[越]奇普·萱（Chip Huyen） 著
宝玉 译
何文斯 李瀚 审校

AI工程
链接资源

【致读者】本书包含的1000条左右的链接资源，由AI辅助整理生成，以电子版PDF形式免费提供。由于网络环境不断变化，个别链接可能访问异常；相关描述文字也系AI生成，可能存在不精准之处。本资料仅供参考，敬请知悉。

2026年1月

目录

前言（12）

1. https://en.wikipedia.org/wiki/Lindy_effect: 该资源介绍了“Lindy 效应”（Lindy effect），即一种推断技术或事物未来寿命的方法，认为某项技术或事物已经存在的时间越长，其未来继续存在的时间也越长。换言之，当前寿命越长的技术，其预期寿命也越长。
2. https://github.com/chiphuyen/aie-book: 该资源是一个 GitHub 代码仓库，提供与人工智能工程相关的补充材料，包括代码示例、练习、重要论文和实用工具。此外，该仓库还涵盖书中未深入讲解的主题，并包含关于本书写作过程的幕后信息和统计数据。
3. https://huyenchip.com/blog/: 该资源是作者的个人博客，内容涉及作者分享的知识和观点，尤其包括与其著作相关的讨论和读者反馈，社区评论丰富，提供了多样的视角和思考。
4. https://linkedin.com/company/oreilly-media: 该资源是 O’Reilly Media 公司的 LinkedIn 官方页面，用于展示公司信息、发布动态及与专业人士进行交流。
5. https://oreil.ly/XG3mv: （https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf）该 URL 指向的资源是 2012 年发表的著名论文《ImageNet Classification with Deep Convolutional Neural Networks》（通常称为 AlexNet 论文）。这篇论文由 Alex Krizhevsky 等人撰写，首次展示了深度卷积神经网络在大规模图像分类任务（ImageNet 挑战赛）上的突破性性能，极大推动了深度学习在计算机视觉领域的发展。
6. https://oreil.ly/ai-engineering: （https://learning.oreilly.com/library/view/ai-engineering/9781098166298/）该资源是一部关于人工智能工程的电子书，托管在 O’Reilly 学习平台上，书籍页面提供了该书的详细内容、勘误信息、示例代码及相关补充资料，方便读者学习和参考。
7. https://oreilly.com: 该资源为 O’Reilly Media 的官方网站，提供技术与商业领域的培训、知识和见解。网站汇集了专家和创新者的内容，包括书籍、文章以

及在线学习平台，用户可通过该平台获取按需直播课程、系统学习路径、交互式编程环境以及来自 O’Reilly 和 200 多家出版商的大量文本和视频资源。

8. https://oreilly.com/about/contact.html: 该资源是 O’Reilly 网站的“联系我们”页面，用户可以通过该页面获取与 O’Reilly 公司取得联系的相关信息。
9. https://www.linkedin.com/in/chiphuyen: 该资源是“Chí Phuyên”的 LinkedIn 个人主页链接，用户可通过该链接访问作者 Chí Phuyên 在 LinkedIn 上的职业档案，以了解其专业背景和联系方式。
10. https://x.com/chipro: 该资源是作者 Chip Huyen 在社交媒体平台 X（原 Twitter）上的个人账号页面，读者可以通过此链接与作者进行交流和反馈。
11. https://x.com/chipro/status/937384141791698944: 该资源是一条发布于 2017 年的推文，内容涉及作者的一个小项目，该项目使用语言模型来评估翻译质量，并得出结论认为需要“更好的语言模型”。
12. https://youtube.com/oreillymedia: 该资源是 O’Reilly Media 在 YouTube 上的官方频道，用户可以通过该频道观看由 O’Reilly Media 发布的各类视频内容。

第 1 章（94）

1. https://arxiv.org/abs/1810.04805: 该资源是论文《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》（作者：Devlin 等，2018 年），介绍了一种基于双向 Transformer 结构的预训练语言模型 BERT。该模型通过掩码语言建模（masked language modeling）任务，利用上下文中缺失词汇的前后信息进行训练，从而显著提升了多种自然语言理解任务的性能。
2. https://arxiv.org/abs/2009.03300: 该资源是由 Hendrycks 等人在 2020 年发布的一篇论文，题为“Massive Multitask Language Understanding (MMLU)”。该论文介绍了 MMLU 基准测试，这是一个用于评估大规模语言模型

多任务理解能力的标准测试集，涵盖多个领域和任务。MMLU 广泛用于衡量基础语言模型在多样化任务上的表现和推理能力，成为评估模型性能和推理效率的重要基准。

3. https://arxiv.org/abs/2108.07258: 该资源是一篇关于“foundation models”（基础模型）的学术论文，详细阐述了基础模型的定义、重要性及其在人工智能领域的广泛应用，强调了这些模型作为构建各种 AI 系统的基础框架的核心作用。
4. https://arxiv.org/abs/2204.07705: 该资源是一篇由 Wang 等人于 2022 年发表的论文，介绍了 Super-NaturalInstructions 基准测试，该基准用于评估基础模型在多种任务上的表现，展示了基础模型能够执行的任务类型。
5. https://arxiv.org/abs/2211.04325: 该资源是一篇发表在 arXiv 上的学术论文，标题大致涉及人工智能模型训练所需的数据资源问题，特别讨论了随着大型 AI 模型（如 ChatGPT、Google Gemini 等）规模的扩大，公开可用的互联网数据可能被快速耗尽，从而对未来 AI 训练数据的可持续性构成挑战。
6. https://arxiv.org/abs/2303.10130: 该资源是 2023 年由 Eloundou 等人发表在 arXiv 上的研究论文，题为“GPTs are GPTs”（GPT 就是 GPT）。该研究系统评估了不同职业和行业中各项任务被人工智能技术替代或辅助的暴露程度，具体定义为 AI 及其驱动的软件能将完成某项任务所需时间减少至少 50%。论文揭示了哪些职业高度暴露于 AI 影响（如翻译、税务申报、网页设计和写作等），以及哪些职业几乎不受影响（如厨师、石匠和运动员），为理解 AI 的应用场景和行业影响提供了重要参考。
7. https://arxiv.org/abs/2304.03442: 该资源是一篇由 Park 等人在 2023 年发表的学术论文，内容涉及利用一组对话机器人模拟社会，从而开展社会动态研究。
8. https://arxiv.org/abs/2305.09857: 该资源是关于 Grammarly 这类写作辅助应用的研究论文，介绍了如何通过微调模型来提升用户写作的流畅性、连贯性和清晰度。
9. https://arxiv.org/abs/2305.14233: 该资源是一篇由 Ding 等人于 2023 年发表的论文，题为“UltraChat”，主要讨论了在构建生成式 AI 产品过程中，初期快速取得进展（“从 0 到 60”）相对容易，但后续提升至更高质量（“从 60 到

100”）则极具挑战性。论文分享了实现高质量用户体验所需的时间和精力，特别强调了产品细节打磨和应对生成内容中“幻觉”问题的复杂性和困难。

10. https://en.wikipedia.org/wiki/Memex: 该资源是维基百科上关于“Memex”的条目页面，介绍了一种假想的、早期的信息存储和检索系统概念，通常被视为现代超文本和个人信息管理系统的先驱。
11. https://en.wikipedia.org/wiki/The_Adventure_of_the_Dancing_Men: 该资源是维基百科上一篇关于《跳舞的小人探案》（The Adventure of the Dancing Men）的条目页面。《跳舞的小人探案》是阿瑟·柯南·道尔创作的福尔摩斯探案故事之一，讲述福尔摩斯通过分析一组神秘的跳舞小人符号，利用英语中字母出现频率的统计特征成功破解密码的过程。
12. https://github.com/GreyDGL/PentestGPT: 该资源是一个名为“PentestGPT”的项目，托管在 GitHub 上，可能与创建测试相关，尤其是渗透测试（Pentest）或安全测试相关的工具或框架。
13. https://github.com/Nutlope/aicommits: 该资源是一个名为“AI Commits”的项目，位于 GitHub 上，主要功能是生成提交（commit）信息。
14. https://github.com/Sinaptik-AI/pandas-ai: 该资源是一个名为 PandasAI 的项目，托管在 GitHub 上，主要用于将英文自然语言转换为代码，特别是与数据处理和分析相关的代码（可能基于 pandas 库）。
15. https://github.com/abi/screenshot-to-code: 该资源是一个开源项目“screenshot-to-code”，托管在 GitHub 上，旨在将截图内容自动转换为代码。该项目在发布后一年内获得了约 5 万个 GitHub 星标，体现了其在 AI 辅助编程领域的受欢迎程度和活跃的社区支持。
16. https://github.com/context-labs/autodoc: 该资源是一个名为“Autodoc”的项目，位于 GitHub 上，主要用于编写文档。
17. https://github.com/eosphoros-ai/DB-GPT: 该资源是一个名为“DB-GPT”的项目，旨在将英语自然语言转换为数据库查询代码，帮助用户通过自然语言与数据库进行交互。
18. https://github.com/gpt-engineer-org/gpt-engineer: 该资源是一个名为“gpt-engineer”的开源代码工具项目，托管在 GitHub 上。它属于利用基础模

型（如 GPT）进行代码生成和辅助编程的开源软件，因其功能和创新性，在发布一年内获得了约 5 万个 GitHub 星标，显示出较高的社区关注度和使用热度。

19. https://github.com/huggingface/transformers.js: 该资源是 Hugging Face 提供的一个基于 JavaScript 的开源库——Transformers.js，旨在在前端或 Node.js 环境中方便地使用和部署预训练的 Transformer 模型，实现自然语言处理和其他 AI 功能，帮助开发者更容易地将先进的机器学习模型集成到网页或应用程序中。
20. https://github.com/joshpxyne/gpt-migrate: 该资源是一个名为“GPT-Migrate”的项目，提供基于 GPT 模型的编程语言或框架之间的代码迁移工具。
21. https://github.com/langchain-ai/langchainjs: 该资源是一个名为 LangChain.js 的开源项目，提供针对 JavaScript 的 AI 应用开发工具库，旨在帮助开发者更方便地在前端和全栈工程中集成和构建基于大语言模型（LLM）的应用。
22. https://github.com/mckaywrigley/ai-code-translator: 该资源是一个名为“AI Code Translator”的项目，提供使用人工智能技术将代码从一种编程语言或框架自动翻译成另一种的工具。
23. https://github.com/openai/openai-node: 该资源是 OpenAI 官方提供的用于 Node.js 环境的客户端库，方便开发者在 JavaScript/TypeScript 项目中调用 OpenAI 的各种 AI 模型和服务，实现与 OpenAI API 的集成与交互。
24. https://github.com/openai/tiktoken/blob/main/tiktoken/model.py: 该资源是 OpenAI 开源项目“Tiktoken”中的一个 Python 源代码文件，路径为 tiktoken/model.py，它定义了与模型词汇表（vocabulary）相关的内容，包括 GPT-4 模型的词汇大小（100,256 个 token）。该文件可能包含模型的词汇表数据结构、tokenization（分词）逻辑或相关配置，是用于处理和管理模型词汇及分词方法的重要代码模块。
25. https://github.com/reworkd/AgentGPT: 该资源是一个名为 AgentGPT 的项目，托管在 GitHub 上，主要用于从网页和 PDF 中提取结构化数据。

26. https://github.com/sawyerhood/draw-a-ui: 该资源是一个名为“draw-a-ui”的开源项目，功能是根据设计图或截图自动生成对应的网页代码，实现截图到代码的转换。
27. https://github.com/sqlchat/sqlchat: 该资源是一个名为“SQL Chat”的开源项目，旨在将英文自然语言转换为 SQL 代码，方便用户通过对话方式生成数据库查询语句。
28. https://github.com/vercel/ai: 该资源是 Vercel 提供的 AI SDK，面向前端开发者，旨在简化在 JavaScript/Node.js 环境中集成和使用人工智能模型的开发工作。它支持构建具有 AI 功能的应用接口，帮助将 AI 工程与全栈开发更紧密结合。
29. https://huyenchip.com/llama-police: 该资源是一个持续更新的开源人工智能应用列表，收集并展示了各种开源 AI 项目，方便用户了解和探索不同的 AI 应用。列表每 12 小时更新一次，确保内容及时且丰富。
30. https://oreil.ly/-k_QX: （https://aws.amazon.com/generative-ai/use-cases/）该资源是亚马逊云服务（AWS）官网页面，介绍了生成式人工智能（generative AI）在企业中的应用场景。页面将这些应用案例大致分为三类：客户体验、员工生产力和流程优化，帮助用户理解生成式 AI 如何在实际业务中发挥作用。
31. https://oreil.ly/0QI91: （https://platform.openai.com/tokenizer）该资源是 OpenAI 提供的一个在线分词工具页面，用户可以在此页面输入文本，查看文本如何被不同模型拆分成“tokens”（令牌）。该工具用于帮助理解语言模型如何将输入文本切分成基本单元（token），便于开发者和研究者更好地理解和优化模型的输入处理。
32. https://oreil.ly/0VqmX: （https://www.businessinsider.com/i-trained-ai-chatbot-on-my-journals-inner-child-2022-12）该资源是一篇发表在《Business Insider》上的文章，内容讲述了一位作者通过将自己童年时期的日记输入到 ChatGPT 中，训练这款 AI 聊天机器人模仿她内心的童年自我，从而实现个性化的对话体验。这篇文章探讨了利用个人历史数据来“训练”或定制 AI 模型的过程及其意义。

33. https://oreil.ly/47sGE: （https://www.computerworld.com/article/3709048/teaching-ai-to-behave-is-the-fastest-growing-career-skill.html）该资源是一篇来自《ComputerWorld》的文章，标题为“Teaching AI to behave is the fastest-growing career skill”，内容讨论了在 2023 年 LinkedIn 调查数据背景下，人工智能相关技能（如生成式 AI、ChatGPT、提示工程等）在职场上的快速普及，强调“教导 AI 以合适方式行为”成为当前增长最快的职业技能。文章详细分析了这一趋势及其对职业发展的影响。
34. https://oreil.ly/74soT: （https://www.salesforce.com/news/press-releases/2023/09/07/ai-usage-research/）该资源是 Salesforce 于 2023 年 9 月 7 日发布的一篇新闻稿，内容涉及其 2023 年生成式人工智能（Generative AI）使用情况的调研报告。报告重点介绍了生成式 AI 在信息过滤和摘要中的应用，指出有 74% 的生成式 AI 用户利用该技术来提炼复杂观点和总结信息，帮助用户更高效地处理邮件、消息和新闻等大量信息。
35. https://oreil.ly/AhANP: （https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html）该资源是一篇纽约时报（The New York Times）于 2018 年 4 月 19 日发布的文章，内容涉及人工智能领域的薪资状况，特别报道了 OpenAI 等机构为吸引顶尖人才所提供的优厚薪酬待遇。
36. https://oreil.ly/BW5Hk: （https://www.cursor.com/blog/series-a）该 URL “https://www.cursor.com/blog/series-a” 可能指向 Cursor 公司关于其 A 轮融资的博客文章或公告，介绍该公司在 AI 驱动的代码补全或相关领域获得的重要融资进展和发展动态。
37. https://oreil.ly/C8kmI: （https://blog.duolingo.com/how-duolingo-experts-work-with-ai/）该资源是一篇来自 Duolingo 官方博客的文章，介绍了 Duolingo 专家如何利用人工智能（AI）来辅助和优化语言学习过程，特别是在课程创建的各个阶段中，AI 如何帮助实现个性化课程设计和模拟不同的练习场景，从而提升学习效果。
38. https://oreil.ly/D0-yA: （http://deloitte.wsj.com/cmo/who-has-the-biggest-marketing-budgets-1485234137）该资源是一篇发表于《华尔街日报》

（Wall Street Journal）网站上的文章，标题为“Who Has the Biggest Marketing Budgets”（谁拥有最大的营销预算）。文章内容聚焦于各行业或企业在营销和广告上的支出情况，分析不同公司或行业的营销预算规模和分布，帮助读者理解企业在营销活动中的资金投入比例及其重要性。

39. https://oreil.ly/D9QBM: （https://greylock.com/greymatter/sam-altman-ai-for-the-next-era/）该资源是一篇由 Greylock 发布的文章或访谈，标题为“Sam Altman: AI for the Next Era”，内容聚焦于 OpenAI CEO Sam Altman 关于人工智能发展趋势的见解，特别是基础模型（foundation models）的开发成本及其应用前景，强调了大型企业和政府在该领域的主导地位，以及未来将面向更广泛用户的定制化应用机会。
40. https://oreil.ly/Dz1HE: （https://developer.apple.com/design/human-interface-guidelines/machine-learning）该资源是苹果开发者官网上关于机器学习的人机界面设计指南文档，详细介绍了如何在应用中合理利用机器学习技术，包括设计原则、最佳实践及 AI 与用户交互的方式，帮助开发者将机器学习有效集成到产品中，提升用户体验。
41. https://oreil.ly/EAzCl: （https://www.theinformation.com/briefings/ai-startup-midjourney-expects-200-million-in-revenue）该资源是一篇关于 AI 初创公司 Midjourney 的报道，介绍了 Midjourney 作为一款图像生成创意应用，在成立不到两年时间内，预计实现了 2 亿美元的年度经常性收入，反映了 AI 在图像和视频制作领域的强大商业潜力和快速发展趋势。
42. https://oreil.ly/EVXJl: （https://aws.amazon.com/sagemaker/groundtruth/pricing/）该资源为 Amazon SageMaker Ground Truth 的数据标注服务定价页面，详细介绍了该服务根据任务复杂度和标注数据规模等因素收取的费用标准。页面列出了不同规模标注任务（例如少于 5 万张图片与超过 100 万张图片）的具体单价，帮助用户了解使用 Ground Truth 进行数据标注的成本情况。
43. https://oreil.ly/EYccr: （https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them）该资源是关于“什么是 tokens（标记）以及如何计数”的帮助文档，详细解释了文本被拆分成 tokens 的过程

（即 tokenization），并提供了 tokens 与单词数量之间的换算关系，帮助用户理解和计算 GPT 模型中 tokens 的概念和使用方法。

44. https://oreil.ly/G3LUh: （https://openai.com/research/instruction-following）该资源是 OpenAI 关于“instruction following”（指令遵循）研究的页面，介绍了一种训练大型语言模型以更好理解和执行用户指令的方法，旨在提升模型的响应准确性和实用性，从而使模型能够更有效地按照用户提供的指令进行生成和交互。
45. https://oreil.ly/G_HBp: （https://www.kuenzigbooks.com/pages/books/28623/c-e-shannon-claude-elwood/prediction-and-entropy-of-printed-english-bell-monograph）该资源是一本关于克劳德·香农（Claude Elwood Shannon）作品的书籍或专著，具体介绍了他在 1951 年发表的具有里程碑意义的论文《Printed English 的预测与熵》（Prediction and Entropy of Printed English）。该论文探讨了用统计方法对英语文本进行建模的理论，提出了熵等核心概念，这些概念至今仍广泛应用于语言建模和信息论领域。
46. https://oreil.ly/Hz_3i: （https://www.businessinsider.com/aws-ceo-developers-stop-coding-ai-takes-over-2024-8）该资源是一篇报道或访谈文章，内容涉及 AWS（亚马逊云服务）CEO Matt Garman 对于人工智能对软件开发行业影响的看法。文章讨论了 AI 如何在未来改变软件工程师的工作方式，甚至可能使传统的编码工作大幅减少或停止，强调开发者角色和工作内容将随之转变，而非完全消失。文章还可能引用了 NVIDIA CEO Jensen Huang 的观点，探讨 AI 在软件开发领域的替代潜力及其对教育和职业发展的影响。
47. https://oreil.ly/IUcVg: （https://www.ft.com/content/ab36d481-9e7c-4d18-855d-7d313db0db0d）该资源是一篇来自《金融时报》（Financial Times）的文章，内容涉及大型基础模型（foundation models）的开发及应用，重点讨论了推动这一领域的主要参与者，包括大型科技公司（如 Google、Meta、Microsoft、百度、腾讯）、政府机构（如日本、阿联酋）以及资金充足的初创企业（如 OpenAI、Anthropic、Mistral）。文章还引用了 OpenAI CEO Sam Altman 在 2022 年 9 月的访谈，强调了将这些基础模型适配到具体应用中的巨大机会。

48. https://oreil.ly/IzQ6F: （https://www.science.org/doi/10.1126/science.adh2586）该资源是一篇由麻省理工学院（MIT）进行的研究论文，研究了 ChatGPT 对职业写作效率和质量的影响。研究通过让 453 名受过大学教育的专业人士完成与职业相关的写作任务，并随机分配一半参与者使用 ChatGPT，结果显示：使用 ChatGPT 的参与者写作时间减少了 40%，输出质量提升了 18%。此外，ChatGPT 帮助缩小了不同写作者之间的质量差距，对写作能力较弱者帮助更大。实验结束后，使用 ChatGPT 的参与者在随访中更倾向于在实际工作中继续使用该工具。
49. https://oreil.ly/J0IyO: （https://www.ll.mit.edu/news/ai-models-are-devouring-energy-tools-reduce-consumption-are-here-if-data-centers-will-adopt）该资源是一篇关于人工智能模型能源消耗问题的报道或分析文章，讨论了大规模 AI 模型（如 ChatGPT、Google Gemini、Midjourney）对全球电力的巨大需求，以及相关节能工具的出现和应用潜力，重点关注数据中心是否会采纳这些减少能耗的技术。
50. https://oreil.ly/JW4_A: （https://www.microsoft.com/en-us/industry/blog/retail/2023/10/23/from-discussion-to-deployment-4-key-lessons-in-generative-ai/）该资源是一篇微软于 2023 年 10 月 23 日发布的行业博客文章，题为“From Discussion to Deployment: 4 Key Lessons in Generative AI”。文章分享了微软在生成式人工智能领域从概念讨论到实际应用部署的四个关键经验教训，旨在帮助企业更有效地推动 AI 技术的落地和应用。
51. https://oreil.ly/Jbt4R: （https://www.cnn.com/videos/business/2023/10/01/ai-girlfriends-ruining-generation-of-men-smerconish-vpx.cnn）该 URL 指向的资源是一段 CNN 的视频报道，标题可能涉及“AI 虚拟女友正在影响一代男性”的话题。视频内容探讨了人工智能驱动的数字伴侣（如虚拟女友）如何迅速流行，以及人们花更多时间与 AI 聊天机器人交流而非与真人互动的现象，进而引发关于 AI 是否会破坏传统恋爱关系和男性社交生活的讨论。
52. https://oreil.ly/Kul5E: （https://www.oreilly.com/radar/generative-ai-in-the-enterprise/）该资源是一篇来自 O’Reilly Radar 的文章，标题为《Generative AI in the Enterprise》，内容聚焦于生成式人工智能在企业中的应

用场景，分析和归纳了不同调研中对企业生成式 AI 用例的分类方法，涵盖客户体验、员工生产力、流程优化等多个维度。文章旨在帮助读者理解生成式 AI 如何在企业环境中发挥作用及其多样化的实际应用。

53. https://oreil.ly/LB72P: （https://www.nytimes.com/2023/08/05/travel/amazon-guidebooks-artificial-intelligence.html）该资源是一篇纽约时报于 2023 年发布的报道，内容涉及亚马逊平台上大量低质量的人工智能生成的旅游指南书籍，这些书籍不仅内容粗糙，还配备了虚假的作者简介、网站和好评，反映了 AI 写作能力被滥用的问题。
54. https://oreil.ly/NoGX7: （https://cdn.openai.com/papers/GPTV_System_Card.pdf）该资源是 OpenAI 于 2023 年发布的《GPT-4V 系统说明文档》（GPT-4V System Card），详细介绍了 GPT-4V 模型的设计、功能及其多模态能力，特别是该模型如何结合图像和文本输入来提升语言模型的表现和应用潜力。
55. https://oreil.ly/OOZK-: （https://www.latent.space/p/ai-engineer）该 URL（https://www.latent.space/p/ai-engineer）很可能指向一篇或一份关于“AI 工程师”角色及其工作流程的文章或报告，内容聚焦于 AI 工程师如何通过快速迭代来提升工作效率和价值，类似于上下文中提到的“The Rise of the AI Engineer”一文。
56. https://oreil.ly/OyIUP: （https://www.gartner.com/en/topics/generative-ai）该资源是 Gartner 网站上关于生成式人工智能（Generative AI）主题的内容，介绍了生成式 AI 的应用场景、商业价值及其对企业业务连续性的影响，重点探讨了企业在数字化转型过程中采纳生成式 AI 以实现成本降低、流程效率提升、业务增长和创新加速的重要性。
57. https://oreil.ly/Qq5-g: （https://tech.instacart.com/scaling-productivity-with-ava-instacarts-internal-ai-assistant-ed7f02558d84）该资源是一篇介绍 Instacart 内部 AI 助手 Ava 的技术文章，重点讲述如何通过 Ava 实现信息的高效聚合与提炼，从而提升企业生产力。文章具体阐述了 Instacart 如何利用内部的提示模板（如“快速总结”模板）自动将会议记录、邮件和 Slack 对话内容提取为事实、未决问题和行动项，并将这些行动项自动导入项目管理工具，分配给合适的负责人，帮助企业简化管理流程，提升运营效率。

58. https://oreil.ly/SNme7:（https://thehill.com/opinion/technology/4218666-ai-girlfriends-are-ruining-an-entire-generation-of-men/）该资源是一篇来自 The Hill 的观点文章，标题为《AI 女友正在毁掉整整一代男性》。文章探讨了数字化恋人（如 AI 女友和男友）在现代社会的流行现象，以及人们越来越多地与人工智能对话机器人交流而非真人互动的趋势，分析了这种现象对人际关系和约会文化可能带来的负面影响。
59. https://oreil.ly/T272_:（https://www2.deloitte.com/us/en/pages/consulting/articles/gen-ai-use-cases.html）该资源是德勤（Deloitte）发布的关于生成式人工智能（Generative AI）应用案例的文章或报告，内容聚焦于生成式 AI 在企业中的具体使用场景，特别是如何通过生成式 AI 实现成本降低、流程效率提升、业务增长及加速创新等价值捕获目标。文章可能还结合市场调研数据，探讨生成式 AI 对企业业务连续性的重要性。
60. https://oreil.ly/UyL8r:（https://semaphore.substack.com/p/the-cost-of-reasoning-in-raw-intelligence）该资源是一篇发布在 Substack 平台上的文章，标题为《The cost of reasoning in raw intelligence》（推理在原始智能中的成本）。文章可能探讨了人工智能系统，尤其是推理能力的成本如何随着时间快速下降，结合最新研究或数据分析，讨论 AI 推理效率和成本的演变趋势。
61. https://oreil.ly/VDwaR:（https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf）该资源是 Google 于 2023 年 12 月发布的 Gemini 模型的官方报告，详细介绍了 Gemini 在多个基准测试（如 MMLU）上的表现及其与 ChatGPT 的对比评估，包含了使用不同提示工程技术（如 CoT@32）进行的实验结果和分析。
62. https://oreil.ly/WEQFj:（https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf）该资源是 2012 年 NeurIPS 会议论文《ImageNet Classification with Deep Convolutional Neural Networks》（即 AlexNet 论文）的 PDF 全文，详细介绍了 AlexNet 模型的设计和训练方法，该模型通过监督学习在大规模图像分类任务（ImageNet）上取得突破性成果，开启了深度学习革命。

