附录C 术语表

大模型领域常用术语中英文对照及解释

C.1 核心概念

中文	英文	缩写	解释
大语言模型	Large Language Model	LLM	参数量在数十亿以上的语言模型
自然语言处理	Natural Language Processing	NLP	使计算机理解和生成人类语言的技术
生成式人工智能	Generative AI	GenAI	能够生成文本、图像等内容的AI
通用人工智能	Artificial General Intelligence	AGI	具有人类级别智能的AI系统
基座模型	Foundation Model	-	在大规模数据上预训练的通用模型
对齐	Alignment	-	使模型行为符合人类价值观和意图

C.2 模型架构

中文	英文	缩写	解释
Transformer	Transformer	-	基于注意力机制的神经网络架构
注意力机制	Attention Mechanism	-	让模型关注输入的不同部分
自注意力	Self-Attention	-	序列内部元素之间的注意力
多头注意力	Multi-Head Attention	MHA	并行的多个注意力头
交叉注意力	Cross-Attention	-	两个序列之间的注意力
前馈网络	Feed-Forward Network	FFN	全连接神经网络层
残差连接	Residual Connection	-	跳跃连接，帮助梯度流动
层归一化	Layer Normalization	LayerNorm	在特征维度归一化
位置编码	Positional Encoding	PE	为序列添加位置信息
旋转位置嵌入	Rotary Position Embedding	RoPE	LLaMA使用的相对位置编码

C.3 训练相关

中文	英文	缩写	解释
预训练	Pre-training	-	在大规模数据上的初始训练
微调	Fine-tuning	FT	在特定任务上继续训练
全参数微调	Full Fine-Tuning	FFT	更新所有模型参数
参数高效微调	Parameter-Efficient Fine-Tuning	PEFT	只更新少量参数

中文	英文	缩写	解释
低秩适配	Low-Rank Adaptation	LoRA	用低秩矩阵近似权重更新
量化LoRA	Quantized LoRA	QLoRA	结合量化的LoRA
指令微调	Instruction Tuning	-	教模型遵循指令
人类反馈强化学习	Reinforcement Learning from Human Feedback	RLHF	用人类偏好优化模型
直接偏好优化	Direct Preference Optimization	DPO	简化的对齐方法
监督微调	Supervised Fine-Tuning	SFT	用标注数据微调
奖励模型	Reward Model	RM	预测人类偏好的模型
近端策略优化	Proximal Policy Optimization	PPO	强化学习算法

C.4 训练技术

中文	英文	缩写	解释
反向传播	Backpropagation	BP	计算梯度的算法
梯度下降	Gradient Descent	GD	优化算法
随机梯度下降	Stochastic Gradient Descent	SGD	使用小批量的梯度下降
自适应矩估计	Adaptive Moment Estimation	Adam	自适应学习率优化器
AdamW	Adam with Weight Decay	AdamW	改进权重衰减的Adam
学习率	Learning Rate	LR	参数更新的步长
学习率预热	Learning Rate Warmup	-	训练初期逐渐增加学习率
余弦退火	Cosine Annealing	-	余弦曲线降低学习率
梯度裁剪	Gradient Clipping	-	限制梯度范数，防止爆炸
梯度累积	Gradient Accumulation	-	累积多步梯度后更新
混合精度训练	Mixed Precision Training	-	使用FP16和FP32混合训练
批量大小	Batch Size	-	一次训练的样本数
轮次	Epoch	-	遍历全部训练数据一次

C.5 分布式训练

中文	英文	缩写	解释
数据并行	Data Parallelism	DP	复制模型，分割数据
模型并行	Model Parallelism	MP	分割模型到多个设备
张量并行	Tensor Parallelism	TP	分割单个张量
流水线并行	Pipeline Parallelism	PP	按层分割模型

中文	英文	缩写	解释
零冗余优化器	Zero Redundancy Optimizer	ZeRO	DeepSpeed的显存优化
分布式数据并行	Distributed Data Parallel	DDP	PyTorch的分布式训练
全归约	All-Reduce	-	聚合所有设备的梯度

C.6 推理优化

中文	英文	缩写	解释
量化	Quantization	-	降低数值精度
后训练量化	Post-Training Quantization	PTQ	训练后量化
量化感知训练	Quantization-Aware Training	QAT	训练时模拟量化
知识蒸馏	Knowledge Distillation	KD	用大模型训练小模型
剪枝	Pruning	-	移除不重要的参数
KV缓存	KV Cache	-	缓存注意力的Key和Value
Flash Attention	Flash Attention	-	内存高效的注意力实现
推测解码	Speculative Decoding	-	小模型生成，大模型验证
批量推理	Batch Inference	-	同时处理多个请求
连续批处理	Continuous Batching	-	动态组合批次

