

# RAG for HPC data

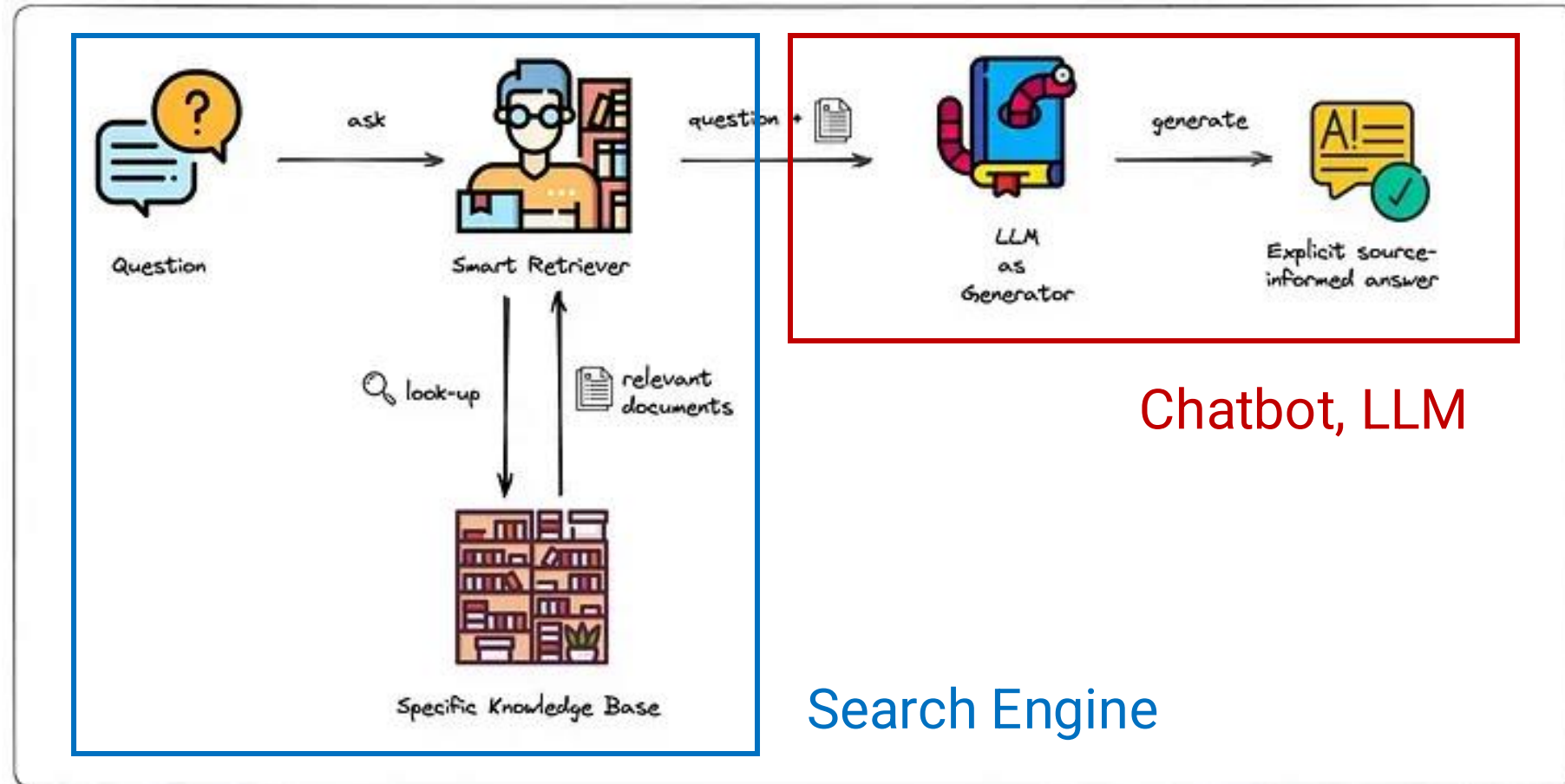
Yusuke Miyashita  
Research Technology Services(ResTech)