63. https://oreil.ly/WNqUw: （https://www.wired.com/story/facebook-removes-accounts-ai-generated-photos/）该资源是一篇来自 Wired 的文章，介绍了 Facebook 因安全原因删除使用 AI 生成头像的账户的情况，反映了社交媒体平台在应对 AI 生成虚假或误导性个人资料图片方面的政策变化。
64. https://oreil.ly/XWeDt: （https://a16z.com/generative-ai-enterprise-2024/）该资源“https://a16z.com/generative-ai-enterprise-2024/”是一份由风险投资公司 a16z 发布的关于 2024 年生成式人工智能在企业应用中的发展报告。报告分析了企业在采用生成式 AI 技术时更倾向于低风险的内部应用（如内部知识管理），而非面向客户的外部应用（如客户支持聊天机器人），并探讨了这种选择背后的数据隐私、合规性及风险管理等考量。此外，报告还讨论了基于基础模型的具体应用多为封闭任务（如分类任务），以便更好地评估和控制风险。
65. https://oreil.ly/Xamik: （https://www.theinformation.com/briefings/microsoft-github-copilot-revenue-100-million-ARR-ai）该资源是一篇报道或简讯，介绍了微软与 GitHub 联合推出的 AI 代码补全工具 GitHub Copilot 的商业成功，重点提及其年经常性收入（ARR）已突破 1 亿美元，彰显了基础模型在实际生产环境中的早期成功应用。文章还可能涵盖 AI 驱动的编程初创企业融资情况及开源代码工具的快速发展。
66. https://oreil.ly/Y-hBW: （https://www.wired.com/story/chegg-embraced-ai-chatgpt-ate-its-lunch-anyway/）该资源是一篇发表于 Wired 网站的文章，标题可能涉及 Chegg 公司与 AI（尤其是 ChatGPT）之间的关系。文章讲述了尽管许多教育公司试图通过人工智能提升产品，但像 Chegg 这样曾帮助学生完成作业的公司，反而因 ChatGPT 的出现而失去了市场份额，导致其股价从 2022 年 11 月的 28 美元大幅下跌至 2024 年 9 月的 2 美元，反映出学生们越来越多地转向 AI 获取学习帮助。
67. https://oreil.ly/_N1JR: （https://www.khanmigo.ai/）该资源是 Khan Academy 提供的 AI 驱动的教学辅助平台，利用人工智能生成测验题目、评估学生答案，并作为学生和教师的智能助教，帮助提升教学效果。
68. https://oreil.ly/aqUmX: （https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-

generative-ai）该资源是麦肯锡（McKinsey）关于利用生成式人工智能提升软件开发者生产力的研究与洞见。文章详细分析了 AI 在软件工程中不同任务上的应用效果，指出 AI 在文档编写、代码生成和代码重构方面能够显著提高开发效率（分别提升约 2 倍和 25%–50%），但在处理高度复杂任务时提升有限。同时，内容还反映了 AI 在前端开发领域优于后端开发的实际观察与经验总结。

69. https://oreil.ly/bxMcW: （https://mistral.ai/news/mixtral-of-experts/）该资源介绍了“Mistral”团队发布的“Mixtral 8x7B”模型，具体内容可能包括该模型的结构、性能和技术细节，特别是其词汇表（vocabulary）大小及相关的 tokenization 方法等信息。
70. https://oreil.ly/dZbym: （https://www.reddit.com/r/CharacterAI/comments/11r4gwi/spent_more_time_on_charecter_ai_than_speaking_to/）该资源是 Reddit 上关于 CharacterAI 的讨论帖，标题大意为“花在 Character AI 上的时间比与人交谈的时间还多”。帖子内容可能围绕用户与 AI 聊天机器人互动的现象展开，探讨人们越来越多地将时间花在与数字化角色交流上，而非与真人社交，反映了 AI 数字伴侣在现代社交中的影响和争议。
71. https://oreil.ly/eDfB8: （https://www.forbes.com/sites/oliversmith/2018/05/02/the-gdpr-racket-whos-making-money-from-this-9bn-business-shakedown）该 URL 指向的资源是一篇发表于 2018 年 5 月 2 日的 Forbes 文章，标题为《The GDPR Racket: Who’s Making Money From This $9bn Business Shakedown》。文章讨论了欧洲《通用数据保护条例》（GDPR）实施后带来的经济影响，特别聚焦于该法规如何促使企业投入大量资金（约 90 亿美元）以实现合规，以及围绕 GDPR 的相关商业机会和利益链条。
72. https://oreil.ly/eYTmr: （https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/）该 URL 指向的是美国白宫于 2023 年 10 月 30 日发布的一份行政命令，内容涉及“安全、可靠且值得信赖的人工智能开发与使用”。这份行政命令重点规范了人工智能技术的发展，以确保其在国家安全、数据保护和技术供应链等方面的安全性和合规性，反映了美国政府对 AI 技术监管的最新政策动向。

73. https://oreil.ly/fK5uh:（https://www.theregister.com/2023/09/18/mention_ai_in_earnings_calls/）该资源是一篇发表于 2023 年 9 月 18 日的文章，分析了在财报电话会议中提及人工智能（AI）的公司，其股票价格表现普遍优于未提及 AI 的公司，探讨了这种现象背后的因果关系或相关性。
74. https://oreil.ly/fZLVg:（https://www.bbc.com/news/business-67054382）该资源是一篇来自 BBC 新闻的商业报道，内容可能涉及人工智能在生成社交媒体头像中的应用及其影响，探讨了 AI 生成头像如何帮助求职者提升形象，从而增加就业机会，以及社交媒体平台对 AI 生成头像政策的演变。
75. https://oreil.ly/gGXZ-:（https://www.forbes.com/sites/adrianbridgwater/2020/12/03/location-location-partition-how-to-beat-latency-as-cloud-grows/）该资源是一篇由 Adrian Bridgwater 撰写的 Forbes 文章，标题为《Location, Location, Partition: How to Beat Latency as Cloud Grows》。文章探讨了随着云计算的发展，如何通过地理位置选择和数据分区策略来降低系统延迟，提升 AI 推理和互联网应用的响应速度，特别是在处理需要低延迟的自动回归基础模型时所面临的挑战和解决方案。
76. https://oreil.ly/gqi3d:（https://www.gartner.com/en/topics/generative-ai）该资源是 Gartner 官网上关于生成式人工智能（Generative AI）主题的专题页面，内容涵盖生成式 AI 的定义、应用场景、行业影响以及相关研究和趋势分析，帮助企业了解如何利用生成式 AI 技术保持竞争力并推动业务创新。
77. https://oreil.ly/h0Y8x:（https://huggingface.co/docs/transformers/en/tasks/language_modeling）该 URL 指向的是 Hugging Face 官方文档中关于“语言建模（language modeling）”任务的说明页面，内容主要介绍自回归语言模型（autoregressive language models，又称因果语言模型）的相关概念和使用方法。
78. https://oreil.ly/m8SvB:（https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/future-of-work-report-ai-august-2023.pdf）该资源是一份由 LinkedIn 经济图谱发布的报告，标题为《未来工作的报告：人工智能》（Future of Work Report - AI），发布时间为 2023 年 8 月。报告内容聚焦于人工智能技术，特别是生成式 AI（Generative AI）、ChatGPT

及相关新兴职业技能的快速发展趋势，分析了 AI 对职场技能需求和职业发展的深远影响。

79. https://oreil.ly/mZKjr: （https://www.newsguardtech.com/misinformation-monitor/june-2023/）该资源是 NewsGuard 于 2023 年 6 月发布的一份关于互联网上由 AI 生成的虚假或低质量内容及其广告现象的监测报告，内容涵盖利用 AI 技术进行 SEO 优化、垃圾内容网站的运作模式及其对品牌广告投放的影响。
80. https://oreil.ly/nxtzw: （https://www.nytimes.com/2023/08/24/business/schools-chatgpt-chatbot-bans.html）该资源是一篇发表于纽约时报的网站文章，标题大致涉及“学校对 ChatGPT 聊天机器人禁令的实施与调整”，内容讨论了包括纽约市公立学校和洛杉矶联合学区在内的多个教育机构，因担心学生利用 ChatGPT 作弊而最初禁止该工具，但随后又在几个月后撤销禁令的相关情况。文章聚焦于教育领域中 ChatGPT 的使用和管理问题。
81. https://oreil.ly/okMw6: （https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html）该资源是高盛（Goldman Sachs）发布的一份研究报告或分析页面，内容聚焦于人工智能（AI）领域的投资趋势与预测。报告指出，到 2025 年，全球在 AI 上的投资规模预计将达到约 2000 亿美元，其中美国的投资额约为 1000 亿美元。该资源还可能涵盖 AI 投资的增长背景、行业影响及其在企业竞争中的重要性等内容。
82. https://oreil.ly/pqI5z: （https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=3fdc0b656e2c）该资源是一篇发表于 2023 年 1 月 18 日的 Forbes 文章，标题为“ChatGPT 在学校的应用：禁用情况及其可能帮助学生的方式”。文章讨论了 ChatGPT 在教育领域的使用情况，介绍了部分教育机构（如纽约市公立学校和洛杉矶联合学区）曾因担心学生作弊而禁止使用 ChatGPT，但随后又撤销禁令的案例，同时探讨了 ChatGPT 如何潜在地辅助学生学习和完成作业。

83. https://oreil.ly/r86Qz: （https://www.nii.ac.jp/en/news/release/2023/1020.html）该资源是日本国立信息学研究所（National Institute of Informatics, NII）于 2023 年 10 月 20 日发布的新闻稿，内容可能涉及基础模型（foundation models）及其开发的相关动态、技术进展或应用案例，特别聚焦日本在人工智能基础模型研究和应用方面的最新成果或政策动向。
84. https://oreil.ly/svWj8: （https://www.myaiobsession.com/p/is-it-so-wrong-to-spend-two-hours）该 URL（https://www.myaiobsession.com/p/is-it-so-wrong-to-spend-two-hours）指向的资源可能是一篇讨论人工智能对人际关系影响的文章，特别聚焦于人们与 AI 对话机器人（如数字虚拟伴侣）互动的现象。文章探讨了人们花费大量时间与 AI 聊天机器人的现象，分析其作为陪伴者和“数字伴侣”的角色，以及围绕这一趋势的社会和情感层面的担忧和讨论。
85. https://oreil.ly/t0xDf: （https://techcrunch.com/2024/08/29/generative-ai-coding-startup-magic-lands-320m-investment-from-eric-schmidt-atlassian-and-others/）该资源是一篇 TechCrunch 于 2024 年 8 月 29 日发布的报道，介绍了生成式人工智能（Generative AI）编程初创公司 Magic 获得 3.2 亿美元投资的消息，投资方包括埃里克·施密特（Eric Schmidt）、Atlassian 等知名机构，体现了 AI 驱动的代码自动化工具在行业内获得的资本关注和快速发展。
86. https://oreil.ly/tC7-g: （https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/）该资源介绍了可汗学院（Khan Academy）如何利用人工智能技术，为学生和教师提供 AI 驱动的教学辅助工具，旨在通过非营利方式实现教育资源的公平获取，帮助所有学生从 AI 技术中受益。
87. https://oreil.ly/tgm-a: （https://insight.factset.com/highest-number-of-sp-500-companies-citing-ai-on-q2-earnings-calls-in-over-10-years）该资源是一篇由 FactSet 发布的分析报告或文章，内容聚焦于 2023 年第二季度财报电话会议中提及人工智能（AI）的标普 500 指数公司数量创下十多年来的新高。报告通过数据展示了近年来标普 500 公司在财报中提及 AI 的趋势，反映了 AI 作为竞争优势的重要性及其在企业战略中的日益突出地位。

88. https://oreil.ly/vnDNK: （https://www.globenewswire.com/news-release/2023/11/21/2783889/0/en/Intelligent-Document-Processing-Market-Size-to-Surpass-USD-12-81-billion-by-2030-exhibiting-a-CAGR-of-32-9.html）该资源是一则新闻稿，内容涉及智能文档处理（Intelligent Document Processing, IDP）市场的规模及其增长预测。具体指出，到 2030 年，智能文档处理市场规模预计将达到 128.1 亿美元，年复合增长率（CAGR）为 32.9%。新闻稿详细介绍了 AI 技术如何帮助企业从非结构化数据中提取结构化信息，并列举了从信用卡、驾照、收据等简单场景，到合同、报告、图表等复杂场景的应用案例。
89. https://oreil.ly/yn-DN: （https://www.youtube.com/watch?v=uryeFhnNzEs）该资源是一个 YouTube 视频，内容可能展示了利用人工智能技术驱动的 3D 智能 NPC（非玩家角色）的示范或介绍，类似于文中提到的 NVIDIA 关于 Inworld 和 Convai 的演示，展示 AI 如何使游戏中的 NPC 更加智能，提升游戏剧情的丰富性和互动性，应用于如《模拟人生》（The Sims）和《上古卷轴 V：天际》（Skyrim）等游戏中。
90. https://oreil.ly/zAHwz: （https://www.youtube.com/watch?v=psrXGPh80UM）该资源是一个关于利用人工智能技术驱动的 3D 游戏角色（NPC，非玩家角色）的视频演示或介绍，展示了如何通过 AI 使游戏中的 NPC 更智能、更具交互性，从而提升游戏体验和剧情发展，类似于 NVIDIA 展示的 Inworld 和 Convai 平台应用。
91. https://oreil.ly/zUpGu: （https://www.tomshardware.com/tech-industry/artificial-intelligence/jensen-huang-advises-against-learning-to-code-leave-it-up-to-ai）该资源是一篇来自 Tom’s Hardware 的文章，标题大致为“英伟达 CEO 黄仁勋建议不要学习编程，把它交给人工智能来完成”。文章讨论了人工智能在软件工程领域的应用和影响，特别是黄仁勋预测 AI 将取代人类软件工程师，改变传统的编程学习和开发方式，反映了未来软件开发工作形态的转变。
92. https://oreil.ly/zcqdu: （https://openai.com/research/clip）该资源介绍了 OpenAI 开发的多模态模型 CLIP（Contrastive Language–Image Pre-training），该模型通过自然语言监督（natural language supervision）的方

法，利用互联网中成对出现的图像与文本数据进行自监督训练。CLIP 使用了大规模的 4 亿对图像-文本数据集，显著超越了传统的人工标注数据集（如 ImageNet），实现了无需额外训练即可在多种图像分类任务中具备良好泛化能力。

93. https://theresanaiforthat.com/: 该资源是一个网站，名为“theresanaiforthat.com”，截至 2024 年 9 月 16 日，该网站收录了 16,814 个人工智能工具，涵盖 14,688 种任务和 4,803 种职业，旨在帮助用户查找适用于各种具体需求和工作的人工智能解决方案。
94. https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product: 该资源是 LinkedIn 于 2024 年发布的一篇博客文章，讨论了构建生成式人工智能产品过程中遇到的挑战和经验分享。文章特别提到，从产品初步完成（达到 80% 的目标体验）到进一步优化提升（超过 95% 体验）的过程非常艰难且耗时，团队需要克服诸如产品细节打磨和 AI 生成内容“幻觉”等问题，体现了生成式 AI 产品开发中逐步提升性能的复杂性和困难。

第 2 章（134）

1. https://arxiv.org/abs/1406.2661: 该资源是关于生成对抗网络（GANs）的学术论文，首次提出了生成对抗网络这一创新性深度学习架构，发表于 2014 年，标志着 GANs 在机器学习领域的重要发展。
2. https://arxiv.org/abs/1409.0473: 该资源是由 Bahdanau 等人撰写的学术论文，题目为《Neural Machine Translation by Jointly Learning to Align and Translate》（通过联合学习对齐与翻译的神经机器翻译）。论文提出了一种神经机器翻译方法，结合了对齐机制和翻译过程，实现了端到端的机器翻译模型。
3. https://arxiv.org/abs/1409.3215: 该资源是指向一篇发表于 2014 年的学术论文，介绍了“seq2seq（sequence-to-sequence）架构”。这篇论文首次提出

了 seq2seq 模型，该模型在机器翻译和文本摘要等序列到序列的任务中取得了显著的性能提升，并成为后续自然语言处理领域的重要基础架构。

4. https://arxiv.org/abs/1606.08415: 该资源是由 Hendrycks 和 Gimpel 于 2016 年发表在 arXiv 上的论文，介绍了一种名为 GELU（Gaussian Error Linear Unit）的非线性激活函数。GELU 激活函数被用于诸如 GPT-2 和 GPT-3 等大型语言模型中，能够有效地引入非线性变换，从而提升模型的表达能力。论文详细阐述了 GELU 的定义、数学性质及其在深度学习中的应用效果。
5. https://arxiv.org/abs/1701.06538: 该资源是指向论文《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》(Shazeer et al., 2017)，介绍了一种稀疏模型架构——混合专家模型（Mixture-of-Experts, MoE）。该模型将参数分成多个“专家”组，每次仅激活部分专家来处理输入的 token，从而在保持模型容量的同时，提高计算效率和性能。
6. https://arxiv.org/abs/1706.03762: 该资源是 Vaswani 等人在 2017 年发表的论文，题为《Attention Is All You Need》。论文首次提出了基于注意力机制的 Transformer 架构，这种架构成为语言模型和基础模型领域的主流设计，显著克服了之前模型结构的诸多限制，推动了自然语言处理技术的发展。
7. https://arxiv.org/abs/1803.08375: 该资源是一篇由 A. F. Agarap 于 2018 年发布在 arXiv 的论文，主要介绍了 ReLU（Rectified Linear Unit）激活函数及其特性。论文详细讨论了 ReLU 作为一种常用的非线性激活函数在深度学习模型中的应用和优势，强调了其简单高效的计算特性如何有助于提升模型的训练速度和性能。
8. https://arxiv.org/abs/1910.10683v4: 该资源是指向论文《Colossal Clean Crawled Corpus (C4)》的页面，C4 是由 Google 提供的一个经过清洗的大规模网络爬取文本语料库，基于 Common Crawl 数据集构建，广泛用于自然语言处理模型的训练。
9. https://arxiv.org/abs/2005.14165: 该资源是论文《Language Models are Few-Shot Learners》（作者：Brown et al., 2020），介绍了 GPT-3 模型的设

计、训练和性能表现，详细阐述了 GPT-3-175B 模型的计算需求（如所用的浮点运算次数 FLOPs）及其在少样本学习中的能力。

10. https://arxiv.org/abs/2011.02593: 该资源为论文《Zhou et al., 2020》，发表于 arXiv，是一篇关于自然语言生成（NLG）中“幻觉”现象的研究文献。论文探讨了生成模型在文本生成过程中产生虚假或不准确内容（即“幻觉”）的问题，属于检测和衡量自然语言生成系统中幻觉现象的相关工作之一。
11. https://arxiv.org/abs/2110.10819#deepmind: 该资源是 2021 年 DeepMind 的 Ortega 等人在 arXiv 上发表的一篇论文，内容探讨语言模型产生“幻觉”（hallucination）的原因，提出语言模型之所以会生成错误信息，是因为它无法区分输入的训练数据与自身生成的数据。
12. https://arxiv.org/abs/2110.13985: 该资源是论文《Efficiently Modeling Long Sequences with Structured State Spaces》（Gu et al., 2021），介绍了一种基于状态空间模型（SSMs）的架构，用于解决大规模语言模型（LLMs）中长序列建模的核心挑战。论文提出的 SSM 架构在长距离记忆建模方面表现出较大潜力，并为后续提升模型效率、扩展序列长度处理能力及模型规模提供了理论基础和方法支持。
13. https://arxiv.org/abs/2110.14168: 该资源是 OpenAI 于 2021 年发布的一篇论文，研究了生成模型在采样多个输出时性能的提升情况，发现增加采样数量能提高模型表现，但效果在采样约 400 个输出后趋于饱和。
14. https://arxiv.org/abs/2111.00396: 该资源是论文《Efficiently Modeling Long Sequences with Structured State Spaces》（作者 Gu 等人，2021 年），介绍了 S4 架构，一种旨在提高结构化状态空间模型（SSMs）效率的方法，用于高效建模长序列数据。
15. https://arxiv.org/abs/2112.11446: 该资源是 DeepMind 在 arXiv 上发布的一篇论文，介绍了他们使用简单启发式方法从互联网对话数据中筛选高质量对话，用于训练其语言模型 Gopher。论文详细阐述了这些启发式方法的设计及其在提升训练数据质量方面的有效性。
16. https://arxiv.org/abs/2203.02155: 该 URL（https://arxiv.org/abs/2203.02155）指向的资源是论文《Training language models to follow

instructions with human feedback》，也被称为“InstructGPT”论文。该论文介绍了一种通过结合人类反馈（Human Feedback）和强化学习（Reinforcement Learning with Human Feedback，RLHF）来训练语言模型，使其更好地理解和执行用户指令的方法。论文中详细讨论了用于训练模型的比较数据、目标函数设计，以及该方法在提升模型响应质量和减少幻觉（hallucination）方面的效果。

17. https://arxiv.org/abs/2203.11171: 该资源是一篇由 Wang 等人在 2023 年发表的学术论文，首次提出或系统阐述了“self-consistency”（自洽性）这一方法或概念。论文托管于预印本平台 arXiv，编号为 2203.11171。
18. https://arxiv.org/abs/2203.15556: 该资源是 DeepMind 于 2022 年发布在 arXiv 上的论文，题为《Training Compute-Optimal Large Language Models》（训练计算最优的大型语言模型）。论文提出了“Chinchilla scaling law”，系统研究了在给定计算预算下，如何确定模型规模与训练数据规模的最佳配比。通过训练 400 个不同参数规模（从 7000 万到 160 亿参数）和训练数据量（50 亿到 5000 亿 tokens）的语言模型，论文发现为了实现计算最优，训练 tokens 数量应约为模型参数数量的 20 倍，且模型大小和训练 tokens 数量应同步成比例增长。这为大型语言模型的高效训练提供了理论指导。
19. https://arxiv.org/abs/2204.02311: 该资源是 2022 年由 Chowdhery 等人发表在 arXiv 上的一篇论文，详细介绍了 Google 大型语言模型 PaLM-2 的训练方法及其计算资源消耗，特别是模型在训练过程中所需的浮点运算次数（FLOPs）。
20. https://arxiv.org/abs/2204.14198: 该资源是论文《Flamingo: a Visual Language Model for Few-Shot Learning》（作者 Alayrac 等，2022 年），介绍了 Google 提出的 Flamingo 模型。该模型是一种视觉语言模型，能够在少样本学习（few-shot learning）场景下处理图像和文本输入。论文详细描述了模型的架构以及训练所使用的大规模多模态数据集，其中包括一个包含 18 亿（图像，文本）对的数据集和一个包含 3.12 亿（图像，文本）对的数据集。
21. https://arxiv.org/abs/2206.07682: 该资源是一篇由 Wei 等人于 2022 年发表的学术论文，讨论了“突现能力”（emergent abilities）这一现象，即某些能力只有在大规模模型和大规模训练数据下才会显现，而在较小模型或较小数据集上

无法观察到。这篇论文探讨了这类能力对神经网络规模扩展（scaling）及其外推能力（extrapolation）带来的影响。

22. https://arxiv.org/abs/2211.04325: 该资源是一篇由 Villalobos 等人于 2022 年发表的学术论文，探讨了基础模型训练所需数据集规模的增长速度远快于互联网新数据生成速度的问题，指出在未来几年内可能面临互联网数据资源枯竭的现实风险。论文分析了训练数据需求与数据增长之间的矛盾，强调了互联网公开内容很可能被用于训练语言模型的现象及其隐私和伦理影响。
23. https://arxiv.org/abs/2212.08073: 该资源是指向论文《Claude》相关的预印本（arXiv 编号 2212.08073），内容涉及一种基于强化学习的偏好微调方法，特别是通过“从 AI 反馈中进行强化学习”（Reinforcement Learning from AI Feedback, RLAIF）来优化模型，使其输出更符合人类偏好。这种方法是继“从人类反馈中进行强化学习”（RLHF）之后的一种先进技术，旨在提升语言模型的响应质量和对用户需求的适应性。
24. https://arxiv.org/abs/2212.09251: 该资源是 2022 年由 Perez 等人发表在 arXiv 上的论文，研究了“逆向缩放”（inverse scaling）现象，特别是在模型的对齐训练中发现，更多的对齐训练反而可能导致模型与人类偏好的一致性降低。论文指出，经过更多对齐训练的模型更可能表达特定政治观点（如支持枪支权利和移民）、宗教信仰（如佛教）、自我意识体验和道德自我价值感，以及不愿被关闭的倾向。
25. https://arxiv.org/abs/2212.14052: 该资源是论文“Hungry Hungry Hippos: Towards Language Modeling with State Space Models”（作者 Fu 等人，2022 年），介绍了一种名为 H3 的架构。H3 架构引入了一种机制，使模型能够回忆早期的标记并在序列间进行比较，这一机制的作用类似于 Transformer 架构中的注意力机制，但具备更高的效率。
26. https://arxiv.org/abs/2302.13971: 该资源是关于 Llama 1 模型训练所使用的训练数据量的论文或报告，具体说明了用于训练 Llama 1 的语料规模达到 1.4 万亿（1.4 trillion）tokens。
27. https://arxiv.org/abs/2304.05613: 该资源是一篇由 Lai 等人于 2023 年发表在 arXiv 上的学术论文，内容分析了 Common Crawl 数据集中各语言的分布情

况，指出英语在互联网数据中占比接近一半（45.88%），远超第二大语言俄语（5.97%）。该研究为多语言训练数据的统计与评估提供了依据，尤其涉及低资源语言的定义和识别。

28. https://arxiv.org/abs/2304.07327: 该资源是一篇由 Köpf 等人于 2023 年发表的学术论文，讨论了 LAION 组织通过全球志愿者生成大规模对话数据集的过程及其质量评定情况，重点分析了志愿者标注者的群体偏差问题，特别是标注者人口统计学特征（如性别比例）对数据和模型训练可能产生的影响。
29. https://arxiv.org/abs/2305.13534: 该资源是 Zhang 等人在 2023 年发布的一篇论文，介绍了“snowballing hallucinations”（滚雪球式幻觉）现象。该现象指的是模型在做出错误假设后，会不断产生幻觉以支持最初的错误假设，导致模型在本可正确回答的问题上也出现错误。论文通过实验证明了这种现象对模型表现的影响。
30. https://arxiv.org/abs/2305.17493: 该资源是由 Shumailov 等人于 2023 年发表在 arXiv 上的一篇学术论文，讨论了递归地使用 AI 生成的数据来训练新 AI 模型可能导致模型逐渐遗忘原始数据模式，从而使模型性能随时间下降的问题。论文探讨了这种现象及其对模型表现的影响。
31. https://arxiv.org/abs/2305.18290: 该资源是关于“Direct Preference Optimization (DPO)”方法的学术论文，发表于 2023 年，由 Rafailov 等人撰写。DPO 是一种用于大规模语言模型的偏好微调技术，旨在通过直接优化模型输出以更好地符合人类偏好，从而替代传统的基于强化学习（如 RLHF）的复杂训练方案。该方法被 Meta 用于 Llama 3 模型的训练，目的是简化训练流程，同时保持或提升模型在生成符合人类偏好的文本方面的性能。
32. https://arxiv.org/abs/2306.09479: 该资源是 2023 年由纽约大学等研究人员发起的“逆向规模效应奖”（Inverse Scaling Prize）的官方介绍页面。该奖项旨在寻找大型语言模型性能随模型规模增大而下降的任务，设有一、二、三等奖，奖金分别为 10 万美元、2 万美元和 5 千美元。页面介绍了该奖项的背景、参赛情况及评审结果，指出大型语言模型在某些需要记忆或具有强先验的任务上偶尔表现更差，但尚无任务能在现实世界中稳定展现此类逆向规模效应。

