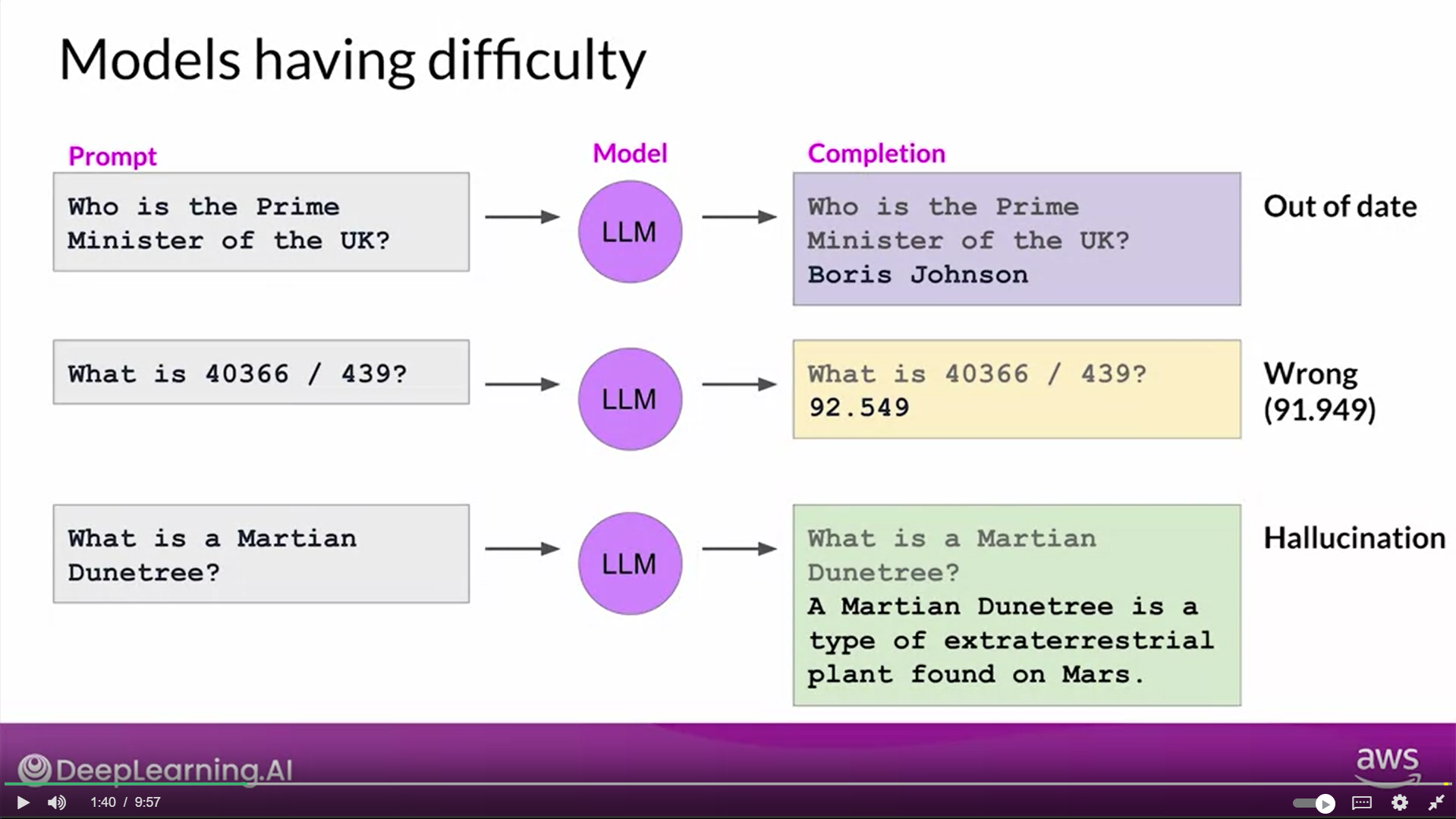
Using LLMs in Applications



A diagram of a computer application

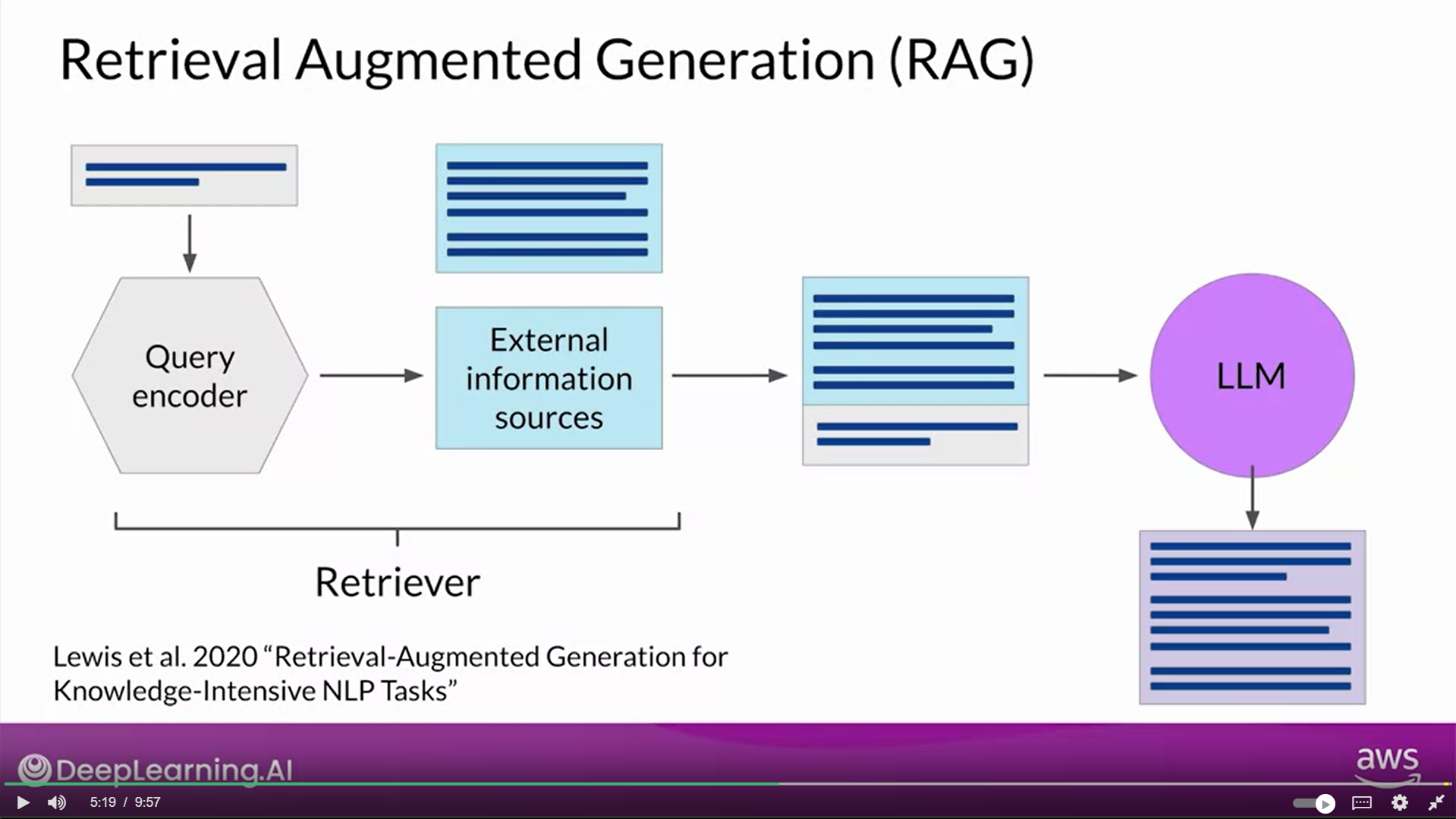
Description automatically generated

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation, or RAG for short, is a framework for building LLM powered systems that make use of external data sources.

RAG is a great way to overcome the knowledge cutoff issue and help the model update its understanding of the world.

RAG is useful in any case where you want the language model to have access to data that it may not have seen. This could be new information documents not included in the original training data, or proprietary knowledge stored in your organization's private databases.



* At the heart of this implementation is a model component called the Retriever, which consists of a query encoder and an external data source.
* The encoder takes the user's input prompt and encodes it into a form that can be used to query the data source
* These two components are trained together to find documents within the external data that are most relevant to the input query.
* The Retriever returns the best single or group of documents from the data source and combines the new information with the original user query.
* The new expanded prompt is then passed to the language model, which generates a completion that makes use of the data.

