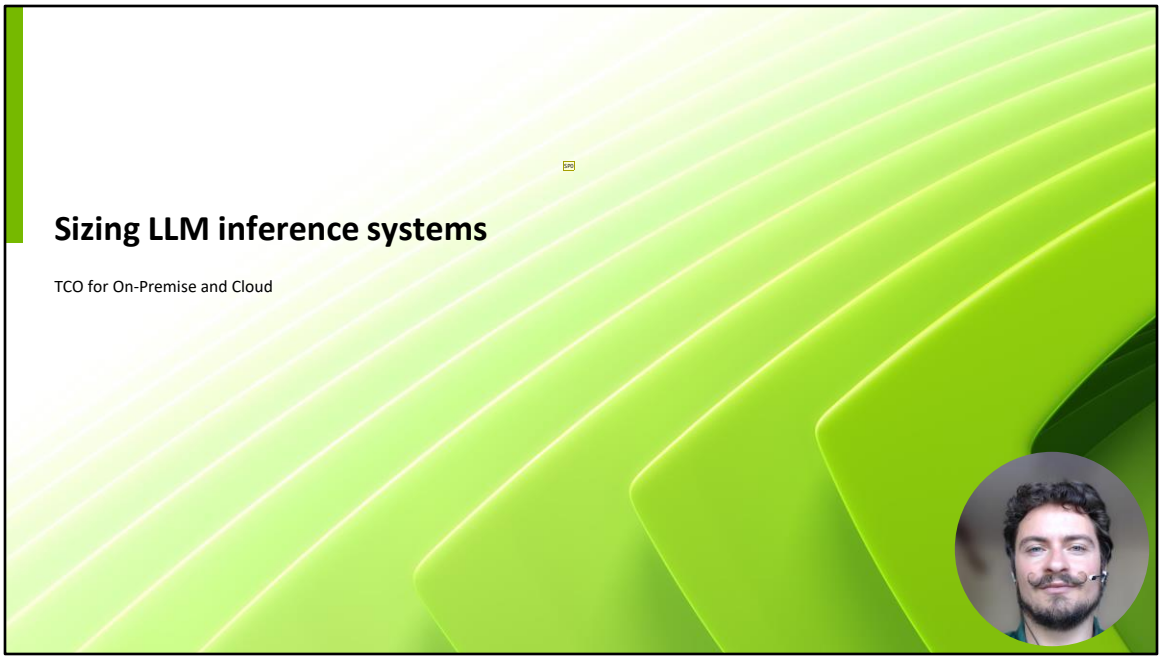


Discuss about the same things that Anshul/Meriem do

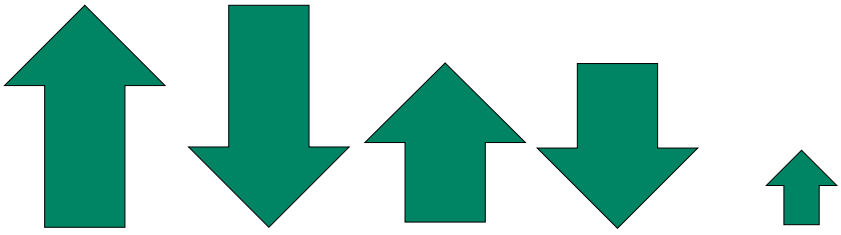
In this course we will learn…

Before we start, we'd like to remark that you don't need a deep expertise of LLMs to benefit from the course
We've prepared ready-to-run notebooks with all the commands and assets for you. Once you click on the play notebooks, you will access the course environment with the notebooks. We recommend you click on "play" now since it takes a few minutes to load.
Before jumping into the notebooks, let me introduce the instructors of this DLI and an introduction to inference and sizing.

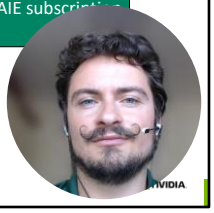
From Speeds To Total Costs Of Ownership

Multiple Factors

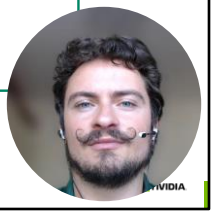
Effect On Final Price

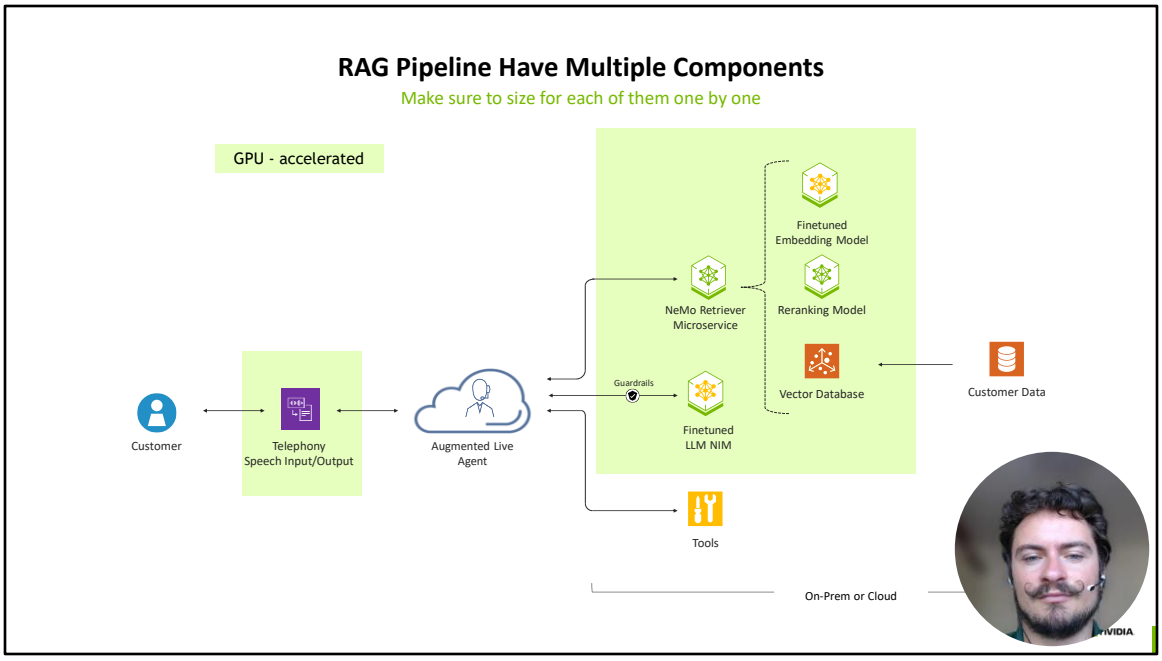


To meet latency requirements, you must use at least 4 GPUs per instance	Model accuracy in FP8 is amazing	You can only rely on pre-approved cloud	They offer value discount	They have no NVAIE subscri



Cloud		On-Prem
	Hardware Costs	
Is autoscaling possible? How long is commitment? Additional storage and networking costs?		Depreciation period? Hosting Costs? Cooling? Electricity? Management?
	Software Costs	
Any custom SW from the cloud provider? Is NVAIE license included in the instance price?		NVAIE license.

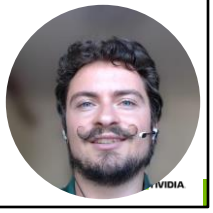




Long context
Tools calling
Embedding/Reranking

Objectives of this notebook

1. Learn how to estimate total TCO of LLM inference for both on-prem and in the cloud
2. Learn about requests peak estimations
3. Be able to select the best deployment option from the available ones




NVIDIA.