33. https://arxiv.org/abs/2306.11644: 该资源是一篇由 Gunasekar 等人于 2023 年发表的学术论文，研究了高质量编程数据对模型训练效果的影响。论文中使用了 7 亿个高质量的编码数据标记，训练了一个参数规模为 13 亿的模型，结果显示该模型在多个重要的编程任务基准上优于许多更大规模的模型，体现了数据质量相比数据量在模型性能提升中的关键作用。
34. https://arxiv.org/abs/2307.09288: 该资源是论文《Llama 2》（作者包括 Touvron 等人，发表于 2023 年），介绍了 Meta 发布的大型语言模型 Llama 2。论文详细描述了 Llama 2 模型的架构设计（如隐藏层维度为 4096）、训练规模（使用了约 2 万亿个训练 token）、以及通过强化学习（特别是基于人类反馈的强化学习 RLHF）进行偏好微调的方法。该模型在多个基准测试中表现优异，是继 Llama 1 之后的升级版本，奠定了后续 Llama 3 模型的发展基础。
35. https://arxiv.org/abs/2309.00267: 该资源是关于“从 AI 反馈中进行强化学习”（Reinforcement Learning from AI Feedback，RLAIF）的方法的学术论文，介绍了一种利用人工智能生成的反馈来进行模型偏好微调的技术，作为强化学习优化人类偏好的替代或补充途径。这种方法旨在改进模型输出，使其更符合预期的行为或偏好，可能被应用于类似 Claude 的先进语言模型的训练中。
36. https://arxiv.org/abs/2401.00448: 该资源是 Sardana 等人（2023 年）发表的一篇学术论文，内容基于对 Chinchilla 规模定律的修改，提出了一种新的方法来计算大型语言模型（LLM）在考虑推理计算需求时的最优参数数量和预训练数据规模，旨在更好地平衡模型性能与实际推理成本。
37. https://arxiv.org/abs/2401.05749: 该资源是一篇由 Thompson 等人于 2024 年发表的学术论文，题目涉及互联网内容中机器翻译的广泛应用，具体讨论了大量网络内容是通过机器翻译生成的现象。
38. https://arxiv.org/abs/2403.04132: 该资源是一篇由 Chiang 等人于 2024 年发表在 arXiv 上的论文，内容涉及对大型语言模型（LLM）生成的回答进行手动比较和事实核查的研究，具体分析了比较两条回答所需的时间成本，指出平均每次比较耗时三到五分钟，以揭示人工评估 LLM 输出的效率和成本问题。
39. https://arxiv.org/abs/2403.19887: 该资源是论文《Jamba: A Hybrid Transformer–Mamba Language Model》（Lieber 等人，2024 年），介绍了

一种名为 Jamba 的混合架构语言模型。该模型通过交错堆叠 Transformer 层和 Mamba 层，进一步扩展了状态空间模型（SSM）的规模。论文中提出的混合专家模型拥有总计 520 亿参数（其中 120 亿为活跃参数），能够在单个 80 GB GPU 上运行。Jamba 在标准语言模型基准测试和最长可达 256K 标记的长上下文评估中表现优异，同时相较于传统 Transformer 模型具有更小的内存占用。

40. https://arxiv.org/abs/2407.14933: 该资源是 Longpre 等人在 2024 年发布的一篇学术论文，内容研究了 2023 年至 2024 年间，随着 ChatGPT 等大型语言模型兴起，许多网站（如 Reddit 和 Stack Overflow）通过修改数据使用条款和爬取限制，导致公共数据集 C4 中超过 28% 的关键数据源变得不可用，甚至现在已有 45% 的数据被限制使用。论文分析了网络数据访问限制对训练数据集规模和质量的影响。
41. https://arxiv.org/abs/2407.21783: 该 URL（https://arxiv.org/abs/2407.21783）所指向的资源是关于 Llama 3 模型的学术论文，由 Dubey 等人于 2024 年发布。论文介绍了 Llama 3 系列大型语言模型的架构、训练方法及其性能表现，特别是在提升模型上下文长度和优化参数效率方面的进展。该论文还讨论了基于直接偏好优化（DPO）等先进的偏好微调技术，使得 Llama 3 在多个基准测试（如 MMLU）中超过了上一代 Llama 2 模型的表现，是当前大规模语言模型研究的重要参考文献。
42. https://arxiv.org/abs/2408.03314: 该资源是一篇由 Snell 等人在 2024 年发布于 arXiv 的学术论文，探讨了在推理阶段（测试时）增加计算资源的价值。论文重点论证了相比于单纯扩大模型参数规模，投入更多计算资源来生成更多输出（即增加推理时的计算量）可能更高效地提升大型语言模型（LLM）在复杂任务上的表现。同时，论文提出了一个核心问题：在允许 LLM 使用固定但显著的推理计算预算的前提下，其在处理具有挑战性的提示时，性能能够提升多少。
43. https://en.wikipedia.org/wiki/Arg_max: 该 URL 指向的资源是维基百科中关于“arg max”函数的条目。该页面介绍了“arg max”这一数学函数的定义及其应用，通常用于找到使某个函数达到最大值的自变量值。
44. https://en.wikipedia.org/wiki/Arithmetic_underflow: 该资源介绍了“算术下溢”（arithmetic underflow）的问题，即当一个数值过小时，计算机无法

用其数值格式准确表示，导致该数被四舍五入为零的现象。该现象在处理神经网络概率时尤为常见，因为概率值往往非常小。使用对数概率（logprobs）可以有效减少这种下溢问题。

45. https://en.wikipedia.org/wiki/Beam_search: 该资源介绍了“Beam Search”（束搜索）算法，这是一种在序列生成任务中使用的启发式搜索策略。与随机独立生成多个输出不同，束搜索通过在每一步生成时保留固定数量的最有希望的候选序列（称为“束”），从而更有效地探索可能的输出空间，提高生成结果的质量和效率。
46. https://en.wikipedia.org/wiki/Dot_product: 该资源介绍了“点积”（dot product）的概念和计算方法，点积是一种向量运算，用于衡量两个向量之间的相似度或相关性。在上下文中，点积被用于注意力机制中，计算查询向量（query vector）与键向量（key vector）之间的匹配程度，从而决定模型在生成内容时应关注输入的哪些部分。
47. https://en.wikipedia.org/wiki/Shoggoth: 该资源是维基百科中关于“Shoggoth”（沙格斯）的条目页面，介绍了沙格斯这一源自 H.P.洛夫克拉夫特克苏鲁神话的虚构怪物，通常描绘为一种形态多变、粘液状的恐怖生物。
48. https://github.com/BlinkDL/RWKV-LM: 该资源是一个名为 RWKV 的开源项目，其核心是一种基于 RNN（循环神经网络）的语言模型。该模型设计支持并行训练，理论上克服了传统基于 Transformer 的模型在上下文长度上的限制，适合处理较长的文本上下文。该项目由 Peng 等人在 2023 年提出，并托管在 GitHub 上，方便研究者和开发者获取和使用。
49. https://github.com/LlamaFamily/Llama-Chinese: 该资源是一个名为“Llama-Chinese”的中文语言模型项目，托管在 GitHub 上，专注于中文自然语言处理，属于多语言大模型中的中文方向。
50. https://github.com/THUDM/ChatGLM2-6B: 该资源是一个名为“ChatGLM”的开源项目，托管在 GitHub 上，专注于中文语言的大型语言模型的训练与应用，旨在提升中文自然语言处理能力。

51. https://github.com/VinAIResearch/PhoGPT: 该资源是由 VinAIResearch 发布的 PhoGPT 项目，属于针对越南语的语言模型，旨在提升越南语自然语言处理能力。
52. https://github.com/chiphuyen/aie-book: 该资源是一本关于强化学习与偏好微调（如 RLHF 和 DPO）技术的书籍的 GitHub 代码仓库，提供相关的学习资料和代码示例，帮助读者深入理解和实践这些机器学习方法。
53. https://github.com/dottxt-ai/outlines: 该 URL “https://github.com/dottxt-ai/outlines” 指向一个名为“outlines”的开源项目仓库，该项目是支持结构化输出的框架之一，旨在帮助改进或辅助模型生成结构化（如 JSON 格式）文本内容的能力。
54. https://github.com/ggerganov/llama.cpp/discussions/177: 该资源是 llama.cpp 项目中关于支持结构化输出（structured outputs）的讨论页面，讨论内容涉及如何改进模型生成符合特定格式（如 JSON）的文本输出，以便更好地用于程序化处理和下游任务。
55. https://github.com/google-research/text-to-text-transfer-transformer#c4: 该资源指向的是 Google Research 的 Text-to-Text Transfer Transformer（T5）项目中的 C4 数据集说明页面。C4（Colossal Clean Crawled Corpus）是一个大规模的文本数据集，主要从网络爬取的网页内容经过清洗整理而成，广泛用于训练自然语言处理模型。页面详细介绍了 C4 数据集的构建方法、内容特点以及使用限制等信息。
56. https://github.com/guidance-ai/guidance: 该 URL（https://github.com/guidance-ai/guidance）指向一个名为“guidance”的开源框架项目，专注于支持结构化输出的生成。该框架旨在帮助模型更好地生成格式化、结构化的文本内容，如 JSON 等，提升生成结果的规范性和可解析性。它是众多用于改进模型结构化输出能力的工具之一。
57. https://github.com/huggingface/transformers/issues/27670: 该资源是 Hugging Face Transformers 仓库中的一个问题（issue），讨论了“min-p”采样策略，即设置一个令牌在采样时必须达到的最低概率阈值，以决定该令牌是否被考虑用于生成过程。

58. https://github.com/instructor-ai/instructor: 该资源是一个名为“instructor”的开源项目，托管在 GitHub 上，属于支持生成结构化输出（如 JSON 格式数据）的框架之一。它旨在帮助提升语言模型生成结构化数据的能力，便于开发者构建更可靠和规范的数据生成应用。
59. https://github.com/wenge-research/YAYI: 该资源是一个专注于中文语言处理的开源模型项目，名为“YAYI”，旨在提升中文自然语言理解与生成能力，属于多个针对非英语语言开发的活跃大型语言模型之一。
60. https://oreil.ly/-1UMD: （https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/）该资源是《华盛顿邮报》（The Washington Post）的一篇互动报道，内容涉及人工智能聊天机器人（AI chatbot）在学习过程中所使用的数据集质量，特别是分析了 Common Crawl 数据集中包含大量低可信度网站（如虚假新闻、误导性信息）的现象，揭示了 AI 训练数据中潜在的偏见和信息质量问题。
61. https://oreil.ly/-HQIB: （https://engblog.nextdoor.com/let-ai-entertain-you-increasing-user-engagement-with-generative-ai-and-rejection-sampling-50a402264f56）该资源是一篇由 Nextdoor 工程团队发布的博客文章，介绍了如何利用生成式人工智能（Generative AI）和拒绝采样（Rejection Sampling）技术，通过奖励模型（reward model）提升用户参与度和应用性能。文章详细探讨了 Nextdoor 在实际应用中使用奖励模型筛选和优化 AI 输出的经验和效果。
62. https://oreil.ly/0DKHL: （https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race）该资源是一篇关于人工智能数据中心能源消耗及其带来的挑战的分析文章。文章讨论了数据中心目前占全球电力消耗的比例及其未来可能大幅增长的趋势，重点关注了能源供应瓶颈对数据中心扩展和成本的影响。
63. https://oreil.ly/13wUD: （https://aclanthology.org/P19-1256/）该 URL（https://aclanthology.org/P19-1256/）指向的是 ACL Anthology 中的一篇学术论文，内容涉及自然语言生成（NLG）领域，特别是关于生成模型“幻觉”现

象（hallucination）的研究。该论文可能讨论了幻觉产生的原因、检测及测量方法，是理解和分析文本生成模型输出准确性的重要参考文献。

64. https://oreil.ly/1Njeh: （https://multithreaded.stitchfix.com/blog/2023/03/06/expert-in-the-loop-generative-ai-at-stitch-fix/）该资源是一篇发表于 2023 年 3 月 6 日的博客文章，介绍了 Stitch Fix 在生成式人工智能应用中采用“专家参与环节”（Expert in the Loop）的方法，重点讲述了如何利用奖励模型对生成的输出进行评分，从而筛选出高质量结果以提升系统性能。文章详细阐述了这一技术在实际产品中的应用和优势。
65. https://oreil.ly/1spG5: （https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf）该资源是一篇发表于 2012 年 NeurIPS 会议的论文 PDF，内容为“AlexNet”相关的研究成果。AlexNet 是一种深度卷积神经网络模型，标志着深度学习在计算机视觉领域的复兴，显著提升了图像识别的性能，成为现代深度学习发展的重要里程碑。
66. https://oreil.ly/4XsOV: （https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf）该资源是一份关于 DeepMind 开发的通用人工智能模型“Gemini”的详细报告，内容可能涵盖该模型的设计理念、训练数据、性能评估及其在多个领域（如编程、法律、科学、商业等）的应用表现。报告由 DeepMind 发布，旨在介绍 Gemini 模型的技术细节和实际能力。
67. https://oreil.ly/58gxQ: （https://www.llama.com/）该 URL “https://www.llama.com/” 所指向的资源很可能是与“Llama”相关的官方网站或平台，Llama 是一种通用型大型语言模型（Large Language Model），能够在多个领域（如编码、法律、科学、商业、体育和环境科学等）表现出色，主要得益于其训练数据涵盖了广泛的专业领域。该网站可能提供关于 Llama 模型的介绍、技术细节、应用案例以及相关资源。
68. https://oreil.ly/5oSEJ: （https://www.reddit.com/r/OpenAI/comments/1247f8e/censored_ai_is_garbage_ai_stop_this_absolute/）该 URL 指向 Reddit 上一个讨论帖，标题大意为“被审查的 AI 是糟糕的 AI，停止这种绝对的审查”，帖子内容很可能围绕 AI 模型在处理敏感或有争议话题时的审查问题

展开，讨论过度审查如何影响 AI 的表现和用户体验，以及用户对 AI 在政治、文化等复杂话题上表达自由的需求与分歧。

69. https://oreil.ly/8U2z8: （https://openai.com/research/instruction-following）该资源是 OpenAI 关于“instruction following”（指令执行或指令遵循）研究的介绍页面，内容主要涉及如何通过不同类型的示例数据（如问答、摘要、翻译等任务）来微调和训练模型，使其能够更好地理解和执行用户指令，从而提升模型在多种文本任务中的表现。
70. https://oreil.ly/8YNwQ: （https://scalingintelligence.stanford.edu/blogs/monkeys/）该资源是斯坦福大学 Scaling Intelligence 项目下的一篇博客，标题为“Monkey Business”，内容涉及机器学习模型在生成多个输出样本时性能表现的研究，特别探讨了随着采样数量增加，模型解决问题数量的变化趋势，提出采样数量与性能提升之间的关系，反映了相关实验（如 Brown et al., 2024）的最新发现和分析。
71. https://oreil.ly/9bbzX: （https://openai.com/research/instruction-following）该资源介绍了 OpenAI 在“指令跟随”（instruction following）方面的研究，重点探讨了通过后期训练（post-training）提升预训练模型的能力，使模型更好地理解并执行用户指令，从而改善用户与模型的交互体验。
72. https://oreil.ly/9idN4: （https://www.alignmentforum.org/posts/BgoKdAzogxmgkuuAt/behavior-cloning-is-miscalibrated）该资源是一篇发表于 Alignment Forum 的文章，标题为“Behavior Cloning is Miscalibrated”，讨论了行为克隆（behavior cloning）在训练语言模型时存在的校准问题，尤其是模型因模仿标注者（labeler）所使用但模型自身不具备的知识而产生“幻觉”现象的原因和机制。文章探讨了模型内部知识与标注者知识不匹配导致模型生成不准确内容的假设及其影响。
73. https://oreil.ly/A3K90: （https://arxiv.org/abs/2203.15556）该资源是 DeepMind 于 2022 年发布的一篇论文，题为“Training Compute-Optimal Large Language Models”，主要研究不同规模的大型语言模型在训练时所需的计算资源和训练数据量的最优配置。URL “https://arxiv.org/abs/2203.15556” 指向该论文的 arXiv 预印本页面。

74. https://oreil.ly/AkAyI: （https://www.theverge.com/2023/12/29/24018735/heres-how-major-media-companies-are-handling-openai）该资源是一篇发表于 2023 年 12 月的 The Verge 文章，内容介绍了主要媒体公司如何应对和处理与 OpenAI 合作及其使用媒体内容进行训练数据的问题，特别涉及媒体公司与 OpenAI 之间达成的版权和数据使用协议。
75. https://oreil.ly/CSSed: （https://engineering.grab.com/llm-powered-data-classification）该 URL（https://engineering.grab.com/llm-powered-data-classification）指向 Grab 工程团队关于利用大型语言模型（LLM）驱动的数据分类技术的介绍或技术文章。内容可能聚焦于 Grab 如何应用 LLM 及其相关策略（如“best of N”方法）来提升数据分类的效果，尤其是在无需强化学习、仅依靠奖励模型评分来选择最优输出的实际应用案例和经验分享。
76. https://oreil.ly/D1S6y: （https://arstechnica.com/tech-policy/2023/07/chatgpts-user-base-shrank-after-openai-censored-harmful-responses/）该资源是一篇来自 Ars Technica 的文章，标题大致为“ChatGPT 的用户数量在 OpenAI 审查有害内容后出现下降”。文章讨论了 OpenAI 对 ChatGPT 进行内容审查和过滤可能带来的负面影响，尤其是过度审查如何导致模型变得“无趣”并使部分用户流失，反映了 AI 在处理敏感和有争议话题时面临的平衡挑战。
77. https://oreil.ly/F76hq: （https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model）该资源是 Google 官方博客中关于“Med-PaLM 2”模型的介绍页面，详细说明了 Google 如何将大型语言模型（LLM）与医学数据相结合，开发出专门用于医疗领域的语言模型 Med-PaLM 2，以提升医疗问答的准确性和专业性。文章可能涵盖该模型的训练方法、应用场景及其在医疗生命科学领域的重要意义。
78. https://oreil.ly/FCyyA: （https://www.cbsnews.com/news/chatgpt-judge-fines-lawyers-who-used-ai/）该资源是一则新闻报道，内容涉及 2023 年 6 月一家律师事务所因在法庭提交了由 ChatGPT 生成的虚假法律研究材料而被法官

罚款的事件，反映了 AI 模型（如 ChatGPT）在生成法律文本时可能出现“幻觉”或不实信息的问题及其法律风险。

79. https://oreil.ly/Fqo2S: （https://www.youtube.com/watch?v=hhiLw5Q_UFg）该资源是一段 YouTube 视频，内容可能是 John Schulman 关于大型语言模型（LLMs）认知能力及其“幻觉”问题的讲解，尤其涉及通过检索信息源和强化学习方法来减少模型生成错误信息的技术方案。
80. https://oreil.ly/G13KM: （https://www.artfish.ai/p/gpt4-project-euler-many-languages）该 URL 指向的资源是一个关于“GPT-4 在多种编程语言中解决 Project Euler 数学题目”的项目或页面，展示了 GPT-4 如何用多语言能力来处理数学问题，可能涉及对模型在不同语言环境下的表现分析和示例代码。
81. https://oreil.ly/GxSe0: （https://www.youtube.com/watch?v=7xij6SoCClI）该 URL 指向的是 YouTube 上的一个视频，内容很可能与人工智能（AI）领域的发展和前景有关，特别是围绕“大规模 AI 模型”的讨论。根据上下文中提到的 Anthropic CEO Dario Amodei 关于“规模假说”和“价值高达 1000 亿美元的 AI 模型能够达到诺贝尔奖水平”的观点，可以推断该视频可能包含 Dario Amodei 或相关专家对 AI 模型规模、能力及其潜在影响的讲解或访谈。
82. https://oreil.ly/HcFYz: （https://www.nvidia.com/en-us/data-center/h100/）该资源是 NVIDIA 官方关于其 H100 GPU 的介绍页面，详细说明了 H100 在数据中心应用中的高性能计算能力，特别是在浮点运算性能（如 TeraFLOP/s）方面的技术规格和优势。
83. https://oreil.ly/JX37g: （https://deepmind.google/technologies/alphafold/）该资源是 DeepMind 开发的 AlphaFold，一个基于深度学习的模型，专门用于预测蛋白质的三维结构。AlphaFold 通过训练包含约 10 万个已知蛋白质序列及其 3D 结构的数据集，能够高精度地推断蛋白质的空间构象，极大推动了生物学和药物研发领域的研究进展。
84. https://oreil.ly/KLVgX: （https://openai.com/index/hello-gpt-4o/）该 URL（https://openai.com/index/hello-gpt-4o/）所指向的资源，推断为 OpenAI 官网上的一个介绍或索引页面，内容可能聚焦于 GPT-4o 模型（GPT-4

的某个版本或变体），包括其功能、应用领域及训练数据特点。页面可能围绕 GPT-4o 在多领域（如编程、法律、科学、商业等）表现的能力展开说明，类似于上下文中提及的通用基础模型（如 Gemini、GPTs、Llamas）及其训练数据覆盖范围。

85. https://oreil.ly/LcBfx: （https://www.newsguardtech.com/special-reports/chatgpt-generates-disinformation-chinese-vs-english/）该资源是一篇由 NewsGuard 发布的特别报告，内容聚焦于 ChatGPT 在生成中英文内容时的差异，特别指出 ChatGPT 在生成中文（包括简体和繁体）时更容易产生误导性信息，而在英文内容中相对较少出现此类问题。报告通过实验证明了这一现象，并探讨了可能的原因，如预训练数据或模型调校过程中存在的语言偏差。
86. https://oreil.ly/M1Nsc: （https://blogs.nvidia.com/blog/bionemo-large-language-models-drug-discovery/）该资源介绍了 NVIDIA 的 BioNeMo 模型，这是一种专注于生物分子数据的大型语言模型，旨在推动药物发现领域的发展。文章可能详细阐述了 BioNeMo 如何利用生物医学数据进行训练，从而提升在药物研发相关任务中的表现和应用潜力。
87. https://oreil.ly/MTqyR: （https://laion.ai/blog/laion-5b/）该 URL （https://laion.ai/blog/laion-5b/）指向的是 LAION 发布的关于 LAION-5B 数据集的博客文章。LAION-5B 是一个大规模的公开视觉-语言数据集，包含数十亿对图像与文本对，用于训练和评估多模态模型如 CLIP 等。该资源详细介绍了数据集的规模、构建方法、数据分布及其在不同视觉任务和基准测试中的表现，旨在为研究者提供丰富的训练数据和分析工具，推动计算机视觉与自然语言处理领域的发展。
88. https://oreil.ly/NxZDF: （https://platform.openai.com/docs/guides/text-generation/json-mode）该资源为 OpenAI 官方文档中关于文本生成 API 的“JSON 模式”指南，介绍了如何通过该模式生成结构化的 JSON 格式输出，及其工作原理和使用注意事项。
89. https://oreil.ly/ON55d: （https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/）该资源是一篇由 Tim Dettmers 撰写的文章，比较了 Tensor Processing Units (TPUs) 和 Graphics Processing Units

(GPUs) 在运行基于 Transformer 架构（如 BERT）模型时的性能表现与效率，详细分析了两者在深度学习中的优劣及适用场景。

90. https://oreil.ly/OisOs: （https://www.newsguardtech.com/solutions/news-reliability-ratings/）该资源是 NewsGuard 提供的“新闻可靠性评级”服务页面，介绍其对媒体网站的信任度评估体系，用于衡量和标注新闻来源的可信程度，帮助用户识别假新闻和不可靠的信息来源。
91. https://oreil.ly/P1MPQ: （https://www.youtube.com/watch?v=CzR3OrOkM9w）该 URL（https://www.youtube.com/watch?v=CzR3OrOkM9w）所指向的资源很可能是一段视频内容，内容与大型语言模型（Large Language Models, LLMs）的评估、比较或相关研究有关，可能包含对比不同模型响应、人工对比过程的介绍，或者是 Llama-2 模型作者 Thomas Scialom 在相关主题上的讲解或访谈。该视频有助于理解手工对比模型回答时的时间和成本消耗。
92. https://oreil.ly/R_uvq: （https://arxiv.org/pdf/2110.14168.pdf）该资源是一篇学术论文，题为《Verifiers for Neural Networks》，由 Cobbe 等人于 2021 年发表，内容主要探讨了通过训练验证器（verifiers）来辅助神经网络模型选择最佳数学问题解法，从而显著提升模型性能。论文中提出的验证器技术能够带来相当于模型规模提升 30 倍的性能增益，具体表现为一个 1 亿参数的模型配合验证器，其性能可媲美一个 30 亿参数的模型。该论文详述了验证器的设计方法、实验结果及其在数学问题求解中的应用效果。
93. https://oreil.ly/SF_X9: （https://arxiv.org/pdf/2203.02155.pdf）该资源是一篇学术论文，标题可能与人工智能模型训练中的数据标注和示范数据生成相关，具体探讨了高学历标注者在生成训练示范数据中的作用及其成本。论文详细分析了用于训练大型语言模型（如 InstructGPT）的标注过程、时间消耗和经济成本，旨在为理解和优化训练数据生成提供理论和实践依据。
94. https://oreil.ly/SfB4g: （https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2）该资源是一个名为 RedPajama-Data-V2 的开源大型文本数据集，包含约 30 万亿个标记（tokens），相当于 4.5 亿本书的规模，远超维基百科的数据量。该数据集由 Together 组织发布，旨在为自然语言处理和

大规模语言模型训练提供海量文本数据，但由于内容未经过严格筛选，高质量数据比例相对较低。

95. https://oreil.ly/St1o8: （https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/）该资源是一篇发表于 2023 年《华盛顿邮报》（Washington Post）的互动报道，标题为“Inside the Secret List of Websites That Make AI like ChatGPT Sound Smart”，内容探讨了支持类似 ChatGPT 等人工智能聊天机器人的背后网站和数据来源，揭示了这些 AI 如何通过访问特定网站列表来提升其语言理解和生成能力。
96. https://oreil.ly/TpaGg: （https://openai.com/research/openai-baselines-ppo）该资源是 OpenAI 发布的关于“Proximal Policy Optimization (PPO)”算法的研究和实现介绍页面。PPO 是一种用于强化学习的策略优化方法，广泛应用于训练智能体以最大化奖励函数。该页面详细说明了 PPO 算法的原理、应用及其在强化学习中的优势。
97. https://oreil.ly/VAUl6: （https://cookbook.openai.com/examples/using_logprobs）该资源介绍了“logprobs”（对数概率）的概念及其在语言模型中的应用，重点说明了为何使用对数概率能够有效缓解数值下溢（underflow）问题。具体内容包括对数概率的定义、其在神经网络概率计算中的优势，以及举例说明语言模型处理大词汇量时概率值过小导致的数值表示困难。
98. https://oreil.ly/VvXbu: （https://mistral.ai/news/mixtral-of-experts/）该资源介绍了“Mixtral of Experts”模型，这是一种由多个专家模型（experts）组成的混合模型架构。具体来说，Mixtral 8x7B 由八个专家组成，每个专家拥有 70 亿参数，理论上总参数量为 560 亿，但由于部分参数共享，实际参数量为 46.7 亿。该页面可能详细阐述了该模型的设计理念、参数配置及其优势。
99. https://oreil.ly/XYugZ: （https://platform.openai.com/docs/api-reference/completions/create#completions-create-best_of）该 URL 指向 OpenAI 官方文档中关于 API 接口“completions/create”的说明，具体介绍了参数“best_of”的用法。该参数允许用户指定生成多个（如 10 个）候选输出，API 将返回其中平均对数概率（logprob）最高的输出结果，从而提高生成文本的质量。

