大语言模型入门

Getting Started with Large Language Models

A Beginner's Guide and Live Demo

Presented by Yi Li

Research Triangle AI

2024-09-21

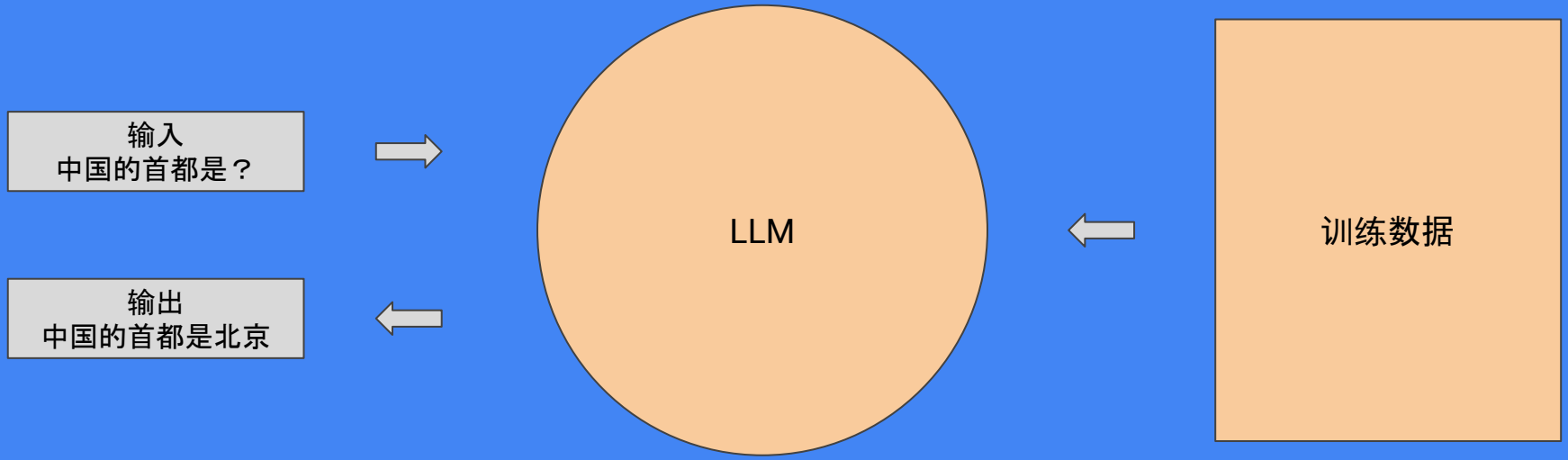
议程概述 Agenda

- 引言与自我介绍 Introduction
- 理解大语言模型 Understanding LLM
- 实际应用 Practical Applications
- 演示：在Google Colab上部署Llama 3.1

  Demo: Deploying Llama 3.1 on Google Colab
- 伦理考量与LLM的未来 Ethical Considerations and Future of LLMs
- 总结与问答 Summary and Q&A

啥是大语言模型 What is Large Language Model (LLM)

- 大语言模型（Large Language Model，LLM）
- 通过大量文本数据训练的一种人工智能模型
- 专门用于理解、生成和处理自然语言



如何理解大语言模型

○ 大规模

大量文本数据训练

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Common Crawl 网络爬虫公开数据集
WebText2 Reddit论坛网页文本
Books1, Books2 互联网书籍语料库
Wikipedia 维基百科知识库

大量参数的AI模型

Model Name	n_{params}	n_{layers}
GPT-3 Small	125M	12
GPT-3 Medium	350M	24
GPT-3 Large	760M	24
GPT-3 XL	1.3B	24
GPT-3 2.7B	2.7B	32
GPT-3 6.7B	6.7B	32
GPT-3 13B	13.0B	40
GPT-3 175B or "GPT-3"	175.0B	96

GPT4 (not official)

- 1.8T parameters
- 120 layers
- 13T tokens

如何理解大语言模型

- 大规模
  - 大量参数的AI模型
  - 大量文本数据训练
- 语言
  - 基于自然语言处理 (Nature Language Processing / NLP)
  - 能够理解、生成和处理人类语言

