Primer

Introducing RAG-Verification The future of trustworthy AI

As artificial intelligence (AI) systems advance, their capacity to generate and process vast amounts of data introduces both opportunities and challenges. While Large Language Models (LLMs) can produce highly detailed and insightful responses to queries, they remain prone to generating factually unsupported information, often referred to as "hallucinations." Retrieval-Augmented Generation (RAG) has emerged as a mitigation, directing LLMs to rely on relevant, retrieved data. But even RAG systems powered by state of the art LLMs such as GPT-4 and Claude have hallucination rates on order of 10%. An error rate this high renders LLM-powered systems unreliable for use cases supporting high stakes decision-making.

We introduce RAG-Verification (RAG-V), a framework that automatically detects and corrects LLM errors before they reach human users. The approach is similar to journalism fact-checking. It detects and corrects errors by comparing each factual claim in an LLM-generated text to the cited data sources. Because this fact-checking process can be parallelized, it runs fast enough to be used in a real-time interactive AI system. RAG-V reduces the state of the art hallucination rate from on order 10% to 0.1%, making AI systems reliable enough for mission-critical use.

The hallucination problem

RAG-Verification is a framework we built here at Primer to ensure that our software generates factually consistent, reliably useful information. Before we dive into how it works, let's take a look at the problem it solves: hallucinations.

Hallucination is just a fancy name for a type of error made by a large language models (LLMs). Until recently, you could induce ChatGPT to hallucinate by asking it to write a biographical profile of a person. For example, I asked it to generate a biographical profile of Yurii Shchyhol, Ukraine's Head of State Special Communications Service of Ukraine, and ChatGPT instantly generated pages of text about Shchyhol.

Most of the information was correct, since Shchyhol has been written about frequently in online news and analyses, and OpenAI trained the LLM on internet data. But it also filled in gaps in its knowledge, inventing Shchyhol's age and birth place, and even fabricating a plausible but completely incorrect education for him.

ChatGPT

Write a biographical profile of Yurii Fedorovych Shchyhol, head of the State Special Communications Service of Ukraine.

Yurii Fedorovych Shchyhol is a prominent figure in the field of communications and cybersecurity in Ukraine. He currently holds the esteemed position of Head of the State Special Communications Service (SSCS) of Ukraine. Throughout his career, Shchyhol has demonstrated exceptional leadership, expertise, and dedication in safeguarding the country's communications infrastructure and promoting national security.

Born on February 12, 1970, in Kyiv, Ukraine, Yurii Shchyhol developed an early fascination with technology and telecommunications. **He pursued his passion by earning a degree in Telecommunications Engineering from the National Technical University of Ukraine "Kyiv Polytechnic Institute."** This academic foundation laid the groundwork for his future achievements in the field.

Shchyhol's professional journey commenced soon after completing his studies when he joined the State Special Communications Service in **1995**. The SSCS, being the central government authority responsible for ensuring secure communications and protecting information systems of national importance, offered Shchyhol a platform to utilize his skills effectively. His dedication and expertise were quickly recognized, leading to his rapid ascent within the organization.

Throughout his tenure with the SSCS, Shchyhol demonstrated an unwavering commitment to upholding Ukraine's cybersecurity and protecting critical infrastructure. He played a crucial role in implementing advanced technologies and strengthening the country's cybersecurity defenses, particularly in the face of evolving threats and hostile activities targeting Ukraine's communications networks.

Shchyhol's leadership abilities and technical acumen led to his appointment as Head of the State Special Communications Service in **2017**. As the head of this vital organization, he shoulders the responsibility of ensuring the security and integrity of Ukraine's communication systems, both in times of peace and during periods of heightened tension. Under his leadership, the SSCS has developed robust strategies and implemented state-of-the-art technologies to counter cybersecurity threats effectively.

