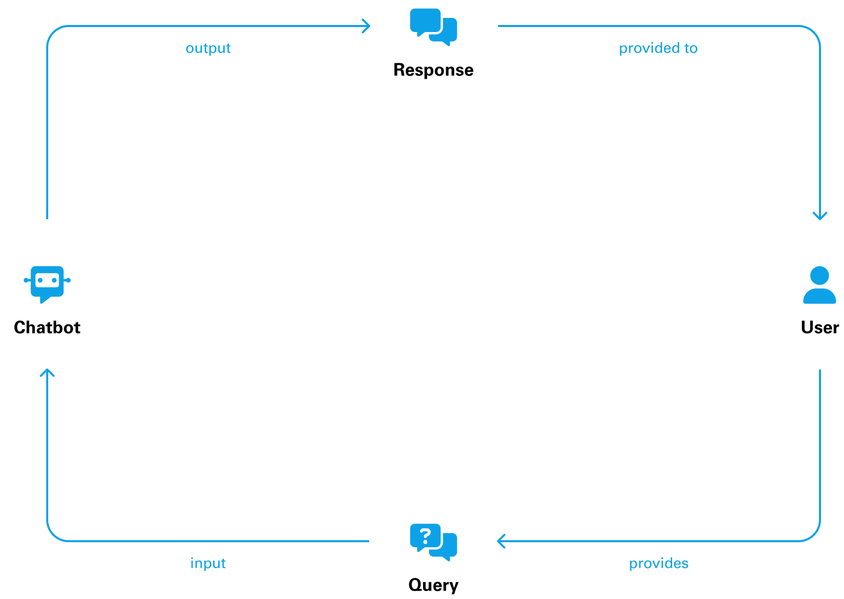
December 20, 2023

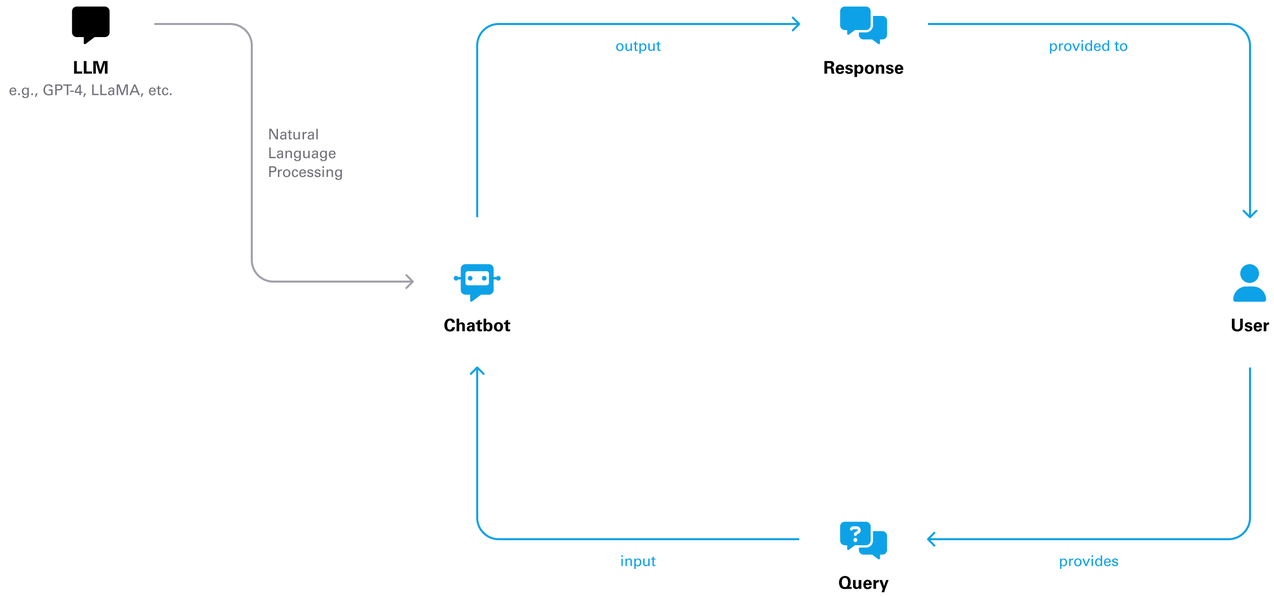
Designing LLM-based Chatbots
An Overview of Configuration, Retrieval-Augmented Generation, and Feedback

