Model Specific Questions:

Common Questions

- What is the name of the AI model you plan to use (e.g., Llama 3)?
  Field type: Similar to Model name drop down that is already present in the UI in simulate performance tab

- Which framework is the model built in (e.g., PyTorch, TensorFlow, JAX)?
  Field type: Similar to Framework drop down that is already present in the UI in simulate performance tab

- What is the primary task for this model (e.g., Inference, Training)?
  Field type: Similar to Task Type drop down that is already present in the UI in simulate performance tab

- What is the underlying architecture of the model (e.g PhiForCausalLM)?
  Field type: Similar to architecture drop down that is already present in the UI in simulate performance tab

- What type of model is it (e.g phi, lama)?
  Field type: Similar to model type drop down that is already present in the UI in simulate performance tab

- What is the size of the model in terms of total parameters (in millions)?
  Field type: Similar to Parameters (millions) text field that is already present in the UI in simulate performance tab

- What is the file size of the model on disk (in MB)?
  Field type: Similar to Model size (MB) text field that is already present in the UI in simulate performance tab

- What numerical precision is required for inference (e.g., FP32, FP16, INT8)?
  Field type: Similar to Precision drop down that is already present in the UI in simulate performance tab

- What is the vocabulary size of the model?
  Field type: Similar to Vocabulary size text field that is already present in the UI in simulate performance tab

- Which activation function is predominantly used in the model (e.g., ReLU, GeLU)?
- **Field type:** Similar to activation function drop down that is already present in the UI in simulate performance tab

- What are the computational requirements of the model in GFLOPs (billions of floating-point operations)?
  Field type: Similar to GFLOPs (Billions)text field that is already present in the UI in simulate performance tab

- How many hidden layers does the model contain?
  Field type: Similar to Number of hidden layers text field that is already present in the UI in simulate performance tab

- How many attention layers are in the model?
- **Field type:** Similar to Number of Attention layers text field that is already present in the UI in simulate performance tab

  -What are the primary business objectives of this new infrastructure?
  Field type: Drop down with 2 options – AI, IoT

- Is high-performance computing (HPC) or GPU acceleration needed?
- **Field type:** Drop down Yes/No

- Number of GPUs that the model will run on?
- **Field type:** text field

Inference Specific Questions (When the Task Type is chosen as Inference, Question number: 3)

- What is the typical number of input size?
  Field Type: Text field

- What is the expected number of output size to be generated per request?
  Field Type: Text field

- What is the primary deployment scenario (e.g., Single or batch)?
  Field type: Similar to Scenario drop down that is already present in the UI in simulate performance tab

- What batch size do you plan to use for inference?
  Field type: text field

- What is your target throughput (e.g., tokens per second, inferences per second)?
  Field type: text field

- What is the maximum acceptable latency for a single inference request (e.g., in milliseconds)?
  Field type: text field

- How many concurrent users will be accessing the model simultaneously?
  Field type: text field

- How many requests per second do you expect to handle at peak load?
- **Field type:** text field

- What is your budget or target cost for inference (e.g., per 1000 inference)?
- **Field type:** text field

- How quickly should the response begin? (Target Time – to – first token (ms))
- **Field type:** text field

Training Specific Questions (When the Task Type is chosen as Training, Question number: 3)

- Are you performing full training from scratch or single pass training.
- **Field type:** Similar to Is Full Training drop down that is already present in the UI in simulate performance tab

- If fine-tuning, which method will be used (e.g., LoRA, QLoRA, Full Fine-Tuning)?
  Field type: Drop Down with 3 options LoRa, QLoRA, Full Training

- What is the size of your training dataset (e.g., number of samples)?
- **Field type:** text field

- What is the input size or sequence length for the training data?
- **Field type:** text field

- What is the expected output size or sequence length?

- **Field type:** text field

- What batch size will be used for training?
- **Field type:** text field

- Which optimizer will you be using (e.g., Adam, SGD)?
- **Field type:** Drop down with AdamW and SGD

- What is the learning rate for the training process?
- **Field type:** text field

- How many epochs do you plan to train for?
- **Field type:** text field

- What is the target time to complete the entire training job?
- **Field type:** text field

- What is your target training throughput (e.g., samples per second, steps per second)?
- **Field type:** text field

- How many concurrent training jobs do you plan to run?
- **Field type:** text field

- What is your budget or target cost for the entire training process?
- **Field type:** text field