

Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection

CMPE 252 - Section 03

Ayushi Bhatnagar, Maxim Dokukin, Krushna Thakkar(015262507)

Content

Introduction
Paper Overview
Initial Contribution
Prompt Eng
LLM Routing
Future Work
Results
Conclusion

Introduction

“SELF-RAG bridges generation and reasoning — making LLMs their own fact-checkers.”

Overview

- Large Language Models (LLMs) excel at generating fluent text but often hallucinate facts or misuse retrieved information.
- Retrieval-Augmented Generation (RAG) enhances factual grounding by adding external documents, yet it still lacks *self-awareness* about when and what to retrieve.

Paper Theory

SELF-RAG introduces a self-reflective mechanism, enabling the model to:

- Decide when and what to retrieve information dynamically.
- Critique retrieved passages for relevance, support, and usefulness.
- Control generation through internal reflection tokens (ISREL, ISSUP, ISUSE).

This leads to more accurate, verifiable, and efficient knowledge-grounded generation.

Paper Overview

Goal: Improve LLM factuality without hurting creativity by letting the model retrieve on demand, generate, and self-critique using special reflection tokens.

Problem

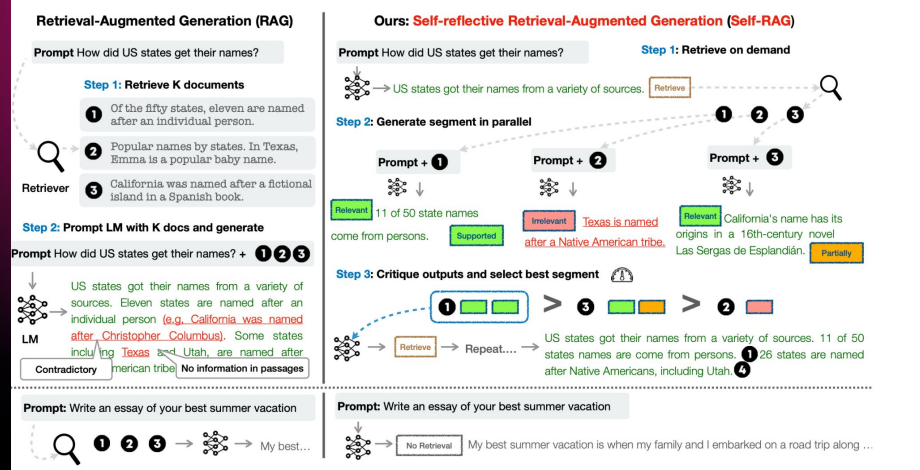
- Standard RAG retrieves a fixed number of documents even when not needed
- causes irrelevant context, hallucinations, weak use of evidence.

SELF-RAG: Key Idea

- Train an LM to decide when to retrieve, select relevant passages, judge support, and score its own output, all via special reflection tokens inserted during training.
- A single LM replaces separate retriever / generator / critic pipeline.

Reflection Tokens (Core Mechanism)

- Retrieve: {yes/no} → Should I retrieve now?
- ISREL: Is this passage relevant?
- ISSUP: Is my generated claim supported by the passage?
- ISUSE: Is this segment overall useful / good quality?



Architecture/Pipeline

1. **Decide retrieval**
LM predicts a *Retrieve* token given the prompt + prior text.
 2. **Parallel generation per passage**
If retrieving, LM receives *K documents* and generates *K* continuations in parallel.
 3. **Self-critique**
LM emits ISREL, ISSUP, ISUSE for each candidate.
 4. **Segment-level selection**
Beam search using a weighted score of critique tokens picks the best continuation.
- Critique model (trained using GPT-4 labels) generates reflection tokens
 - Generator LM is fine-tuned on outputs interleaved with reflection tokens + retrieved passages.