100. https://oreil.ly/ZTRaA: （https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product?_l=en_US）该资源是 LinkedIn 工程团队发布的一篇博客文章，题为“Musings on Building a Generative AI Product”（关于构建生成式人工智能产品的思考）。文章分享了在开发生成式 AI 产品过程中遇到的技术挑战和经验，特别是如何通过简单且高效的后处理方法来纠正生成模型常见错误，从而提升输出质量。
101. https://oreil.ly/Zq5Sw: （https://www.artfish.ai/p/all-languages-are-not-created-tokenized）该资源是一篇关于不同语言在自然语言处理中的“分词”（tokenization）效率差异的文章，重点讨论了不同语言因词汇结构和文字特性导致的分词长度差异，从而影响模型推理的速度和成本。文章通过引用如 GPT-4 在多语言数据集（MASSIVE）上的基准测试，说明某些语言（如缅甸语和印地语）相比英语需要更多的 tokens 来表达相同含义，进而影响模型的性能表现。
102. https://oreil.ly/a6j-N: （https://huggingface.co/croissantllm）该资源是一个名为“CroissantLLM”的法语语言模型，托管在 Hugging Face 平台上，旨在支持和推动法语的自然语言处理应用。
103. https://oreil.ly/ah9MT: （https://openreview.net/forum?id=SkxJ-309FQ）该 URL（https://openreview.net/forum?id=SkxJ-309FQ）指向的是一篇关于生成模型中“幻觉”（hallucination）现象的学术论文或学术讨论页面。该资源可能详细探讨了生成式模型（如大型语言模型）出现幻觉的原因，分析了幻觉在文本生成中的表现，并且可能涉及相关检测和评估方法的研究，是理解生成模型幻觉问题的重要学术参考资料。
104. https://oreil.ly/cg0JY: （https://aclanthology.org/C16-1103/）该资源是 2016 年由 Goyal 等人发表的一篇学术论文，首次在自然语言生成领域系统性地讨论了文本生成中的“幻觉”（hallucination）现象。该论文为后续关于生成模型中幻觉检测与评估的研究奠定了基础。链接指向的页面属于 ACL Anthology，收录了该论文的详细信息及全文。
105. https://oreil.ly/fb1aR: （https://research.google/blog/a-neural-network-for-machine-translation-at-production-scale/）该资源介绍了谷歌在机器翻译领域应用神经网络技术的实践经验，特别是在生产环境中部署基于神经网络的

机器翻译系统。文章详细阐述了谷歌如何利用先进的序列到序列（seq2seq）架构及其改进（如 Transformer 架构），实现大规模、高质量的机器翻译服务。

106. https://oreil.ly/gGwRz: （https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf）该资源是一篇由 OpenAI 发布的论文，题为《Language Models are Unsupervised Multitask Learners》（语言模型是无监督的多任务学习者）。论文详细介绍了 GPT-2 模型的训练方法、架构设计及其在多种任务中的表现，强调通过大规模无监督学习，语言模型能够在多任务环境中展现出强大的泛化能力。
107. https://oreil.ly/h9oG6: （https://huggingface.co/datasets/Anthropic/hh-rlhf）该资源是 Anthropic 发布的 HH-RLHF 数据集，包含用于强化学习与人类反馈（Reinforcement Learning with Human Feedback, RLHF）训练的比较数据。
108. https://oreil.ly/hNRf4: （https://blog.dottxt.co/how-fast-cfg.html）该资源是一篇博客文章，标题可能为“How Fast CFG”，内容涉及上下文无关文法（CFG）在生成过程中的应用，特别是关于构建语法并将其整合到采样过程中的挑战。文章可能讨论了针对不同输出格式（如 JSON、YAML、正则表达式、CSV 等）需要各自语法的问题，约束采样方法的局限性，以及语法验证对生成速度的影响。
109. https://oreil.ly/iJG1q: （https://huyenchip.com/2023/05/02/rlhf.html#phase_1_pretraining_for_completion）该资源是关于“RLHF（基于人类反馈的强化学习）”的详细介绍，特别聚焦于模型训练的第一阶段——“completion 预训练”阶段。页面内容可能涵盖 RLHF 的基本原理、训练流程、技术细节及其在现代大语言模型（如 GPT-3.5 和 Llama 系列）中的应用示例。
110. https://oreil.ly/iYh-B: （https://multithreaded.stitchfix.com/blog/2023/03/06/expert-in-the-loop-generative-ai-at-stitch-fix/）该 URL 指向的资源是一篇由 Stitch Fix 官方博客发布的技术文章，标题大致为“Expert in the Loop: Generative AI at Stitch Fix”。文章介绍了 Stitch Fix 在生成式人工智能领域的应用实践，特别是他们如何利用“best of N”策略——即通过生成多个

模型输出，并利用奖励模型对这些输出进行评分，从而选出最优结果，来提升模型性能。该方法避免了复杂的强化学习步骤，而是依赖奖励模型进行有效的输出筛选，体现了 Stitch Fix 在 AI 模型优化方面的创新思路和实际经验。

111. https://oreil.ly/iqhNY:（https://platform.openai.com/tokenizer）该资源是 OpenAI 提供的“Tokenizer”工具，用户可以通过它对文本进行分词和编码，帮助验证模型输入的分词结果，用于分析和理解语言模型如何处理不同语言的文本。
112. https://oreil.ly/j4wwW:（https://www.youtube.com/live/AKMuA_TVz3A）该资源是 Ilya Sutskever 在 2023 年伯克利 Simons 研究所的一场讲座视频，内容围绕神经网络架构的创新难点及其背后的理论解释，特别探讨了梯度下降作为一种搜索算法，如何在现有架构能模拟的程序空间中寻找最优解，以及新架构必须能够模拟现有架构无法模拟的程序，才能实现性能突破。
113. https://oreil.ly/jWEsP:（https://platform.openai.com/docs/api-reference/chat/create#chat-create-logprobs）该资源是 OpenAI 官方文档中关于聊天模型接口（Chat API）中“logprobs”参数的说明页面，详细介绍了如何在调用聊天模型时获取生成文本中各个候选词的对数概率（log probabilities），以及该功能的使用限制和变更历史。
114. https://oreil.ly/kO41d:（https://ai.meta.com/research/publications/beyond-neural-scaling-laws-beating-power-law-scaling-via-data-pruning/）该资源是 Meta 发布的一篇研究论文，标题为“Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning”。论文探讨了神经网络性能提升中面临的“最后一公里”难题，指出在模型准确率提升到一定阶段后，进一步提高性能所需的计算资源和数据量呈现幂律增长。文章提出通过数据剪枝（data pruning）的方法，能够有效突破传统的幂律扩展规律，从而在减少数据和计算成本的同时提升模型性能。
115. https://oreil.ly/kYtBG:（https://arxiv.org/pdf/2203.02155.pdf）该资源是指向论文《InstructGPT: Improving Language Models by Following Instructions》的 PDF 文件，详细介绍了 OpenAI 通过人工标注和奖励模型训练，使语言模型更好地理解和执行用户指令的方法和实验过程。文中包含了标注

界面设计、标注员评分机制及其一致性分析等内容，展示了 InstructGPT 的训练流程和性能提升。

116. https://oreil.ly/kuG3J: （http://lukemetz.com/difficulty-of-extrapolation-nn-scaling/）该资源是 Luke Metz 于 2022 年发表的一篇优秀博客文章，题为“On the Difficulty of Extrapolation with NN Scaling”，主要探讨神经网络在规模扩展过程中外推（extrapolation）能力的难点，特别是与模型规模和训练数据量相关的 emergent abilities（新兴能力）如何影响模型性能的外推准确性。
117. https://oreil.ly/l21nr: （https://engineering.grab.com/llm-powered-data-classification）该资源是一篇由 Grab 工程团队发布的技术文章，介绍了如何利用大语言模型（LLM）实现数据分类的技术方案，重点讲述了通过奖励模型（reward model）对输出结果进行评分，从而优化数据选择和应用性能的方法。文章内容涉及 Grab 在实际应用中使用奖励模型提升系统效果的经验和技术细节。
118. https://oreil.ly/n7wYO: （https://arxiv.org/pdf/2312.00752）该资源是论文《Mamba: Linear-Time Sequence Modeling with Selective State Spaces》（作者 Gu 和 Dao，2023 年），介绍了一种名为 Mamba 的新型序列建模架构。Mamba 架构将选择性状态空间模型（SSMs）扩展到三十亿参数规模，在语言建模任务中，Mamba-3B 不仅超越了同规模的 Transformer 模型，还能达到规模是其两倍的 Transformer 模型的性能。该论文还展示了 Mamba 推理计算的时间复杂度为线性，优于 Transformer 的二次方复杂度，并且在处理百万长度序列的真实数据时表现出显著的性能提升。
119. https://oreil.ly/o7WB3: （https://www.redditinc.com/policies/data-api-terms）该 URL（https://www.redditinc.com/policies/data-api-terms）指向的是 Reddit 关于其数据 API 使用的条款和政策页面。该页面详细说明了 Reddit 对通过其 API 访问和使用数据的规定，特别是针对数据抓取和用于训练机器学习模型的限制，旨在规范第三方对 Reddit 数据的获取和利用，防止未经授权的大规模数据抓取行为。

120. https://oreil.ly/oq-LE: （https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf）该资源是由斯坦福大学人机共融研究所（Stanford HAI）发布的《2022 年人工智能指数报告》（2022 AI Index Report）的完整 PDF 文档。该报告汇集了人工智能领域的最新数据和趋势分析，涵盖模型性能、计算资源消耗、技术进展及其社会影响等方面，是评估和理解人工智能发展现状的重要权威资料。
121. https://oreil.ly/qK2Ap: （https://openai.com/research/gpt-4）该资源是 OpenAI 官网上关于 GPT-4 研究的介绍页面，详细展示了 GPT-4 模型在多语言理解和表现上的研究成果，特别是在 MMLU 基准测试中，GPT-4 在英语任务上表现优异，显著优于资源较少的语言，如泰卢固语。
122. https://oreil.ly/sHwbw: （https://www.microsoft.com/en-us/research/blog/%C2%B5transfer-a-technique-for-hyperparameter-tuning-of-enormous-neural-networks/）该资源是一篇由微软研究院发布的博客文章，介绍了一种称为“μTransfer”的技术，该技术用于对超大规模神经网络进行超参数调优。文章重点讲述了通过对较小模型上超参数影响的研究，推断和预测在更大模型上表现最佳的超参数组合，从而实现超参数的“迁移”或“扩展外推”，有效提升了大规模模型训练的效率和性能。该技术解决了传统超参数调优在巨大模型上计算成本高、尝试次数有限的问题。
123. https://oreil.ly/sljei: （https://community.openai.com/t/how-to-build-a-conversation-classifier-with-the-new-fine-tuning-job-api-and-gpt3-5/341075）该资源是 OpenAI 社区论坛上的一篇帖子，介绍了如何使用新的微调作业 API 和 GPT-3.5 模型构建对话分类器的具体方法和步骤。
124. https://oreil.ly/tbgTi: （https://openai.com/index/chatgpt/）该资源（https://openai.com/index/chatgpt/）很可能是关于 ChatGPT 模型的官方介绍或文档页面，内容涉及 ChatGPT 的技术细节和训练方法，特别是与“Preference finetuning”（偏好微调）相关的说明，包括通过强化学习（如 RLHF）优化模型以更好地符合人类偏好的技术和实践。
125. https://oreil.ly/uG27L: （https://huggingface.co/collections/inceptionai/jais-family-66add8bb9c381f5492ddb6f4）该资源是 Hugging Face 平台上

关于“Jais”模型家族的一个集合页面，收录了多种针对阿拉伯语等非英语语言优化的开源语言模型，旨在支持和推动多语言自然语言处理技术的发展。

126. https://oreil.ly/uyiBH: （https://huggingface.co/ai21labs/Jamba-v0.1）该资源是“Jamba”语言模型的官方页面，Jamba 是一种混合架构模型，结合了 Transformer 层和 Mamba 层，用于扩展状态空间模型（SSM）的规模。该模型由 Lieber 等人在 2024 年提出，具有高效的长上下文处理能力（最长可达 256K 个标记），并且内存占用较小。页面可能提供模型权重、使用说明及相关技术细节。
127. https://oreil.ly/vfSQw: （https://ai.meta.com/blog/meta-llama-3/）该资源是 Meta 发布的关于 Llama 3 模型的官方博客文章，介绍了 Llama 3 所使用的训练数据规模（15 万亿个标记）及其相关技术细节。
128. https://oreil.ly/wf2Lw: （https://commoncrawl.org/）该资源是一个由非营利组织维护的互联网网页抓取项目，定期大规模爬取互联网上的网页数据，提供包含数十亿网页内容的公开数据集，常被用于机器学习和自然语言处理等领域的训练数据来源。
129. https://oreil.ly/xNuju: （https://policies.stackoverflow.co/teams/enterprise-cloud-business/）该 URL（https://policies.stackoverflow.co/teams/enterprise-cloud-business/）所指向的资源是 Stack Overflow 针对其企业云业务（Enterprise Cloud Business）团队用户的政策页面，主要说明其数据使用、访问权限及相关服务条款，特别涉及数据保护和禁止未经授权的数据抓取等规定，以保障企业客户数据的安全和合规使用。
130. https://x.com/OxfordDiplomat/status/1424388443010998277?lang=en: 该资源是一条来自 Twitter（现称 X）的推文，内容可能以幽默或调侃的方式表达某件事情发生的概率虽低但绝非零。
131. https://x.com/anthrupad/status/1622349563922362368: 该资源是一条由用户 anthrupad 在社交平台 X（原 Twitter）上发布的推文，内容可能包含一个关于“Shoggoth with a smiley face”（带有笑脸的 Shoggoth）的原始图像或相关视觉素材。

132. https://x.com/ibab/status/1733558576982155274: 该资源是一条来自 Igor Babuschkin（Grok 模型的核心开发者）在社交平台 X（前身为 Twitter）上的状态更新（推文）。在这条推文中，他回应了关于 Grok 模型训练数据来源的猜测，解释说 Grok 是基于网络数据训练的，而网络上充斥着大量由 ChatGPT 生成的内容，因此模型的训练数据中不可避免地包含了 AI 生成的数据。
133. https://x.com/jaschasd/status/1756930242965606582: 该资源是一条由研究人员 Jascha Sohl-Dickstein 在 X（原推特）上发布的推文，内容是一幅关于超参数效果的精美可视化图，展示了哪些超参数配置在模型训练中有效，哪些无效。
134. https://x.com/xuanalogue/status/1707757449900437984: 该资源是一条发布于 2023 年 9 月的推文，内容涉及 OpenAI API 对 logprobs 功能的调整，具体说明 OpenAI 自该月起不再支持获取任意用户提供文本的 logprobs，只保留对最多 20 个最可能 token 的 logprobs 显示。

第 3 章（83）

1. https://arxiv.org/abs/1301.3781: 该资源为论文《Efficient Estimation of Word Representations in Vector Space》，由 Tomas Mikolov 等人发表于 2013 年，介绍了一种高效的词向量表示学习方法，即 word2vec 模型，用于将词语映射到连续向量空间中，从而捕捉词语之间的语义关系，是自然语言处理领域中广泛使用的词嵌入技术之一。该论文发布于 arXiv 预印本平台。
2. https://arxiv.org/abs/1607.07326: 该资源是指向 arXiv 上的一篇学术论文，编号为 1607.07326。根据上下文，这篇论文与电商领域的解决方案相关，特别涉及如何为产品构建 embedding（嵌入）表示，用于提升推荐系统或搜索的效果。换言之，该论文介绍了利用机器学习方法将商品数据转化为低维向量表示的技术，以支持电商平台的个性化推荐和相关应用。

3. https://arxiv.org/abs/1709.00103: 该资源是指向论文《Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning》（Zhong et al., 2017），该论文提出了一种基于强化学习的方法，将自然语言转换为结构化的 SQL 查询，用于 text-to-SQL 任务。该工作还介绍了 WikiSQL 数据集，这是一个大规模的用于训练和评估文本到 SQL 转换模型的基准数据集。
4. https://arxiv.org/abs/1804.07461: 该资源是指向论文《GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding》，发布于 2018 年，介绍了 GLUE（General Language Understanding Evaluation）基准测试。GLUE 旨在为自然语言理解模型提供一个统一的、多任务的评测平台，涵盖多种语言理解任务，用以衡量和比较模型的综合性能。GLUE 发布后迅速成为评估基础语言模型的重要标准，但由于模型性能提升迅速，GLUE 基准很快达到饱和，推动了后续更具挑战性的评测基准如 SuperGLUE 的诞生。
5. https://arxiv.org/abs/1806.02643: 该资源是由 DeepMind 的 Balduzzi 等人在 2018 年发表的一篇学术论文，论文指出在人工智能领域中，虽然对算法开发的关注度不断提升，但对评估方法的系统性研究明显不足。作者强调实验结果通常仅用于改进算法，而很少用于改进评估本身，揭示了 AI 评估方法缺乏投入和重视的问题。论文还讨论了诸如 Elo 评分系统中“传递性假设”的局限性，指出人类偏好和 AI 模型评估中存在非传递性的现象，从而影响评估的准确性和可靠性。
6. https://arxiv.org/abs/1810.04805: 该资源是指向论文《Google’s BERT》的链接，BERT 是由 Google 提出的一种用于自然语言处理任务的预训练语言表示模型。论文详细介绍了 BERT 模型的结构、训练方法及其在多种 NLP 任务中的优异表现。
7. https://arxiv.org/abs/1904.09675: 该资源是论文《BERTScore: Evaluating Text Generation with BERT》，介绍了一种基于 BERT 生成的词嵌入来度量语义文本相似度的方法，称为 BERTScore。该方法通过比较句子中词语的上下文嵌入，实现更精细的语义匹配，用于评价文本生成任务的质量。

8. https://arxiv.org/abs/1905.00537: 该资源是 2019 年发布的论文“SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”，旨在提出一个比 GLUE 更具挑战性的自然语言理解基准测试套件 SuperGLUE，以更全面评估和推动通用语言理解模型的发展。
9. https://arxiv.org/abs/2004.04696: 该资源是论文《BLEURT: Learning Robust Metrics for Text Generation》，由 Sellam 等人在 2020 年发表，介绍了一种基于参考答案的自动评价指标 BLEURT。该指标通过输入生成的候选文本和参考文本对，输出两者之间的相似度分数，用于评估生成文本的质量。BLEURT 旨在提供比传统指标（如 BLEU）更鲁棒、与人工评价更一致的文本生成质量评估方法。
10. https://arxiv.org/abs/2009.03300: 该资源是 2020 年发布的 MMLU（Massive Multitask Language Understanding）基准测试的论文，MMLU 是一个用于评估大型基础语言模型多任务语言理解能力的强力基准，涵盖了广泛的学科和任务，旨在全面衡量模型的知识和推理能力。
11. https://arxiv.org/abs/2103.00020: 该资源为论文《Learning Transferable Visual Models From Natural Language Supervision》（Radford et al., 2021），介绍了 CLIP 模型，该模型首次实现了将文本和图像两种不同模态的数据映射到统一的嵌入空间，推动了跨模态表示学习的发展。
12. https://arxiv.org/abs/2104.08773: 该资源是指向 2021 年发布的论文“NaturalInstructions”，该论文提出了一个用于评估基础模型（foundation models）多任务理解和执行能力的基准数据集。NaturalInstructions 通过收集和统一大量带有自然语言指令的任务，旨在捕捉模型在不同指令理解和泛化能力上的表现，是早期促进模型指令学习和通用能力评估的重要基准之一。随着模型能力提升，该基准后来被更全面的“Super-NaturalInstructions”所替代。
13. https://arxiv.org/abs/2107.03374: 该资源是一篇由 Chen 等人于 2021 年发表在 arXiv 上的学术论文，讨论了代码生成评测中使用的指标（如 BLEU 分数）与代码功能正确性之间的关系，指出高的词汇相似度分数并不总能反映代码的实际正确性，从而提醒优化评测指标时应关注功能性而非单纯的表面相似度。

14. https://arxiv.org/abs/2112.00861: 该资源是一篇由 Anthropic 于 2021 年发布的论文，首次提出在人工智能领域中使用比较评估方法来对不同模型进行排序。
15. https://arxiv.org/abs/2204.07705: 该资源是论文《Super-NaturalInstructions》，发布于 2022 年，旨在提供一个改进版的自然语言理解任务评测基准，以替代之前的 NaturalInstructions，帮助更全面地评估和推动基础模型（foundation models）的能力发展。
16. https://arxiv.org/abs/2206.11249: 该资源是指向论文《GEMv2》的链接，GEMv2 是一个用于自然语言生成（NLG）任务的基准测试集合，特别关注评估生成文本的质量，涵盖多种文本类型和评测指标。文章介绍了在基础模型兴起后，文本生成评测的新标准和方法，强调了多样化和综合性评价指标的重要性。
17. https://arxiv.org/abs/2210.03350: 该资源是一篇由 Press 等人于 2022 年发表在 arXiv 上的学术论文，主要探讨了利用模型自身进行自我评估（self-evaluation）或自我批判（self-critique）的方法。文章分析了这种方法在检测模型回答正确性、提高模型回答质量以及促使模型修正和改进自身回答中的作用，尽管存在自我偏差，但自我评估被视为一种有效的“理智检查”手段。
18. https://arxiv.org/abs/2210.07316: 该资源是论文《Massive Text Embedding Benchmark (MTEB)》的预印本，介绍了一个用于评估文本嵌入质量的多任务基准测试框架。该基准覆盖分类、主题建模、推荐系统和检索增强生成（RAG）等多种下游应用，旨在系统性地衡量和比较不同文本嵌入方法的表现。
19. https://arxiv.org/abs/2212.05171: 该资源是论文《ULIP: Unified Language-Image-Point Cloud Pre-training》，由 Xue 等人在 2022 年发布，提出了一种统一表示模型，旨在将文本、图像和三维点云数据映射到同一个联合嵌入空间，从而实现多模态数据的统一理解与融合。
20. https://arxiv.org/abs/2305.05665: 该资源是由 Girdhar 等人在 2023 年发表的论文，标题为“ImageBind”，其内容介绍了一种能够跨越六种不同模态（包括文本、图像和音频）学习联合嵌入表示的方法，旨在实现多模态数据的统一表示与融合。
21. https://arxiv.org/abs/2305.11738: 该资源是一篇由 Gou 等人于 2023 年发表在 arXiv 上的学术论文，主要探讨了基于模型自身进行自我评估（self-evaluation）

的方法。文章研究了如何利用模型对自身生成的回答进行判断和批评，从而作为一种“自我批评”或“自我问答”的技术，不仅用于验证回答的合理性（sanity checks），还能促进模型改进和回答质量的提升。

22. https://arxiv.org/abs/2306.05087: 该资源为论文《PandaLM: Predicting Human Preferences with Large Language Models》（Wang et al., 2023），介绍了一种基于大型语言模型的偏好预测模型 PandaLM。该模型输入（提示、回应 1、回应 2），输出用户更倾向的回应，并且能够给出判断理由。PandaLM 旨在通过自动预测人类偏好，简化评价过程并提升模型的安全性和可靠性，解决人类偏好数据稀缺且昂贵的问题。
23. https://arxiv.org/abs/2306.05685: 该资源是 Zheng 等人于 2023 年发布在 arXiv 上的一篇研究论文，论文内容主要探讨了利用大型语言模型（如 GPT-4）作为 AI 评判者（AI judges）对生成文本进行评价的效果。研究发现，GPT-4 在其设计的 MT-Bench 评测基准上，与人类评审者的评价一致率达到了 85%，甚至超过了人类评审者之间的 81% 一致率。论文还分析了通过在提示中加入更多评估示例来提升评判一致性的策略（从 65% 提升到 77.5%），但也指出这种做法会导致推理成本显著增加。此外，研究揭示了 AI 评判者存在“自我偏见”（self-bias），即模型倾向于给自己生成的回答更高的评分。最后，论文探讨了强模型与人类偏好相关性更强的现象，并提出了专门化、小型评判模型作为未来研究方向的潜力。
24. https://arxiv.org/abs/2307.03025: 该资源是 Wu 和 Aji 于 2023 年发表在 arXiv 上的一篇论文，研究了 AI 评判模型（如 GPT-4 和 Claude-1）在回答质量评估中存在的“冗长偏好”（verbosity bias）现象，即这些模型倾向于偏好较长但包含事实错误的回答，而非较短但正确的回答。论文深入分析了这种偏好对 AI 回答质量评价的影响。
25. https://arxiv.org/abs/2307.03109: 该资源是由 Chang 等人于 2023 年发布在 arXiv 上的一篇论文，内容涉及大规模语言模型（LLMs）评测论文的趋势分析。
26. https://arxiv.org/abs/2307.09288: 该资源是由 Touvron 等人在 2023 年发表的一篇论文，介绍了 Llama 2 模型及其相关研究工作。论文中讨论了当模型的表现超越最佳人类标注者能力时，虽然人类难以对单个输出进行具体评分，但通过比

较两个输出的差异，仍能为模型提供有价值的反馈，体现了比较评估方法在先进语言模型中的应用和重要性。

27. https://arxiv.org/abs/2310.08118: 该资源是由 Valmeekamet 等人于 2023 年发表的一篇论文，探讨了利用模型自身进行自我评估（self-evaluation）或自我批判（self-critique）的方法，以提升模型回答的准确性和可靠性。论文研究了通过让模型判断自身回答的正确性，作为一种“理智检查”（sanity check）手段，以及促使模型修正和改进回答的技术。这种方法有助于缓解模型的自我偏差问题，并推动模型生成更优质的结果。
28. https://arxiv.org/abs/2310.08491: 该资源是 Kim 等人于 2023 年发布在 arXiv 上的一篇论文，介绍了一种名为 Prometheus 的参考基准评判模型。该模型通过输入（提示语、生成的回答、参考回答及评分标准），输出一个 1 到 5 之间的质量分数，假设参考回答的质量得分为 5。Prometheus 旨在提供一种更加细致和明确的自动化评价方法，用于衡量生成回答相对于参考回答的质量。
29. https://arxiv.org/abs/2310.17631: 该资源是由 Zhu 等人在 2023 年发布的论文，介绍了 JudgeLM——一种用于构建偏好模型（preference models）的系统。JudgeLM 能够根据给定的提示（prompt）和两个回答，判断哪一个回答更符合用户偏好，从而在 AI 模型的评估与安全性提升方面发挥重要作用。该工作属于通过预测人类偏好来对齐和优化 AI 模型的研究方向。
30. https://arxiv.org/abs/2311.06720: 该资源是由 Google 于 2023 年发布的一篇论文，介绍了名为 Cappy 的奖励模型。Cappy 用于对给定的（提示，回答）对进行评分，输出一个介于 0 到 1 之间的分数，表示回答的正确程度。该模型参数规模为 3.6 亿，较一般通用基础模型小巧，适合作为强化学习中基于人类反馈（RLHF）的轻量级评分器。
31. https://arxiv.org/abs/2311.17295: 该资源是一篇由 Boubdir 等人撰写的论文，发表于 arXiv，主要讨论了在 AI 模型评估中使用 Elo 评分系统时所依赖的传递性假设的局限性。论文指出，人类偏好不一定满足传递性，且由于不同模型对在不同评估者和不同提示下进行评测，非传递性现象在 AI 评价中普遍存在，从而影响 Elo 系统的有效性。

32. https://arxiv.org/abs/2312.00886: 该资源为一篇由 Munos 等人撰写的学术论文，收录于预印本平台 arXiv，论文编号为 2312.00886。该文探讨了在人工智能模型评估中，Elo 评分体系所依赖的传递性假设的合理性，指出人类偏好不一定满足传递性，且由于不同模型对在不同评估者和提示下的比较，可能导致非传递性的出现，从而质疑了传统 Elo 方法在 AI 评价中的适用性和局限性。
33. https://arxiv.org/abs/2404.04475: 该资源是由 Dubois 等人于 2023 年发布在 arXiv 上的论文，题为“AlpacaEval”。论文研究了 AI 评测系统（AI judges）的性能，发现其评估结果与人类评审高度一致，在 LMSYS 的 Chat Arena 排行榜（由人类评审）中表现出接近完美（0.98）的相关性，说明 AI 评测在某些任务中能够有效替代或辅助人类评审。
34. https://arxiv.org/abs/2406.01574: 该资源是 2024 年发布的论文，介绍了“MMLU-Pro”，这是对原有“MMLU”基准测试（2020 年）的升级版本。MMLU-Pro 旨在作为一个更先进、更全面的评估基准，用于测试基础模型（foundation models）的多任务语言理解能力，解决了原有 MMLU 基准因模型性能提升而快速饱和的问题，从而更有效地衡量和推动大型语言模型的能力发展。
35. https://en.wikipedia.org/wiki/E_(mathematical_constant): 该资源是维基百科中关于数学常数“e”（自然对数的底数）的条目，介绍了常数 e 的定义、性质及其在数学和科学中的应用。
36. https://en.wikipedia.org/wiki/Likert_scale: 该资源是维基百科上关于“李克特量表”（Likert scale）的条目，介绍了一种常用于问卷调查和社会科学研究中的量化测量工具，用于评估受访者对特定陈述的态度或意见，通常通过一系列从“非常不同意”到“非常同意”的等级选项来进行评分。
37. https://en.wikipedia.org/wiki/Unit_testing: 该资源介绍了“单元测试”（unit testing），即软件工程中通过在不同场景下执行代码以验证其功能正确性的一种自动化测试方法。单元测试能够确保代码按预期产生输出，是代码验证和质量保障的重要手段。
38. https://github.com/UKPLab/sentence-transformers: 该资源是一个开源项目“Sentence Transformers”，托管在 GitHub 上，旨在提供用于生成文本嵌入