Dubberly Design Office

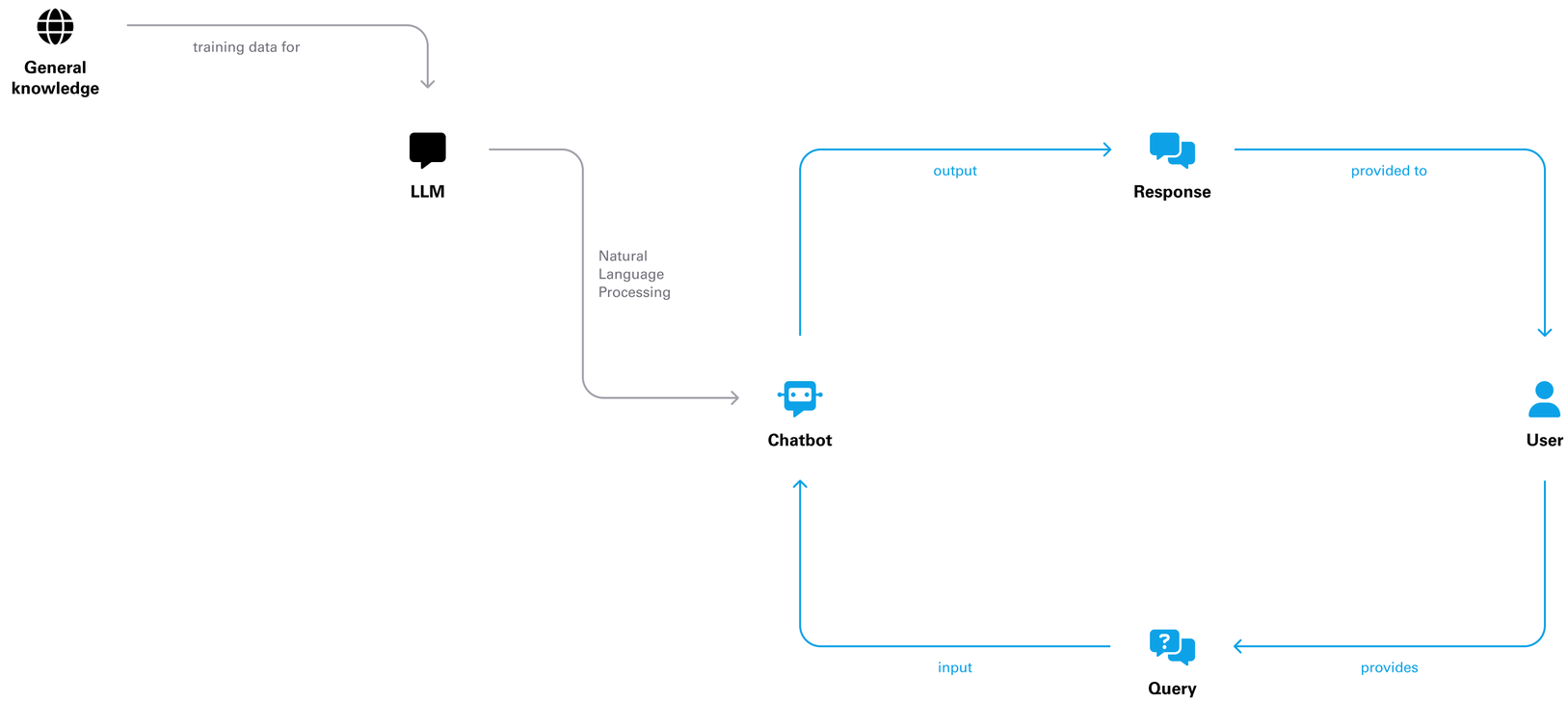
A chatbot provides a response to a user’s query.



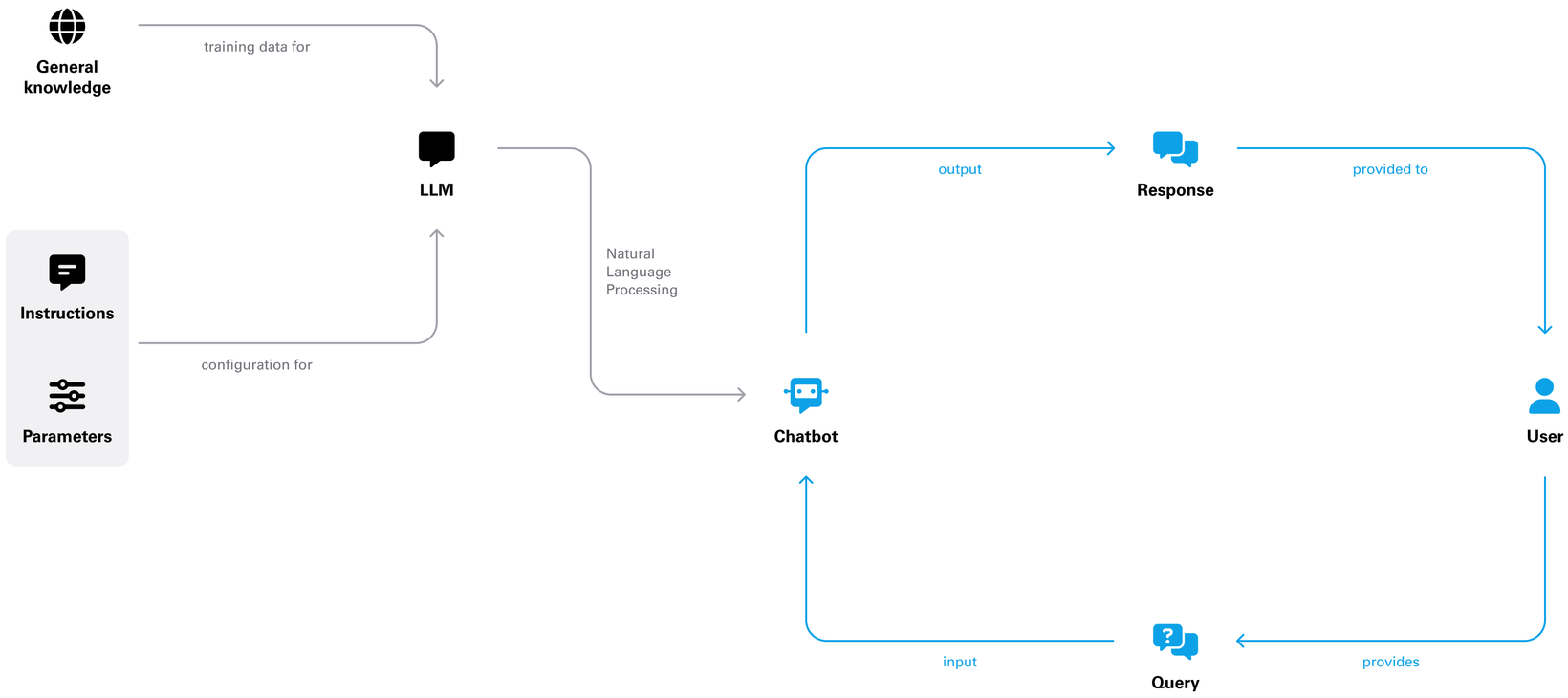
An LLM provides the chatbot with the capability to understand and respond in natural language.