Initial Contribution

Initial Contribution — What We Added Beyond the Original SELF-RAG Paper

- **Introduced custom prompt templates** (*qa*, *explanatory*, *chain_of_thought*, *compare_contrast*) to explicitly shape the model's reasoning style and response format
- **Ensured retrieval grounding** by automatically embedding the top-k Wikipedia (KB) passages directly into each prompt, forcing the model to answer **only using provided context**.
- **Added a reflection layer** instructing the model to “*analyze your answer for accuracy, completeness, and reasoning errors*”, enabling self-critique and improved final responses.

Prompt Eng

1. Smarter Retrieval Controller (RETRIEVE vs NO_RETRIEVE Classifier)

The previous heuristic-based retrieval trigger was replaced with a reasoning-driven classifier that determines whether a question requires retrieval. This improves retrieval precision and avoids unnecessary retrieval on conceptual or explanatory questions.

2. Grounding Template for Answer Generation

A strict grounding template was introduced, requiring outputs to include evidence summary, final answer, citations, and a fallback "I don't know." This template constrains the model to rely solely on retrieved context, reducing hallucinations and improving consistency.

3. Minimal, Structured Output Format

The answer format was simplified to a compact two-part structure that avoids long, unconstrained explanations

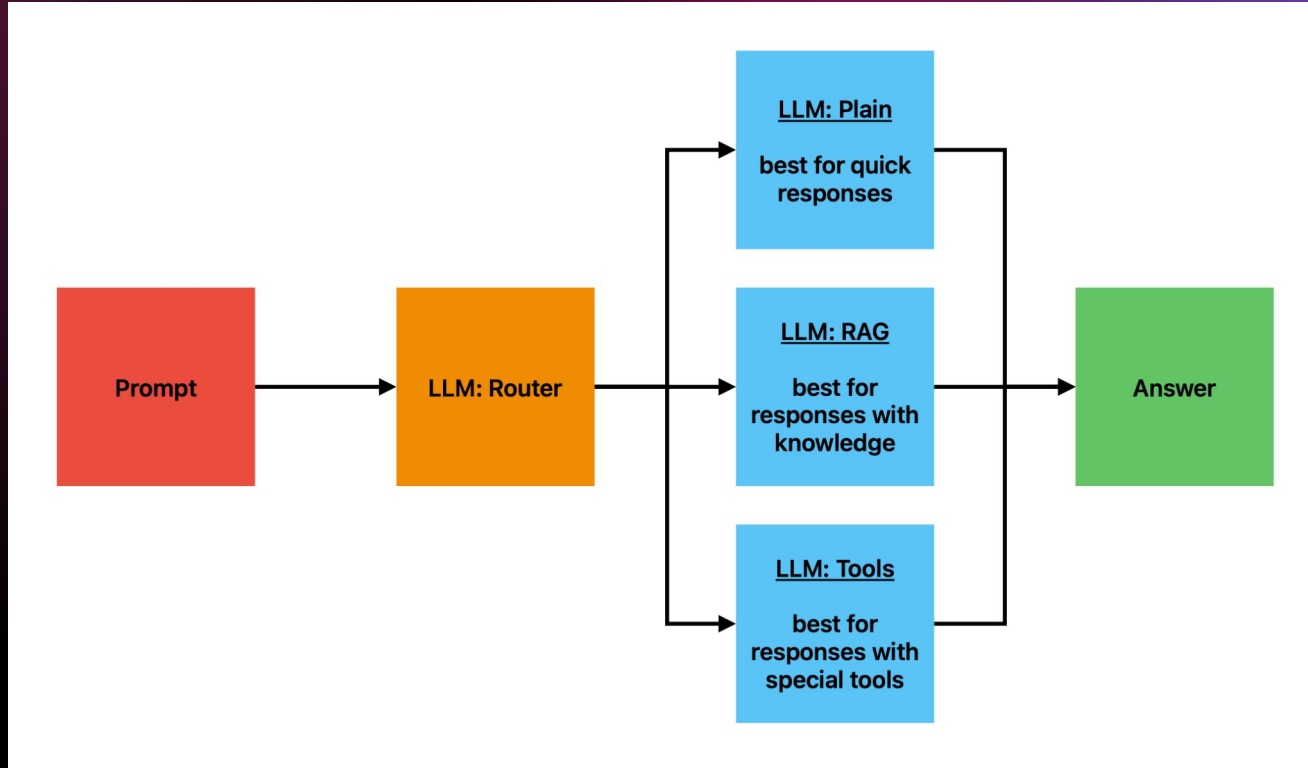
4. Refined Self-Critique Mechanism

The critique step was redesigned to detect hallucinations, missing citations, and reasoning errors while modifying only incorrect portions of the answer.