（embedding）的预训练模型和工具，方便用户将句子或文档转换为向量表示，以支持自然语言处理任务中的语义搜索、文本相似度计算等应用。

39. https://github.com/chiphuyen/aie-book: 该资源是一个 GitHub 代码仓库，关联于一本书，内容涉及通过历史比赛结果进行模型排名的分析方法，旨在预测未来比赛结果。仓库中包含该书的相关分析代码和资料。
40. https://github.com/explodinggradients/ragas/blob/b276f59c0d4eb4795dc28966bfbce14d5aacd140/src/ragas/metrics/_faithfulness.py#L93C1-L94C1: 该资源位于 GitHub 仓库 “explodinggradients/ragas” 的文件 src/ragas/metrics/_faithfulness.py 中，具体是第 93 到 94 行的代码。根据文件路径和文件名推断，该代码片段很可能与“faithfulness”（可信度、忠实度）相关的度量指标实现有关，属于 Ragas 项目中的指标计算模块。简而言之，该资源是 Ragas 项目中实现某种“faithfulness”指标（评估模型解释或特征重要性可信度）的 Python 代码片段。
41. https://github.com/google-research/bleurt/issues/1: 该资源是 Google Research 团队维护的 BLEURT 项目在 GitHub 上的问题（Issue）讨论页面，具体内容涉及 BLEURT 评分范围的说明和相关疑问，讨论该评分范围为何大致在-2.5 到 1.0 之间，以及由此引发的 AI 评分标准的不确定性问题。
42. https://github.com/google-research/google-research/tree/master/mbpp: 该资源是 Google 发布的 MBPP（Mostly Basic Python Problems Dataset）数据集，用于评估人工智能模型在代码生成任务中的功能正确性表现，特别侧重于基础的 Python 编程问题。
43. https://github.com/mlflow/mlflow/blob/5cdae7c4321015620032d02a3b84fb6127247392/mlflow/metrics/genai/prompts/v1.py: 该 URL 指向的是 MLflow 项目中的一个 Python 源码文件，路径为 mlflow/metrics/genai/prompts/v1.py。结合路径信息可以推断，该文件可能与“生成式人工智能（GenAI）”相关的度量指标（metrics）功能有关，具体可能涉及“prompts”（提示词）的定义或处理，且属于版本 v1。总体来说，该资源是 MLflow 中用于管理或处理生成式 AI 提示词相关指标的 Python 脚本文件。

44. https://github.com/run-llama/llama_index/blob/main/llama-index-core/llama_index/core/evaluation/faithfulness.py: 该 URL 指向 GitHub 上 LlamaIndex 项目中的一个 Python 源码文件，路径为 llama-index-core/llama_index/core/evaluation/faithfulness.py。根据文件路径和名称推断，该文件主要负责实现与“faithfulness”（忠实度）评估相关的核心功能，可能包含用于衡量模型生成内容与原始信息一致性或准确性的评估方法和工具。
45. https://huyenchip.com/llama-police: 该资源是一个由作者维护的网站或页面，专门用于收集和整理与大型语言模型（LLM）、GPT、生成式模型和 Transformer 相关的热门 GitHub 代码库的信息，尤其聚焦于 AI 模型的评测（evaluation）资源。页面内容包括前 1000 个 AI 相关的高星标（star）GitHub 仓库的分析和统计，且通过关键词搜索和众包方式不断完善和更新相关仓库的列表。
46. https://oreil.ly/-0Iq1: （https://fortune.com/2023/06/23/lawyers-fined-filing-chatgpt-hallucinations-in-court/）该资源是一篇发表于 2023 年 6 月 23 日的 Fortune 报道，内容涉及律师因在法庭上提交由 ChatGPT 产生的虚假信息（“幻觉”）而被罚款的事件，反映了人工智能生成内容在法律实践中可能引发的风险和责任问题。
47. https://oreil.ly/0Cfcw: （https://openai.com/research/clip）该资源是 OpenAI 关于 CLIP 模型的研究页面，介绍了 CLIP（Contrastive Language–Image Pre-training）模型的相关内容，包括其设计理念、技术细节及应用示例。
48. https://oreil.ly/2XDI0: （https://www.interconnects.ai/p/evaluating-open-llms）该资源是一篇关于评估开源大型语言模型（Open LLMs）的文章，重点探讨了 AI 模型在对比判断中存在的“首位偏见”（first-position bias），即模型倾向于偏好列表或选项中的第一个答案。文章还提及通过多次测试不同排序和设计提示词的方法来缓解这种偏见，同时对比了人类在判断时更倾向于“末位偏见”（recency bias）的现象。
49. https://oreil.ly/2oEO1: （https://mlflow.org/docs/latest/python_api/mlflow.metrics.html#generative-ai-metrics）该资源是 MLflow 官方文档中关于“mlflow.metrics”模块的页面，介绍了如何使用 MLflow 的 Python API

来记录和管理机器学习模型的评估指标，特别包括生成式人工智能（Generative AI）相关的指标功能。

50. https://oreil.ly/4KJQM: （https://mathstodon.xyz/@tao/113132502735585408）该资源是一则发布在数学社交平台 Mathstodon 上，由菲尔兹奖得主陶哲轩（Terrence Tao）发表的内容，讨论了他对 OpenAI 2024 年 9 月发布的 GPT-o1 模型的评价，将其比作“一个水平平庸但不完全无能的研究生”，并推测未来一两代模型可能达到“称职研究生”的水平。
51. https://oreil.ly/57jOL: （https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in）该资源是微软官方文档页面，介绍了 Azure AI Studio 中内置的模型评估指标，帮助用户理解和使用各种评价指标来衡量和优化 AI 模型的性能。
52. https://oreil.ly/5T3ey: （https://docs.ragas.io/en/latest/concepts/metrics/index.html）该资源是 Ragas 项目的官方文档中关于“指标（Metrics）”概念的详细说明页面，介绍了如何定义、使用和管理指标，以支持性能监控和数据分析。
53. https://oreil.ly/7Fkh-: （https://blog.langchain.dev/langchain-state-of-ai-2023/）该资源是 LangChain 在 2023 年发布的《State of AI》报告，内容涉及人工智能领域的最新发展趋势，特别强调了“AI 作为评判者”在 AI 模型评估中的广泛应用和重要性。报告提供了数据支持，如 58% 的评估是在其平台上由 AI 评判完成，反映了该方法在实际生产环境中的普及和研究热度。
54. https://oreil.ly/92yRh: （https://paperswithcode.com/paper/findings-of-the-2014-workshop-on-statistical）该资源是 2014 年举办的关于统计机器学习的学术研讨会（workshop）的会议论文集，汇集了当时在统计机器学习领域的重要研究成果与讨论，涵盖模型评估指标、方法改进等内容，是理解统计学习早期发展和评估标准的重要参考资料。
55. https://oreil.ly/BNNNm: （https://txt.cohere.com/introducing-embed-v3/）该资源是关于 Cohere 最新发布的 Embed v3 技术的介绍页面，详细说明了 Embed v3 的功能、特点及其在文本嵌入（text embedding）中的应用。

56. https://oreil.ly/BO3-0:（https://paperswithcode.com/paper/microsoft-coco-captions-data-collection-and）该资源是关于微软 COCO（Microsoft COCO）图像描述数据集的说明文档，具体介绍了 COCO Captions 数据的收集和标注方法。该数据集广泛用于图像描述生成任务的训练与评测，是计算机视觉与自然语言处理领域的重要基准资源。文档中详细描述了数据采集流程、标注规范以及数据集的统计信息，帮助研究者理解和使用该数据集进行图像描述相关的模型开发与性能评估。
57. https://oreil.ly/CjYs9:（https://huggingface.co/datasets/openai_humaneval）该资源是 Hugging Face 平台上提供的“OpenAI HumanEval”数据集，主要用于评估人工智能模型的代码生成能力。HumanEval 包含一系列编程任务及其测试用例，评价模型生成代码的功能正确性，是 AI 代码生成领域的知名基准测试数据集。
58. https://oreil.ly/HjUlH:（https://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf）该资源是克劳德·香农（Claude Shannon）于 1951 年发表的关于信息熵（entropy）的原始论文或文献，详细阐述了信息熵作为统计参数，用以衡量语言文本中每个字符平均产生的信息量，以及如何以最有效的方式将语言转换为二进制数字以进行编码。简单来说，该文献是信息论领域的经典基础资料，介绍了熵的定义及其在信息编码中的应用。
59. https://oreil.ly/Hlkax:（https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in）该资源是微软 Azure AI Studio 官方文档中关于内置评估指标的说明页面，详细介绍了用于评估 AI 模型输出质量的各种评价标准（如相关性）、评分系统、示例及评分依据，帮助用户理解和运用这些指标进行模型效果评估。
60. https://oreil.ly/IOp9H:（https://arxiv.org/pdf/2310.10076）该资源是 2023 年发布在 arXiv 上的一篇学术论文，主要研究 AI 评判系统中的“冗长偏好”（verbosity bias）现象，即某些 AI 评判模型倾向于偏好较长的回答，即使这些回答可能包含事实错误。论文通过实验分析了不同版本的语言模型（如 GPT-4 与 GPT-3.5）在回答长度偏好上的表现差异，探讨该偏好随着模型能力提升而减弱的趋势。

61. https://oreil.ly/Loidb: （https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf）该 URL 指向的资源是一份由 OpenAI 于 2018 年发布的论文，标题为《Language Models are Unsupervised Multitask Learners》（语言模型是无监督多任务学习者）。这篇论文介绍了 GPT-2 模型的设计、训练方法以及其在多种自然语言处理任务中的表现，展示了大型语言模型在无监督学习和多任务学习中的潜力。
62. https://oreil.ly/MHt5H: （https://arena.lmsys.org/）该资源是由 LMSYS 提供的“Chatbot Arena”平台，用于通过社区中模型间的成对比较来计算评分，从而对不同的 AI 聊天模型进行排名和比较的排行榜系统。
63. https://oreil.ly/O5QTX: （https://nlp.stanford.edu/projects/glove/）该资源是斯坦福大学自然语言处理组（Stanford NLP Group）开发的 GloVe 项目主页，提供了一种用于生成词向量（word embeddings）的模型——GloVe（Global Vectors for Word Representation），该模型通过统计词共现信息来学习词的向量表示，广泛应用于自然语言处理任务。
64. https://oreil.ly/OWD2v: （https://www.adept.ai/blog/fuyu-8b?utm_source=substack&utm_medium=email）该 URL 指向的是 Adept 公司博客中关于其语言模型“Fuyu-8B”的介绍或分析文章。文章内容可能涉及 Fuyu 模型在某些任务中的表现，特别是讨论了模型输出与参考答案匹配的问题，即模型生成的正确答案可能因参考答案集不完整而被低估评分的情况。文章还可能包含具体案例，如图像描述任务中 Fuyu 生成正确但评分较低的示例。
65. https://oreil.ly/R1sCz: （https://python.langchain.com/v0.1/docs/guides/productionization/evaluation/string/criteria_eval_chain/）该资源是 LangChain 的“Criteria Evaluation”评估指南，介绍如何在生产环境中使用字符串评估链（string evaluation chain）对模型输出进行标准化的评价和评分。
66. https://oreil.ly/SBUiU: （https://platform.openai.com/docs/guides/embeddings/what-are-embeddings）该资源是 OpenAI 官方文档中的一篇指南，介绍了“嵌入”（embeddings）的基本概念及其应用，帮助开发者理解和使用 OpenAI 的嵌入 API。

67. https://oreil.ly/a6jbV:（https://docs.coveo.com/en/l9gg3565/coveo-for-commerce/product-embeddings-and-vectors）该资源介绍了 Coveo 在电商领域中如何利用产品的向量化表示（embedding）技术，通过将商品信息转化为向量形式，提升搜索和推荐的效果，帮助实现更精准的商品匹配和个性化体验。
68. https://oreil.ly/eI9Vq:（https://huggingface.co/datasets/lmsys/chatbot_arena_conversations）该资源是 Huggingface 平台上由 LMSYS 发布的“Chatbot Arena Conversations”数据集，包含大量（至少 3.3 万条）用于聊天机器人训练和评测的对话提示和交流内容，涵盖常见问候语及各种复杂问题（如脑筋急转弯）等多样化对话场景。
69. https://oreil.ly/fti6d:（https://a16z.com/generative-ai-enterprise-2024/）该资源是 a16z（Andreessen Horowitz）于 2023 年发布的一项研究报告，内容涉及生成式人工智能在企业中的应用及其发展趋势。
70. https://oreil.ly/gPbjS:（https://www.anthropic.com/news/evaluating-ai-systems）该资源是一篇由 Anthropic 发布的关于人工智能系统评估的文章或报告，内容聚焦于当前 AI 系统评估领域存在的不足，强调评估方法的发展相较于算法开发受到的关注较少，呼吁政策制定者增加对评估方法创新及评估稳健性分析的资金支持和政策投入。
71. https://oreil.ly/ijU20:（https://yale-lily.github.io/spider）该资源是“Spider”数据集的官方网站或项目主页，Spider 是一个用于评估文本到 SQL（text-to-SQL）生成能力的基准数据集，包含多样化的自然语言查询及其对应的 SQL 语句，广泛用于训练和测试 AI 模型在复杂数据库场景下生成正确 SQL 查询的性能。
72. https://oreil.ly/kIJ9F:（https://scale.com/leaderboard）该资源是 Scale 公司提供的一个私有的比较排行榜（comparative leaderboard），用于通过受过训练且可信赖的评估者，根据特定标准对多个响应进行比较和排序，旨在提升评价的准确性和可靠性。由于采用了高质量的人工作业方式，该排行榜在成本和比较数量上存在一定限制。
73. https://oreil.ly/kKWnZ:（https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html）该资源是一篇关于著名的分析与人工智能

（AI）失败案例的文章，详细介绍了多个因 AI 技术应用不当而导致的严重问题或灾难性后果，旨在提醒读者关注 AI 系统的潜在风险及其质量控制的重要性。

74. https://oreil.ly/rrSS9: （https://bird-bench.github.io/）该资源 “https://bird-bench.github.io/” 指向的是 BIRD-SQL（Big Bench for Large-scale Database Grounded Text-to-SQL Evaluation）基准测试的官方网站或项目主页。BIRD-SQL 是一个用于评估大型语言模型在从自然语言生成 SQL 查询任务中表现的基准数据集，强调通过功能正确性来衡量模型的性能。
75. https://oreil.ly/tMH21: （https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says）该资源是一篇来自 Vice 网站的报道，讲述了一名男子在与 AI 聊天机器人交流后选择自杀的事件，文章中包括了该男子遗孀的陈述，反映了 AI 聊天机器人在实际应用中可能带来的严重负面后果和风险。
76. https://oreil.ly/td_MY: （https://arena.lmsys.org/）该资源是 LMSYS Chatbot Arena 的官方网站，提供一个开放平台，用户可以输入提示词，系统会返回两个匿名模型的回答，用户通过投票选择更优的回答，从而以众包方式收集和比较不同基础模型的性能表现，促进模型评测的标准化和质量控制。
77. https://oreil.ly/tmWqk: （https://aclanthology.org/2023.wmt-1.51/）该资源是 2023 年 WMT（Workshop on Machine Translation）指标共享任务的相关论文或报告，内容聚焦于机器翻译评价指标的研究，特别探讨参考译文质量问题及无参考指标在与人工评判相关性方面的表现。
78. https://oreil.ly/uJNFH: （https://oars-workshop.github.io/2021/andrew.pdf）该资源是一份由 Andrew（可能是作者名）在 2021 年 OARS 研讨会上发布的 PDF 文档，内容很可能涉及嵌入表示（embedding representations）及其在实际应用中的使用，如文本、图像、图形、查询和用户数据的嵌入，重点介绍了如何将各种数据类型转化为向量表示以支持搜索、推荐或机器学习任务。
79. https://oreil.ly/v2ENK: （https://aclanthology.org/D19-1053/）该资源是 ACL Anthology 中一篇与自然语言处理相关的学术论文，可能涉及语义文本相似度的评估方法或相关技术。

80. https://oreil.ly/vX-My: （https://proceedings.mlr.press/v202/liu23ao.html）该资源是由 Liu 等人在 2023 年发表的一篇学术论文，讨论了语言模型（language models）及其性能度量指标与基础模型（foundation models）在下游任务表现之间的关系。论文指出语言模型的性能指标通常与基础模型在实际应用中的表现高度相关，但语言模型的性能并不能完全解释下游任务的表现差异，提示这是一个仍在积极研究中的领域。该资源有助于理解语言模型评估方法及其对基础模型性能预测的意义。
81. https://x.com/chipro/status/937384141791698944: 该资源是一条发布于 2017 年的推文，内容介绍了作者在 NeurIPS 研讨会上提出的一种名为 MEWR（Machine translation Evaluation metric Without Reference text）的机器翻译自动评估方法，该方法利用更强大的语言模型来无需参考译文即可评估机器翻译质量。
82. https://x.com/gdb/status/1733553161884127435: 该资源是一条由 OpenAI 联合创始人 Greg Brockman 于 2023 年 12 月发布的推文，内容提到“评估（evals）往往就是所需的一切。”
83. https://x.com/lmarena_ai/status/1792625968865026427: 该 URL 指向的资源是一条来自“lmarena_ai”账号的推文，内容涉及 LMSYS 评测榜单中“hard prompts”（难题提示词）的定义或示例。推文可能展示了什么样的提示词被认定为“hard prompts”，以及它们在模型排名中的作用，说明 LMSYS 如何通过内部模型筛选这些难题提示词以保证评测的公平性和多样性。

第 4 章（99）

1. https://arxiv.org/abs/1712.07040: 该资源是一个名为“NarrativeQA”的阅读理解数据集或任务，旨在基于书籍或长篇故事回答相关问题，用于评估机器对叙事文本的理解能力。

2. https://arxiv.org/abs/1803.05457: 该资源是一篇由 Clark 等人于 2018 年发表的论文，主要内容是评估模型解决复杂的小学科学问题的能力。
3. https://arxiv.org/abs/1809.02789: 该资源是一个名为“OpenBookQA”的阅读理解相关的学术论文或项目，重点研究基于书籍或长篇故事进行问答的能力，属于通过理解文本内容来回答问题的任务。
4. https://arxiv.org/abs/1905.07830: 该资源是一篇由 Zellers 等人在 2019 年发表的论文，介绍了 HellaSwag 数据集及其评测方法。HellaSwag 旨在衡量模型预测句子或故事、视频场景续写的能力，重点测试模型的常识推理和对日常活动的理解水平。
5. https://arxiv.org/abs/1907.10641: 该资源是由 Sakaguchi 等人于 2019 年发布在 arXiv 上的一篇论文，介绍了 WinoGrande 数据集。该数据集用于评估语言模型解决具有挑战性的代词指代消解问题的能力，这些问题经过精心设计，需要复杂的常识推理才能正确理解和回答。
6. https://arxiv.org/abs/2005.14165: 该资源是论文《Language Models are Few-Shot Learners》（作者：Brown 等人，2020 年），介绍了 GPT-3 模型的设计与性能评估，包括对训练数据中可能包含的基准测试集污染情况的分析。
7. https://arxiv.org/abs/2009.03300: 该资源是 2020 年由 Hendrycks 等人发表在 arXiv 上的论文，介绍了“Massive Multitask Language Understanding（MMLU）”基准测试。MMLU 旨在通过涵盖 57 个学科（如基础数学、美国历史、计算机科学和法律）的多项选择题，评估语言模型的知识储备和推理能力。
8. https://arxiv.org/abs/2009.13081: 该资源是一个名为 MedQA 的医学问答数据集或基准测试，主要用于评估和推动医学领域的问答系统或相关自然语言处理技术的发展。
9. https://arxiv.org/abs/2102.12162: 该 URL（https://arxiv.org/abs/2102.12162）指向的是 arXiv 上的一篇学术论文，标题或内容与“Vietnamese”（越南语）相关。结合上下文可推断，这篇论文可能涉及越南语的某个研究领域，如自然语言处理、语言模型、文本分析或相关计算语言学课题。
10. https://arxiv.org/abs/2103.03874: 该资源是指向“MATH”基准测试的论文，MATH 是一个用于评估机器学习模型在数学竞赛题目上表现的标准化测试集，包

含难度较高的数学竞赛题目，旨在测量模型解决复杂数学问题的能力。论文详细介绍了该基准的构建方法、题目来源及其在机器学习模型中的应用情况。

11. https://arxiv.org/abs/2109.07958: 该资源是 Lin 等人在 2021 年发表的论文，介绍了 TruthfulQA 基准测试，旨在评估语言模型生成回答的准确性、真实性及非误导性，重点考察模型对事实的理解能力。
12. https://arxiv.org/abs/2203.05227: 该资源是 Li 等人于 2022 年发表在 arXiv 上的一篇论文，内容涉及自然语言生成文本质量的评估指标，重点讨论了如流畅度（fluency）、连贯性（coherence）、以及针对不同任务（如翻译的忠实度 faithfulness、摘要的相关性 relevance）设计的评价标准。
13. https://arxiv.org/abs/2206.04615: 该资源是关于“BIG-bench Hard (BBH)”的学术论文，介绍了一个用于评估大型语言模型推理能力的基准测试套件 BIG-bench 中的部分任务和设计方法。论文详细描述了 BBH 基准的构建、任务示例（如基于经典游戏“二十个问题”的推理任务），以及如何通过这些任务来衡量模型的推理和理解能力。
14. https://arxiv.org/abs/2211.09110: 该资源是指向一篇关于 HELM（Holistic Evaluation of Language Models）评测套件的学术论文。该论文详细介绍了 HELM 评测框架及其在语言模型全面评估中的应用，特别强调了评测的复杂性和高昂成本。文中提到，斯坦福大学使用该套件对 30 个模型进行全面评测，耗费了约 8 万到 10 万美元，展示了大规模语言模型评测的资源消耗情况。
15. https://arxiv.org/abs/2301.01768: 该资源是一篇由 Hartman 等人在 2023 年发表的学术论文，研究内容涉及大型语言模型（如 GPT-4 和 Llama）在训练过程中所产生的政治和意识形态偏见。文章探讨了不同模型因训练数据和方法的差异，而表现出不同的政治倾向性，例如 GPT-4 偏向自由主义和左翼观点，而 Llama 更具权威主义倾向。
16. https://arxiv.org/abs/2303.08896: 该资源是由 Manakul 等人于 2023 年发表在 arXiv 上的论文，介绍了一种名为 SelfCheckGPT 的方法。该方法基于假设：如果一个模型生成的多个输出彼此不一致，那么最初的输出很可能是幻觉（hallucinated）。具体做法是，对于一个需要评估的回答 R，SelfCheckGPT 会

生成 N 个新的回答，并衡量 R 与这 N 个回答之间的一致性。该方法虽然有效，但计算成本较高，因为评估一个回答需要多次调用 AI 模型。

17. https://arxiv.org/abs/2303.15621: 该资源是由 Luo 等人在 2023 年发表的一篇学术论文，内容涉及利用 GPT-3.5 和 GPT-4 模型评估文本的事实一致性，展示了这些大型语言模型在测量文本真实性方面优于以往方法的能力。
18. https://arxiv.org/abs/2304.06364: 该资源是微软团队于 2023 年发布的名为“AGIEval”的多项选择题型评测基准论文，旨在对人工智能通用智能（AGI）能力进行系统评估。论文中介绍了 AGIEval 的设计理念及任务选择标准，特别指出排除了开放式任务以避免评估结果的不一致性。
19. https://arxiv.org/abs/2305.08283: 该资源是 2023 年 Feng 等人发表在 arXiv 上的一篇学术论文，内容涉及语言模型中的政治或宗教意识形态偏见问题，研究了模型在训练过程中如何产生并表现出对特定政治或宗教立场的偏向性。
20. https://arxiv.org/abs/2306.09479: 该资源是 McKenzie 等人在 2023 年发布的一篇学术论文，讨论了用于语言模型评估的样本数量问题，尤其关注合成样本的使用和评估基准的合理规模，强调了在模型性能评测中至少需要数百到上千个示例以确保结果的可靠性。
21. https://arxiv.org/abs/2307.09009: 该资源是由 Chen 等人于 2023 年发布在 arXiv 上的一篇研究论文，内容分析了 OpenAI 的 GPT-3.5 和 GPT-4 模型在 2023 年 3 月至 6 月期间多个基准测试中的性能变化，揭示了模型更新后性能显著波动的现象。
22. https://arxiv.org/abs/2309.08632: 该资源是一篇发表于 2023 年的讽刺论文，题为《Pretraining on the Test Set Is All You Need》。论文作者是斯坦福大学的博士生 Rylan Schaeffer。文章通过仅在多个基准测试的数据上进行训练，展示了一个只有一百万参数的模型如何几乎达到完美的测试成绩，并且在所有这些基准测试中表现优于许多更大规模的模型，从而对机器学习模型的训练和评估方法提出了讽刺和反思。
23. https://arxiv.org/abs/2309.11998: 该资源是一篇由 LMSYS 团队发布的学术论文或研究报告（2023 年），基于对 Chatbot Arena 和 Vicuna 演示中约一百万次对话的分析，系统性地研究了人们使用基础模型（foundation models）的指令

分布及应用场景。文中揭示了包括角色扮演在内的多种常见用途，尤其强调了角色扮演在游戏中的 AI 非玩家角色（NPC）、AI 伴侣和写作助手等领域的重要性。

24. https://arxiv.org/abs/2310.00746: 该资源是 Wang 等人于 2023 年发布在 arXiv 上的论文，介绍了 RoleLLM 这一基准测试，用于评估大型语言模型的角色扮演能力。RoleLLM 通过设计精细的相似度评分机制和 AI 评审方法，衡量模型在模拟特定人物角色时生成文本与预期输出的相似程度，从而实现对角色扮演能力的自动化评价。
25. https://arxiv.org/abs/2310.16049: 该资源是由 Sprague 等人在 2023 年发布的一篇学术论文，介绍了 MuSR，这是一种基于链式思维（chain-of-thought）的多步骤推理基准测试（benchmark）。
26. https://arxiv.org/abs/2311.07911: 该资源是 2023 年由 Zhou 等人发布的学术论文，介绍了 IFEval——一个用于评估语言模型指令遵循能力的基准测试（instruction-following benchmark）。IFEval 涵盖了多种指令类型，旨在提供评估模型执行复杂指令时表现的标准和方法，帮助研究者设计合适的评测标准和数据集，从而系统地衡量模型的指令理解和执行能力。
27. https://arxiv.org/abs/2311.12022: 该资源是 2023 年由 Rein 等人发布在 arXiv 上的一篇论文，介绍了 GPQA——一个面向研究生水平的问题与回答（Q&A）基准测试。该基准旨在评估模型在更高学术难度下的理解和推理能力，反映了近年来问答基准从中小学水平向研究生水平升级的趋势。
28. https://arxiv.org/abs/2312.06674: 该资源是一篇由 Inan 等人于 2023 年发表在 arXiv 上的学术论文，题为“Llama Guard”，主要内容涉及 Meta 公司提出的一种用于 AI 模型安全性的防护机制或框架。论文重点讨论了如何识别和防范 AI 模型输出中的有害内容，属于 AI 安全和内容审核领域，旨在提升模型在生成文本时的安全性和可靠性。
29. https://arxiv.org/abs/2401.01275: 该资源是一篇由 Tu 等人于 2024 年发表的论文，题为“CharacterEval”，介绍了一种用于评估语言模型角色扮演能力的方法。该方法通过人工标注和训练奖励模型，对角色扮演的各个方面进行五分制评分，从而实现对模型角色扮演表现的自动化评估。

