squirro

Squirro‘s Retrieval-Augmented Generation Technology

Retrieval-Augmented Generation (RAG) for Large Language Models (LLMs) represents a significant advancement in the field of Artificial Intelligence (AI). By merging Squirro's Insight Engine's information retrieval (IR) stack with the capabilities of LLMs, Squirro blends conventional knowledge discovery with state-of-the-art generative AI. RAG overcomes the token limit of LLMs and empowers the chat assistant to:

Include the latest and potential real-time data

Uphold access control lists (ACLs)

Use classification signals for context selection

Incorporate traditional filtering based on categories

Overcome hallucinations by providing evidence

It begins when the user asks a question. Squirro's prompt reformulator interprets and restates this question, aiming to condense it into a format optimized for retrieving the user's most relevant context pieces from Squirro's Insight Engine's knowledge base.

The IR component fetches relevant information from the connected data sources of structured and unstructured data. Squirro semantically searches through the user's entitled content, retrieving contextually relevant paragraphs for the LLM to answer questions accurately.

On the LLM side, once it receives the targeted input from Squirro's Insight Engine, it uses generative question answering to construct accurate responses tailored to the user's initial intent.

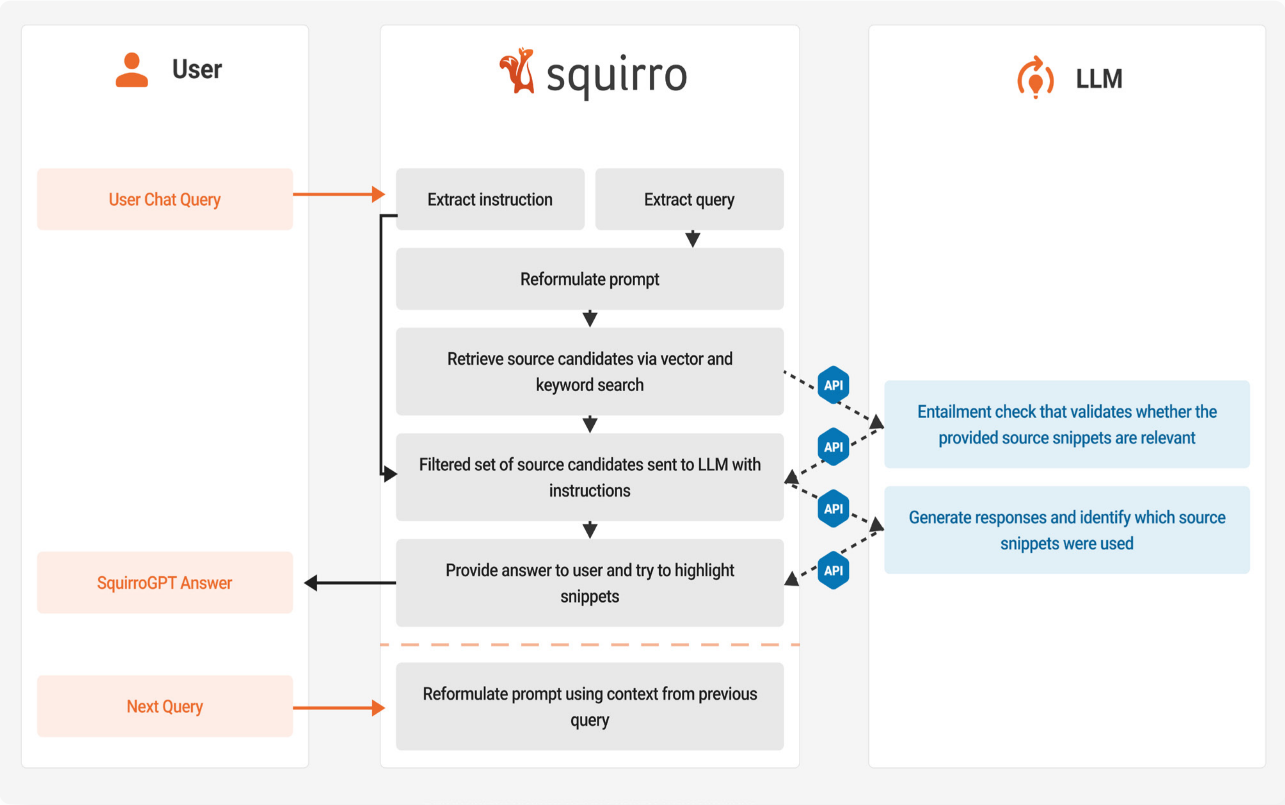
The provided evidence with each answer ensures that each response by the LLM can be traced back to a source, enhancing transparency and trust in the generative AI's outputs.

Next, if the conversation continues, the question reformulator adjusts the subsequent query by integrating the context from the previous question, thus maintaining a coherent and contextually enriched dialogue.

How it Works

This sophisticated interplay between Squirro's Insight Engine and the LLM's generative capabilities leads to a seamless and intelligent conversation flow, boosting efficiency by providing quicker, more accurate answers and fostering a more personalized and intuitive user experience.

The following diagram shows how user chat queries are converted to answers via the interaction between SquirroGPT and the underlying LLM that generates responses based upon provided evidence candidates.



User

squirro

LLM

User Chat Query

Extract instruction

Extract query

Reformulate prompt

Retrieve source candidates via vector and keyword search

Filtered set of source candidates sent to LLM with instructions

Provide answer to user and try to highlight snippets

SquirroGPT Answer

Entailment check that validates whether the provided source snippets are relevant

Generate responses and identify which source snippets were used

Next Query

Reformulate prompt using context from previous query

Note: This shows the general logic of the interaction between SquirroGPT and the LLM, it is not a technically precise process-flow diagram.

squirro

Generative AI for Enterprise Search & Insights

+41 44 586 98 98
contact@squirro.com
www.squirro.com

Zürich | London | New York | Singapore

About Squirro

Squirro is the generative AI-enabled semantic enterprise search and insight cloud company for searching, analyzing, and interpreting unstructured data. With that, Squirro empowers organizations to transform enterprise data into AI-driven insights and tailored recommendations.