C.7 提示工程

中文	英文	缩写	解释
提示	Prompt	-	输入给模型的文本
提示工程	Prompt Engineering	-	设计有效提示的技术
零样本学习	Zero-Shot Learning	-	无示例直接推理
少样本学习	Few-Shot Learning	-	提供少量示例
上下文学习	In-Context Learning	ICL	从上下文中的示例学习
思维链	Chain-of-Thought	CoT	展示推理过程
思维树	Tree of Thoughts	ToT	树状探索推理路径
自我一致性	Self-Consistency	-	多次采样选择一致答案
角色扮演	Role-Playing	-	给模型分配角色
提示注入	Prompt Injection	-	恶意操纵模型行为

C.8 RAG相关

中文	英文	缩写	解释
检索增强生成	Retrieval-Augmented Generation	RAG	结合检索和生成
向量数据库	Vector Database	-	存储和检索向量的数据库
嵌入	Embedding	-	文本的向量表示
语义搜索	Semantic Search	-	基于语义的检索
余弦相似度	Cosine Similarity	-	向量相似度度量
分块	Chunking	-	将文档分割成小片段
重排序	Reranking	-	对检索结果重新排序
混合检索	Hybrid Search	-	结合关键词和向量检索
稀疏检索	Sparse Retrieval	-	基于关键词的检索
稠密检索	Dense Retrieval	-	基于向量的检索

C.9 Agent相关

中文	英文	缩写	解释
智能体	Agent	-	能感知环境并行动的系统
工具调用	Tool Calling	-	模型调用外部工具
函数调用	Function Calling	-	模型调用预定义函数
ReAct	Reasoning and Acting	ReAct	推理和行动结合的框架
记忆	Memory	-	存储历史信息
规划	Planning	-	制定行动计划
多Agent系统	Multi-Agent System	MAS	多个Agent协作

C.10 评估指标

中文	英文	缩写	解释
困惑度	Perplexity	PPL	语言模型的评估指标
BLEU	Bilingual Evaluation Understudy	BLEU	机器翻译评估
ROUGE	Recall-Oriented Understudy for Gisting Evaluation	ROUGE	摘要评估
F1分数	F1 Score	F1	精确率和召回率的调和平均
准确率	Accuracy	ACC	正确预测的比例
精确率	Precision	-	预测为正中真正为正的比例
召回率	Recall	-	真正为正中被预测为正的比例
人类评估	Human Evaluation	-	人工评判模型输出

C.11 常见任务

中文	英文	缩写	解释
语言建模	Language Modeling	LM	预测下一个词
因果语言建模	Causal Language Modeling	CLM	单向语言建模
遮罩语言建模	Masked Language Modeling	MLM	BERT的预训练任务
序列到序列	Sequence-to-Sequence	Seq2Seq	将一个序列转换为另一个
文本分类	Text Classification	-	给文本分配类别
命名实体识别	Named Entity Recognition	NER	识别文本中的实体
问答	Question Answering	QA	回答问题
文本摘要	Text Summarization	-	生成文本摘要
机器翻译	Machine Translation	MT	将文本翻译成另一种语言
情感分析	Sentiment Analysis	-	判断文本情感倾向
文本生成	Text Generation	-	生成连贯文本
对话系统	Dialogue System	-	进行多轮对话

C.12 数据处理

中文	英文	缩写	解释
分词	Tokenization	-	将文本分割成token
词元	Token	-	文本的基本单位
词表	Vocabulary	Vocab	所有token的集合
字节对编码	Byte Pair Encoding	BPE	子词分词算法
WordPiece	WordPiece	-	BERT使用的分词
SentencePiece	SentencePiece	-	语言无关的分词
子词	Subword	-	词的一部分
特殊标记	Special Token	-	如[CLS]、[SEP]等
填充	Padding	-	将序列补齐到相同长度
截断	Truncation	-	裁剪过长的序列
数据增强	Data Augmentation	-	生成更多训练数据