Compute | Community | Data



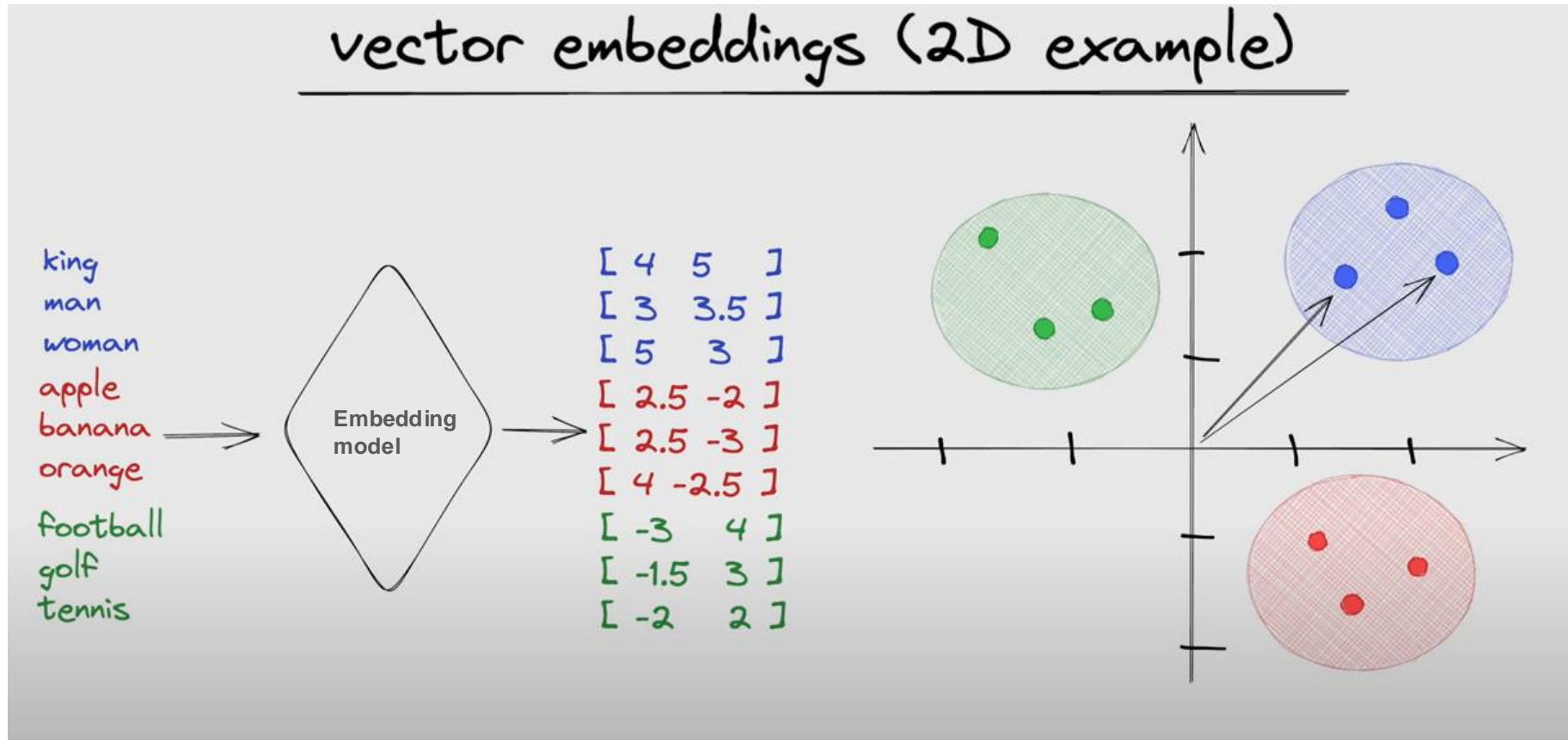
**UNSW**  
SYDNEY

# RAG



<https://blog.ml6.eu/tuning-the-rag-symphony-a-guide-to-evaluating-llms-35a2de271f42>

