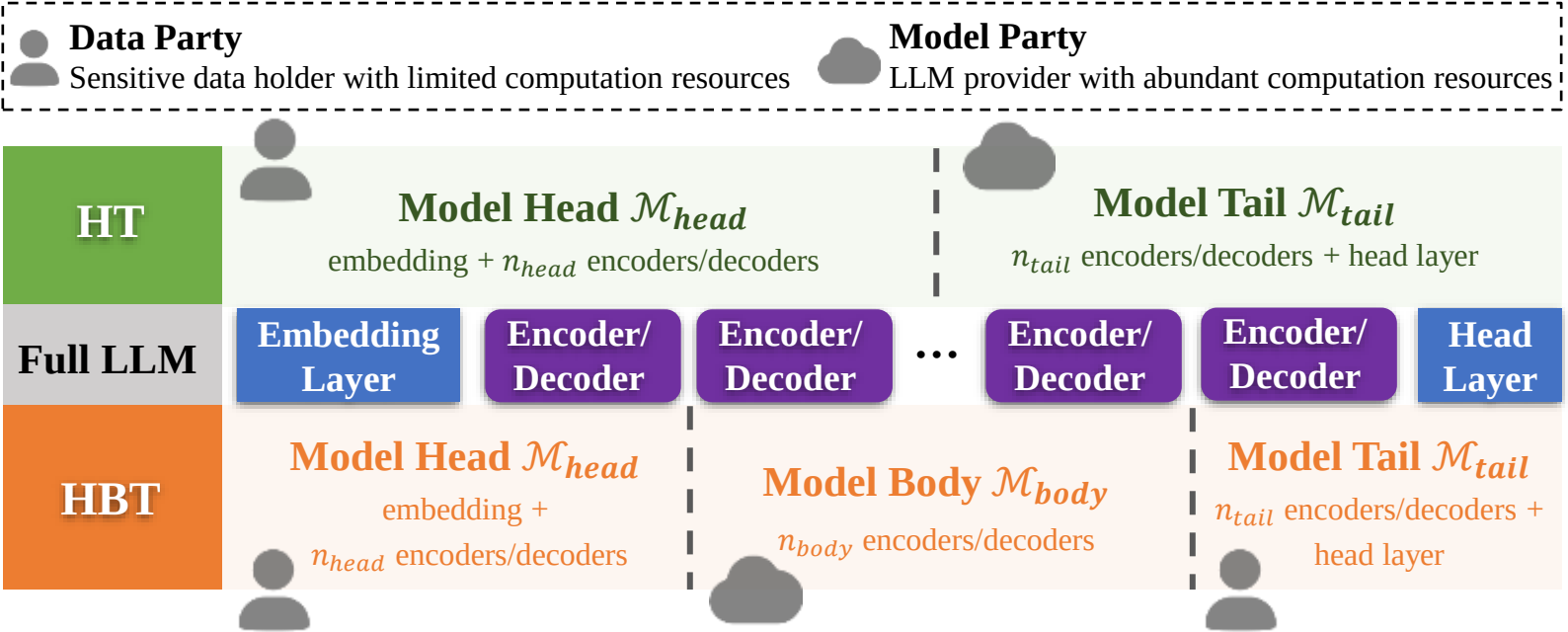
Data Party
Sensitive data holder with limited computation resources
Model Party
LLM provider with abundant computation resources
HT
Model Head $\mathcal{M}_{head}$
embedding + n_{head} encoders/decoders
Model Tail $\mathcal{M}_{tail}$
n_{tail} encoders/decoders + head layer
Full LLM
Embedding Layer
Encoder/ Decoder
Encoder/ Decoder
...
Encoder/ Decoder
Encoder/ Decoder
Head Layer
HBT
Model Head $\mathcal{M}_{head}$
embedding +
n_{head} encoders/decoders
Model Body $\mathcal{M}_{body}$
n_{body} encoders/decoders
Model Tail $\mathcal{M}_{tail}$
n_{tail} encoders/decoders +
head layer