LLMs are built by processing massive amounts of text data (e.g., scraped from the internet), which provides the chatbot with broad, general knowledge.

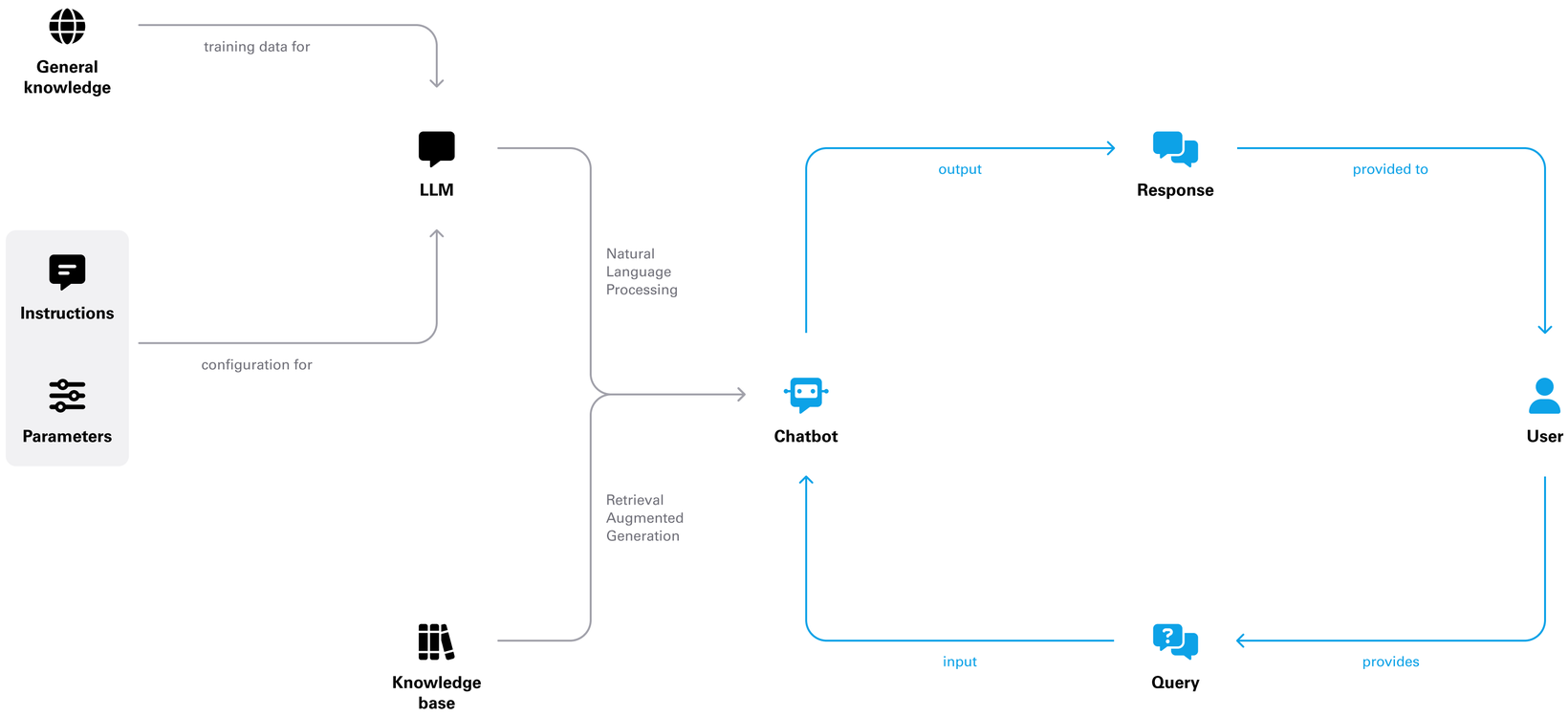


The LLM is also configured with instructions[1] and other parameters[2] which affect its responses.