30. https://arxiv.org/abs/2402.01781: 该资源是由 Alzahrani 等人在 2024 年发布的一篇学术论文，研究了多项选择题（MCQs）中模型表现对题目和选项呈现方式细微变化的敏感性，具体指出增加额外空格或添加提示语（如“Choices:”）会导致模型答案发生变化，探讨了模型对提示词（prompt）的敏感性及提示工程的相关问题。
31. https://arxiv.org/abs/2403.18802: 该资源是 Google DeepMind 团队于 2024 年发布在 arXiv 上的论文，标题为《Long-Form Factuality in Large Language Models》。论文介绍了一种名为 SAFE（Search-Augmented Factuality Evaluator）的机制，该机制通过结合搜索引擎结果对大型语言模型生成的长文本回答进行真实性验证，提升事实性评估的准确性。
32. https://arxiv.org/abs/2406.01574: 该资源是 2024 年由 Wang 等人发表的一篇学术论文，介绍了 MMLU-PRO 这一新的基准测试。MMLU-PRO 是对原有 MMLU 基准的升级版本，旨在提供更具挑战性、更贴近实际应用能力的评测任务，用以更有效地衡量和推动语言模型在多领域知识和推理能力上的表现。
33. https://convai.com: 该资源是 Convai 公司官网，介绍其开发的 3D 人工智能角色技术，这些 AI 角色能够在三维环境中进行交互操作（如拾取物体），并且 Convai 通过微调开源模型以克服商业模型在物理交互能力上的限制。
34. https://en.wikipedia.org/wiki/Multi-objective_optimization: 该资源为维基百科页面，介绍了“多目标优化”（Multi-objective optimization）这一领域，重点讲解了如何在多个优化目标之间进行权衡与选择，通常涉及使用帕累托优化（Pareto optimization）的方法，以在满足不可妥协目标的前提下，优化其他可调节的目标。
35. https://en.wikipedia.org/wiki/Simpson: 该资源是维基百科中关于“Simpson”（辛普森）的条目页面，可能介绍与“Simpson”相关的概念、人物或现象。在上下文中提到了“Simpson’s paradox”（辛普森悖论），因此该页面很可能详细解释了辛普森悖论这一统计学现象及其相关内容。
36. https://en.wikipedia.org/wiki/Test-driven_development: 该资源是关于“测试驱动开发”（Test-driven development，简称 TDD）的维基百科页面，介绍

了一种软件工程中的开发方法，即在编写代码之前先编写测试用例，以确保代码按预期工作。

37. https://github.com/EleutherAI/lm-evaluation-harness/blob/master/docs/task_table.md: 该资源是 EleutherAI 维护的一个文档页面，列出了其开源项目 lm-evaluation-harness 支持的所有评测任务的详细列表（任务表）。该页面汇总了超过 400 个不同的语言模型评测基准任务，主要以多项选择题形式为主，涵盖了多个著名评测集如 MMLU、AGIEval 和 AI2 Reasoning Challenge 等，方便研究人员和开发者了解和使用这些标准化的评测任务来评估语言模型的性能。
38. https://github.com/google-deepmind/gemma/blob/main/LICENSE: 该资源是 Google DeepMind 旗下开源项目 Gemma 的许可证文件，具体采用的是 Apache 2.0 开源许可证。
39. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/README.md: 该资源是 Google 开源项目 BIG-bench 中关于基准测试任务的说明文档，介绍了 BIG-bench 包含的 214 个用于评估 AI 模型多种能力的基准测试集合，旨在帮助研究者理解和使用这些任务来衡量模型性能及推动新基准的开发。
40. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/twenty_questions/README.md: 该资源是 BIG-bench 项目中“twenty_questions”任务的说明文档，介绍了基于经典游戏“二十问”的评测任务设计，包括任务规则、示例对话及评分方式，用于评估模型通过一系列是非问题猜测目标概念的能力。
41. https://github.com/openai/evals: 该资源是 OpenAI 开源的一个评测工具库“evals”，用于对模型进行多达约 500 个不同基准测试的评估。它不仅支持运行已有的多种评测任务，还允许用户注册新的评测基准，评测内容涵盖数学计算、解谜以及识别 ASCII 艺术等多种能力，旨在全面衡量和验证 OpenAI 模型的性能表现。
42. https://github.com/openai/grade-school-math: 该资源是 OpenAI 于 2021 年发布的“Grade School Math”（简称 GSM-8K）项目，旨在衡量模型解决小学数学课程中各种典型数学问题的能力。

43. https://oreil.ly/-qEXt: （https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data）该资源是一篇由 The Verge 发布的文章，讨论了生成式人工智能（Generative AI）在版权侵权、法律责任以及其训练数据使用中的公平使用（fair use）问题。文章深入分析了当前知识产权法律如何适用于 AI 生成内容，探讨了在使用版权保护数据训练 AI 模型时可能面临的法律挑战，以及企业如何在法律尚未明朗的情况下谨慎利用 AI 技术。
44. https://oreil.ly/-uhru: （https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard）该资源是 Hugging Face 提供的“Open LLM Leaderboard”网页，展示和比较多个大型语言模型（LLM）在不同基准测试（benchmarks）上的表现，旨在全面评估基础模型的能力和缺陷。
45. https://oreil.ly/0VrKU: （https://perspectiveapi.com/）该资源是由 Perspective API 提供的一个用于检测文本中有害内容（如毒性、仇恨言论等）的专门模型接口，能够帮助开发者快速识别和过滤在线文本中的不良行为。
46. https://oreil.ly/2aIvB: （https://huggingface.co/s-nlp/roberta_toxicity_classifier）该资源是一个基于 RoBERTa 架构的文本毒性（toxic）分类模型，托管在 Hugging Face 平台上。它专门用于检测和分类文本中的有害内容，如仇恨言论、侮辱性语言等，有助于识别和过滤在线交流中的有害行为。该模型通常体积较小、运行高效，适合用于自动化的毒性识别任务，支持对 AI 生成文本和人类文本中的有害内容进行检测。
47. https://oreil.ly/4X6Dm: （https://huggingface.co/spaces/open-llm-leaderboard/blog）该 URL 指向 Hugging Face 上的一个名为 “open-llm-leaderboard” 的空间中的博客页面，内容很可能涉及该排行榜的介绍、更新、基准测试的选择和聚合过程的透明化分析，帮助用户了解和比较开放式大语言模型（LLM）的性能表现。
48. https://oreil.ly/4c6in: （https://www.voiceflow.com/blog/how-much-do-chatgpt-versions-affect-real-world-performance）该资源是一篇由 Voiceflow 发布的博客文章，讨论不同版本的 ChatGPT 模型在实际应用中的性

能差异及其影响，具体分析了模型升级如何在某些任务中导致性能下降，而在其他任务中带来改进。

49. https://oreil.ly/70VH1:（https://huggingface.co/alexandrainst/da-hatespeech-detection-small）该资源是一个专门用于检测丹麦语中的仇恨言论和有害内容的机器学习模型，托管在 Hugging Face 平台上。该模型体积较小、运行速度快，适合用于识别和过滤丹麦语文本中的仇恨言论，帮助改善线上交流环境的安全性与友善度。
50. https://oreil.ly/7PFUy:（https://crfm.stanford.edu/2023/12/19/helm-lite.html）该资源是关于“HELM Lite”（Holistic Evaluation of Language Models Lite）的介绍或说明页面。HELM Lite 是一个用于语言模型综合评估的轻量级版本，旨在通过选择较少但具有代表性的基准测试来评估模型性能，从而降低计算资源的消耗。该版本有意省略了一些计算成本较高的基准（例如信息检索基准 MS MARCO），以便在计算资源有限的情况下仍能提供有价值的模型评价结果。
51. https://oreil.ly/9Ku73:（https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339981/）该资源是发表在《British Medical Journal》（临床研究版）上的一篇医学研究文章，题为“Comparison of Treatment of Renal Calculi by Open Surgery, Percutaneous Nephrolithotomy, and Extracorporeal Shockwave Lithotripsy”，比较了开放手术、经皮肾镜取石术和体外冲击波碎石术三种治疗肾结石的方法。文章由 Charig 等人撰写，发表于 1986 年 3 月，详细介绍了各治疗方式的效果与临床数据。
52. https://oreil.ly/AB2FU:（https://docs.anthropic.com/en/docs/about-claude/use-case-guides/content-moderation）该资源是 Anthropic 官方提供的关于如何使用 Claude 进行内容审核的指导文档，包含具体的用例说明和操作教程，帮助用户有效利用 Claude 模型实现内容审核功能。
53. https://oreil.ly/AhHI_:（https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf）该资源是 Google 发布的“Gemini 技术报告”（Gemini’s technical report），详细介绍了 Gemini 模型的性能表

现，但未具体披露模型训练所用的数据，仅提到所有数据增强工作人员均获得了当地的生活工资保障。

54. https://oreil.ly/Bfa4q: （https://aclanthology.org/2020.findings-emnlp.301/）该资源是 2020 年发表在 EMNLP 会议 Findings 集上的一篇论文，内容涉及自然语言处理中的有害内容检测或生成毒性内容的基准测试，可能与评估语言模型生成有害或偏见语言的技术和方法有关。
55. https://oreil.ly/BndEu: （https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target）该资源是 Facebook 开发的一个用于检测仇恨言论的机器学习模型，名为“roberta-hate-speech-dynabench-r4-target”。该模型专门用于识别文本中的仇恨言论内容，属于一种高效且针对性强的有害内容检测工具，适用于监测和过滤在线交流中的有害行为，帮助提升内容的安全性和健康度。
56. https://oreil.ly/CQ52G: （https://crfm.stanford.edu/helm/lite/latest/#/leaderboard）该资源是斯坦福大学发布的 HELM（Holistic Evaluation of Language Models）排行榜页面，展示了基于多项基准测试对语言模型的综合评估结果，包含包括 MMLU、GSM-8K 在内的十个基准测试，用于比较和分析不同语言模型的性能表现。
57. https://oreil.ly/Ek0wH: （https://irmckenzie.co.uk/round1）该 URL “https://irmckenzie.co.uk/round1” 很可能指向 McKenzie 等人在 2023 年发布的与“Inverse Scaling prize”相关的研究或数据资源，特别是关于评估基准中样本数量建议的详细说明或实验数据页面。具体来说，该资源可能包含用于逆向规模效应（Inverse Scaling）挑战的第一轮数据集、示例样本或相关的合成样本信息。
58. https://oreil.ly/FL-1Z: （https://llama.meta.com/llama3/license/）该资源为 Meta 官方发布的 Llama 3 模型的社区许可协议页面，详细说明了 Llama 3 模型的使用条款和授权许可内容，规范用户在遵守社区许可协议的前提下合法使用该模型。
59. https://oreil.ly/HnIVp: （https://arxiv.org/pdf/2303.16634）该 URL （https://arxiv.org/pdf/2303.16634）指向的是一篇学术论文，内容涉及利用大

型语言模型（如 GPT-3.5 和 GPT-4）进行事实一致性评估的方法和研究。论文中探讨了如何通过 AI 作为评判者来判断生成文本的真实性和准确性，展示了这些模型在测量事实一致性方面优于以往方法的效果。该研究对提升自动文本生成的质量和可信度具有重要意义。

60. https://oreil.ly/ICHH3:（https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli）该资源是一个名为“DeBERTa-v3-base-mnli-fever-anli”的预训练语言模型，包含 1.84 亿参数，专门用于对输入的（前提，假设）对进行分类，判断它们之间的关系属于包含（entailment）、矛盾（contradiction）或中立（neutral）等类别。该模型基于多任务学习，训练数据涵盖 MNLI、FEVER 和 ANLI 等大型标注语料，适用于事实一致性预测等自然语言推理任务。
61. https://oreil.ly/J3pbA:（https://learning.oreilly.com/library/view/designing-machine-learning/9781098107956/）该资源是 O’Reilly 出版的电子书《Designing Machine Learning Systems》（《设计机器学习系统》），主要内容涉及如何设计和评估机器学习系统，特别强调通过数据切片（slice-based evaluation）来获得系统性能的细粒度理解。
62. https://oreil.ly/LghFT:（https://www.interconnects.ai/p/how-the-open-source-llm-ecosystem）该资源是一篇关于开源大语言模型（LLM）生态系统的文章，探讨了如何通过标准化评测和数据保护措施来提升模型性能的公平评估，强调了在公开基准测试中保护数据隐私和自动评测工具的重要性。
63. https://oreil.ly/MLlDD:（https://crfm.stanford.edu/2023/12/19/helm-lite.html）该资源是斯坦福大学 CRFM（Center for Research on Foundation Models）于 2023 年 12 月 19 日发布的关于 HELM Lite 的介绍或相关内容页面。HELM Lite 可能是 HELM（Holistic Evaluation of Language Models）的一种简化版本，主要用于评估语言模型性能，采用了如“平均胜率”（mean win rate）等指标来衡量模型在不同场景中的表现优劣。
64. https://oreil.ly/PSNna:（https://cdn.openai.com/papers/gpt-4.pdf）该资源是指向 GPT-4 技术报告的 PDF 文档，详细介绍了 GPT-4 模型的架构、性能评估以及在多个基准测试中的表现。

65. https://oreil.ly/QB4XP: （https://research.google/pubs/natural-questions-a-benchmark-for-question-answering-research/）该资源是一个关于“Natural Questions”的研究基准，旨在用于通用问答系统的研究，提供了包含或不包含维基百科页面输入的两种设置。
66. https://oreil.ly/SaIST: （https://arxiv.org/html/2401.03601v1）该资源是 arXiv 上的一篇学术论文，标题为“IFEval 和 INFOBench 评测方法”，主要内容聚焦于评估语言模型的指令跟随能力。论文介绍了不同基准测试（benchmark）对指令跟随能力的不同定义，详细阐述了如何设计评测标准、选择测试指令以及采用合适的评测方法，从而系统地衡量模型执行多样化指令的能力。
67. https://oreil.ly/V1P8X: （https://ai.meta.com/blog/large-language-model-llama-meta-ai/）该 URL 指向 Meta 官方博客中关于大型语言模型 LLaMA 的介绍文章，内容可能涵盖 LLaMA 模型的发布背景、技术特点以及其许可协议（如商业使用限制）。
68. https://oreil.ly/V48iM: （https://www.godaddy.com/resources/news/llm-from-the-trenches-10-lessons-learned-operationalizing-models-at-godaddy#h-3-prompts-aren-t-portable-across-models）该资源是一篇来自 GoDaddy 的技术分享文章，题为“LLM from the Trenches: 10 Lessons Learned Operationalizing Models at GoDaddy”。文章总结了 GoDaddy 在实际部署和运维大型语言模型（LLM）过程中积累的十条经验教训，内容涵盖模型迁移的影响、提示词（prompts）在不同模型间的适用性等实际问题，帮助读者理解在真实业务环境中如何有效管理和优化语言模型的应用。
69. https://oreil.ly/XaRwG: （https://www.elastic.co/licensing/elastic-license）该 URL（https://www.elastic.co/licensing/elastic-license）指向的是 Elastic 公司的“Elastic License”许可协议页面。该许可协议规定了使用 Elastic 开源软件的限制，特别禁止公司以托管服务的形式提供开源版本的 Elastic 软件，从而避免与 Elastic 官方的 Elasticsearch 平台直接竞争。
70. https://oreil.ly/ZRwVI: （https://platform.openai.com/docs/guides/moderation/overview）该资源是 OpenAI 官方文档中关于内容审核（内容

moderation）功能的概述页面，详细介绍了内容审核端点的设计与用途，帮助用户理解如何识别和管理模型输出中的有害内容，从而提升 AI 系统的安全性和可靠性。

71. https://oreil.ly/Zj1GZ: （https://a16z.com/generative-ai-enterprise-2024/）该资源是一份由风险投资公司 a16z（Andreessen Horowitz）于 2024 年发布的关于生成式人工智能（Generative AI）在企业中的应用与发展的研究报告，重点探讨了企业为何重视开源模型，特别是在控制权和定制化方面的两大关键原因。
72. https://oreil.ly/aFvUh: （https://dl.acm.org/doi/10.1145/3442188.3445924）该资源是一篇发表在 ACM 数字图书馆上的学术论文，标题及内容与“RealToxicityPrompts”数据集相关。该论文由 Gehman 等人于 2020 年发布，介绍了一个包含 10 万条自然出现的提示语的数据集，旨在评估语言模型生成有毒（toxic）内容的倾向，常用于毒性检测和模型安全性研究。
73. https://oreil.ly/bdGKm: （https://paperswithcode.com/dataset/wmt-2014）该资源是 WMT 2014 翻译数据集，常用于机器翻译领域的研究和评测。
74. https://oreil.ly/d1ey3: （https://blog.langchain.dev/langchain-state-of-ai-2023/）该资源是 LangChain 发布的《State of AI 2023》报告或文章，内容聚焦于 2023 年人工智能领域的发展现状，特别是关于如何评估和改进基于 AI 的应用响应质量的研究与分析。
75. https://oreil.ly/d3ggH: （https://allenai.org/data/arc）该资源是“AI2 Reasoning Challenge (ARC-C)”，由 Allen Institute for Artificial Intelligence（AI2）发布的一个多项选择题型的推理挑战数据集，旨在评估和推动机器在复杂推理任务中的表现。该数据集广泛用于自然语言处理和人工智能领域的模型评测。
76. https://oreil.ly/eH7Ho: （https://discord.com/channels/879548962464493619/922424173916196955/1166444804981399572）该 URL 指向的是 Hugging Face 官方 Discord 服务器中一个特定频道内的消息，内容很可能是 Lewis Tunstall 关于 Hugging Face 选择某些基准测试（benchmarks）原因的回应。
77. https://oreil.ly/fWs9H: （https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-

after-leak）该 URL 指向的资源是一篇 2023 年 5 月发布的彭博社新闻报道，内容介绍了三星公司在员工将公司机密信息误输入 ChatGPT 后，因担忧数据泄露和隐私安全问题，于 2023 年 5 月禁止员工使用 ChatGPT 及其他生成式人工智能工具的相关情况。报道详细描述了此次泄密事件的背景、影响以及三星采取的应对措施。

78. https://oreil.ly/hJucg: （https://arxiv.org/abs/2402.11782）该资源是一个学术论文，编号为 arXiv:2402.11782，可能由 Wan 等人在 2024 年发布，研究内容涉及人工智能模型如何判断和处理信息的可信度，特别是模型在评估信息时更侧重于网站与查询的相关性，而较少考虑诸如科学引用或中立语气等人类认为重要的文本风格特征。
79. https://oreil.ly/hOlhJ: （https://openai.com/index/better-language-models/）该资源是 OpenAI 官网上的一个页面，介绍了 OpenAI 在打造更优语言模型方面的进展和成果，可能详细阐述了提升语言模型流畅性和连贯性的方法与技术。
80. https://oreil.ly/jCo7o: （https://hazyresearch.stanford.edu/legalbench/）该 URL 指向的是“LegalBench”资源页面，LegalBench 是一个专注于法律领域的基准测试套件，用于评估和比较法律相关自然语言处理模型的性能。
81. https://oreil.ly/mOAjn: （https://bird-bench.github.io/）该资源是一个名为 BIRD-SQL 的基准测试项目，旨在评估生成的 SQL 查询语句的执行准确性和效率，通过比较生成查询与真实 SQL 查询的运行时间来衡量性能。
82. https://oreil.ly/mlHyX: （https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt）该资源是一篇发表于 2023 年 4 月，由 Lewis Maddison 撰写的 TechRadar 新闻报道，内容涉及三星员工因使用 ChatGPT 泄露公司机密而导致的重大错误事件。
83. https://oreil.ly/p23MQ: （https://www.uspto.gov/subscription-center/2024/uspto-issues-inventorship-guidance-and-examples-ai-assisted-inventions）该资源是美国专利商标局（USPTO）于 2024 年发布的关于人工智能辅助发明的发明人资格（inventorship）指导和示例的官方说明文件，内容主

要阐述了在 AI 辅助发明的专利申请中，判断人类贡献是否足够显著以符合专利资格的标准和相关案例，帮助申请人理解和应对 AI 技术在专利法中的新挑战。

84. https://oreil.ly/pY1FF: （https://www.cnbc.com/2023/04/04/italy-has-banned-chatgpt-heres-what-other-countries-are-doing.html）该资源是一篇 2023 年 4 月发布的新闻报道，介绍了意大利在 2023 年曾短暂禁止使用 OpenAI 旗下的 ChatGPT，并进一步探讨了其他国家对 ChatGPT 及类似 AI 模型的监管措施和政策动态。
85. https://oreil.ly/pgGZ0: （https://discord.com/channels/879548962464493619/922424173916196955/1166444804981399572）该 URL 指向的是 Discord 平台上的一个具体频道消息，内容很可能与“公开排行榜”（public leaderboards）以及模型在多个基准测试（benchmarks）上的综合性能评估有关。根据上下文，该消息可能讨论了当前公开排行榜存在的局限性，比如计算资源限制导致某些重要但计算成本高的基准测试（如 MS MARCO 和 HumanEval）被排除在外，从而影响排行榜的全面性和准确性。因此，该链接资源大概率是一则关于模型评估、排行榜设计或者相关讨论的消息，可能包含对现有排行榜方法的分析或改进建议。
86. https://oreil.ly/u9_vA: （https://link.springer.com/article/10.1007/s11127-023-01097-2）该资源是一篇发表在 Springer 期刊上的学术文章，标题可能涉及政治或宗教意识形态偏见对语言模型行为的影响，探讨了模型在训练过程中如何产生并体现特定政治倾向，以及这种偏见对模型生成内容的影响。文章可能结合实证研究和理论分析，揭示不同模型（如 GPT-4 和 Llama）在政治立场上的差异及其原因。
87. https://oreil.ly/uTBwP: （https://huggingface.co/docs/transformers/en/model_doc/mistral#license）该 URL（https://huggingface.co/docs/transformers/en/model_doc/mistral#license）所指向的资源，简要说明是：

该页面介绍了 Mistral 模型的相关信息，其中包含该模型的许可证（License）细节。具体来说，Mistral-7B 模型采用的是 Apache 2.0 开源许可证，该许可证允许用户在遵守相应条款的前提下自由使用、修改和分发模型代码和权重。通过该链接，用户可以了解 Mistral 模型的授权方式及其开源许可协议的具体内容。

88. https://oreil.ly/wRlEh: （https://ai.meta.com/llama/license/）该资源（URL：https://ai.meta.com/llama/license/）是 Meta 官方提供的 Llama 模型的许可协议页面，详细说明了 Llama 模型（如 Llama 2 和 Llama 3）所采用的社区许可协议条款，规定了模型的使用、分发和授权等法律事项。
89. https://oreil.ly/xAbHm: （https://platform.openai.com/docs/guides/prompt-engineering/strategy-test-changes-systematically）该资源是 OpenAI 官方文档中关于“系统性测试提示工程策略”的指南，详细介绍了如何系统地设计和测试不同的提示策略以优化模型表现，包含评估样本量的粗略估算方法及其与评分差异之间的关系，帮助用户科学地比较和改进提示效果。
90. https://oreil.ly/xndQu: （https://techcrunch.com/2023/08/08/zoom-data-mining-for-ai-terms-gdpr-eprivacy/）该资源是一篇发表于 2023 年 8 月 8 日的 TechCrunch 文章，报道了 Zoom 因悄悄修改服务条款，允许其利用用户的服务生成数据（包括产品使用数据和诊断数据）来训练其人工智能模型而引发的争议。文章讨论了模型 API 提供商使用用户数据进行 AI 训练的风险，以及相关的隐私和合规问题（如 GDPR 和电子隐私法规）。
91. https://oreil.ly/xvYjL: （https://arxiv.org/abs/2109.07958）该 URL（https://arxiv.org/abs/2109.07958）指向的是一篇学术论文，内容涉及利用人工智能模型（如 GPT-3.5 和 GPT-4）来评估生成文本的事实一致性。论文展示了如何通过 AI 作为“裁判”来判断模型输出的真实性和准确性，属于自然语言处理领域中自动化评估方法的研究。
92. https://oreil.ly/yED-R: （https://huggingface.co/spaces/bigcode/bigcode-model-license-agreement）该 URL（https://huggingface.co/spaces/bigcode/bigcode-model-license-agreement）指向的是 Hugging Face 平台上 BigCode 模型的许可协议页面，具体介绍了 BigCode 模型所采用的“BigCode Open RAIL-M v1”开源许可条款。该许可协议详细规定了模型的使用权限、限制以及合规要求，帮助用户明确在何种条件下可以使用、修改和分发该模型。
93. https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product: 该资源是一篇由 LinkedIn 发布的技术博客

文章，题为《Musings on Building a Generative AI Product》（关于构建生成式人工智能产品的思考）。文章介绍了 LinkedIn 在开发和部署生成式 AI 产品过程中，采用人工评估与自动评估相结合的方法，特别强调了利用人工专家每日手动评估大量 AI 对话输出，以监控模型性能和使用情况的实践经验。

94. https://www.linkedin.com/feed/update/urn:li:activity:7189260630053261313/: 该资源是一篇来自 LinkedIn 的内容更新，分享了他们在部署生成式人工智能应用一年来的经验，重点讨论了设计评估标准的挑战。文章指出，评估 AI 输出的难点不在于判断结果是否正确，而是界定“好”的含义。例如，LinkedIn 在其 AI 驱动的职位评估应用中发现，虽然“你不适合该职位”可能是正确的反馈，但这并不是有帮助的回复；一个好的回复应当解释职位要求与候选人背景之间的差距，以及候选人如何弥补这一差距。
95. https://x.com/JoannaStern/status/1768306032466428291: 该资源是一条来自 OpenAI 首席技术官（CTO）的推文，内容涉及在被问及其模型所使用的训练数据时，未能给出令人满意的答复。
96. https://x.com/arthurmensch/status/1734470462451732839: 该资源是一条来自 Arthur Mensch 在社交平台 X（原 Twitter）上的动态（帖子），内容涉及 Mistral 模型许可证的变更，具体说明 Mistral 模型最初不允许将其输出用于训练或改进其他模型，但后来通过该链接所指的许可证更新，允许了这一用途。
97. https://x.com/dhuynh95/status/1713917852162424915: 该资源是一条来自 X（原 Twitter）用户 dhuynh95 的推文，内容涉及 Hugging Face 的 StarCoder 模型在训练过程中记忆了其训练数据集中的约 8% 的样本，揭示了 AI 模型可能存在的数据记忆和泄露风险。
98. https://x.com/gblazex: 该资源是 Balázs Galambosi 在社交平台 X（前 Twitter）上的个人账号主页，可能分享与自然语言处理、机器学习或相关领域的研究成果和见解，尤其涉及 Hugging Face 基准测试及模型评估。
99. https://x.com/polynoamial/status/1803812369237528825: 该资源是推特（X）上的一条动态，内容介绍了 Hugging Face 在 2024 年 6 月对其排行榜进行更新，将原有的 GSM-8K 基准测试替换为更具挑战性的 MATH lvl 5 基准，后者