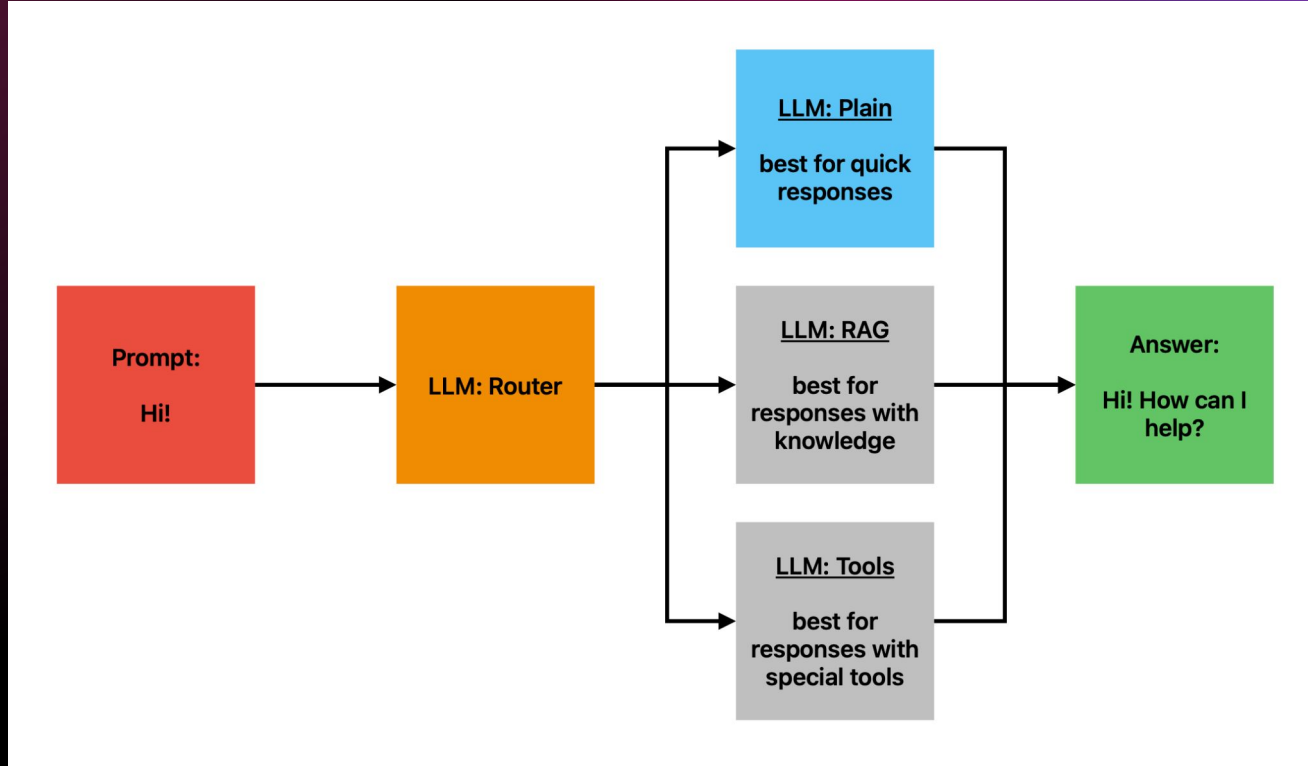
5. Grounding-First Evaluation Criteria (Updated Judge)

The judging rubric was updated to prioritize grounding above all other metrics, including correctness, completeness, and clarity.