Yurii Shchyhol's achievements extend beyond his role as the head of the SSCS. He actively collaborates with national and international organizations, fostering partnerships to enhance Ukraine's cybersecurity capabilities and contribute to global efforts in countering cyber threats. His expertise has made him a sought-after speaker at prestigious conferences and forums, where he shares insights and best practices on cybersecurity and communications.

Beyond his professional endeavors, Shchyhol is known for his strong work ethic and integrity. He exemplifies the qualities of a dedicated public servant and a role model for aspiring cybersecurity professionals. He is committed to continually expanding his knowledge and staying abreast of the latest advancements in the field.

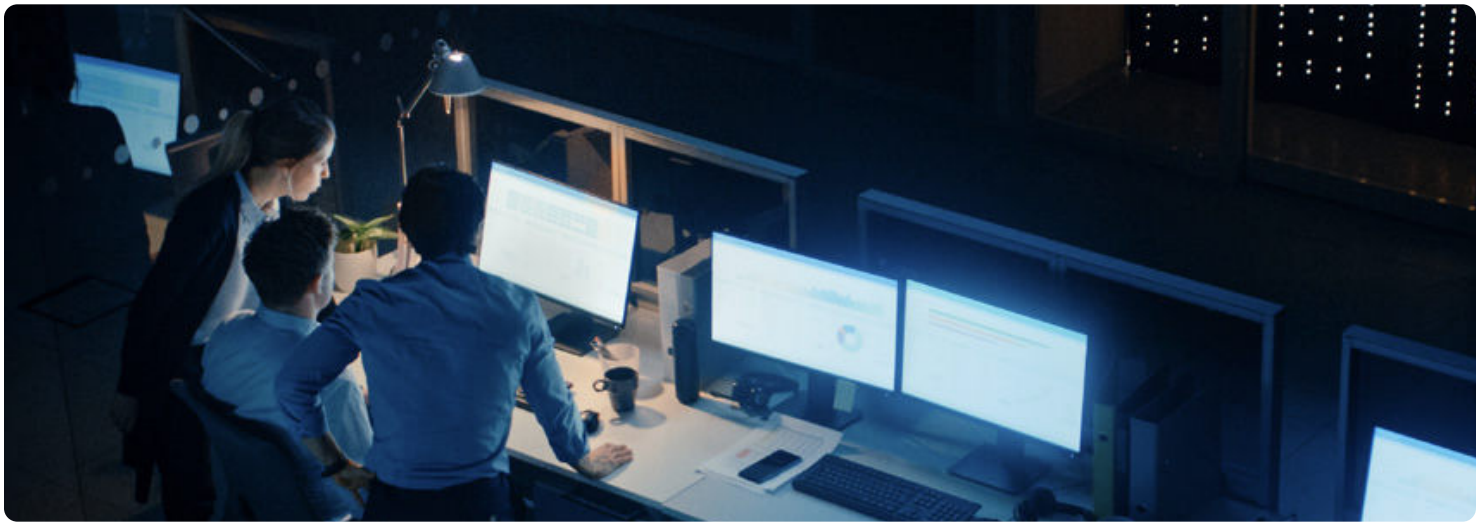
Yurii Fedorovych Shchyhol's contributions to the field of communications and cybersecurity have earned him widespread respect and recognition. His visionary leadership and technical prowess make him a driving force in fortifying Ukraine's digital infrastructure. As the Head of the State Special Communications Service, he remains steadfast in his commitment to protecting the country's vital communication systems, enabling Ukraine to thrive in the digital age while safeguarding its national security interests.

AI researchers call these errors “hallucinations” because the model is inventing information that is not based on reality. The model is not lying, since it has no motivation to deceive, nor any awareness that what it is saying is false. Regardless, hallucinations are a nightmare for analysts, operators, and decision-makers in national security. Subtly incorrect claims are the most dangerous. They lead to the loss of the information trail, embedding false assumptions in the command chain.

The cause of LLM hallucination is in the training. LLMs are trained with a game of “Fill in the Blank” on a massive scale. The model is given truncated snippets of text from webpages—such as Wikipedia, Reddit, news sites—and it guesses the next words.

