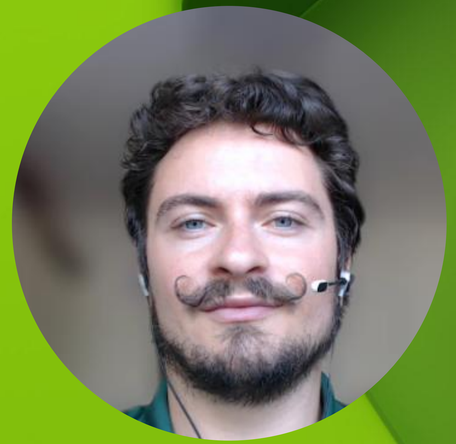
Sizing LLM inference systems

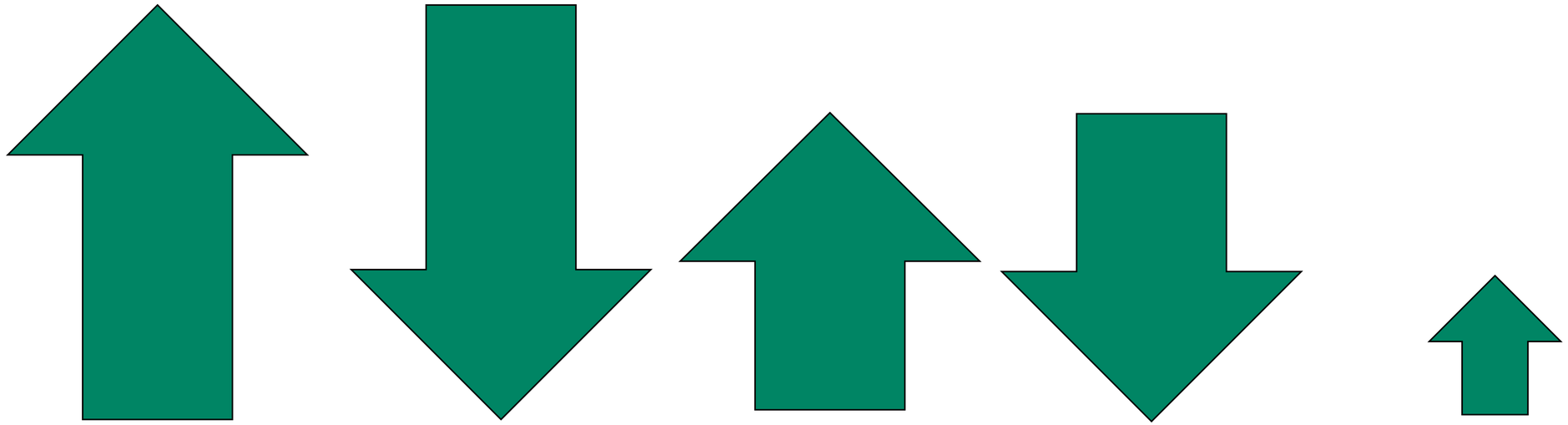
TCO for On-Premise and Cloud



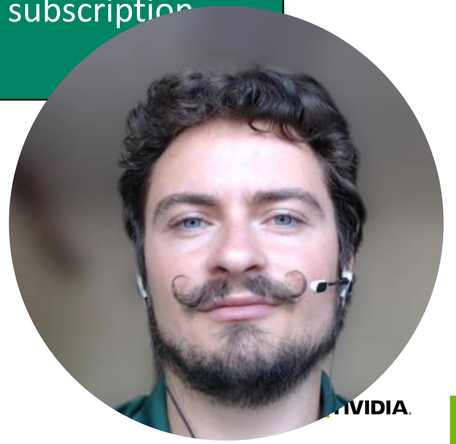
From Speeds To Total Costs Of Ownership

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 subscription



Cloud

On-Prem

Hardware Costs

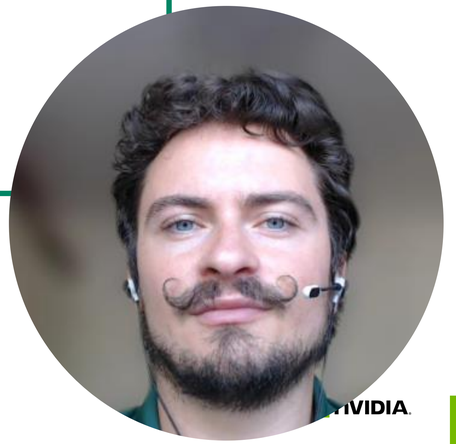
Is autoscaling possible?
How long is commitment?
Additional storage and networking costs?