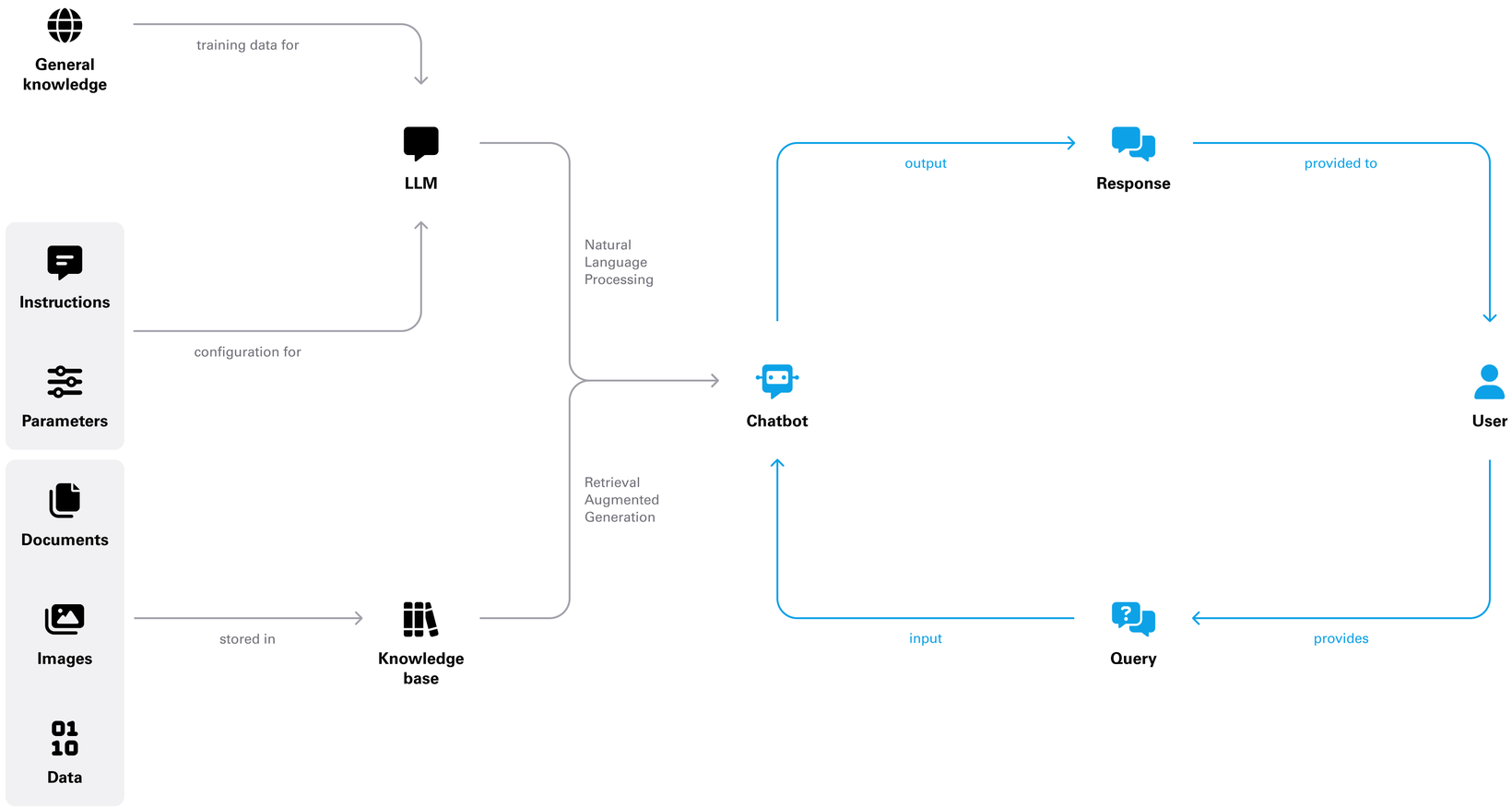


1. Instructions guide the chatbot's personality and define its goals — instructions can be used to assign the chatbot with a role, dictate the tone of its responses, etc. In a virtual assistant, the chatbot's goals should relate to the user's goals.
2. In GPT-3 and GPT-4 for example, the 'temperature' parameter controls the randomness of the model's responses. Lower temperature decreases randomness and makes responses more predictable. Other parameters enable control over the length of responses, repetition of words or topics, etc.

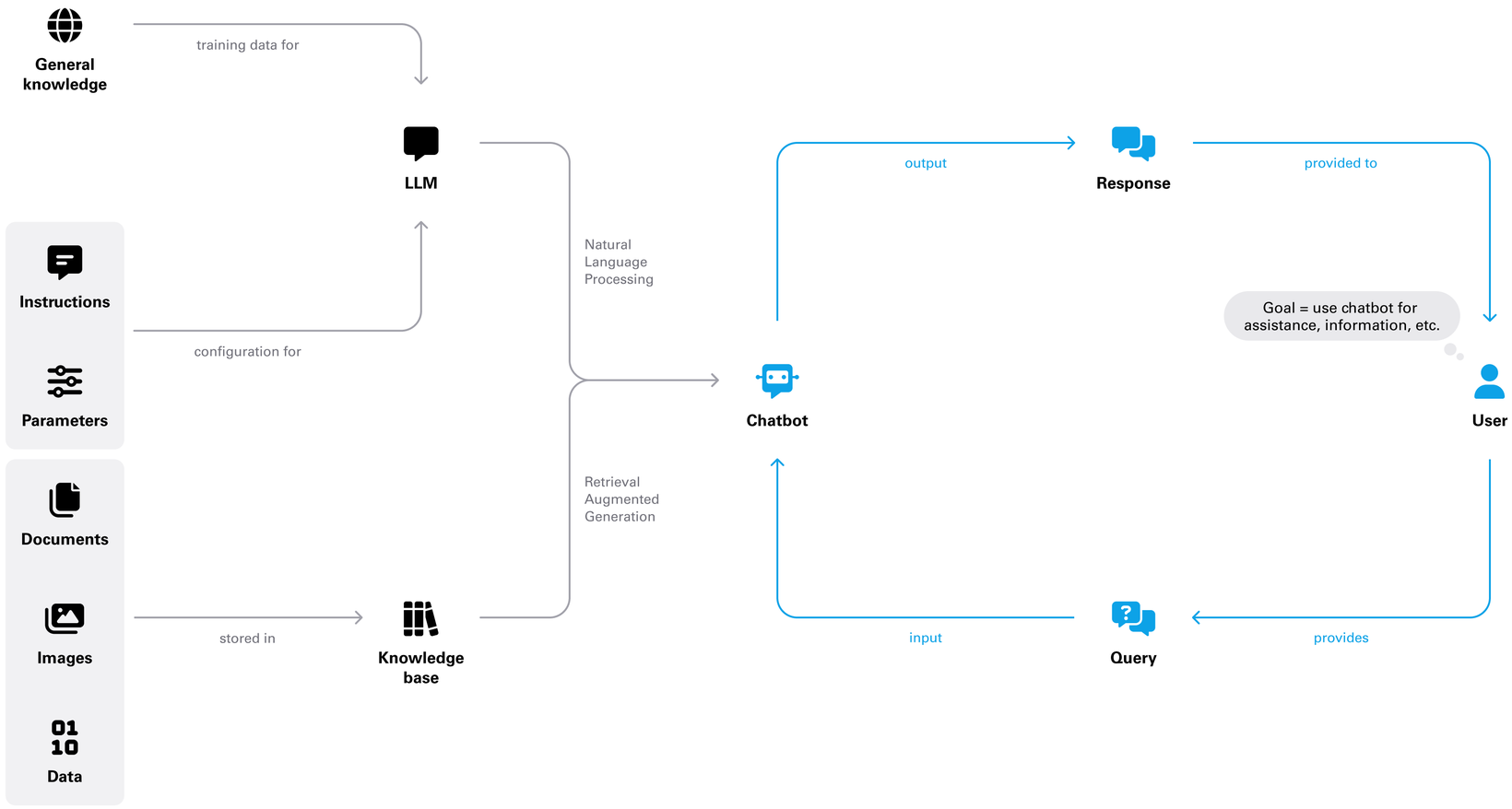
Retrieval-Augmented Generation (RAG) enables a chatbot to access a knowledge base with specific information it can reference directly for more accurate and reliable responses.