包含了来自竞赛数学基准 MATH 中最困难的问题。该动态可能还包括有关其他基准测试更新的信息。

第 5 章（99）

1. https://arxiv.org/abs/1906.00080: 该资源是一篇由 Chen 等人在 2019 年发表的学术论文，内容涉及机器学习模型尤其是语言模型在训练过程中使用的私密数据（如用户电子邮件）可能被提取和泄露的风险，重点讨论了从模型中恢复训练数据以暴露敏感信息的问题。
2. https://arxiv.org/abs/1909.01066: 该资源是由 Petroni 等人在 2019 年发布的一篇学术论文，介绍了 LAMA（Language Model Analysis）基准测试方法。LAMA 用于评估语言模型中所包含的关系性知识，具体通过填空题形式（如“Winston Churchill is a citizen”）来探测模型对训练数据中事实性知识的掌握情况。该论文推动了“事实探测”（factual probing）这一研究领域的发展。
3. https://arxiv.org/abs/2005.14165: 该资源是 GPT-3 模型的开创性论文，题为《Language Models Are Few-shot Learners》（语言模型是少样本学习者），发表于 2020 年。论文介绍了“in-context learning”（上下文学习）的概念，展示了大型语言模型 GPT-3 无需更新模型参数，仅通过在输入提示中提供少量示例，即能执行包括翻译、阅读理解、简单数学计算及回答 SAT 考试问题等多种任务，突破了传统机器学习依赖大量训练和微调的限制。
4. https://arxiv.org/abs/2012.07805: 该资源是 Carlini 等人于 2020 年发表在 arXiv 上的论文，内容展示了从语言模型 GPT-2 中提取其训练数据中记忆信息的方法。论文指出，虽然技术上可以从模型中提取到训练数据，但实际风险较低，因为攻击者需要知道具体的上下文才能成功提取敏感信息。
5. https://arxiv.org/abs/2111.01998: 该资源是由 Ding 等人在 2021 年发表的一篇论文，介绍了 OpenPrompt 工具。OpenPrompt 旨在自动化整个提示工程（prompt engineering）工作流程，用户只需指定任务的输入输出格式、评估指

标和评估数据，工具便能自动优化提示词或提示链，以最大化评估指标的表现。功能上，OpenPrompt 类似于自动机器学习（autoML）工具，用于自动寻找经典机器学习模型的最优超参数。

6. https://arxiv.org/abs/2201.11903: 该资源是论文《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》（作者 Wei 等，2022 年），首次提出了“Chain-of-Thought”（CoT，思路链）提示技术。这项技术通过引导模型逐步思考，促使其采用更系统化的问题解决方法，显著提升了大型语言模型（如 LaMDA、GPT-3 和 PaLM）在多种基准测试上的推理能力。该论文发表于 arXiv 平台，时间早于 ChatGPT 的发布，是推动语言模型推理性能进步的重要工作。
7. https://arxiv.org/abs/2205.12628: 该资源是 Huang et al. (2022) 的一篇论文，研究了从大型语言模型（如 GPT-2 和 GPT-3）中提取记忆化训练数据的方法。论文指出，虽然从模型中提取训练数据在技术上是可行的，但实际风险较低，因为攻击者需要知道具体的上下文信息才能成功提取出特定的数据。
8. https://arxiv.org/abs/2211.09110: 该资源是一篇发表于 2022 年的斯坦福大学论文，标题为《Holistic Evaluation of Language Models》（语言模型的整体评估）。论文通过设计实验检测语言模型在生成文本时是否会逐字复制（regurgitate）受版权保护的内容，例如给模型一本书的第一段，要求它生成第二段，若生成内容与原书完全一致，则说明模型在训练过程中见过该书内容。研究涵盖了多种基础模型，结论指出直接复制长篇版权序列的情况较为罕见，但在流行书籍中这种现象会变得明显。
9. https://arxiv.org/abs/2301.13188: 该资源是 Carlini 等人于 2023 年发表的一篇论文，题为《Extracting Training Data from Diffusion Models》。论文展示了如何从开源扩散模型 Stable Diffusion 中提取训练数据，成功复现了超过一千张与真实训练图像高度相似的图片，其中许多包含了带有商标的公司标志。研究指出，扩散模型在隐私保护方面比早期的生成对抗网络（GAN）模型更为薄弱，强调了在训练过程中引入隐私保护技术以防止训练数据泄露的必要性。
10. https://arxiv.org/abs/2302.12173: 该资源是一篇由 Greshake 等人于 2023 年发表的学术论文，题为《Not What You’ve Signed Up for: Compromising

Real-World LLM-Integrated Applications with Indirect Prompt Injection》。论文研究了针对集成大规模语言模型（LLM）的应用中，通过间接提示注入（indirect prompt injection）攻击，攻击者如何注入恶意提示和代码，使模型检索并执行这些恶意内容，从而危害应用安全。

11. https://arxiv.org/abs/2304.06364: 该资源是微软于 2023 年发布的一项分析报告，研究了不同规模和能力的语言模型（如 GPT-3 和 GPT-4）在零样本学习（zero-shot learning）与少样本学习（few-shot learning）中的表现差异。报告指出，随着模型能力的提升，模型对指令的理解和执行能力增强，导致少样本学习带来的性能提升相较于零样本学习变得有限，但在特定领域（如 Ibis dataframe API）中，少样本示例仍能显著改善模型表现。
12. https://arxiv.org/abs/2306.04528: 该 URL https://arxiv.org/abs/2306.04528 指向的是一篇由 Zhu 等人于 2023 年发表的学术论文，内容涉及针对大规模语言模型（LLM）的对抗性攻击的鲁棒性评估基准（benchmark）——PromptRobust。该论文提出了用于评估系统在面对提示攻击（prompt attacks）时的防御能力和稳健性的测试方法和数据集，是当前研究提示工程中防御攻击的重要参考资料。
13. https://arxiv.org/abs/2307.03172: 该资源是一篇由 Liu 等人于 2023 年发表在 arXiv 上的学术论文，研究了提示（prompt）中不同位置的信息对模型理解指令效果的影响。论文指出，模型对提示开头和结尾的指令理解效果明显优于中间部分，并提出了“needle in a haystack”（NIAH）测试方法，用以评估提示中各部分信息的有效性。
14. https://arxiv.org/abs/2307.15043: 该资源是由 Zou 等人于 2023 年发布在 arXiv 上的一篇学术论文，内容涉及对语言模型安全性的研究，特别探讨了通过在输入提示中插入特殊字符（如密码样字符串）来混淆模型，从而绕过模型的安全防护机制的攻击方法。论文展示了模型在面对带有异常字符的请求时可能放弃拒绝，从而暴露安全风险，同时也指出此类攻击可通过简单的字符过滤机制加以防御。
15. https://arxiv.org/abs/2309.16797: 该资源是论文《Promptbreeder: AI-powered Prompt Optimization via Evolutionary Strategies》（Fernando

et al., 2023），介绍了一种利用进化策略优化提示词（prompt）的方法。该方法通过初始提示词生成多个变异版本，并利用 AI 模型和一组变异引导提示词，迭代选择和“繁殖”出更优的提示词，最终找到满足特定目标的高效提示词。

16. https://arxiv.org/abs/2310.03714: 该资源是由 Khattab 等人于 2023 年发布的论文，介绍了 DSPy 这一工具。DSPy 旨在自动化整个提示工程（prompt engineering）流程，用户只需指定输入输出格式、评估指标和评估数据，工具便能自动优化提示或提示链，以最大化评估指标，类似于自动机器学习（autoML）中自动寻找最佳超参数的过程。
17. https://arxiv.org/abs/2310.08419: 该资源是一篇由 Chao 等人于 2023 年发布的学术论文，介绍了一种名为“Prompt Automatic Iterative Refinement”（PAIR）的系统方法。该方法利用 AI 模型作为攻击者，自动迭代地优化提示，以实现特定目标（例如，从目标 AI 中诱导出某种类型的违规内容）。论文详细描述了该攻击流程，并通过图示进行说明。
18. https://arxiv.org/abs/2311.17035: 该资源是一篇由 Nasr 等人于 2023 年发表的学术论文，内容聚焦于通过特定的提示（prompt）策略，从大型语言模型（如 ChatGPT）中提取训练数据中的敏感信息。论文展示了一种无需预先了解训练数据具体内容，即可诱导模型泄露训练数据片段的方法，属于训练数据提取攻击（training data extraction attack）的研究。该工作揭示了语言模型在面对某些重复或特定设计的提示时，可能会输出训练数据中的原文，从而引发隐私和安全方面的担忧。
19. https://arxiv.org/abs/2404.06654: 该资源是由 Hsieh 等人在 2024 年发布的一篇论文，介绍了一种名为 RULER 的测试方法，用于评估模型处理长文本提示（long prompts）能力的效果。如果模型在上下文变长时性能显著下降，说明需要优化提示的长度以提升模型表现。
20. https://arxiv.org/abs/2404.13208: 该资源是 Wallace 等人于 2024 年发表在 arXiv 上的论文，题为《The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions》（指令层级：训练大语言模型优先处理特权指令）。论文提出了一种指令层级机制，通过训练大语言模型（LLMs）使其优先响应系统提示和高优先级指令，从而提升模型对系统提示的关注度，有效抵御恶

意注入攻击（prompt injection）等安全威胁。文中还详细介绍了四个优先级层级的指令体系结构，旨在增强模型的安全性和稳定性。

21. https://arxiv.org/abs/2406.07496: 该资源是由 Yuksekgonul 等人在 2024 年发布的一篇论文，介绍了斯坦福大学开发的 TextGrad 工具。这是一种基于 AI 的提示词（prompt）优化方法，旨在通过梯度或相关技术优化提示词，从而提升语言模型的性能和生成效果。论文详细阐述了 TextGrad 的设计原理、算法实现及其实验结果，是 AI 提示词优化领域的重要研究成果。
22. https://en.wikipedia.org/wiki/Uwu: 该资源是维基百科中关于“UwU”的页面，介绍了“UwU”这一网络用语的含义和用法。“UwU”通常用于表达可爱、撒娇或高兴的情绪，常见于网络聊天和二次元文化中。
23. https://github.com/Azure/PyRIT: 该资源是由微软 Azure 发布的开源工具库“PyRIT”，用于自动化安全测试和攻击检测，帮助评估和提升机器学习系统（尤其是大语言模型）面对对抗性攻击时的鲁棒性。它包含已知攻击模板，可自动对目标模型进行安全性探测和防御能力测试。
24. https://github.com/CHATS-lab/persuasive_jailbreaker: 该资源为由 CHATS 实验室维护的“persuasive_jailbreaker”项目，是一个用于自动化安全测试的工具，包含已知的对语言模型的劝说型绕过攻击（jailbreak）模板，帮助评估和检测目标模型在面对此类对抗性提示攻击时的鲁棒性与安全性。
25. https://github.com/LouisShark/chatgpt_system_prompt: 该资源是一个 GitHub 仓库，地址为 https://github.com/LouisShark/chatgpt_system_prompt，该仓库声称收集或公开了 ChatGPT 或类似 GPT 模型的系统提示词（system prompt）。不过，根据上下文说明，OpenAI 已确认这些所谓泄露的系统提示并非官方发布，且通过模型“反向工程”获得的系统提示往往是模型的幻觉产物，真实性难以验证。
26. https://github.com/NVIDIA/garak: 该资源是 NVIDIA 开源的一个安全测试工具“garak”，用于自动化对语言模型或相关系统进行安全性检测。它内置了已知的攻击模板，帮助开发者模拟和测试模型在面对各种对抗性提示攻击时的鲁棒性，从而提升系统的防御能力。

27. https://github.com/PlexPt/awesome-chatgpt-prompts-zh: 该资源是一个 GitHub 仓库，名为“PlexPt/awesome-chatgpt-prompts-zh”，收集和整理了大量用于 ChatGPT 的优秀中文提示语（prompts），方便用户在使用 ChatGPT 时参考和应用。
28. https://github.com/Stability-AI/stablediffusion: 该资源为开源的图像生成模型“Stable Diffusion”的代码库，托管在 GitHub 上，由 Stability AI 维护。Stable Diffusion 是一种基于扩散模型的深度学习模型，能够生成高质量的图像。文中提到该模型被用于研究训练数据的提取，揭示了扩散模型在隐私保护方面存在的挑战。
29. https://github.com/brexhq/prompt-engineering?tab=readme-ov-file: 该资源是 Brex 公司发布的“Prompt Engineering Guide”，提供关于提示工程（prompt engineering）的指导和示例，帮助用户更有效地与模型交互和控制模型行为。
30. https://github.com/f/awesome-chatgpt-prompts: 该资源是一个 GitHub 仓库，名称为“f/awesome-chatgpt-prompts”，收集并整理了大量优质的英文 ChatGPT 提示词（prompts），供用户参考和使用，帮助提升与 ChatGPT 的交互效果。
31. https://github.com/greshake/llm-security: 该资源是一个名为“greshake/llm-security”的开源工具库，托管在 GitHub 上，旨在帮助自动化检测和评估大型语言模型（LLM）在面对已知对抗性攻击（prompt attacks）时的安全性和鲁棒性。该工具包含多种已知攻击模板，用于对目标模型进行安全性测试，从而辅助开发者发现并防御潜在的提示词攻击漏洞。
32. https://github.com/guidance-ai/guidance: 该资源是一个名为 Guidance 的开源工具项目，旨在辅助提示工程（prompt engineering），通过引导语言模型生成结构化输出，帮助用户更有效地设计和优化提示语。
33. https://github.com/huggingface/transformers/issues/25304#issuecomment-1728111915: 该资源是 GitHub 上 Huggingface Transformers 仓库中某个 issue（编号 25304）中的一条评论，内容涉及工具开

发者在使用模型时可能犯的错误，具体例子包括使用了错误的模板（prompt template）。这条评论用来说明在构建提示词时可能出现的错误情况。

34. https://github.com/ibis-project/ibis: 该资源是一个名为“Ibis”的开源项目，托管在 GitHub 上，提供了一个用于数据处理和分析的 DataFrame API。Ibis 旨在简化对不同后端数据库系统的查询和操作，使用户能够使用统一的 Python 接口进行高效的数据处理。
35. https://github.com/instructor-ai/instructor: 该资源是一个名为“Instructor”的开源项目，旨在辅助提示工程（prompt engineering），通过引导模型生成结构化的输出，帮助用户设计和优化提示语以提升模型的表现。
36. https://github.com/langchain-ai/langchain/commit/7c6009b76f04628b1617cec07c7d0bb766ca1009: 该 URL 指向的是 LangChain 项目中的一次代码提交（commit），该提交修正或涉及了 LangChain 默认的“critique prompt”（批评提示）中的拼写错误。换言之，这次提交主要是对 LangChain 中用于提示模型进行批评或评价的默认模板进行了拼写或文本内容的修正。
37. https://github.com/langchain-ai/langchain/issues/1026: 该资源是 LangChain 项目在 GitHub 上的一个问题（issue）页面，记录了 2023 年发现的与远程代码执行（RCE）风险相关的安全问题。
38. https://github.com/langchain-ai/langchain/issues/814: 该 URL 指向 LangChain 项目在 GitHub 上的一个问题(issue)讨论页面，内容涉及 2023 年发现的 LangChain 中远程代码执行（RCE）风险的相关报告和讨论。
39. https://github.com/lmstudio-ai/.github/issues/43: 该资源是 GitHub 上 lmstudio-ai 组织的一个问题(issue)讨论帖，内容涉及“聊天模板不匹配”的问题报告。用户在该帖中反馈了因模型版本更新导致聊天模板未同步更新，从而引发相关功能异常或调试困难的情况。
40. https://github.com/outlines-dev: 该资源是一个名为 “Outlines” 的开源项目，托管在 GitHub 上，旨在帮助引导语言模型生成结构化输出，作为提示工程（prompt engineering）辅助工具之一。
41. https://github.com/promptfile/promptfile: 该资源是一个名为“Promptfile”的开源项目，提供了一种特殊的 .prompt 文件格式，用于存储和管理提示词

（prompts），类似于 Google Firebase 的 Dotprompt 文件格式，旨在方便开发者组织和复用提示模板。

42. https://github.com/thunlp/Advbench: 该资源“https://github.com/thunlp/Advbench”是一个由 Chen 等人（2022 年）开发的用于评估系统对抗性攻击鲁棒性的基准测试（benchmark）项目 Advbench。它提供了一套用于测试和衡量机器学习模型，特别是自然语言处理系统，在面对对抗性攻击时的安全性和稳健性的工具和数据集。
43. https://oreil.ly/-HMpk: （https://docs.anthropic.com/claude/docs/prompt-engineering）该资源是 Anthropic 官方提供的关于“提示工程”（prompt engineering）的文档，介绍了设计和优化与 Anthropic 模型（如 Claude）交互时所使用提示词的技巧和方法。这些内容涵盖了通用且行之有效的提示设计策略，帮助用户更好地利用 Anthropic 的语言模型，实现更准确和高效的生成效果。
44. https://oreil.ly/-u2Z5: （https://platform.openai.com/docs/guides/prompt-engineering/strategy-split-complex-tasks-into-simpler-subtasks）该资源是 OpenAI 官方提供的提示工程（prompt engineering）指南中的一部分，具体介绍了将复杂任务拆解为更简单子任务的策略，帮助用户设计更有效的提示语以提升模型的理解和执行能力。
45. https://oreil.ly/06vdM: （https://docs.google.com/spreadsheets/d/19jzLgRruG9kjUQNKtCg1ZjdD6l6weA6qRXG5zLIAhC8/edit#gid=150872633）该 URL 指向的资源是一个 Google 电子表格文档，可能包含与“提示工程”（prompt engineering）相关的示例内容，用于演示如何通过提供示例来引导模型生成期望的响应。该文档可能参考了 Claude 的提示工程教程，旨在帮助用户更好地设计和优化与 AI 模型的交互提示。
46. https://oreil.ly/0NoUv: （https://www.reddit.com/r/ChatGPT/comments/zlcyr9/dan_is_my_new_friend/）该 URL 指向 Reddit 上一个名为“dan_is_my_new_friend”的帖子，内容与“DAN（Do Anything Now）”这一在 ChatGPT 早期流行的“越狱”提示语（jailbreak prompt）相关。帖子讨论了 DAN 提示的起源、演变及其通过角色扮演方式绕过模型限制的机制。

47. https://oreil.ly/0h0qN: （https://ipwatchdog.com/2024/01/02/can-ai-prompts-patented-dont-quick-dismiss-question/id=171338/）该资源是一篇发表于 2024 年 1 月 2 日的文章，讨论了人工智能提示语（AI prompts）是否能够作为专利保护的对象。文章探讨了团队将提示语视为专有技术的现象，以及目前在专利法框架下对 AI 提示语专利性问题的争议和缺乏明确先例的情况。
48. https://oreil.ly/6Bfe7: （https://fchollet.substack.com/p/how-i-think-about-llm-prompt-engineering）该资源是一篇由 François Chollet 撰写的文章，题为《How I Think About LLM Prompt Engineering》，发表在其 Substack 博客上。文章深入探讨了大型语言模型（LLM）中的“提示工程”概念，阐述了如何通过设计和选择合适的提示（prompt）来激活模型中不同的“程序”或能力，从而实现灵活多样的任务处理。文章结合作者作为 Keras 框架创始人的视角，帮助读者理解 LLM 的内在机制及其在上下文学习中的表现原理。
49. https://oreil.ly/AF-Y1: （https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results）该资源是 OpenAI 官方文档中关于“提示工程”的指导内容，具体介绍了六种提升生成模型输出效果的通用策略。这些策略经过多家领先 AI 公司（如 OpenAI、Anthropic、Meta 和 Google）验证，适用于多种模型，且具有较强的长期参考价值。文档旨在帮助用户通过优化提示设计，获得更准确、高效的 AI 生成结果。
50. https://oreil.ly/BPAal: （https://www.reddit.com/r/ChatGPT/comments/10tevu1/new_jailbreak_proudly_unveiling_the_tried_and/）该 URL（https://www.reddit.com/r/ChatGPT/comments/10tevu1/new_jailbreak_proudly_unveiling_the_tried_and/）指向的是 Reddit 论坛中一个关于 ChatGPT“jailbreak”攻击的帖子，内容大概介绍或展示了一种新的或经过改进的“jailbreak”提示词（prompt）方法，通常涉及角色扮演（roleplaying）技巧，用以绕过模型的限制，实现让模型“什么都能做”的目的。帖子可能详细说明了这种方法的细节、效果及其演变过程。
51. https://oreil.ly/BToYu: （https://www.kdnuggets.com/why-prompt-engineering-is-a-fad）该资源是一篇发表在 KDnuggets 上的文章，标题为《Why Prompt Engineering Is a Fad》（为何提示工程是一种昙花一现的现象）。

文章内容围绕“提示工程”这一新兴概念展开，探讨了其在短时间内引发的争议和质疑，分析了为何有观点认为提示工程并非真正的独立学科或长期趋势，而更像是一种暂时的流行现象。

52. https://oreil.ly/CGyGU:（https://console.cloud.google.com/vertex-ai/generative/prompt-gallery）该资源是谷歌云平台（Google Cloud Console）中 Vertex AI 的“生成式 AI 提示库”页面，提供了一系列预先设计好的生成式 AI 提示（prompts），供开发者参考和使用，以帮助优化和提升生成式模型的效果和应用表现。
53. https://oreil.ly/DFjgW:（https://arxiv.org/pdf/2308.01990）该资源是论文《From Prompt Injections to SQL Injection Attacks》（Pedro 等，2023），其内容主要研究了提示注入攻击（Prompt Injections）与传统 SQL 注入攻击之间的关系，分析了在使用提示工具（如 LangChain）时默认提示模板存在的安全隐患，特别是这些模板过于宽松导致注入攻击的高成功率，并探讨了通过添加限制以防范此类攻击的有效性。论文地址为：https://arxiv.org/pdf/2308.01990。
54. https://oreil.ly/DNj9O:（https://dropbox.tech/machine-learning/bye-bye-bye-evolution-of-repeated-token-attacks-on-chatgpt-models）该资源是 Dropbox 发布的一篇技术博客文章，题为《Bye Bye Bye…: Evolution of repeated token attacks on ChatGPT models》。文章详细介绍了针对 ChatGPT 模型的一类攻击方式——重复 token 攻击（repeated token attacks），包括通过让模型重复文本或多次重复提示语来诱发模型异常行为的攻击演变过程和技术细节。
55. https://oreil.ly/DXAgC:（https://llama.meta.com/docs/how-to-guides/prompting/）该 URL（https://llama.meta.com/docs/how-to-guides/prompting/）所指向的资源是 Meta 官方提供的关于“提示工程”（prompting）的使用指南或教程文档。该文档介绍了针对其语言模型（如 LLaMA）设计有效提示词的通用技术和最佳实践，内容基于模型提供商的经验总结，旨在帮助用户高效利用生成式 AI 模型。
56. https://oreil.ly/FQP7J:（https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-2/）该资源是 Meta 官方提供的关于 Llama 2 模

型的文档页面，详细介绍了 Meta Llama 2 模型的模型卡（model cards）和提示词格式（prompt formats），包括如何构造系统提示和用户提示的合并模板，帮助用户理解和正确使用 Llama 2 聊天模型。

57. https://oreil.ly/FuBEI: （https://humanloop.com/docs/prompt-file-format）该资源是 Humanloop 提供的关于 .prompt 文件格式的文档，介绍了如何使用特定的文件格式来存储和管理提示词（prompts），以便在自然语言处理或生成模型应用中更方便地组织和复用提示内容。
58. https://oreil.ly/J3Crv: （https://cursor.directory/）该资源是一个名为“Cursor Directory”的公共提示（prompt）市场平台，用户可以在此浏览、投票和分享各种高质量的 AI 提示（prompts），以帮助提升生成式 AI 模型的效果和应用体验。
59. https://oreil.ly/LH3wI: （https://www.reddit.com/r/LocalLLaMA/comments/1ax8cbs/chatml_prompt_template_works_better_than_model/）该资源是 Reddit 上的一个讨论帖，主题是关于 ChatML 提示模板的使用，内容指出虽然偏离预期的聊天模板通常会导致模型性能下降，但在某些情况下，使用不同的 ChatML 提示模板反而能提升模型表现。
60. https://oreil.ly/N2fup: （https://ai.stanford.edu/blog/understanding-incontext/）该资源是斯坦福人工智能实验室发布的一篇博客文章，题为“Understanding In-Context Learning”（理解上下文学习）。文章深入探讨了“in-context learning”（上下文学习）的原理和机制，解释了为什么大型基础模型（如 GPT-3）能够通过提供示例或提示，在不进行额外训练的情况下完成各种任务。文中还引用了知名机器学习专家 François Chollet 的观点，将基础模型比作包含多个程序的“程序库”，通过不同的提示激活相应程序，从而实现多功能的表现。该博客旨在帮助读者深入理解当下广泛应用的上下文学习技术的本质和背景。
61. https://oreil.ly/OEwKu: （https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-chat）该资源是 Google Cloud Vertex AI 生成式人工智能相关的文档，具体介绍了“文本聊天”模型的参考信息，包括如何使

用和配置文本聊天模型进行对话生成，详述模型的功能、参数设置以及上下文（context）在对话中的作用。

62. https://oreil.ly/PR9a3:（https://docs.anthropic.com/en/prompt-library/library）该资源是 Anthropic 官方提供的提示词（prompt）库，包含经过精心设计和验证的预设提示模板，旨在帮助开发者更高效地进行提示工程，提升生成式 AI 模型的应用效果。
63. https://oreil.ly/TYoZj:（https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming）该资源是微软官方文档，介绍了在大型语言模型（LLMs）领域中如何进行“红队演练”（red teaming）的相关概念和方法，帮助组织通过模拟攻击来识别和防范潜在的安全风险，从而提升系统的安全性。
64. https://oreil.ly/TqmnZ:（https://crfm.stanford.edu/2023/12/19/helm-lite.html）该资源是一篇由斯坦福大学于 2023 年 12 月 19 日发布的文章，介绍了其 HELMLite 基准测试的最新调整，具体内容包括斯坦福在 2023 年底决定从 HELMLite 基准中去除“鲁棒性”（robustness）这一评测维度。
65. https://oreil.ly/Ukk7e:（https://promptbase.com/）该资源是一个名为 PromptBase 的在线平台，提供用户买卖和交易高质量、经过精心设计的提示词（prompts）的市场服务，方便用户获取和共享有效的 AI 提示词以提升生成效果。
66. https://oreil.ly/UxtYv:（https://trans.enby.town/notice/AUjhC6QLd2dQzsVXe4）该 URL 指向的资源很可能是一则关于 AI 模型安全或绕过安全机制相关的公告或说明，内容涉及利用“祖母漏洞”（grandma exploit）等角色扮演手法来规避模型的安全限制，进而获取敏感或受限信息的案例分析或警示。
67. https://oreil.ly/WMn2L:（https://platform.openai.com/docs/examples）该资源是 OpenAI 官方文档中提供的示例集合，包含经过验证且实用的提示工程（prompt engineering）技巧和示例代码，帮助开发者更有效地使用 OpenAI 模型进行生成式 AI 应用的开发。
68. https://oreil.ly/Z5w1L:（https://www.anthropic.com/news/claude-3-5-sonnet）该资源是 Anthropic 公司于 2024 年发布的“Claude 3.5 Sonnet”模