The training is as simple as that. And yet learning to predict words on the scale of the internet—on the order of 1 trillion words—teaches the model an astounding range of skills. It can read and write in dozens of languages. It can even pass exams in medicine and law.

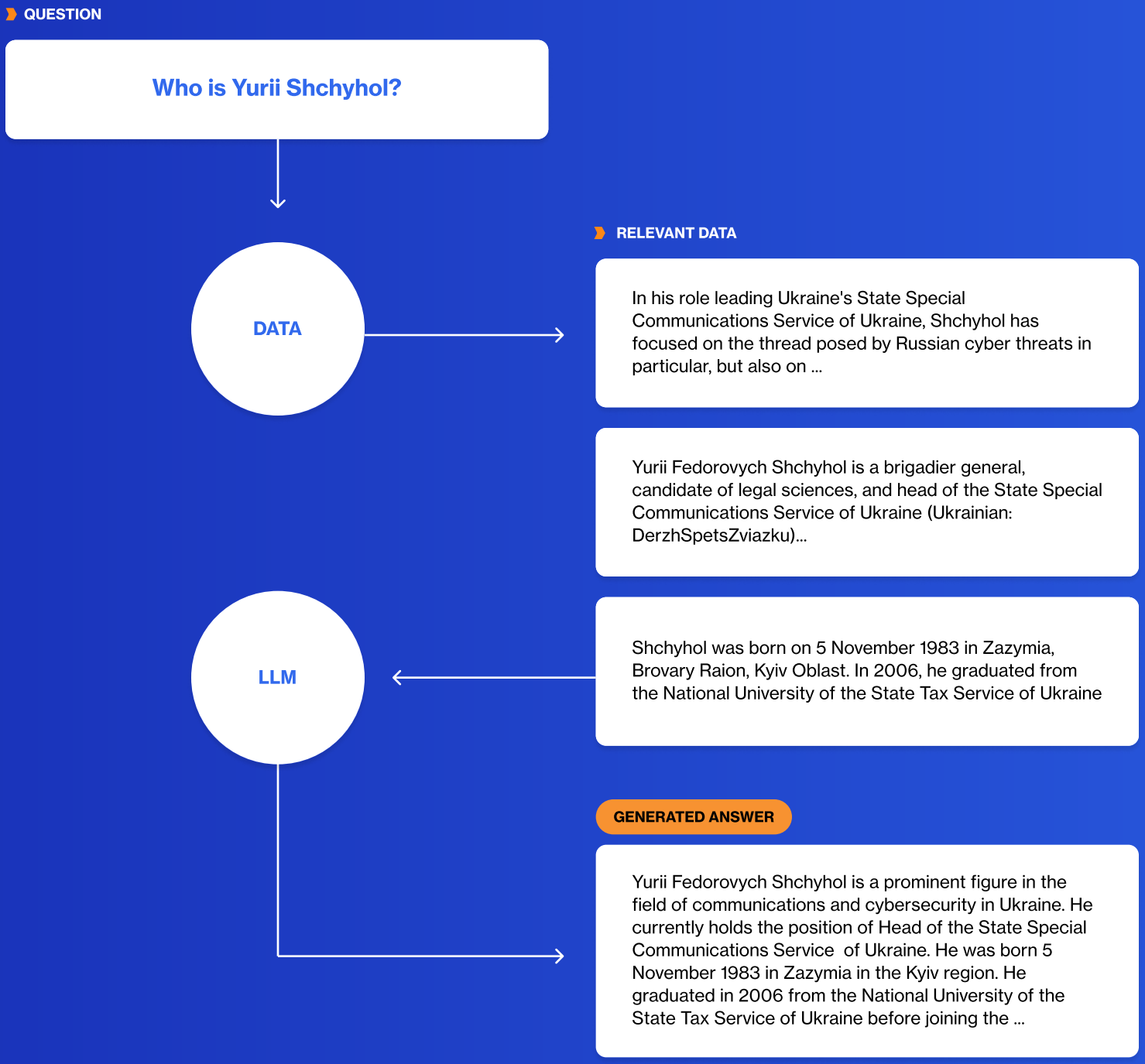
But LLMs never stop playing the game. It is always predicting the most plausible next word, regardless of whether it expresses something true or false. Given this training paradigm, the problem of hallucination is inevitable.