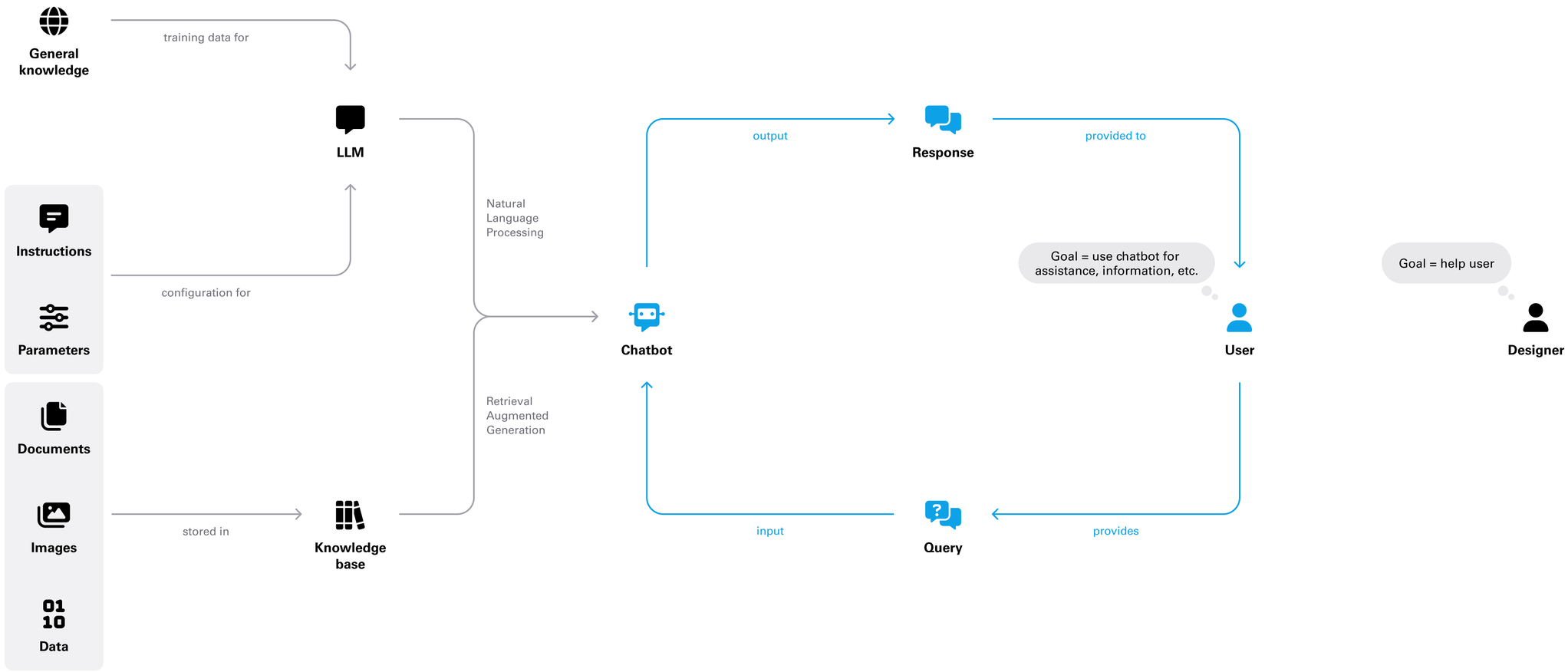
The knowledge base might contain documents, images, or other data.



