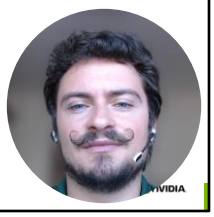

In this notebook, you will explore the provided results of the NIM speed benchmarks.
You'll learn, what metrics are represented there, why is batching so important, and what tradeoffs one has to navigate, to choose the deployment configuration for their model.

## Dataset Provided

## Dataset Overview

- Models
  - llama3-8b
  - llama3-70b
- GPUs: A100 and H100, both with NVLink
- Precision: 'fp16', 'fp8', 'bf16’
- Tensor Parallelism: 1, 2, 4, 8
- Input and Output lengths:
  - '200 in → 200 out’, '200 in → 1000 out', '200 in → 2000 out’
  - '2000 in → 200 out', '2000 in → 2000 out’
  - '7000 in → 1000 out'



## Objectives of this notebook

1. Understand, how to analyze benchmarking results
2. Observe the effect on the speeds from changing the key parameters
3. Learn, how to estimate the required size of the deployment


NVIDIA.