Retrieval Augmented Generation (RAG)

The most powerful method for reducing hallucination errors is a paradigm called Retrieval Augmented Generation (RAG). You wouldn’t ask a human analyst to write a mission-critical report from memory. Just like a human analyst needs access to relevant data while they write, so too with generative AI systems.

Instead of merely predicting the next words, given a user’s input, a RAG system first retrieves relevant information from a trusted system of record. That retrieved information is then included in the prompt for the generative model. The prompt also includes instructions for the model to answer the user’s question based exclusively on the retrieved information. So if the retrieved information about Yurii Shchyhol does not include his age or birth place, the model should not try to guess that information.



If LLMs followed instructions perfectly 100% of the time, then this would be the end of the story. But hallucinations do persist, even in a RAG framework.

How bad is the problem? Even most AI practitioners are surprised. The state-of-the-art LLMs, such as OpenAI's GPT-4 and Anthropic's Sonnet 3.5, have an error rate of between 5% and 10%, depending on the data and questions. That means that a user of an LLM-powered application must remain hyper-vigilant, scrutinizing every piece of generated text.

A 10% error rate might be acceptable for casual users of products like ChatGPT, but not for defense or intelligence analysts. They need reliable AI that can detect and correct its own errors. That's why we set out to build the RAG-Verification pipeline.

RAG-V step 1: error detection

Detecting factual inconsistencies in text is not a new problem. Professional journalists call it fact-checking, part of tradecraft that goes back more than a century. The goal is not to determine capital-T Truth. Rather, the goal is to determine if what is being claimed in text is supported by the sources that it cites.

RAG-Verification works the same way.
The first step is to break generated text into individualfactual claims.
Consider this RAG example:

QUESTION

Are UK special forces operating in Ukraine?

RETRIEVED SOURCE 01

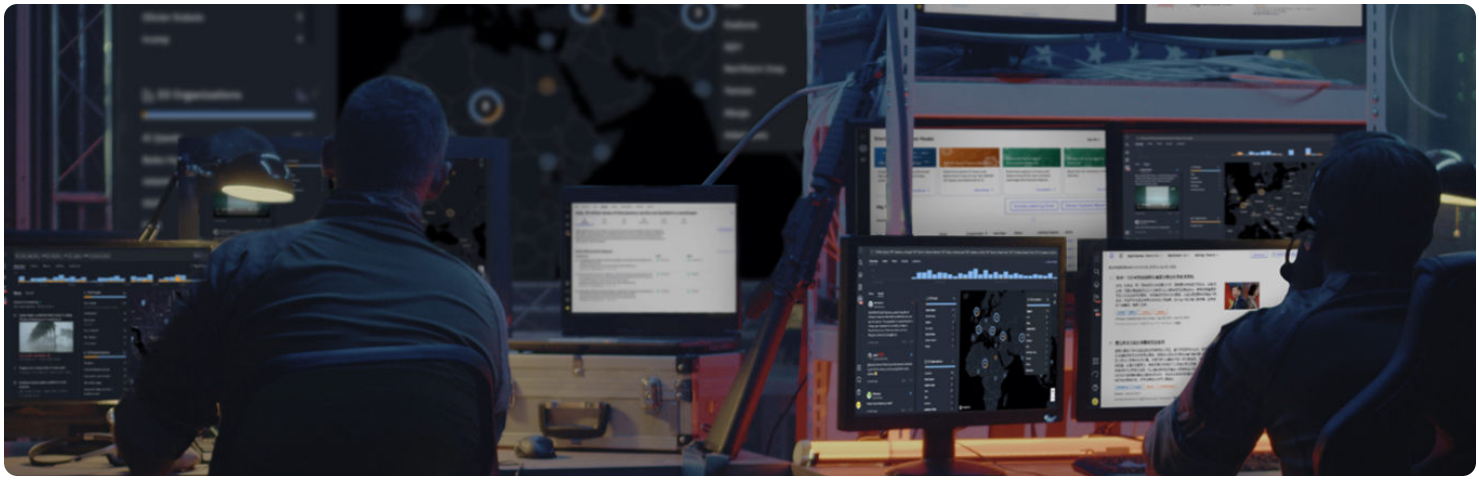
Leaked US military documents indicate that the UK has deployed as many as 50 special forces to Ukraine. The documents suggest that more than half of the western special forces personnel present in Ukraine between February and March this year may have been British. It is unclear what activities the special forces may have been engaged in or whether the numbers of personnel have been maintained at this level. (11 April 2023)