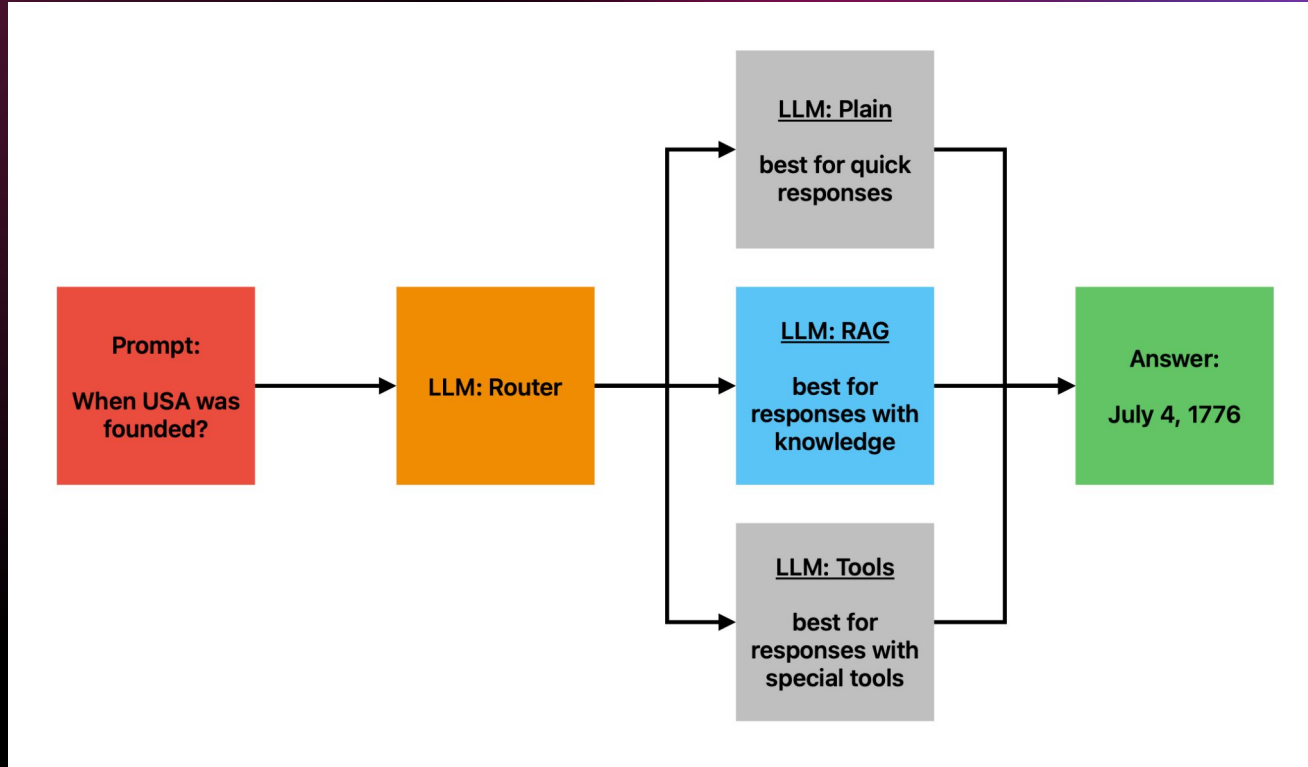
Max: LLM Routing



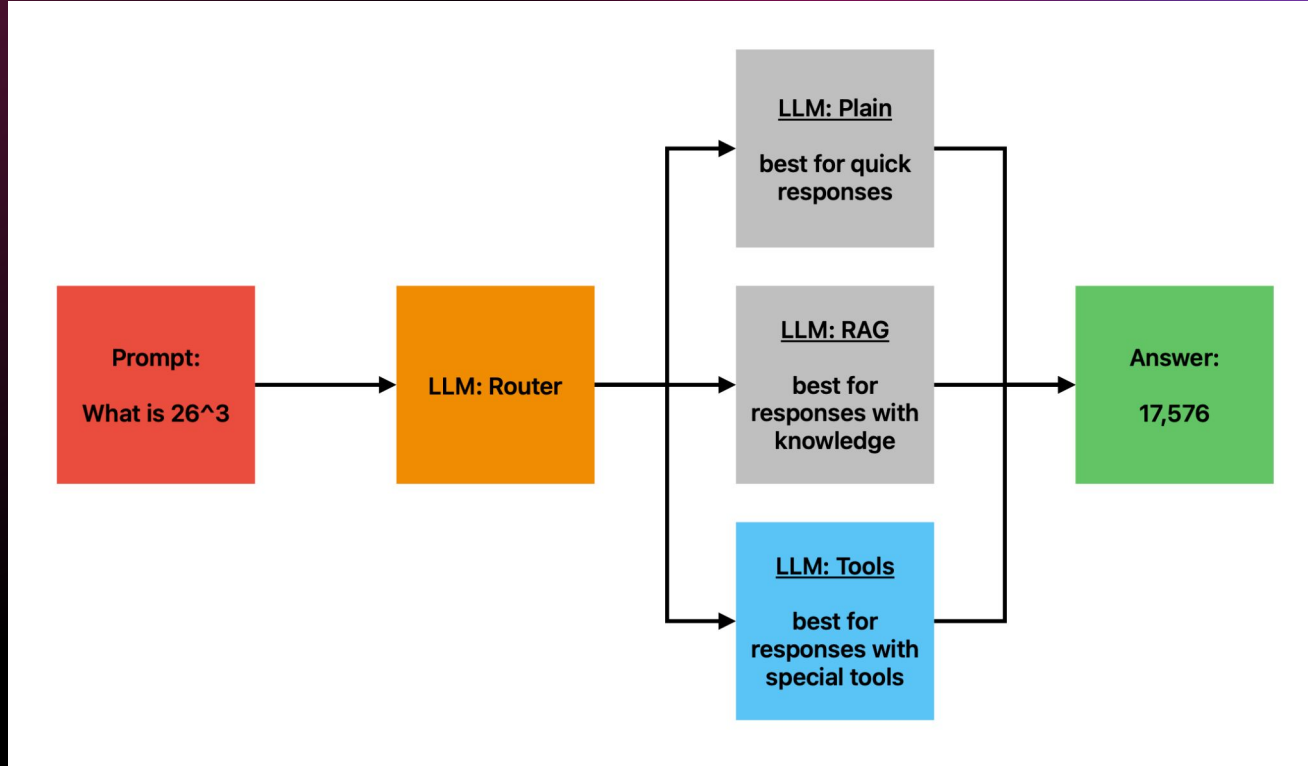
Max: LLM Routing



Max: LLM Routing



Max: LLM Routing



Future Work

An agentic layer that verifies at each step. An agent that can plan, verify, revise and justify—not just guess.

Normal agents hallucinate more than standard LLMs because they:

- break tasks into uncontrolled multi-step reasoning
- make implicit assumptions
- misuse tools or retrieve irrelevant data
- drift away from the original query

Our Agentic Extension adds a Self-RAG Safety Loop at Every Step:

- Retrieve? → *Should I fetch evidence for this step?*
- IsRelevant? → *Are the retrieved documents actually related to the plan?*
- IsSupported? → *Does my intermediate action or reasoning match the evidence?*
- IsUseful? → *Did this step produce information that helps move toward the goal?*