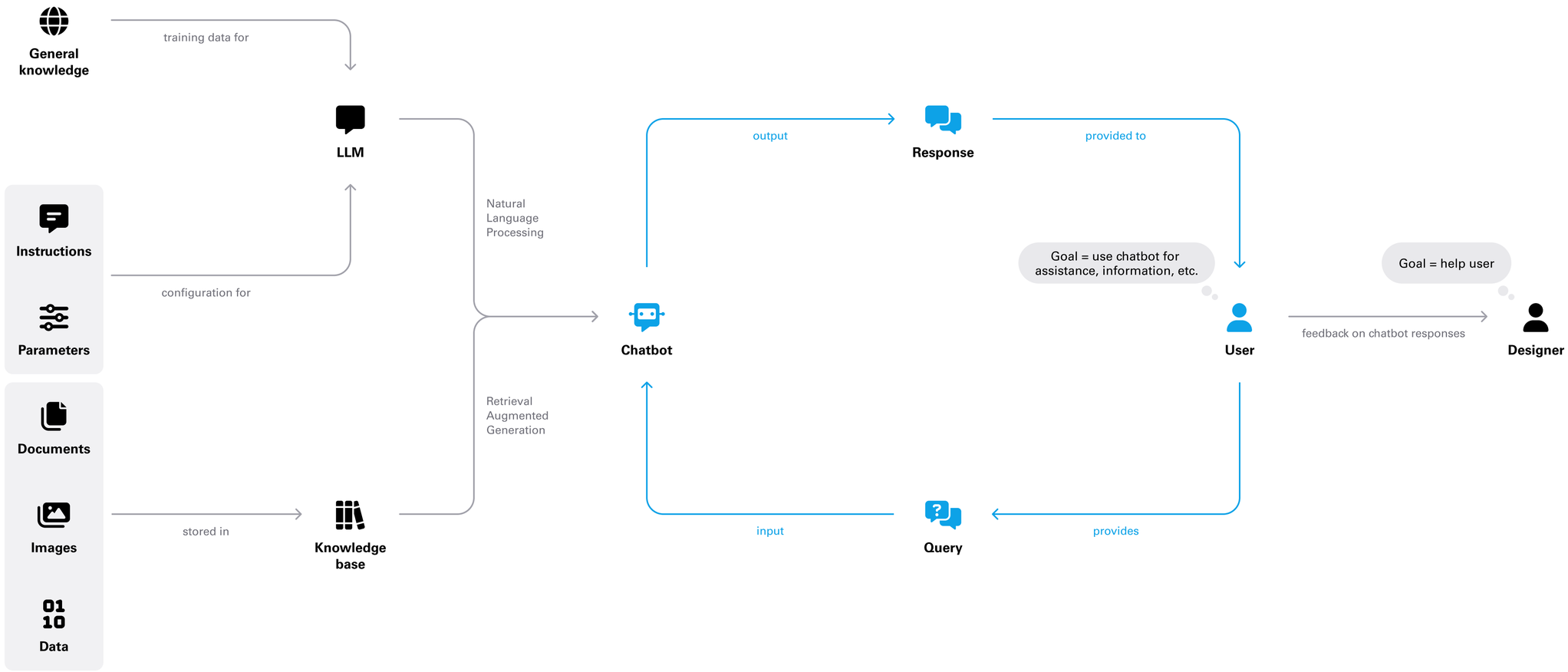
A user’s goal might be to use the chatbot for assistance with a task, retrieving and digesting information, etc.



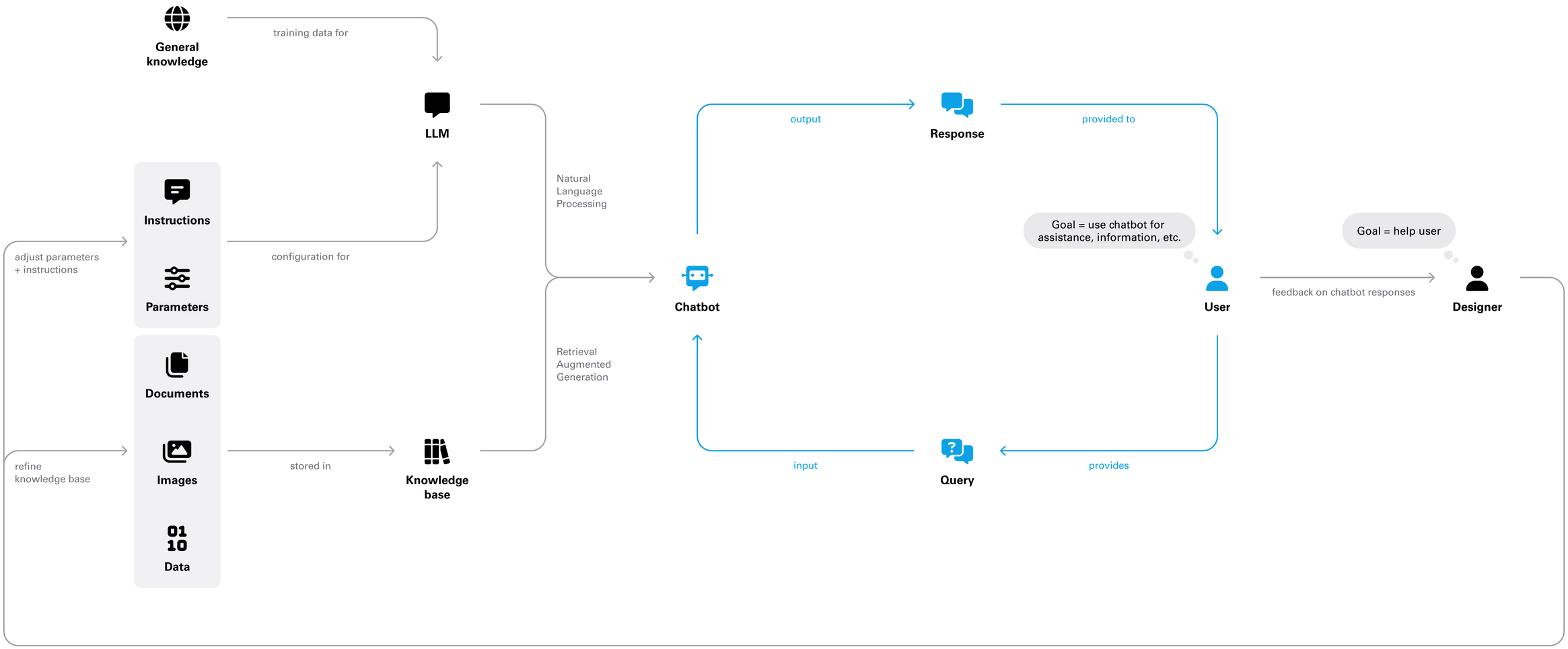
The chatbot’s designer (i.e., whoever creates it) might be to help the user.



