How Retrieval-Augmented Generation Can Supercharge Automation Tools

Personal Notes:

What is RAG:

- Retrieval-Augmented Generation (RAG) is an architecture implemented in generative AI models that enhances the reliability of the model by searching for relevant information in an external database.
- Standard LLM models, such as GPT-4 are trained on static datasets such as provided websites and codebases up to a fixed cutoff date depending on when the model is implemented.
- RAG provides dynamic searching -- allowing a system to send a query to an external retriever, such as the web or other knowledge bases (wikipedia).

Traditional LLM Limitations:

- Hallucinations: LLM "hallucination" describes the phenomena of LLM's generating incorrect information by relying on patterns in trained data instead of real-time facts.
- LLM only "sees" a limited number of tokens and thus cannot "remember" earlier parts of conversations and cannot directly process large documents or images.

RAG + LLM Solution:

- Dynamic allocation of information prevents hallucinations. Utilizes search engines and other web tools to get up-to-date information.
- RAG allows the retrieval of relevant snippets from large documents/PDF's instead of having to process the entire provided input which prevents token overflow.

How RAG Supercharges Automation:

- Context-Aware Understanding: Automation relies on understanding the context behind the task needed to be performed. RAG enables the system to respond intelligently even if the LLM was not trained on the specific given input.
- Generating Next-Steps: RAG goes beyond summaries -- it can generate entire new steps and take action by itself. Given a request from a user, RAG enables LLM processing to take decisions on its own and create its own task.
    - This happens through the retriever. When the user sends in a request, RAG enables the retriever to pull contextual-data, outside of the LLM's framework, to generate actions.
- Scalability Across Departments:
    - Sales: Auto-draft replies based on previous deals or proposals
    - HR: Answer employee queries about policy by retrieving from handbooks
    - IT Support: Respond to requests using documentation and past tickets
    - Finance: Parse invoice-related queries and cross-reference accounting databases