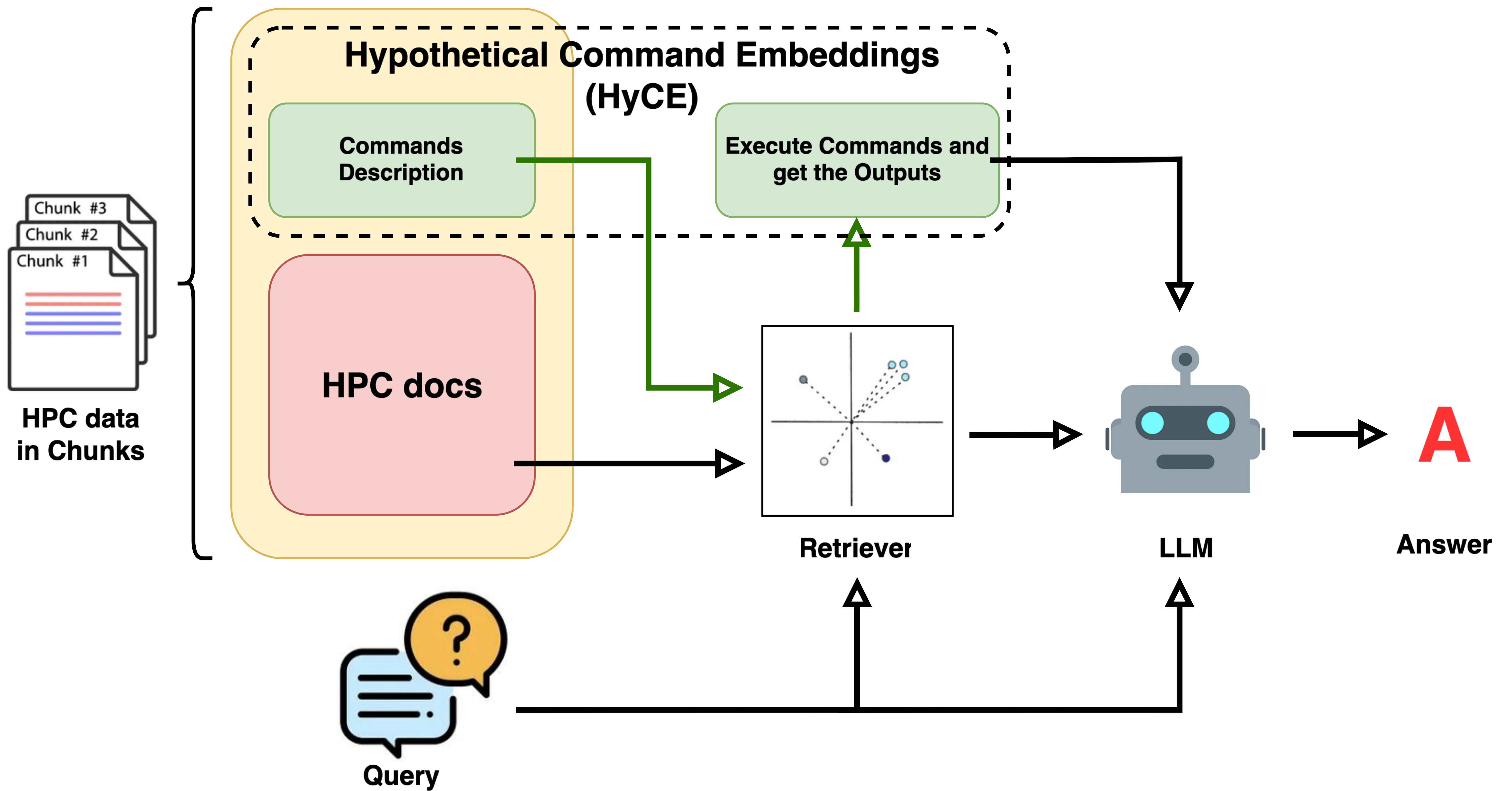


Problem? How do we combine real-time and user specific HPC data to RAG?
Solution! Embed command descriptions instead of command names. Retrieve relevant docs and **commands, execute** commands. Then include the command output as prompt for LLM.



Inspect HyCE closely

Query: What Is the status of my job?

Command Name: `qstat -u $USER`

Command Description: The command `qstat -u $USER` shows the status of jobs submitted by the current user in HPC environments.

Comparing Text Similarity (higher the better)

Cross-Encoders	Avg Sim Score Query vs Command Name	Avg Sim Score Query vs Command Description
cross-encoder/stsb-roberta-large	0.0906	0.5285
cross-encoder/ms-marco-MiniLM-L-12-v2	-9.0231	-5.1792
dangvantuan/CrossEncoder-camembert-large	0.2501	0.5718
yunyu/cross-encoder-stsb-deberta-v3-large	0.1472	0.3981

Example Responses

Q1: What GPUs are available for me?

RAG with just HPC docs: The documents do not provide specific information about the types or models of GPUs available. They mention the high demand for GPU nodes and the process for accessing them, such as waiting in a common queue or purchasing a GPU node for priority access. If you need precise details about specific GPU models or availability, you may need to consult additional resources or contact the provider directly.

RAG with HPC docs and HyCE: **You have V100 and A100 GPUs available.**

Q2: What is the status of my job?

RAG with just HPC docs: To determine the status of your job, you should run the ``qstat`` command and look for the "Q" in the second last column in the output:
 - If you see an "R", it means your job is currently running.
 - If you see an "H", it means your job is on hold and cannot currently run.

You can also use the command with your ZID or the environment variable ``$USER`` to list your jobs and have you more detailed status information.

RAG with HPC docs and HyCE: Based on the provided documents, **it seems your job is currently queued, waiting for a suitable compute node to become available.** You can check its status using the ``qstat`` command and look for a "Q" in the second last column to confirm it is in the queue.

✓ Command description aligns more semantically with the query

✓ Responses are more specific and accurate