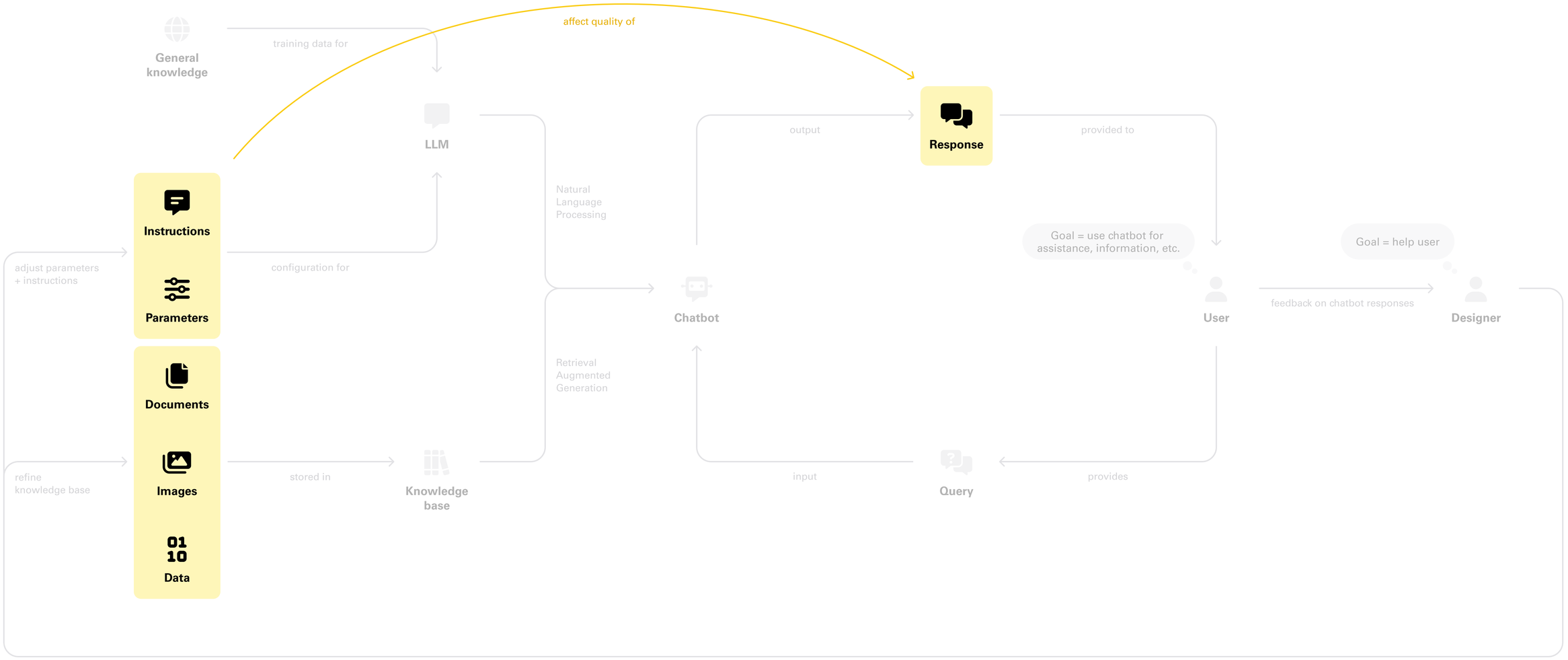
The user can provide feedback on how helpful the chatbot’s responses are.



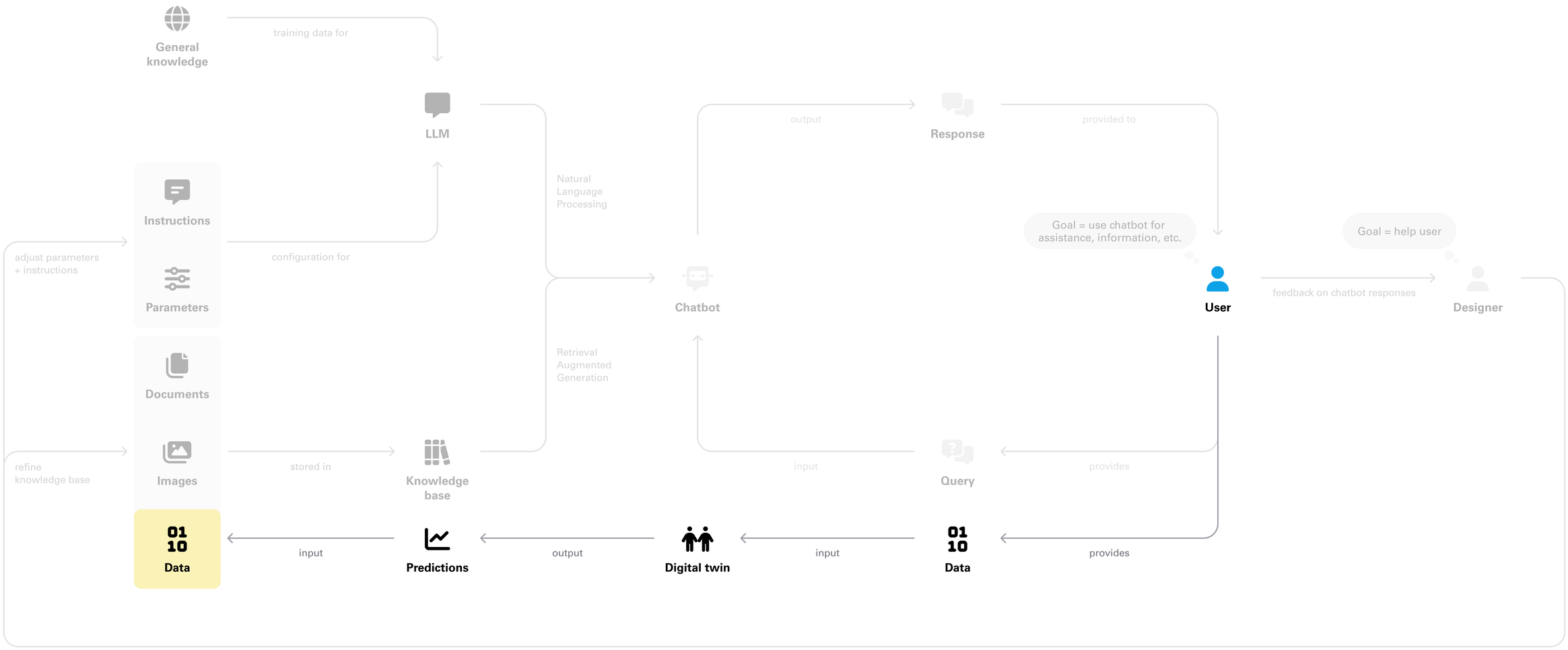
The designer can then refine the chatbot's knowledge base and adjust its configuration based on user feedback.