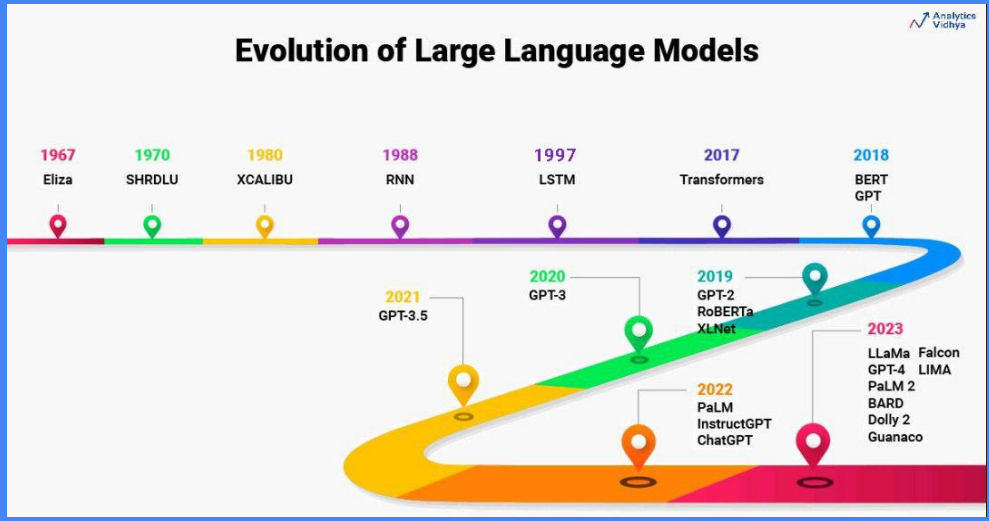
如何理解大语言模型

- 大规模
  - 大量参数的AI模型
  - 大量文本数据训练
- 语言
  - 基于自然语言处理 (Nature Language Processing / NLP)
  - 能够理解、生成和处理人类语言
- **模型**
  - 模拟语言规则和模式的数学框架
  - 训练的模型可以预测语言中的下一个词或者句子

LLM的简史与演变

- 早期发展
  - 早期的NLP主要依赖于规则系统和基本的统计模型
  - 神经网络的引入极大地改变了NLP领域

- LLM的崛起:
  - Attention is all you need (2017)
  - 2022年11月30日 ChatGPT发布



常见的大语言模型

OpenAI ChatGPT / o1

Google Gemini

Meta Llama

Anthropic Claude

Mistral Mistral/Mixtral

百度 文心一言

阿里云 通义千问

华为 盘古

腾讯 混元

科大讯飞 星火

零一万物 Yi

大模型是如何炼成的 How LLM is Trained

Step 1: Pre-training (unsupervised)

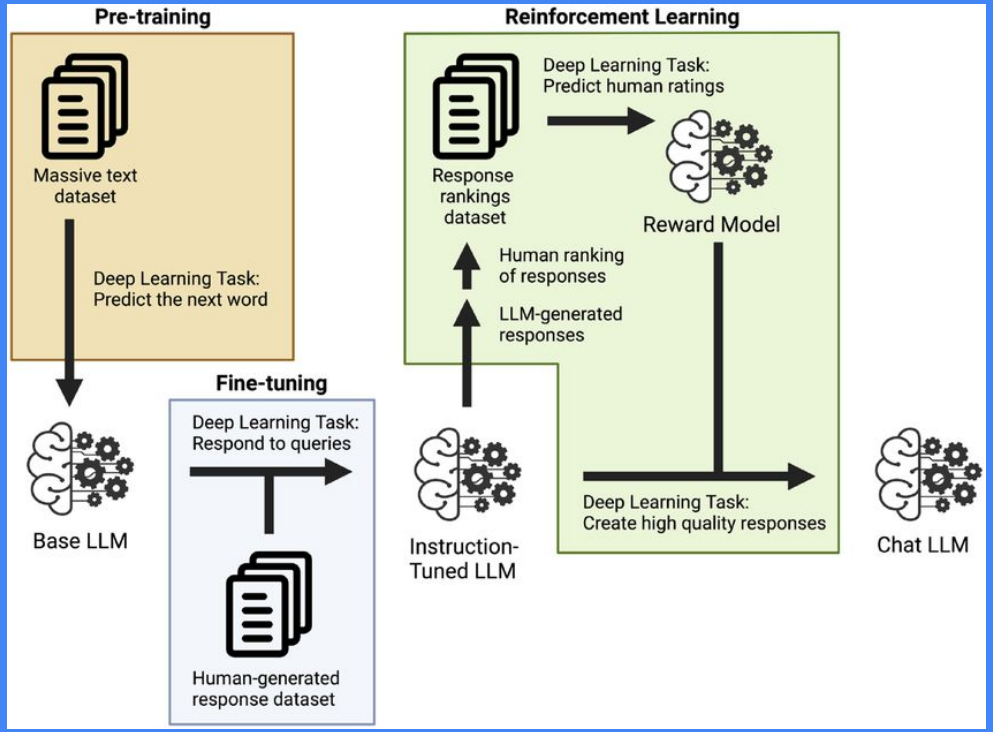
Base LLM 文字接龙机

Step 2: Supervised Fine-tuning (SFT)

