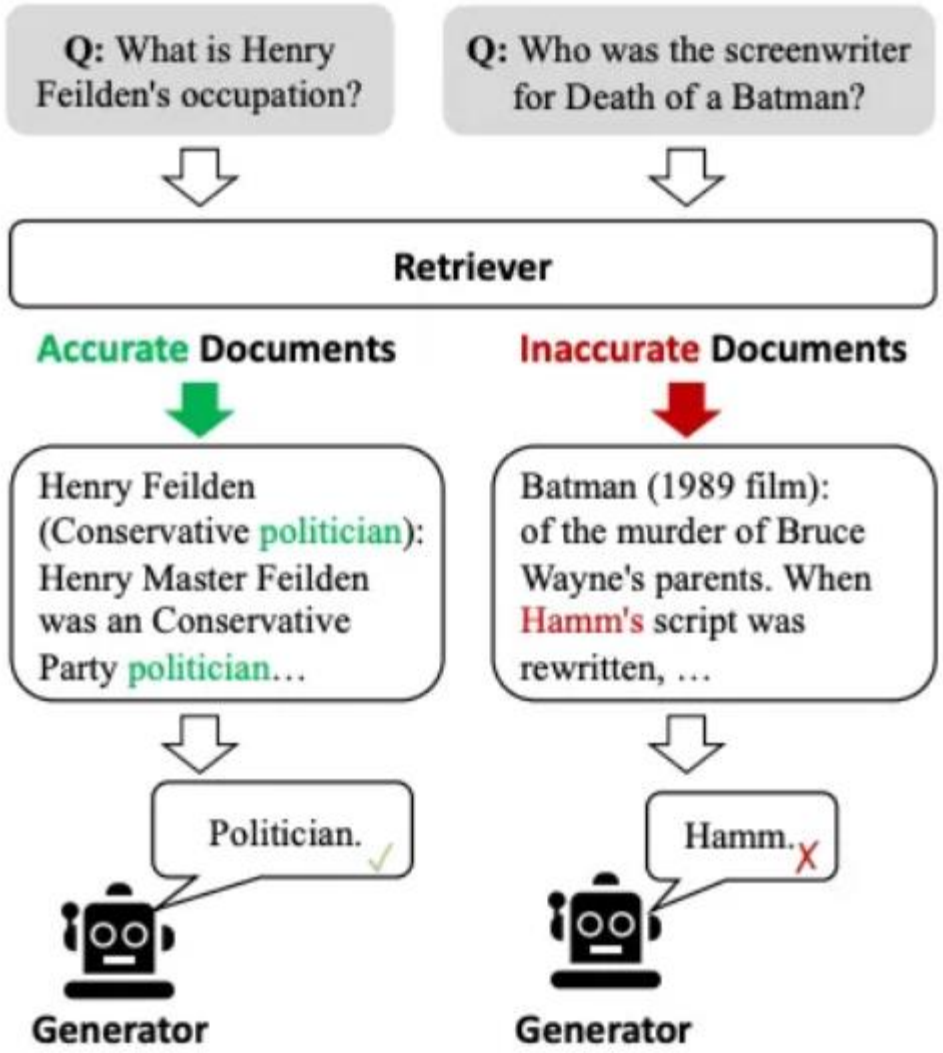
# Corrective RAG

https://arxiv.org/abs/2401.15884

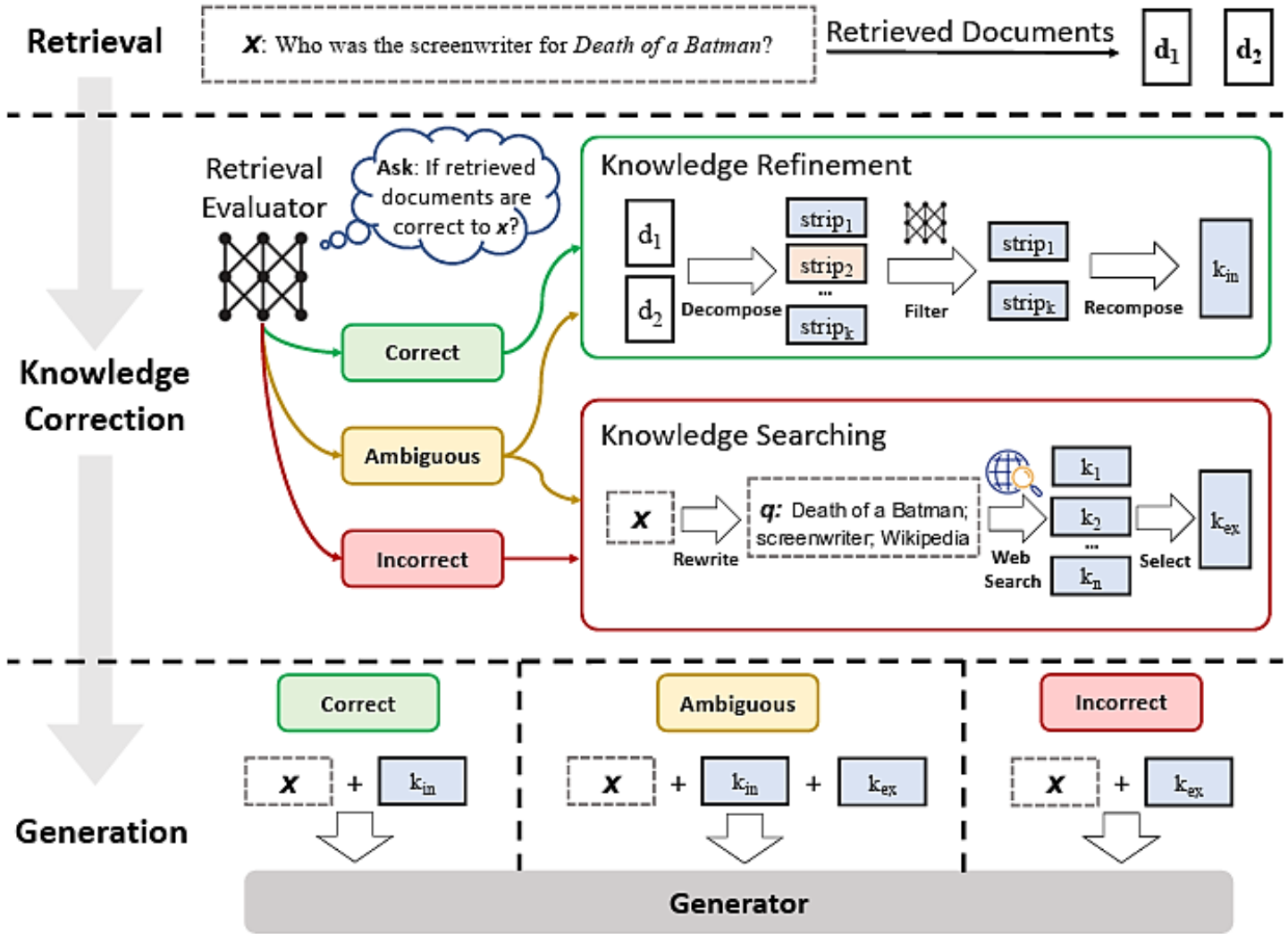
# Why need of CRAG?

- What if a retriever selects documents that are irrelevant or incorrect?
- These errors can lead to the propagation of misinformation in the generated content.
- Another significant challenge is hallucination, where the AI model generates content that is factually incorrect or misleading, even when based on retrieved documents.




The limitations of traditional RAG approaches underscore the importance of advanced methods like CRAG**, which aim to enhance the accuracy and robustness of the generation process** by incorporating self-correction mechanisms.

- Building on the foundation of traditional retrieval-augmented generation (RAG), CRAG introduces a **self-correction mechanism** to address the limitations of RAG, particularly the issues of **retrieval errors** and **hallucinations**.
- Enhance the quality of AI-generated responses by evaluating and refining the retrieved knowledge before using it in the generation process.
- This is achieved through **a series of corrective actions** that assess the relevance and reliability of the retrieved documents, filter out irrelevant or inaccurate information, and incorporate additional external knowledge when necessary.


Retrieval
x: Who was the screenwriter for Death of a Batman?
Retrieved Documents
d1
d2
Knowledge Correction
Retrieval Evaluator
Ask: If retrieved documents are correct to x?
Correct
Ambiguous
Incorrect
Knowledge Refinement
d1
d2
Decompose
strip1
strip2
...
stripk
Filter
strip1
stripk
Recompose
kin
Knowledge Searching
x
Rewrite
q: Death of a Batman; screenwriter; Wikipedia
Web Search
k1
k2
...
kn
Select
kex
Generation
Correct
x + kin
Ambiguous
x + kin + kex
Incorrect
x + kex
Generator

**Algorithm 1:** CRAG Inference

**Require :** $E$ (Retrieval Evaluator), $W$ (Query Rewriter), $G$ (Generator)
**Input :** $x$ (Input question), $D = \{d_1, d_2, ..., d_k\}$ (Retrieved documents)
**Output :** $y$ (Generated response)