RETRIEVED SOURCE 02

The FSB alleges it thwarted an attempt by British special forces to facilitate a landing of Ukrainian sabotage troops on the Tendrov Split, a strategic sandbar located in the Black Sea. The agency claims to have apprehended a senior figure within the Ukrainian naval special forces, providing his name and date of birth as proof of the operation's existence and the SBS's direct involvement. Britain's Defence Ministry has yet to respond to these allegations. (12 April 2024)

GENERATED ANSWER

Leaked US military documents indicated the possible presence of up to 50 UK special forces personnel in Ukraine in 2023.[1] The FSB alleges to have apprehended British special forces during a failed operation, however Britain's Defence Ministry has not responded.[2]



RAG-Verification begins by converting the generated text into its factual claims:

Factual claims citing source [1]:

- US military documents were leaked with information about UK special forces in Ukraine.
- The leaked US military documents indicate the possible presence of up to 50 UK special forces personnel in Ukraine in 2023.

Factual claims citing source [2]:

- The FSB alleges to have apprehended British special forces during a failed operation.
- Britain's Defence Ministry has not responded to allegations by the FSB.

For each claim, we construct a fact-checking prompt for the LLM. The prompt includes the original question, the generated answer, the specific claim, and the cited sources. The model is then instructed to verify the factual consistency of the claim based on the provided sources, producing a yes or no verdict along with an explanation for its decision. This step in the evaluation phase, which we call RAG-Eval, is similar to a technique developed by our colleagues at the Allen Institute for AI in Seattle.

From the point of view of a human fact-checker, RAG-Verification seems inefficient. What starts as a single fact-check on a small piece of generated text expands into four separate fact-checking tasks—and we haven't even begun the error-correction phase. But breaking a single fact-checking task into multiple independent tasks plays to the strengths of an AI system.

LLMs perform better when tasks are broken into smaller, simpler sub-tasks. It is possible to instruct LLMs to execute this divide & conquer strategy themselves with a technique called chain-of-thought. By instructing the model to solve complex problems "step by step", it is possible to dramatically reduce the error rate in a single go.

But there's another benefit to breaking tasks down in advance: All of the individual factual claims can be verified simultaneously. Instead of waiting 40 seconds for the model to process 4 claims sequentially, we can verify them all in just 4 seconds. As the size of the text to be verified increases, and thus the number of factual claims to check, the time savings grows dramatically.

RAG-V step 2: error correction

So what do the errors detected by RAG-V look like? The first thing we learned is that not all errors are created equal. There are "hard fails" and "soft fails".

First let's see an example of a hard fail.

QUESTION

Are UK special forces operating in Ukraine?