Instruct LLM 初级客服

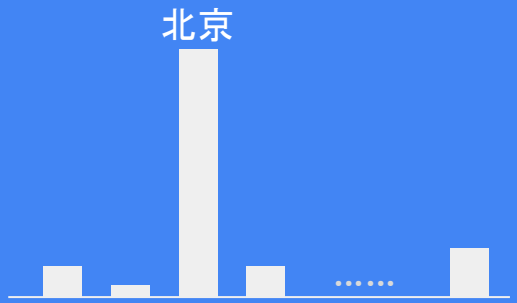
Step 3: Reinforcement Learning from Human Feedback (RLHF)

Chat LLM 专业顾问



大语言模型如何工作

中国的首都是 → LLM →



Transformer

Tokenization → Embedding → Attention → Feed Forward → Output

文字转成 token

中国的首都是

理解 语义/位置

$$\begin{bmatrix} 1.2 \\ -3.5 \\ \vdots \\ -1.9 \\ 7.2 \end{bmatrix}$$

考虑 上(下)文

综合思考

反复思考

输出

RAG / PAL / ReAct

借助外部数据/工具

人要是行

要是不行

下面这段话如何发音，如何理解
人要是行，干一行，行一行。一行行，行行行；
要是不行，干一行，不行一行。一行不行，行行不行。

这段话通过重复"行"和"干一行"的不同组合，巧妙地表达了两种人的工作能力和态度。发音上，由于"行"字在不同语境中有不同的读音和含义，需要注意正确的读音。

发音：

- 第一个"行"是"有能力、能干"的意思，读作 **xíng**。
- 第二个"行"是指"行业、职业"，读作 **háng**。

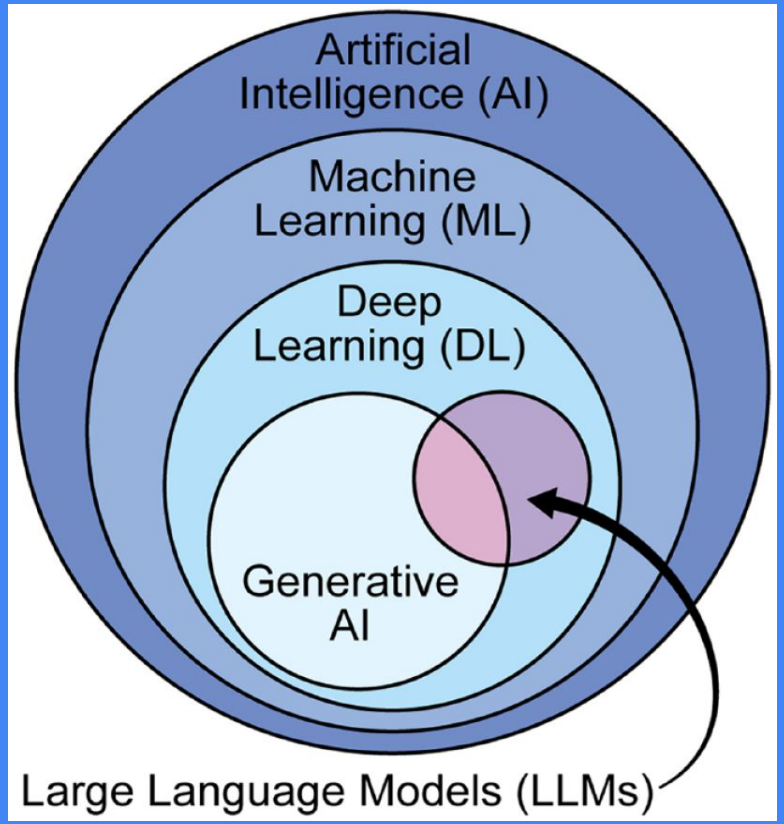
下面分解每句话的读音和含义：

1. **人要是行，干一行，行一行。**
   读作：**rén yào shì xíng, gàn yī háng, xíng yī háng.**
   含义：如果一个人有能力，做哪一行工作，就能胜任哪一行。
2. **一行行，行行行。**
   读作：**yī háng háng, háng háng xíng.**
   含义：每一行工作都能做好，行行都能胜任。
3. **要是不行，干一行，不行一行。**
   读作：**yào shì bù xíng, gàn yī háng, bù xíng yī háng.**
   含义：如果一个人没有能力，做哪一行工作，都做不好。
4. **一行不行，行行不行。**
   读作：**yī háng bù xíng, háng háng bù xíng.**
   含义：做一行不行，行行都做不好。

