大型语言模型（LLM）技术详解

一、LLM 的发展历程

1. 早期统计语言模型（n-gram）
2. 神经网络语言模型（Word2Vec, GloVe）
3. Transformer 架构革命（2017 年）
4. 预训练-微调范式（BERT, GPT）
5. 大模型时代（GPT-3, ChatGPT, GPT-4）

二、Transformer 核心组件

1. 自注意力机制（Self-Attention）
   - 查询（Query）、键（Key）、值（Value）
   - 多头注意力（Multi-Head Attention）
2. 位置编码（Positional Encoding）
   - 绝对位置编码
   - 相对位置编码
3. 前馈神经网络（Feed-Forward Network）
4. 层归一化（Layer Normalization）
5. 残差连接（Residual Connection）

三、模型训练方法

1. 预训练（Pre-training）
   - 自监督学习
   - 掩码语言建模（MLM）
   - 下一句预测（NSP）
2. 微调（Fine-tuning）
   - 指令微调（Instruction Tuning）
   - 人类反馈强化学习（RLHF）
3. 提示工程（Prompt Engineering）
   - 零样本学习（Zero-shot）
   - 少样本学习（Few-shot）
   - 思维链（Chain-of-Thought）

四、主要技术挑战

1. 计算资源需求大
2. 训练数据质量要求高
3. 模型幻觉问题
4. 推理速度慢
5. 部署成本高

五、未来发展方向

1. 多模态融合
2. 更高效的架构
3. 小样本学习能力提升

4. 推理能力增强
5. 个性化与定制化