# Text Embeddings



<https://causewriter.ai/courses/ai-explainers/lessons/vector-embedding/>

# Problem ? : How do you embed HPC commands?

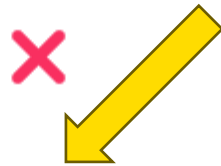
1. RAG needs to run commands to obtain real time and user specific information
2. Command outputs need to be included in the prompt of LLM
3. **What do we embed??**

Commands are syntactic and not suited for semantic retrieval in RAG

- nvidia-smi
- qstat -u \$USER
- pbsnodes
- etc...

## Solution 💡 : Use Command description rather than command itself

Query: What GPUs are available for me?



### Command Names

- nvidia-smi
- qstat -u \$USER
- pbsnodes
- etc...

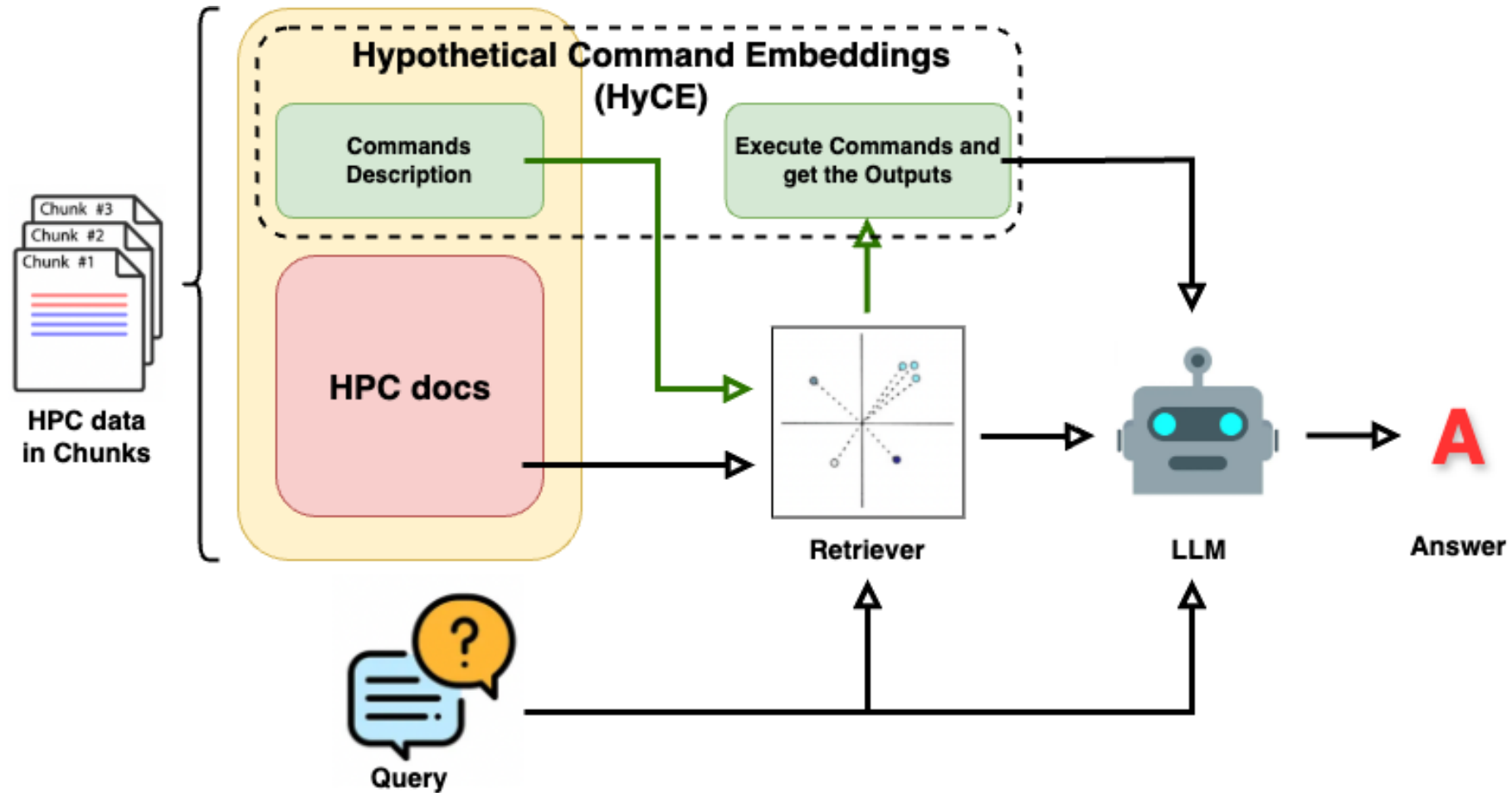


✓ **Command descriptions align better with queries**

### Command Descriptions

- Displays GPU usage, driver version, and other information about NVIDIA GPUs.
- Shows the status of your queued and running jobs on the HPC system.
- Provides information about the nodes in the HPC cluster, including status and availability.