理解：

这段话生动地说明了如果一个人有能力，他在任何职业中都会表现出色；相反，如果没有能力，无论做什么都无法成功。

大语言模型和AI / ChatGPT的关系



- LLM 是基础技术
- ChatGPT是LLM技术的一种应用实现

实际应用 Practical Applications

- 内容生成 Content Generation
- 知识库问答 Knowledge Base Answering
- 文本分类 Text Classification
- 情感分析 Sentiment Analysis
- 搜索 Search
- 计算机安全 Cybersecurity

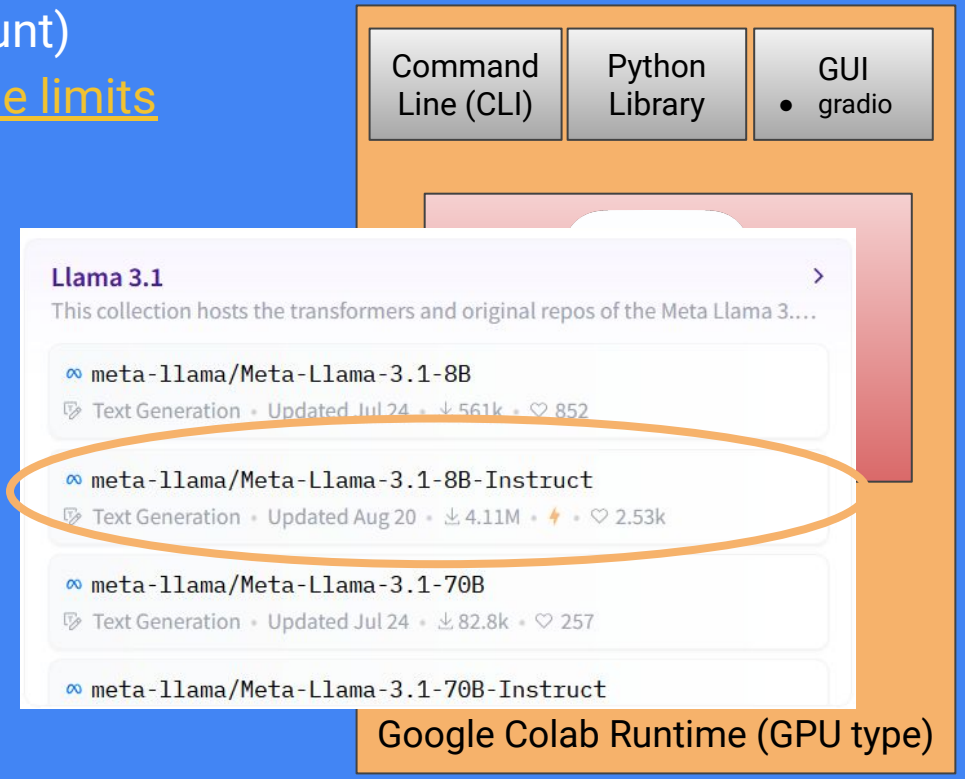

LLM APPLICATIONS IN TOP INDUSTRIES
Marketing & Advertising
Customized Consumer Outreach
Virtual Assistants
Content Generation
Ad Campaign Precision
Efficacy Assessment
Multifaceted Text Constructs
Retail & ECommerce
Trend Analysis
Intelligent Customer Interaction
Guided Buying Processes
Fraudulent Activity Recognition
Education Industry
Curated Learning
Intelligent Query Resolution
Automated Generation
Supplemental Applications
Healthcare
New Wave In Diagnostics
Health Monitoring
Turbocharging Drug Development
Tailoring Treatments
Practicing Medicine Virtually
Finance
Unlocking Syntax Of Finance
Dynamic Risk Profiling
Tactical Market Entries
Fraud Landscape Cartography
Mastering Regulatory Fluidity
Legal Industry
Advanced Legal Research
Automated Drafting
Legal Analytics
Litigation Support
Compliance Monitoring
Media
Tailor-Made Content
Smarter Content
Next-Level Engagement
Data-Driven Advertising
Defense & Strategic Services
Data Augmentation
Semantic Text Mining
Multilingual NLP
Dialog Systems
Anomaly Detection
Human Resources
Candidate Procurement
Accelerated Integration
Performance Metrics Analysis
Customized Skill Augmentation
Employee Sentiment Assessment
Fashion
Customized Apparel Guidance
Customized Apparel Guidance
Design Processes Optimized
Virtual Apparel Fitting
Customer Relationship Management