Depreciation period?
Hosting Costs?
Cooling? Electricity?
Management?

Software Costs

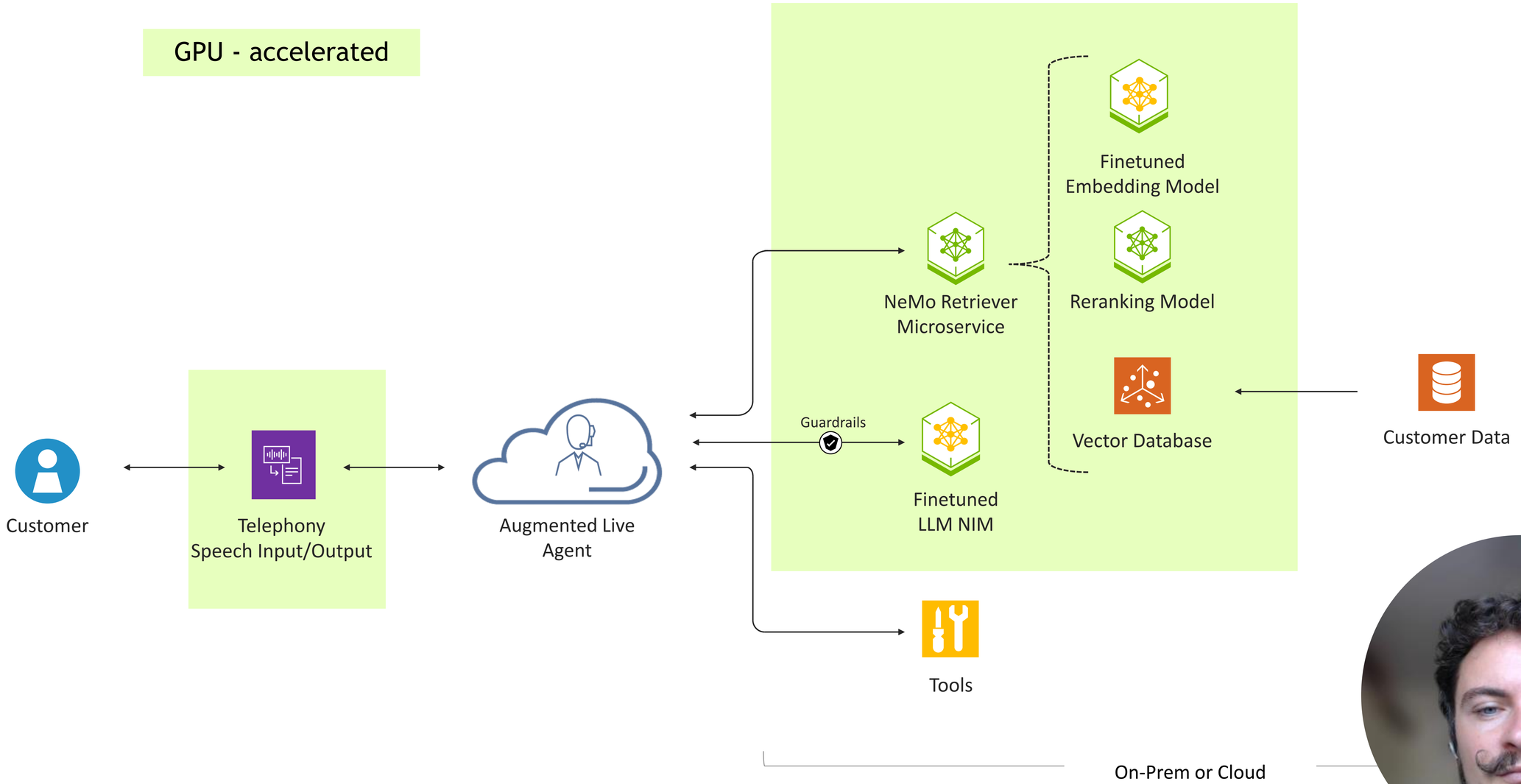
Any custom SW from the cloud provider?
Is NVAIE license included in the instance price?

NVAIE license.



RAG Pipeline Have Multiple Components

Make sure to size for each of them one by one



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