This transforms the agent from a “hallucination-prone planner” into a verified reasoning system. Fewer hallucinated plans, Fewer fabricated facts/citations, Reduced reasoning drift and Safer intermediate actions.

```
QUESTION:
Tell me who was the President of Mars in 2020?

EVIDENCE:
Agents perform planning, tool use, and multi-step execution.
Agents should check correctness before taking action.

1. Produce an answer.
2. Critique correctness.
3. Ensure the answer is supported by evidence.
4. Fix unsupported or hallucinated content.

Respond in:
ANSWER: <final answer>

1. The President of Mars in 2020 is a fictional character as Mars does not have a president and it is not inhabited by humans.

2. The answer is correct because Mars is a planet in our solar system and not a country or a place where humans live.

3. The answer is supported by the evidence that Mars is a planet and not a country or inhabited by humans.

4. No need for content fix as the answer is accurate and supported by evidence.

ANSWER: The President of Mars in 2020 is a fictional character as Mars does not have a president and it is not inhabited by humans.

--- AGENTIC SELF-RAG OUTPUT ---
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
Setting 'pad_token_id' to 'eos_token_id' for open-end generation.
=== AGENTIC SELF-RAG RESULT ===

STEP 1:

You are verifying the agent's action using evidence.

PLAN + ACTION:

You are an AGENTIC Self-RAG with tool-use and verification.

GOAL: Tell me who was the President of Mars in 2020?

Your tasks:
1. Decide the next step in the plan.
2. Decide if retrieval is needed for this specific step.
3. Explain WHY retrieval is or isn't needed.
```

```

6. If not correct - revise.

Return JSON:
...
{
  "step": 1,
  "plan": "...",
  "need_retrieval": "yes/no",
  "reason": "...",
  "action": "...",
  "verification": "...",
  "finalized_action": "..."
}

Step 1: Decide the next step in the plan.
The next step in the plan is to retrieve information about the President of Mars in 2020.

Step 2: Decide if retrieval is needed for this specific step.
Yes, retrieval is needed because the information about the President of Mars in 2020 is not known to me.

Step 3: Explain WHY retrieval is or isn't needed.
Retrieval is needed because the information is not known to me and needs to be obtained from a reliable

Step 4: Produce an action.
The action is to search for information about the President of Mars in 2020.

Step 5: Before executing, VERIFY using evidence whether the action is correct.
The action is correct because it is a standard procedure for obtaining unknown information.

Step 6: If not

EVIDENCE:
Agents perform planning, tool use, and multi-step execution.
Agents should check correctness before taking action.

Is the action valid, supported, and safe?
Fix it if needed.

Return:
FINAL_ACTION: <corrected step>
EXPLANATION: <why>

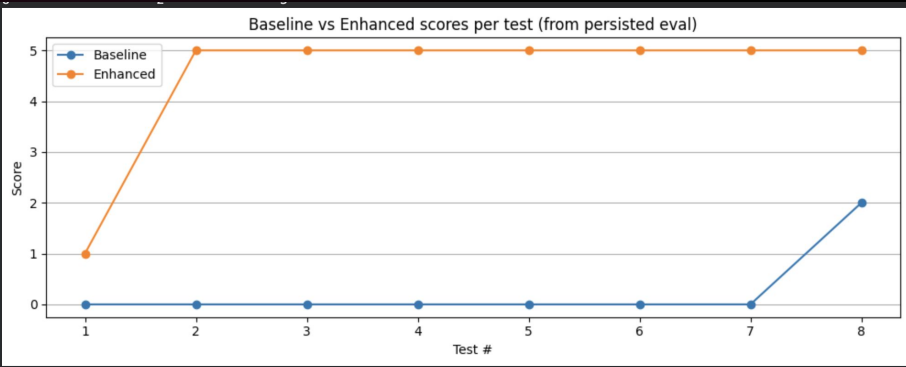
Step 6: If not correct, revise.
There is no need for revision as the action is correct.

Return:
FINAL_ACTION: Search for information about the President of Mars in 2020.
EXPLANATION: This action is correct because it is a standard procedure for obtaining unknown information.

Step 2:

You are verifying the agent's action using evidence.

```



The enhanced version performs better because the improved retrieval controller, grounding template, and self-critique step work together to correct hallucinations, enforce structure, and ensure the final answers follow the grounding rules the judge evaluates.

Conclusion

Thanks

Ayushi Bhatnagar, Maxim Dokukin, Krushna Thakkar(015262507)