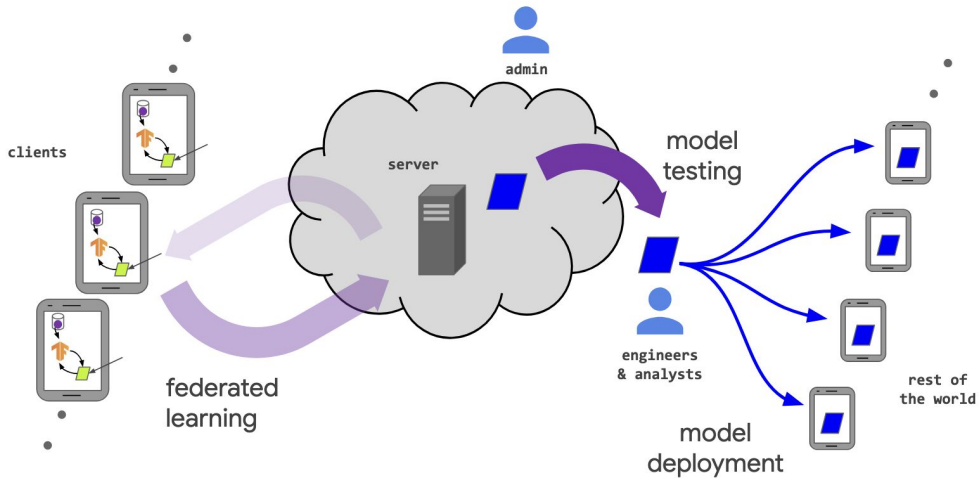
Communication-Efficient Federated Learning Using Embeddings

Communication-Efficient Federated Learning Using Embeddings

- Communication cost is often the bottleneck in Federated Learning

- Reducing the cost of communication increases the feasibility of training larger models such as LLMs



Communication-Efficient Federated Learning Using Embeddings

- Motivations:
  - LoRA is empirically effective
  - It is possible to build hypernetworks that generate performant LoRA
  - Embeddings can be made parameter-efficient

- Idea: Apply FL to train embeddings as input to a hypernetwork that generates LoRA
  - Index of layer: 1~12/24/32/48 in transformer-based architectures
  - Type of weights to generate: W_q, W_k, W_v, W_o1, W_o2 in transformers