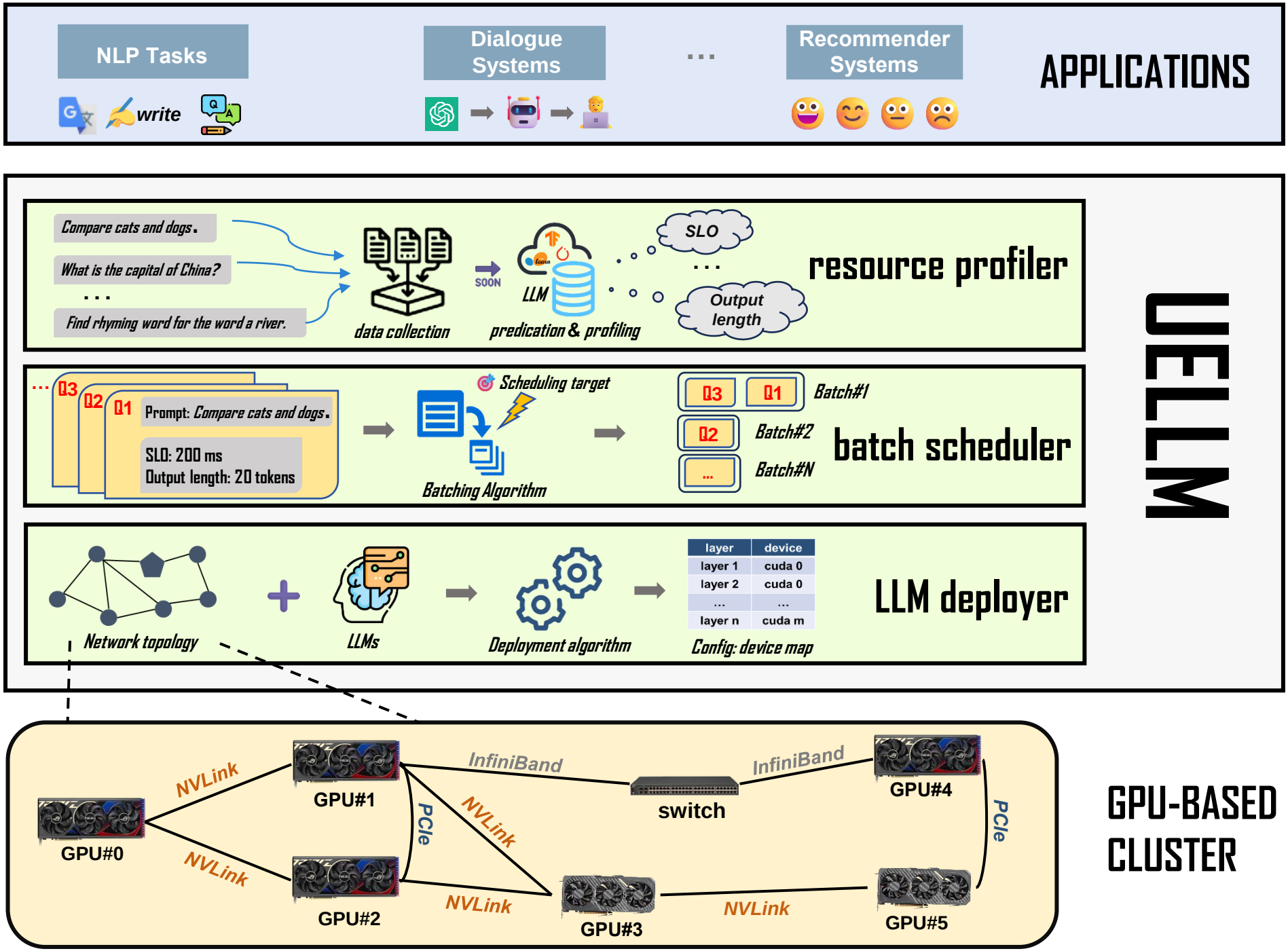
NLP Tasks
write
Dialogue Systems
...
Recommender Systems
APPLICATIONS
Compare cats and dogs.
What is the capital of China?
...
Find rhyming word for the word a river.
data collection
SOON
LLM
predication & profiling
SLO
...
Output length
resource profiler
Q3
Q2
Q1
Prompt: Compare cats and dogs.
SLO: 200 ms
Output length: 20 tokens
Scheduling target
Batching Algorithm
Q3
Q1
Batch#1
Q2
Batch#2
...
Batch#N
batch scheduler
UELLM
Network topology
LLMs
Deployment algorithm
layer
device
layer 1
cuda 0
layer 2
cuda 0
...
...
layer n
cuda m
Config: device map
LLM deployer
GPU#0
NVLink
NVLink
GPU#1
GPU#2
PCIe
InfiniBand
NVLink
NVLink
switch
GPU#3
InfiniBand
GPU#4
GPU#5
PCIe
NVLink
GPU-BASED CLUSTER