0001年1月1日 1 分钟阅读

[1hr Talk] Intro to Large Language Models

https://www.youtube.com/watch?v=zjkBMFhNj_g 该视频讲稿系统地介绍了大型语言模型（LLMs）。首先，它从基本概念入手，解释了LLM的构成（参数文件和运行代码），并以Llama 2为例进行了说明，强调了其开放权重的特点。接着，深入探讨了LLM的训练过程，分为预训练（海量互联网文本、高昂算力成本）和微调（高质量人工标注数据，塑造助手模型）两个阶段，并提及了可选的**通过人类反馈强化学习（RLHF）**进行性能提升。

随后，讲稿展示了LLM的强大能力，例如工具使用（浏览器、计算器、代码执行、图像生成），以及多模态特性（处理文本、图像、音频等）。展望未来，它探讨了LLM的发展方向，包括模拟人类的系统二思维、自我改进的可能性，以及定制化的应用前景，并提出了LLM可能成为新兴操作系统内核的类比。

最后，讲稿也强调了LLM带来的安全挑战，通过越狱攻击、提示注入攻击和数据投毒/后门攻击等实例，揭示了LLM安全领域的攻防博弈。总而言之，该视频旨在为听众提供一个关于LLM的全面入门，既展现了其潜力，也指出了其面临的挑战。

分享这篇文章

目录

文章信息

字数

阅读时间

发布时间