$score_i$ = $E$ evaluates the relevance of each pair $(x, d_i)$, $d_i \in D$
**Confidence** = Calculate and give a final judgment based on $\{score_1, score_2, ...score_k\}$
`// Confidence has 3 optional values: [CORRECT], [INCORRECT] or [AMBIGUOUS]`
**if** *Confidence* == `[CORRECT]` **then**
    Internal_Knowledge = Knowledge_Refine($x$, $D$)
    $k$ = Internal_Knowledge
**else if** *Confidence* == `[INCORRECT]` **then**
    External_Knowledge = Web_Search($W$ Rewrites $x$ for searching)
    $k$ = External_Knowledge
**else if** *Confidence* == `[AMBIGUOUS]` **then**
    Internal_Knowledge = Knowledge_Refine($x$, $D$)
    External_Knowledge = Web_Search($W$ Rewrites $x$ for searching)
    $k$ = Internal_Knowledge + External_Knowledge
**end**
$G$ predicts $y$ given $x$ and $k$

# CRAG Working

- **Input and Initial Retrieval**
- **Retrieval Evaluator**
  - Retrieval evaluator assesses the relevance and quality of each document concerning the input query.
  - This evaluator, often a fine-tuned language model (T5), assigns a **relevance score** to each document. These scores are used to determine the overall quality of the retrieved information.
  - They use a relatively small T5 model (compared to the main LLM used for generation) fine-tuned specifically for this evaluation task.

•**How it works:** The **evaluator** takes the query and *each retrieved document individually* as input and outputs a relevance score. This score is not a probability, but a value between -1 and 1, indicating how relevant the document is to the query.

•**Why a small LLM?** Using a smaller LLM for **evaluation** is a design choice for efficiency.

•Evaluating every retrieved document with a massive LLM like GPT-4 would be computationally expensive, especially in real-time applications.

•They show that their smaller T5 model performs comparably to ChatGPT for this specific task. The idea is that judging relevance doesn't require the full generative power of a large LLM.

# CRAG Working (Evaluator)

**Training Setup:**

•For each question/query, you retrieve *k* documents (e.g., 10).

•You create (query, document) pairs.

•Each pair is labeled depending on whether it helped produce the correct answer.

**Label Type:**

In **Self-RAG (and CRAG by extension)**, the label is **binary**:

- Yes, if the document helped generate a correct answer.
- No, if it did not.

Output: Yes or No → converted to score in [-1, 1]

Table 8: The direct prompt to GPT-3.5 Turbo as the evaluator.

| |
|---|
| Given a question, does the following document have exact information to answer the question? Answer yes or no only.<br>Question: [question]<br>Document: [document] |

Table 9: The prompt to GPT-3.5 Turbo with Chain-of-Thought as the evaluator.

| |
|---|
| Given a question, does the following document have exact information to answer the question?<br>Question: [question]<br>Document: [document]<br>Think Step by step, and answer with yes or no only. |

**How Does T5 Give a Score although it provides you answers with Yes or No?**

It takes the raw probabilities of these generated Yes or No.

# Self-RAG

**Basic RAG:**

- Retrieve top-k documents → concatenate with query → pass to generator.
- No feedback loop. It assumes retrieved documents are all good.

**Self-RAG:**

- Adds *self-supervision* by checking whether a document helped in generating the correct answer.
- It runs generation, then does *answer comparison* (e.g., via semantic similarity with the gold answer).
- It then labels which documents were helpful or not.
- These labels are used to train a reranker (like T5) to score future documents better.
- **CRAG builds on Self-RAG, but introduces web-based corrective retrieval when current retrieved docs are poor.**

# CRAG Working (Evaluator)

**How a small LLM effectively evaluates the retrieved documents?**

It's trained on a dataset where query-document pairs are labeled with relevance scores.

So, it learns to predict this relevance based on the relationship between the query and the document.

It's not about "understanding" in a human sense, but about learning patterns of relevance.

Table 8: The direct prompt to GPT-3.5 Turbo as the evaluator.

| |
|---|
| Given a question, does the following document have exact information to answer the question? Answer yes or no only.<br>Question: [question]<br>Document: [document] |

Table 9: The prompt to GPT-3.5 Turbo with Chain-of-Thought as the evaluator.

| |
|---|
| Given a question, does the following document have exact information to answer the question?<br>Question: [question]<br>Document: [document]<br>Think Step by step, and answer with yes or no only. |

# CRAG Working

- **Action Trigger**

Based on the relevance scores from the retrieval evaluator, the system triggers one of three actions:

- **Correct:** If at least one document has a high relevance score, indicating it is relevant and accurate,.
- I**ncorrect:** If all documents have low relevance scores, indicating they are irrelevant or incorrect,.
- **Ambiguous:** If the relevance scores are intermediate, indicating uncertainty about the overall quality,.

- **Knowledge Refinement (if Correct)**

When the documents are deemed correct, a knowledge-refinement process is applied. This involves:

- 1. **Decomposition:** Breaking down each document into smaller, fine-grained *knowledge strips*.
- Each strip has one or two sentences.

