RAG Retrieval Augmentation and Model Training Samples

1. Course Content

1. This section introduces the concepts of large language model hallucinations, RAG retrieval augmentation, and model training samples, providing a theoretical foundation for subsequent practical operations and facilitating understanding.

2. What are Large Language Model Hallucinations? Why do Large Language Models Produce Hallucinations?

RAG retrieval augmentation and model training samples are primarily used to reduce large language model hallucinations. Let's first understand what large language model hallucinations are.

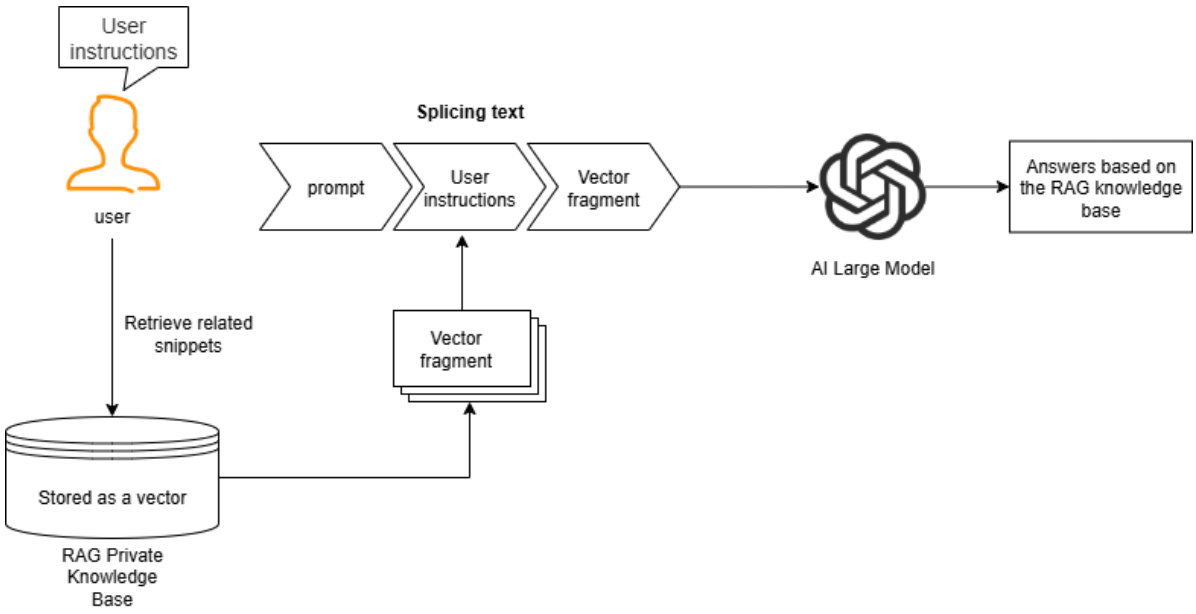
Large language model hallucinations refer to the model's output of content that is inconsistent with real-world facts and logic when understanding the environment, generating instructions, or planning actions. This leads to the robot performing incorrect actions or the user's expected results deviating significantly.

Why do large language models produce hallucinations? **Large language models are essentially probabilistic prediction systems based on massive amounts of text data.** In certain specific domains or scenarios, the training data may lack sufficient samples, causing the model to "fabricate" content to fill the gaps.

3. RAG Retrieval Augmentation and Training Examples

3.1 RAG Retrieval Augmentation

RAG is an architecture that combines a **retrieval system with a generative model**, aiming to address the limitations of traditional generative models in scenarios such as open-domain knowledge question answering and real-time information retrieval (e.g., outdated knowledge, factual hallucinations, and long-text dependency). Its core logic is: **to obtain relevant knowledge through retrieval and integrate it into the prompt, allowing the large language model to refer to the corresponding knowledge**, achieving a "retrieve first, then generate" approach.



3.2 Training Examples

Training examples are included in the RAG knowledge base. The robot is pre-configured with two knowledge bases at the factory: an action function library and training examples. The training examples contain training samples from the course case scenarios, providing ideas for the large language model's decision-making and planning in specific scenarios. The subsequent section [2. AI Large Language Model Fundamentals - 5. Configuring AI Large Language Models] will explain how to configure and extend proprietary knowledge bases and training examples. ## 4. RAG Retrieval Augmentation and the Role of Training Samples in Robots

4.1 Reducing Large Language Model Hallucinations

When large language models encounter completely unfamiliar domains and specific user requirements, they may provide results based on existing training data. However, these results may differ from what the user expects or may not meet the requirements of a particular scenario.

4.2 Increasing the Robot's Scenario Generalization Capability

Robots are pre-configured with training examples related to course cases (explained in detail in the large language model course configuration). These training examples provide the large language model with reference information in specific scenarios. Users can also add their own training samples to customize their robots to adapt to different application areas.

4.3 Reducing Model Prompting

If the number of training samples and knowledge bases is small, they can be directly included in the large language model prompt. The prompt is the pre-set role and requirements that the user provides to the large language model. If you need to add your own training scenarios and robot action function libraries, as the size of the knowledge base increases, a longer prompt will consume more tokens. Therefore, RAG retrieval augmentation is introduced to uniformly manage the robot action function library and training examples.

4.4 Facilitating the Expansion and Management of Robot Capabilities

The Alibaba Cloud Bailian Large Language Model Platform (referred to as the Tongyi Qianwen platform) provides online functions for managing knowledge bases and training examples, allowing users to easily manage and expand data. Domestic users default to using the Alibaba Cloud Bailian Large Language Model Platform to manage the action function library and training samples, while international users use locally deployed Dify to manage knowledge base content.