### **1. What should you do if the two models have different tokenizers?**

If the two models have different tokenizers, you should:

1] Check for tokenizer compatibility: Some models may use similar tokenization schemes, allowing you to share a tokenizer.

[2] Use the larger model's tokenizer: The expert model typically has a richer vocabulary, so using its tokenizer ensures proper tokenization.

[3] Manually align tokenization: If tokenizers are significantly different, you may need to tokenize separately for each model and handle mapping between tokenized outputs.

4] Perform detokenization adjustments: Ensure that decoded outputs remain meaningful by aligning tokenization strategies.

### **2. Do you think contrastive decoding is used in practice?**

Contrastive decoding is an interesting approach but is not widely used in production for several reasons:

1] Higher computational cost: Running two models simultaneously increases inference time and resource consumption.

2] Limited practical advantage: While contrastive decoding helps filter low-quality outputs, methods like logit biasing, nucleus sampling, and reinforcement learning are often more efficient and controllable.

3] Fine-tuning is preferred: Instead of using two models, many organizations fine-tune a single large model to improve output quality.