C.13 模型行为

中文	英文	缩写	解释
幻觉	Hallucination	-	模型生成虚假信息

中文	英文	缩写	解释
涌现能力	Emergent Abilities	-	大模型突然出现的能力
灾难性遗忘	Catastrophic Forgetting	-	微调时遗忘预训练知识
过拟合	Overfitting	-	在训练集上过度拟合
欠拟合	Underfitting	-	模型能力不足
泛化	Generalization	-	在新数据上的表现
偏见	Bias	-	模型的偏向性
公平性	Fairness	-	对不同群体的公平对待
可解释性	Interpretability	-	理解模型决策的能力
鲁棒性	Robustness	-	对扰动的抵抗能力

C.14 硬件相关

中文	英文	缩写	解释
图形处理器	Graphics Processing Unit	GPU	并行计算硬件
张量处理器	Tensor Processing Unit	TPU	Google的AI芯片
显存	Video Memory	VRAM	GPU的内存
浮点运算	Floating Point Operations	FLOPs	计算量单位
FP32	32-bit Floating Point	FP32	单精度浮点
FP16	16-bit Floating Point	FP16	半精度浮点
BF16	Brain Floating Point 16	BF16	Google的16位格式
INT8	8-bit Integer	INT8	8位整数
高带宽内存	High Bandwidth Memory	HBM	高速GPU内存
静态随机存储器	Static Random-Access Memory	SRAM	片上高速缓存

C.15 框架和工具

中文	英文	缩写	解释
PyTorch	PyTorch	-	深度学习框架
TensorFlow	TensorFlow	TF	Google的深度学习框架
HuggingFace	Hugging Face	HF	开源模型平台
LangChain	LangChain	-	LLM应用开发框架
LlamaIndex	LlamaIndex	-	RAG框架
vLLM	vLLM	-	高效推理引擎
DeepSpeed	DeepSpeed	-	大规模训练框架

中文	英文	缩写	解释
Megatron	Megatron	-	NVIDIA的大模型训练

C.16 商业模型

中文	英文	缩写	解释
GPT	Generative Pre-trained Transformer	GPT	OpenAI的模型系列
Claude	Claude	-	Anthropic的模型
Gemini	Gemini	-	Google的多模态模型
LLaMA	Large Language Model Meta AI	LLaMA	Meta的开源模型
ChatGLM	ChatGLM	-	智谱AI的对话模型
文心一言	ERNIE Bot	-	百度的对话模型
通义千问	Qwen	-	阿里巴巴的模型

C.17 规模法则

中文	英文	解释
规模法则	Scaling Laws	模型性能随规模的变化规律
Chinchilla规模法则	Chinchilla Scaling Laws	参数和数据应等比例增长
计算最优	Compute-Optimal	给定计算预算的最优配置
迁移学习	Transfer Learning	将预训练知识应用到新任务

C.18 安全和伦理

中文	英文	解释
AI安全	AI Safety	确保AI系统安全可控
AI对齐	AI Alignment	使AI目标与人类一致
红队测试	Red Teaming	主动寻找系统漏洞
有害内容	Harmful Content	不当或危险的输出
隐私保护	Privacy Protection	保护用户数据隐私
负责任的AI	Responsible AI	考虑伦理的AI开发

C.19 使用说明

如何使用本术语表：

1. **快速查找**：使用Ctrl+F搜索中文或英文术语
2. **系统学习**：按章节顺序学习相关术语

3. **面试准备**：重点记忆核心概念和常用缩写
4. **阅读论文**：遇到陌生术语时查阅

记忆技巧：

- 理解概念本质，而非死记硬背
- 关联记忆：相关术语一起记
- 实践中学习：在代码和项目中使用
- 制作卡片：Anki等间隔重复工具

常见缩写速记：

- **LLM** = Large Language Model（大语言模型）
- **RLHF** = RL from Human Feedback（人类反馈强化学习）
- **LoRA** = Low-Rank Adaptation（低秩适配）
- **RAG** = Retrieval-Augmented Generation（检索增强生成）
- **CoT** = Chain-of-Thought（思维链）

C.20 本附录小结

本术语表汇总了大模型领域的核心术语：

✅ **完整覆盖**：从基础概念到高级技术 ✅ **中英对照**：便于阅读英文文献 ✅ **清晰解释**：每个术语都有简明说明
✅ **分类整理**：按主题组织，易于查找

更新说明：

- 本术语表会持续更新
- 关注领域最新发展
- 添加新出现的术语

建议：

- 打印或保存为速查手册
- 面试前快速浏览核心术语
- 阅读论文时对照使用

至此，《大模型面试宝典》全部内容编写完成！

祝你面试顺利，斩获理想Offer！🎉