2. **Filtering:** The retrieval evaluator (T5) re-assesses each strip, filtering out irrelevant or redundant information.
3. **Recomposition:** The filtered, high-quality knowledge strips are recomposed to form a coherent and precise knowledge base.

- **Web Search (if Incorrect)**

If the documents are deemed incorrect, a web search is conducted to retrieve additional relevant information from the internet. This involves:

- 1. **Query Rewriting:** The input query is rewritten into search-friendly keywords, often using a language model like ChatGPT.
2. **Web Retrieval:** A search engine API (openai API) retrieves relevant web pages based on the rewritten query.
3. **Content Extraction:** The content of these web pages is extracted and refined using the same decomposition, filtering, and recomposition process as in knowledge refinement.

Dialogue/Question -> Keyword Extraction (via ChatGPT) -> Query Formulation (using extracted keywords) -> Web Search.

# Web Search

- **Reformulation** can feel **task-specific**, but in CRAG, they make it **prompt-based and query-aware**, not domain-trained.
- The inputs are rewritten into **queries** composed of **keywords by ChatGPT** to mimic the daily usage of search engine.
- Use those **keywords** to formulate a **full sentence query**
- In CRAG, a public and accessible commercial web search API is adopted to generate a series of URL links for every query.
- Authoritative and regulated web pages like Wikipedia are preferred
- URL links used to navigate web pages, transcribe their content, and employ the same **knowledge refinement** method, to derive the relevant web knowledge, namely **external knowledge.**

Table 7: The few-shot prompt to GPT-3.5 Turbo for generating knowledge keywords as web search queries.

Extract at most three keywords separated by comma from the following dialogues and questions as queries for the web search, including topic background within dialogues and main intent within questions.

question: What is Henry Feilden's occupation?
query: Henry Feilden, occupation

question: In what city was Billy Carlson born?
query: city, Billy Carlson, born

question: What is the religion of John Gwynn?
query: religion of John Gwynn

question: What sport does Kiribati men's national basketball team play?
query: sport, Kiribati men's national basketball team play

question: [question]
query:

**Prompt for rewriting**

- Dialogue = background context
- Question = current user question
- Query = what we generate (via GPT or keyword extraction) for web search

**Combining Knowledge (if Ambiguous)**

When the system is uncertain (ambiguous), it combines both internal knowledge (from the initial retrieval) and external knowledge (from the web search). This hybrid approach ensures a broader and more reliable information base.

- **1. Internal Knowledge:** Refined from the initial retrieval process.
  **2. External Knowledge:** Gathered and refined from the web search.
- CRAG uses **both** the initial results (after refinement) and **external knowledge** from the web (also refined) to provide a broader range of information to the LLM.

- **Generation**
- Finally, the generator component uses the **input query X** and the **optimized knowledge *k*** (derived from the correct, incorrect, or ambiguous actions) to produce the final output Y .
- This output is now informed by the most relevant and accurate information available, significantly enhancing the reliability and accuracy of the AI-generated content.

**Combining Knowledge (if Ambiguous)**

When the system is uncertain (ambiguous), it combines both internal knowledge (from the initial retrieval) and external knowledge (from the web search). This hybrid approach ensures a broader and more reliable information base.

- **1. Internal Knowledge:** Refined from the initial retrieval process.
  **2. External Knowledge:** Gathered and refined from the web search.
- CRAG uses **both** the initial results (after refinement) and **external knowledge** from the web (also refined) to provide a broader range of information to the LLM.

- **Generation**
- Finally, the generator component uses the **input query X** and the **optimized knowledge *k*** (derived from the correct, incorrect, or ambiguous actions) to produce the final output Y .
- This output is now informed by the most relevant and accurate information available, significantly enhancing the reliability and accuracy of the AI-generated content.

**Advanatges and Limitations of CRAG:** Read the link

https://medium.com/@sahin.samia/crag-corrective-retrieval-augmented-generation-in-llm-what-it-is-and-how-it-works-ce24db3343a7

# Activity

- **Considering the following query and retrieved documents:**
- **Query**: “What are the treatment options for long COVID in 2024?”
- **Doc1**: “COVID-19 case counts peaked globally in early 2021 with the Delta variant contributing to high transmission rates.”
- **Doc2**: “A recent survey showed vaccine hesitancy remains a challenge in pediatric populations, especially in rural areas.”
- **Doc3**: “Chronic fatigue syndrome affects many adults, especially women between the ages of 30–50. Treatment typically includes rest and cognitive therapy.”

**Label each document** as:
•Correct
•Ambiguous
•Incorrect
**Decide CRAG Action**:
•Are the documents sufficient?
•If not, mark the whole set as "Incorrect" and proceed to keyword suggestion.
**Suggest 2–3 keywords** for a better web search.

If any document is Ambiguous or Correct, highlight the most relevant part.