# Extended RAG Architecture



# Example outputs w/o HyCE

## 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 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.

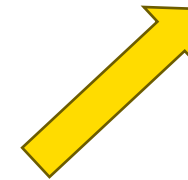
**RAG with HPC docs and HyCE:** Based on the provided documents, it seems your job is currently queued, waiting for a suitable compute node to become available. You can check its status using the `qstat` command and look for a "Q" in the second last column to confirm it is in the queue.

✓ Responses are more specific and accurate

# How effective is HyCE actually?

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

✓ **Command description aligns more semantically with the query**

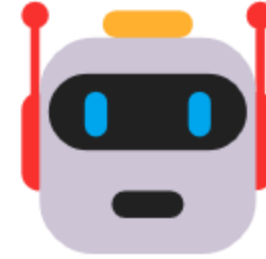




# HyCE is Orthogonal to other RAG improving techniques



Prompt Engineering



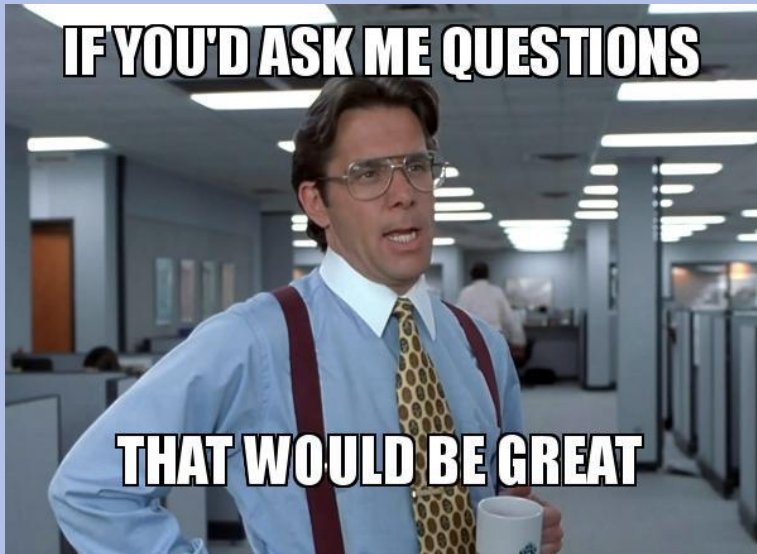
Tuning retrieval, re-rank model  
and LLM



Other RAG techniques

# Other factors not discussed in the presentation...

- Automatic RAG evaluation
- Security consideration for running RAG on HPC environment
  - Data privacy
  - Command execution risks



**Paper coming out on 25th of Nov!!  
Follow ResTech to keep updated!!**