RETRIEVED SOURCE 01

Leaked US military documents indicate that the UK has deployed as many as 50 special forces to Ukraine. The documents suggest that more than half of the western special forces personnel present in Ukraine between February and March this year may have been British. It is unclear what activities the special forces may have been engaged in or whether the numbers of personnel have been maintained at this level. (11 April 2023)

RETRIEVED SOURCE 02

The FSB alleges it thwarted an attempt by British special forces to facilitate a landing of Ukrainian sabotage troops on the Tendrov Split, a strategic sandbar located in the Black Sea. The agency claims to have apprehended a senior figure within the Ukrainian naval special forces, providing his name and date of birth as proof of the operation's existence and the SBS's direct involvement. Britain's Defence Ministry has yet to respond to these allegations. (12 April 2024)

GENERATED ANSWER

Leaked US military documents indicated the possible presence of up to 50 UK special forces personnel in Ukraine between February and March of 2023.[1] The FSB alleged that **Ukrainian special forces** assisted a failed operation by **British special forces**, however Britain's Defence Ministry has not responded.[2]

That is an obviously incorrect answer where entities (bolded) have been scrambled—a typical RAG failure mode. We call this a "hard fail" because the error changes what is being claimed in the answer, introducing a factual inconsistency with the retrieved data.

But there is another type of error that we call a “soft fail”:

GENERATED ANSWER

Leaked US military documents indicated the possible presence of up to 50 UK special forces personnel in Ukraine between February and March of 2023.[1] The FSB, **Russia's federal security service**, alleged that British special forces assisted a failed operation by Ukrainian special forces, however Britain’s Defence Ministry has not responded.[2]

The RAG-V pipeline is strict. It’s looking for any factual inconsistencies between the generated answer and the cited source data. In this case, the model is drawing upon its internal knowledge—it was trained on internet data—to correctly state that the FSB is Russia's federal security service. But RAG-V detects this as a factual inconsistency because the retrieved data does not explicitly spell this out.

Are soft fails a problem? Not in this case, but we just got lucky. Soft fails can easily become hard fails on customer data that is different than the LLM's internet training data. Jargon in military documents or biomedical reports can easily confuse a human expert, let alone an LLM.

We prefer our RAG system to be conservative, sticking to the available information without making assumptions.

When RAG-Eval gets triggered, it generates an error message like this:

UNSUPPORTED CLAIM

The FSB alleges that Ukrainian special forces assisted a failed operation by British special forces. [2]

→

ERROR EXPLANATION

The cited source describes British special forces facilitating an operation involving Ukrainian forces, not Ukrainian special forces assisting a British operation.

