Prompting & Mixture-of-Experts (MoE) — Detailed Presentation

This document provides a detailed explanation of prompting techniques and the Mixture■of■Experts architecture.

1. Fundamental Prompting Techniques

Clear Instructions: The model understands best when the instructions are simple, direct, and unambiguous. Clearly stating what you want reduces confusion and improves accuracy.

Provide Context: Large Language Models generate better responses when they know the background of the task. Context helps the model follow the correct direction.

Specify Format: Telling the model what format you want—such as a list, paragraph, or table—makes the response structured and useful.

Give Examples: Demonstrating the expected output helps the model understand patterns. This is called few■shot prompting.

Step■by■Step Prompts: Asking the model to think step■by■step improves reasoning, avoids errors, and makes the output more reliable.

2. Advanced Prompting Strategies

Chain■of■Thought Prompting: Encourages the model to explain its reasoning. Useful for math, logic, and problem■solving tasks.

Self■Consistency: Instead of generating one answer, the model generates multiple reasoning paths and selects the most consistent answer. This increases accuracy.

Role■Based Prompting: Assigning a role (e.g., 'You are a senior developer') helps the model generate expert■level responses.

Few■Shot and Zero■Shot Prompting: Few■shot means giving examples; zero■shot means the model performs the task without examples. Both techniques improve flexibility.

Error Analysis & Refinement: After receiving an answer, you can ask the model to recheck, refine, or improve the response. This loop produces higher■quality results.

3. Mixture■of■Experts (MoE)

What is MoE? Mixture■of■Experts is a neural network architecture where multiple specialized models (called experts) are trained for different tasks or types of data.

How MoE Works: A routing network decides which expert should handle each input. Only the selected experts activate, making the system efficient.

Why MoE is Powerful: It allows models to increase capacity without increasing computation. Instead of one giant model, many smaller experts contribute selectively.

Benefits of MoE:
- Higher model capacity
- Lower compute cost
- Faster inference
- Better specialization and accuracy

Real■World Applications: Search engines, translation systems, AI assistants, and large■scale reasoning models all use MoE to improve quality and efficiency.