为什么要部署自己的大语言模型

- 数据隐私和安全 Data Privacy and Security
- 无需依赖互联网连接 Independence from Internet Connectivity
- 降低运营成本 Reduced Operational Costs
- 更高的模型灵活性与控制 Greater Flexibility and Control

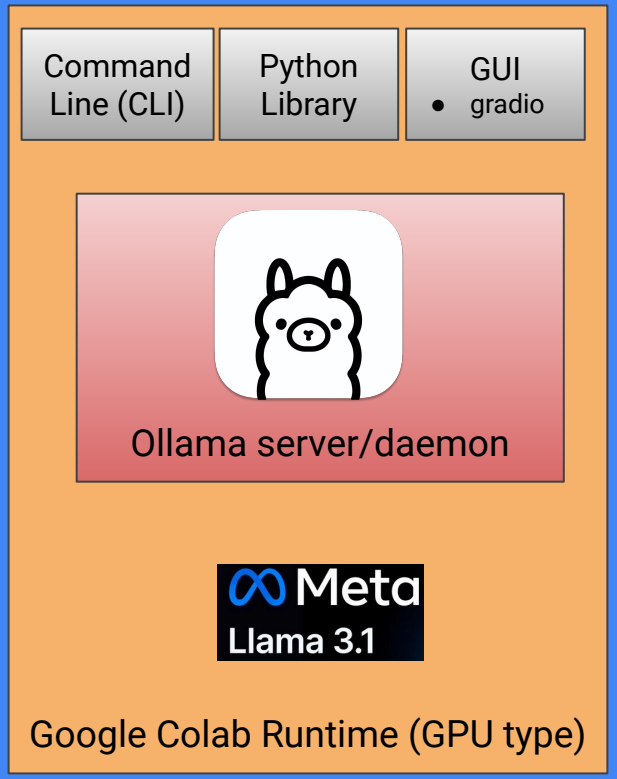
部署需求

- Google Colab (google account)
  - 可用 free tier, with usage limits
  - 基于 Jupyter Notebook
- LLaMA 3.1
  - 开源大模型 (by Meta)
  - 2024年4月发布
  - 多种模型
  - 95%的训练数据是英文
- Ollama
  - LLM 管理平台
  - 开源



部署步骤

- 环境设置
  - 打开Google Colab
  - 选择有GPU的runtime
- 安装Ollama
- 下载并运行Llama 3.1模型
- 与模型交互



演示：在 Google Colab上部署 Llama 3.1

Demo: Deploying Llama 3.1 on Google Colab

Colab Notebook 链接

下一步

- 硬件条件允许可以尝试 本地部署
- 与ngrok结合 随时访问自己的大模型
- Fine tune 拥有了解自己的大模型
- API 调用，与其它应用结合

LLM的未来趋势

- 模型规模与计算能力的提升
- 多模态模型的发展
- 逻辑推理 openai o1 (self replay RL)
- 自主模型 (autonomous models)
- 更高效的微调与个性化及深入与人类的协作
- 减少模型偏见与提高模型透明度
- 伦理与法规的发展

伦理考量

- 数据隐私 Ethical and Privacy Concerns
- 误导信息与虚假内容 Misinformation and False Content
- 偏见与歧视 Biases and Fairness Problems
- 对工作岗位的影响 Jobs/Society Impact
- 责任归属与道德困境 Accountability and Ethical Dilemmas
- 环境问题 Energy Consumption and Carbon Footprints

总结

- LLM 的基础知识
- Llama 3.1 在 Google Colab 上的部署过程
- LLM 的伦理问题和未来趋势

深入了解学习

Large language models, explained with a minimum of math and jargon

The Practical Guides for Large Language Models

Hugging Face - NLP Course

DeepLearning.ai

参考资料

Introduction to Large Language Models

LLM Training - A Simple 3-Step Guide You Won't Find Anywhere Else!

Run Llama 3.1 8B with Ollama on Free Google Colab

AI大语言模型科普课

图片来源

LLM与AI关系 / LLM如何炼成的
https://www.researchgate.net/publication/378394229_Large_language_models_a_primer_and_gastroenterology_applications

LLM 各行应用 https://markovate.com/blog/llm-applications-and-use-cases/

数据来源

GPT3 https://arxiv.org/pdf/2005.14165v4

GPT4 https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

Q & A