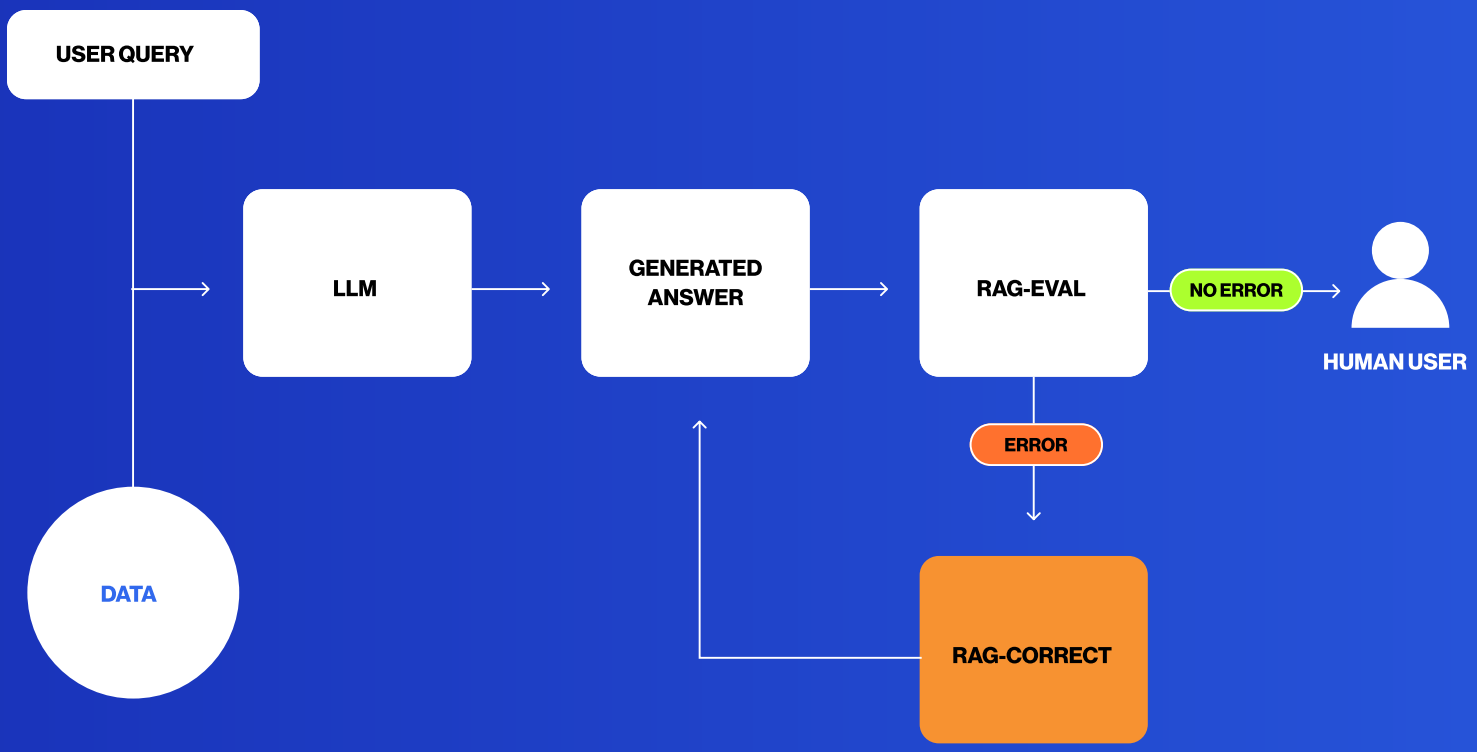
That error message is then bundled together with the information generated so far in the RAG-V pipeline—the user's question, the retrieved sources, the generated answer, and the error message—into an error report. With that, the LLM now has all of the information needed to correct itself. But what exactly do you do with it?

One option is to simply delete the inaccurate sentence. (The RAG-V pipeline maps generated sentences to factual claims, making this possible.) However, while this is both efficient and fast, there are two significant drawbacks. First, this degrades the generated text, often rendering it into an incomplete answer. Second, and worse, removing a sentence can change the overall meaning of a paragraph.

Another option is to fully rewrite the generated answer. This does generally work, but it is inefficient. In our analysis of RAG errors, we find that the vast majority only require a correction to the wording of a single sentence.

We've settled on a compromise solution. When RAG-V detects one or more errors, it rewrites only those sentences. Then we send the full text back through the pipeline. If any error is detected, the process repeats, giving the system another shot on goal.

When RAG-Eval gets triggered, it generates an error message like this:



How many times does RAG-V have to rewrite RAG text before it passes? If the distribution had a fat tail, with a high frequency of high-multiple retries, then this wouldn't work. But luckily, multiple retries are rare. The vast majority of detected errors are successfully corrected after a single pass. The most we have ever seen is 6 retries. So there is a long tail, and minimizing those outliers is our focus now.

But our most important goal has already been achieved. A RAG system using today's state of the art LLMs has an unacceptable error rate of roughly 10%. With RAG-V hooked up, that drops to 0.1%, with an error slipping through only once every 1000 times. With better trained LLMs, that will keep dropping.

RAG is a core AI component, powering critical functions such as search, question-answering, and automatic report generation. To ensure its reliability for mission-critical applications, RAG cannot remain a 'black box.' The feedback we've received so far from our defense and intelligence customers confirms that RAG-V is the way forward for earning that trust.

Learn more: www.primer.ai