By improving the LLM's configuration and the chatbot's knowledge base, they can improve the quality of its responses.



In a system that includes a digital twin, predictions generated by the digital twin would be integrated with the chatbot's knowledge base.



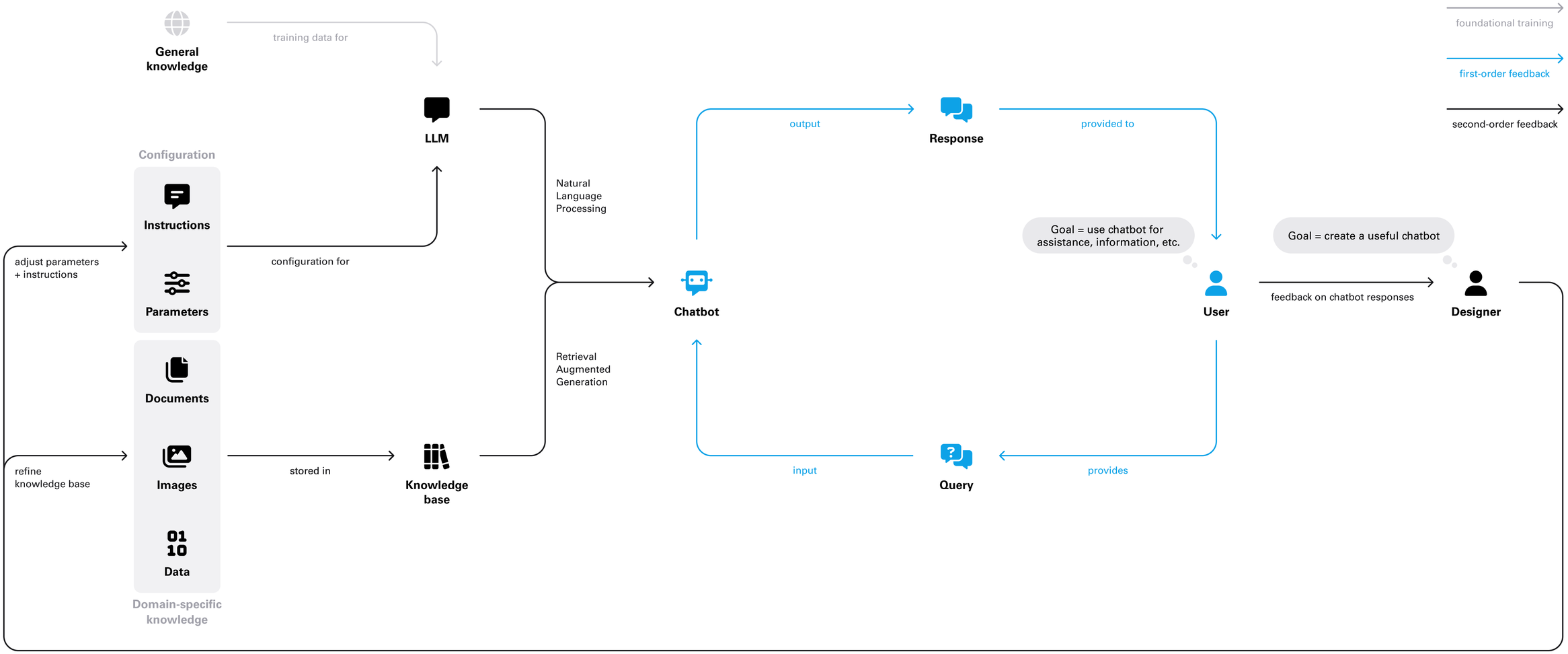
LLM Chatbot design as second-order feedback.

Large Language Models (LLMs) are trained on text data from large and diverse datasets (e.g., scraped from the internet), enabling them to understand and respond in natural language and providing them with broad, general knowledge.

A chatbot built with a LLM can be configured with instructions (which guide the chatbot's personality, define its goals, etc.) as well as other parameters that affect its responses.

Retrieval-augmented generation (RAG) enables a chatbot to access a knowledge base with information from outside its model that it can reference for more accurate and reliable responses.

A designer can improve the chatbot's responses based on user feedback by iteratively refining its knowledge base and configuration.



Designing LLM-based Chatbots

An Overview of Configuration, Retrieval-Augmented Generation, and Feedback

December 20, 2023

Dubberly Design Office