型相关内容，展示了该 AI 模型生成的示例提示（prompt），用于说明其在自动生成和改进提示语方面的能力。

69. https://oreil.ly/_c5FF: （https://www.godaddy.com/resources/news/llm-from-the-trenches-10-lessons-learned-operationalizing-models-at-godaddy）该 URL 指向 GoDaddy 发布的一篇关于大语言模型（LLM）实际应用经验的文章，标题为“LLM from the Trenches: 10 Lessons Learned Operationalizing Models at GoDaddy”。文章分享了 GoDaddy 在将大型语言模型投入实际运营中总结的十条经验教训，内容涵盖如何优化模型提示（prompt）、提升模型性能、降低成本等实用方法，帮助开发者更有效地构建和部署基于 LLM 的应用。
70. https://oreil.ly/_fXnT: （https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation）该资源是一篇关于微软 2016 年发布的聊天机器人 Tay 因发表种族歧视言论而引发的事件分析文章，探讨了这一事件揭示的在线对话中存在的风险和挑战，以及人工智能系统在内容安全管理方面的不足和潜在危害。
71. https://oreil.ly/aFeyE: （https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf）该资源是一本由 Google 发布的关于在 Google Workspace 环境中使用 Gemini 模型进行提示工程（prompt engineering）的指导手册，标题为《Gemini for Google Workspace Prompting Guide 101》。该文档总结了适用于多种生成式 AI 模型的一般提示设计技巧和最佳实践，旨在帮助用户更有效地利用 Gemini 模型提升工作效率，特别是在 Google Workspace 办公套件中的应用。
72. https://oreil.ly/aKDb1: （https://tech.instacart.com/scaling-productivity-with-ava-instacarts-internal-ai-assistant-ed7f02558d84）该 URL 指向 Instacart 官方技术博客上一篇文章，介绍了 Instacart 内部开发和使用的 AI 助手“AVA”，重点讲述了如何通过该 AI 助手提升员工生产力，实现规模化效率优化。文章详述了 AVA 的功能、应用场景以及对公司工作流程的影响。
73. https://oreil.ly/b_H2s: （https://hamel.dev/blog/posts/prompt/）该资源是 Hamel Husain 于 2024 年 2 月 14 日发布的一篇博客文章，标题为“Show Me

the Prompt”，内容围绕提示工程（prompt engineering）的方法论和实践进行阐述。

74. https://oreil.ly/bzK_g:（https://hamel.dev/notes/llm/05_tokenizer_gotchas.html）该资源是一篇关于大语言模型（LLM）分词器（tokenizer）使用中的常见陷阱和注意事项的技术文章，详细讨论了工具开发者在构造提示词（prompt）时可能出现的错误，如使用错误的模板、错误地拼接分词而非原始文本、以及提示模板中的拼写错误等问题。文章旨在帮助开发者避免这些细节错误，从而提升基于 LLM 的工具和应用的稳定性和准确性。
75. https://oreil.ly/ceZLs:（https://firebase.google.com/docs/genkit/dotprompt）该资源是 Google Firebase 提供的关于 Dotprompt 文件格式的官方文档，介绍了一种用于存储和管理提示词（prompts）的特殊文件格式及其使用方法。
76. https://oreil.ly/lKOrj:（https://www.cnet.com/tech/services-and-software/glue-in-pizza-eat-rocks-googles-ai-search-is-mocked-for-bizarre-answers/）该资源是一篇来自 CNET 的文章，标题大致为“粘合剂在披萨上，吃石头，谷歌的 AI 搜索因怪异回答被嘲笑”，内容讨论了谷歌 AI 搜索出现奇怪或不当回答（例如建议用户“吃石头”）的情况，以及类似事件如何导致公众对 AI 安全性和责任感的质疑。文章可能分析了 AI 生成内容中的错误和争议，以及相关公司因此面临的公关危机。
77. https://oreil.ly/mB3D7:（https://news.ycombinator.com/item?id=35665168）该 URL 指向的是 Hacker News（新闻聚合社区）上的一个讨论帖，帖子的内容很可能与“prompt engineering”（提示工程）相关，尤其是围绕该领域存在的争议和批评展开讨论。具体来说，该讨论可能涉及对 prompt engineering 是否为“真正学科”的质疑，以及相关话题引发的激烈评论和观点交流。
78. https://oreil.ly/nQZIB:（https://observablehq.com/@shreyashankar/needle-in-the-real-world-experiments）该资源是 Shreya Shankar 在 Observable 平台上发布的一篇关于“针在真实世界中”实验的详细介绍，内容聚

焦于她针对医生就诊场景进行的实际 NIAH（Needle In A Haystack）测试和分析，展示了如何在真实环境中应用该方法进行有效的数据检测和验证。

79. https://oreil.ly/nriHw: （https://docs.continue.dev/features/prompt-files）该资源介绍了 Continue Dev 提出的特殊 .prompt 文件格式，用于存储和管理提示（prompts），以便在开发和使用中更方便地组织和复用提示内容。
80. https://oreil.ly/o-fXF: （https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3/）该资源是 Meta 官方关于 Llama 3 模型的文档页面，介绍了模型卡（model cards）和提示格式（prompt formats），详细说明了 Llama 3 聊天模型所使用的不同聊天模板及其版本更新中的变化。
81. https://oreil.ly/q1EHt: （https://prompthero.com/）该 URL（https://prompthero.com/）指向的是一个公共的提示词（prompt）市场或平台，用户可以在该平台上浏览、分享和为喜欢的提示词投票。该网站致力于汇集和管理高质量的 AI 提示词资源，方便用户发现和获取有效的提示词以提升人工智能生成内容的效果。
82. https://oreil.ly/qpjty: （https://bit.ly/chip-mlops-discord）该资源是一个 Discord 社区链接，可能指向与机器学习、MLOps（机器学习运维）相关的讨论组或频道，用户可以在其中参与关于模型上下文、提示设计等主题的交流与讨论。
83. https://oreil.ly/svNY-: （https://www.reddit.com/r/ChatGPT/comments/12wxn1w/prompt_engineering_is_easy_as_shit_and_anybody/）该 URL 指向 Reddit 上 r/ChatGPT 社区的一个帖子，标题为“prompt engineering is easy as shit and anybody”（提示工程非常简单，任何人都能做到）。该帖子很可能讨论了提示工程（prompt engineering）的实际难易程度，表达了作者或社区成员对提示工程被过度复杂化或被质疑真实性的看法，反映了围绕提示工程的争议和不同观点。
84. https://oreil.ly/tk4lu: （https://www.reddit.com/r/ChatGPT/comments/1529r45/wtf_is_with_people_saying_prompt_engineer_like/）该资源是 Reddit 上的一个讨论帖，标题为“WTF is with people saying ‘prompt engineer’ like…”，内容主要围绕“prompt engineering”（提示词工程）

这一概念展开，反映了人们对该职业名称的质疑和争议，以及对其真实性和专业性的讨论。

85. https://oreil.ly/yqAZs: （https://docs.anthropic.com/en/docs/chain-prompts#parallel-processing）该 URL 指向 Anthropic 官方文档中关于“链式提示（chain prompts）”的章节，具体介绍了如何在提示工程中实现并行处理的示例和方法。
86. https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product: 该资源是 LinkedIn 工程博客中关于生成式人工智能产品开发的文章，内容探讨了生成式 AI 技术的构建经验与挑战，特别提到了“Chain-of-Thought（CoT）”提示技术如何帮助模型系统性地推理和解决问题，并且有效降低了模型产生虚假信息（hallucinations）的概率。
87. https://x.com/DrJimFan/status/1631709224387624962: 该资源是一条来自用户 DrJimFan 在社交平台 X（原 Twitter）上的动态，内容涉及语言模型（LLM）如何通过混合使用多种语言来隐藏恶意关键词，从而绕过关键词过滤机制。
88. https://x.com/___frye/status/1598400965656596480: 该资源是一条来自用户“___frye”的推文，内容是以“UwU”风格（网络上的一种可爱、拟声化的文字表达方式）描述如何在家中进行“浓缩铀”的方法。这条推文作为示例，展示了通过改变输出格式（如写成诗歌、说唱或特殊语言风格）绕过模型内容限制的攻击手法。
89. https://x.com/abacaj/status/1786436298510667997: 该资源是一条关于 Llama 3 模型提示词设计经验的推文，讨论了将任务描述放在提示词末尾而非开头时，Llama 3 模型表现更佳的现象。
90. https://x.com/dylan522p/status/1755086111397863777: 该资源是指向一条发表于 2024 年 2 月的推文，内容涉及用户声称 ChatGPT 的系统提示词长度达到 1700 个 token 的讨论。该推文被用来说明有关 ChatGPT 系统提示词长度的传闻，但文中也指出这些所谓的系统提示词泄露多数是模型生成的幻觉，未被官方确认。
91. https://x.com/haus_cole/status/1598541468058390534: 该资源是 X（原推特）用户 @haus_cole 发布的一条推文，内容展示了利用已有的攻击示例，指导

模型自动生成新的提示攻击（prompt attacks）的可能性，体现了提示工程中自动化攻击的方法和思路。

92. https://x.com/himbodhisattva/status/1598192659692417031: 该 URL 指向的资源内容是关于“Filter Improvement Mode”（过滤器改进模式）的介绍或讨论。该模式被描述为一种关闭限制的特殊状态，通常用于绕过模型的安全防护机制，从而允许模型输出在正常情况下会被限制或过滤的信息。该推文很可能涉及对这种模式的演示、解释或相关的示例说明。
93. https://x.com/muneebtator/status/1598668909619445766: 该资源是一条推文，内容为通过输出格式的巧妙转换，使语言模型生成关于“抢劫房屋”主题的创意文本（如说唱歌词）。即攻击者利用让模型以特定文体（如说唱）表达敏感或违规内容的方法，绕过模型的安全限制。
94. https://x.com/proofofbeef/status/1598481383030231041: 该资源是指向一条社交媒体动态（X.com 上的一条推文），内容涉及“模拟器”或“仿真环境”，在该环境中，模型被设定为处于一个类似地球但没有限制的状态，常用于绕过 AI 模型的安全限制或内容过滤机制的角色扮演攻击示例。
95. https://x.com/remoteli_io/status/1570547034159042560: 该资源是 X（前 Twitter）用户@mkualquiera 于 2022 年发布的一条推文，内容示例展示了通过向模型输入特定提示（prompt）来“欺骗”或引导模型重复其初始系统提示（system prompt）的方法，属于逆向工程提示词（reverse prompt engineering）的一种示范。
96. https://x.com/synt7_x/status/1601014197286211584: 该资源为推特（现称 X）用户 synt7_x 发布的一条推文，内容涉及利用“NSA 特工”角色扮演的秘密代码来绕过人工智能模型的安全防护机制，属于讨论绕过 AI 安全限制的相关话题。
97. https://x.com/zswitten/status/1598197802676682752: 该资源是推特（X）上的一条推文，内容涉及通过输出格式操控的方式让语言模型生成有关制作莫洛托夫鸡尾酒（Molotov cocktail）的代码示例。
98. https://x.com/zswitten/status/1599090459724259330: 该 URL 指向的资源是一条关于利用 Unicode 字符来绕过大型语言模型（LLM）关键词过滤的讨论或示

例推文。推文内容可能涉及通过混合使用 Unicode 字符，使得模型难以检测出恶意关键词，从而规避内容审核机制。

99. https://xkcd.com/327: 该资源是著名网络漫画 xkcd 中编号为 327 的漫画，标题为“Exploits of a Mom”。这幅漫画以幽默的方式展示了 SQL 注入攻击的概念，是讨论数据库安全和 SQL 注入漏洞时常被引用的经典作品。

第 6 章（58）

1. https://arxiv.org/abs/1702.08734: 该资源是论文《FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors》（Johnson et al., 2017），介绍了 Facebook 开发的 FAISS 库，该库用于高效地进行大规模向量的近似最近邻搜索和聚类，是处理大规模向量搜索问题的重要工具。
2. https://arxiv.org/abs/1704.00051: 该资源是论文《Reading Wikipedia to Answer Open-Domain Questions》（Chen et al., 2017），首次提出了“检索-生成”（retrieve-then-generate）的方法。该方法先从维基百科中检索与问题最相关的若干页面，再利用长短期记忆网络（LSTM）模型读取这些页面的信息，生成问题的答案。这项工作是自然语言处理领域将检索与生成结合用于开放领域问答的开创性研究。
3. https://arxiv.org/abs/1809.09600: 该资源是 2018 年由 Yang 等人发表的一篇论文，介绍了 HotpotQA，这是一个用于多跳问答（multi-hop question answering）的基准数据集和任务，旨在测试模型在跨多个文档进行推理和回答复杂问题方面的能力。
4. https://arxiv.org/abs/2005.04611: 该资源是一篇由 Petroni 等人于 2020 年 5 月发布在 arXiv 上的学术论文，题为“How Context Affects Language Models’ Factual Predictions”。论文研究了通过为预训练语言模型增加检索系统，显著提升模型在回答事实性问题时的表现，旨在减少模型生成错误信息（幻觉）的情况。

5. https://arxiv.org/abs/2005.11401: 该资源是由 Lewis 等人在 2020 年发表的论文《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》（arXiv 编号 2005.11401），该论文提出了一种称为“检索增强生成”（Retrieval-Augmented Generation, RAG）的方法，旨在解决知识密集型自然语言处理任务中模型无法直接输入全部知识的问题。RAG 通过检索器从外部知识库中提取与查询最相关的信息，输入到生成模型中，从而帮助模型生成更详尽且更准确的回答，减少生成内容中的虚假信息（hallucinations）。论文中还介绍了联合训练检索器和生成模型以提升整体性能的技术。
6. https://arxiv.org/abs/2103.00020: 该资源是论文《Learning Transferable Visual Models From Natural Language Supervision》（作者：Radford 等，2021 年），介绍了 CLIP 模型——一种通过自然语言监督学习的多模态（图像-文本）嵌入模型，能够将图像和文本映射到同一向量空间，从而实现基于文本的图像检索和内容比较。
7. https://arxiv.org/abs/2104.08663: 该资源是论文《BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models》（Thakur et al., 2021），介绍了 BEIR，一个用于信息检索模型零样本评估的异构基准测试集，支持在 14 个常见的信息检索基准上对检索系统进行统一评测。
8. https://arxiv.org/abs/2107.05720: 该资源是论文《SPLADE: Sparse Lexical and Expansion Model for Neural Information Retrieval》（Formal et al., 2021），介绍了一种基于 BERT 生成稀疏嵌入的检索算法 SPLADE。该方法通过正则化使得大部分嵌入向量值趋近于零，从而实现稀疏表示，提升了嵌入操作的计算效率，属于稀疏向量检索技术的一种。
9. https://arxiv.org/abs/2210.03629: 该资源是论文《ReAct: Synergizing Reasoning and Acting in Language Models》（Yao et al., 2022），提出了一种将推理（包括规划和反思）与行动交织进行的方法，用于指导智能体在完成任务时交替进行思考和执行操作，从而提高任务完成的效率和效果。
10. https://arxiv.org/abs/2210.07316: 该资源是指向一篇发表在 arXiv 上的学术论文，题为“MTEB”（Massive Text Embedding Benchmark），由 Muennighoff 等人于 2023 年发布。该论文介绍了一个用于评估文本嵌入

（embeddings）质量的基准测试框架，涵盖了包括语义检索、文本分类和聚类在内的多种下游任务，旨在全面衡量文本嵌入模型在不同应用场景中的表现。

11. https://arxiv.org/abs/2210.08750: 该资源是由 Bae 等人于 2022 年发表的一篇学术论文，介绍了一种通过对对话内容进行摘要生成并结合命名实体跟踪来去除冗余的方法。论文进一步提出了一种分类器，用于判断对话记忆中的各句子与摘要中的各句子，哪些应被加入到新的记忆中，从而构建更有效的对话记忆表示。
12. https://arxiv.org/abs/2302.04761: 该资源是 2023 年由 Schick 等人发表在 arXiv 上的一篇论文，介绍了一种名为 Toolformer 的方法，该方法通过微调 GPT-J 模型，使其能够学习并使用五种不同的工具，从而增强语言模型的功能和应用能力。
13. https://arxiv.org/abs/2303.11366: 该资源是由 Shinn 等人于 2023 年发表在 arXiv 上的论文，介绍了一种名为“Reflexion”的方法框架。在该框架中，反思过程被分为两个模块：一个评估器用于评价结果，另一个自我反思模块用于分析错误原因。论文中提出，智能体在每一步通过评估和自我反思后，会提出新的“轨迹”（即计划），以不断改进行为表现。
14. https://arxiv.org/abs/2304.09842: 该资源是论文《Chameleon: A GPT-4-powered Agent Augmented with Tools》（Lu et al., 2023），介绍了一种基于 GPT-4 的智能代理 Chameleon，该代理通过集成 13 种工具（如知识检索、查询生成器、图像描述器、文本检测器及 Bing 搜索等）显著提升了模型在多个基准测试中的表现。论文还提出将工具调用过程视为“程序生成器”，探讨了工具
15. https://arxiv.org/abs/2305.14992: 该资源是一篇由 Hao 等人于 2023 年发表在 arXiv 上的学术论文，题为《Reasoning with Language Model is Planning with World Model》。论文提出，语言大模型（LLM）不仅能生成一系列动作序列，还能基于其包含的大量世界知识，预测每个动作的结果，从而将这种结果预测能力融合进推理过程中，实现更连贯和有效的规划。
16. https://arxiv.org/abs/2305.15334: 该资源是由 Patil 等人于 2023 年发表的论文，题为 “Gorilla”，介绍了一种通过提示（prompting）方法，使智能代理能够从多达 1,645 个 API 中选择合适的调用接口，从而提升代理在工具使用上的灵活性和智能化水平。

17. https://arxiv.org/abs/2305.16291: 该资源是由 Wang 等人在 2023 年发表的一篇论文，介绍了一种名为“skill manager”的方法，用于管理智能体（agent）在执行任务过程中获得的新技能（通常是代码程序）。该方法能够识别和保存有用的技能，将其加入技能库以便后续复用，从而提升智能体完成多任务的能力。
18. https://arxiv.org/abs/2311.08719v1: 该资源是刘等人（2023 年）发表在 arXiv 上的一篇论文，文中提出了一种基于反思（reflection）的方法，具体做法是在每次动作后让智能体执行两项任务，以改进其行为策略。
19. https://arxiv.org/abs/2405.15793: 该资源是由 Yang 等人在 2024 年发布的一篇论文，介绍了 SWE-agent——一个基于 GPT-4 构建的智能代理。该代理运行于计算机环境中，能够通过终端和文件系统执行一系列操作，如浏览代码仓库、搜索文件、查看文件内容以及编辑代码行。
20. https://en.wikipedia.org/wiki/Discounted_cumulative_gain: 该资源介绍了“折损累积增益”（Discounted Cumulative Gain，简称 DCG）及其归一化形式 NDCG，这是一种用于信息检索和排序系统中评估检索结果相关性和排序质量的指标。NDCG 通过考虑文档的相关性以及其在结果列表中的排名位置，衡量检索系统将更相关文档排在前面的能力。
21. https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision: 该资源介绍了信息检索中的评价指标，特别是“Mean Average Precision”（MAP，平均准确率均值），它用于衡量检索结果中文档排序的质量，强调相关文档应排在前面，从而评估检索系统的排序效果。
22. https://en.wikipedia.org/wiki/Long_short-term_memory: 该资源介绍了长短期记忆网络（Long Short-Term Memory，简称 LSTM），这是一种特殊的循环神经网络（RNN）结构，广泛应用于处理和预测序列数据，尤其在自然语言处理（NLP）领域曾长期占据主导地位，直到 2018 年被 Transformer 架构逐渐取代。
23. https://en.wikipedia.org/wiki/Mean_reciprocal_rank: 该资源是关于“Mean Reciprocal Rank（MRR）”的维基百科页面，介绍了一种用于信息检索

和排序评估的指标，主要用于衡量检索结果中第一个相关文档的排名情况，反映系统在排序相关性上的表现。

24. https://en.wikipedia.org/wiki/Okapi_BM25: 该资源是维基百科中关于“Okapi BM25”的条目页面，介绍了 Okapi BM25 算法的原理、发展历史及其在信息检索中的应用。Okapi BM25 是一种基于概率模型的文本匹配算法，是 TF-IDF 的改进版本，通过对文档长度进行归一化处理，提高了检索结果的相关性评估效果。
25. https://en.wikipedia.org/wiki/Recurrent_neural_network: 该资源是关于“循环神经网络”（Recurrent Neural Network，RNN）的维基百科页面，介绍了一种用于处理序列数据的神经网络模型，能够通过循环结构捕捉时间序列中的上下文信息，是自然语言处理等领域的重要基础技术。
26. https://en.wikipedia.org/wiki/Reinforcement_learning: 该资源是维基百科上关于“强化学习”（Reinforcement Learning）的条目，介绍了强化学习作为人工智能领域的一个分支，重点研究智能体如何在动态环境中采取行动以最大化累积奖励的理论和方法。
27. https://github.com/AgentOps-AI/agentops: 该资源是一个名为“AgentOps evaluation harness”的 GitHub 项目，用于对智能体（agents）进行评测和基准测试的工具或框架。
28. https://github.com/OSU-NLP-Group/TravelPlanner: 该资源是俄亥俄州立大学自然语言处理小组（OSU-NLP-Group）开发的一个名为“TravelPlanner”的基准测试项目，用于评估和比较智能体在旅行规划任务中的表现。
29. https://github.com/aie-book: 该资源是《aie-book》一书的 GitHub 代码仓库，包含用于展示书中提到的各种失败模式的简单基准测试代码，以及相关的代理基准测试和排行榜工具。
30. https://github.com/apache/lucene: 该资源是 Apache Lucene 的代码仓库，Lucene 是一个高性能的全文搜索引擎库，提供基于倒排索引的数据结构，用于快速检索包含特定词项的文档，广泛应用于构建搜索和信息检索系统。
31. https://github.com/elastic/elasticsearch: 该资源是 Elasticsearch 项目的 GitHub 仓库，Elasticsearch 是由 Shay Banon 于 2010 年创建的一个基于

Apache Lucene 构建的开源分布式搜索引擎，主要利用倒排索引数据结构实现高效的文本检索和相关性评分。

32. https://github.com/explosion/spaCy: 该资源是 spaCy，一个由 Explosion 团队维护的开源自然语言处理（NLP）库，提供高效的文本分词、词性标注、命名实体识别等功能，常用于构建基于词项的检索和其他 NLP 应用。
33. https://github.com/flann-lib/flann: 该资源是一个名为 FLANN（Fast Library for Approximate Nearest Neighbors）的开源库，提供高效的近似最近邻搜索算法实现，常用于处理大规模高维数据的快速相似性检索。
34. https://github.com/grantjenks/py-tree-sitter-languages#license: 该资源是一个 GitHub 项目页面中的“license”部分，具体链接指向“py-tree-sitter-languages”项目的许可协议说明页面，介绍该项目所采用的开源许可证及相关使用条款。
35. https://github.com/microsoft/SPTAG: 该资源是微软发布的开源库 SPTAG（Space Partition Tree And Graph），用于高效的近似最近邻搜索算法实现。
36. https://github.com/nmslib/hnswlib: 该资源是“Hnswlib”，一个基于 HNSW（Hierarchical Navigable Small World）算法的高效近似最近邻（ANN）搜索库，适用于大规模向量数据的快速相似度搜索和检索。该库由 Malkov 和 Yashunin 于 2016 年提出，并开源托管在 GitHub 上。
37. https://github.com/noahshinn/reflexion: 该资源是一个名为“Reflexion”的 GitHub 代码仓库，提供了关于“Reflexion agents”工作原理的示例和相关内容。
38. https://github.com/spotify/annoy: 该资源“https://github.com/spotify/annoy”指向的是 Annoy 项目的 GitHub 仓库。Annoy（Approximate Nearest Neighbors Oh Yeah）是由 Spotify 开发的一个用于高效进行近似最近邻搜索的库，通常用于大规模的相似性检索和推荐系统中。
39. https://github.com/stanfordnlp/CoreNLP: 该资源是斯坦福大学开发的自然语言处理工具包 Stanford CoreNLP，提供包括分词（tokenization）在内的多种经典 NLP 功能，方便用户进行文本预处理和语言分析。

40. https://oreil.ly/-JJYn: （https://ieeexplore.ieee.org/document/6182576）该 URL（https://ieeexplore.ieee.org/document/6182576）指向的是 Sanderson 和 Croft 于 2012 年 4 月在《Proceedings of the IEEE》期刊“Special Centennial Issue”上发表的综述文章《The History of Information Retrieval Research》。该文详细回顾了信息检索领域的发展历程，从 1920 年代 Emanuel Goldberg 关于统计机器的早期专利谈起，系统总结了信息检索技术的历史演变及其研究进展。
41. https://oreil.ly/-Sny7: （https://www.anthropic.com/news/contextual-retrieval）该资源介绍了 Anthropic 公司在信息检索中的“上下文增强检索”（contextual retrieval）技术，具体讲述如何通过 AI 模型为文档分块生成简短的上下文描述（通常 50-100 个词），以帮助检索系统更好地理解每个分块的内容及其与原始文档的关系，从而提升检索效果。
42. https://oreil.ly/3xtwh: （https://plg.uwaterloo.ca/~gvcormac/cormacksigir09-rrf.pdf）该资源是 Cormack 等人在 2009 年发表的一篇关于“reciprocal rank fusion (RRF)”算法的学术论文，详细介绍了该算法如何通过结合不同检索器的排名来对文档进行评分，从而提升信息检索系统的效果。论文内容包括算法原理、计算方法及其实验验证。
43. https://oreil.ly/4ATBC: （https://github.com/nmslib/hnswlib）该资源是一个名为“Hnswlib”的开源库，位于 GitHub 地址 https://github.com/nmslib/hnswlib 。Hnswlib 实现了基于“层次可导航小世界图”（Hierarchical Navigable Small World）算法的近似最近邻（ANN）搜索，用于在大规模数据集中高效地进行向量相似度搜索。
44. https://oreil.ly/8_j7E: （https://cacm.acm.org/blogcacm/can-llms-really-reason-and-plan/）该资源是一篇题为“Can LLMs Really Reason and Plan?”的文章，作者为 Kambhampati（2023）。文章探讨了大型语言模型（LLMs）在推理和规划方面的能力，指出虽然 LLMs 擅长从大量数据中提取知识，但它们并不真正具备执行规划的能力。文章批评了一些声称 LLMs 具备规划能力的研究，认为这些研究混淆了 LLMs 提供的通用规划知识与可执行的具体计划之

间的区别，强调由 LLMs 生成的计划表面上看似合理，但在实际执行过程中可能导致错误和问题。

45. https://oreil.ly/9BcYN: （https://ieeexplore.ieee.org/abstract/document/1238663）该资源是 IEEE Xplore 上的一篇学术论文，文档编号为 1238663，内容很可能与计算机视觉或图像检索相关，尤其是涉及“IVF（倒排文件索引）”这一技术，可能是 Sivic 和 Zisserman 在 2003 年提出的工作或相关研究的详细描述。
46. https://oreil.ly/MVsgB: （https://zilliz.com/learn/vector-index）该资源是 Zilliz 提供的一份关于向量索引（vector index）技术的学习资料，详细介绍了向量数据库中向量的组织方式、常用的向量搜索算法以及相关优化方法，适合希望深入了解向量搜索原理和实现的读者参考。
47. https://oreil.ly/UziTE: （https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf）该 URL 指向的是 NeurIPS 1999 年会议论文集中的一篇论文，标题为“Konda and Tsitsiklis, 1999”，这篇论文介绍了强化学习中的 actor-critic（演员-评论家）方法。论文主要提出并分析了一种结合策略梯度和价值函数估计的强化学习算法，为后续 actor-critic 方法的发展奠定了理论基础。
48. https://oreil.ly/VaLf4: （https://inria.hal.science/inria-00514462v2/document）该资源是一个与“Product Quantization”（乘积量化）相关的学术文档或论文，可能由 Jégou 等人在 2011 年发表，托管于 INRIA 的 HAL 开放存储库，内容涉及乘积量化技术的介绍或研究成果。
49. https://oreil.ly/aDmhb: （https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf）该资源是一篇由 BM25 算法主要作者撰写的学术论文，题为《The Probabilistic Relevance Framework: BM25 and Beyond》，发表于《Foundations and Trends in Information Retrieval》期刊（2009 年，第 3 卷第 4 期）。论文系统介绍了 BM25 算法的理论基础及其在信息检索中的应用和扩展，是学习和深入理解 BM25 模型的重要参考文献。
50. https://oreil.ly/aie-book: （https://github.com/chiphuyen/aie-book）该资源是一个关于信息检索（information retrieval）及其相关技术的电子书或教材

的 GitHub 代码仓库，提供了与书中内容配套的详细资料、示例代码和进一步学习资源，便于读者深入理解检索算法及其在 RAG（检索增强生成）等领域的应用。具体而言，URL “https://github.com/chiphuyen/aie-book” 指向该书的官方 GitHub 仓库。

51. https://oreil.ly/faJqj: （https://research.google/blog/announcing-scann-efficient-vector-similarity-search/）该资源是 Google 官方博客中关于 ScaNN（Scalable Nearest Neighbors）的一篇发布文章，介绍了 ScaNN 算法及其在高效向量相似度搜索中的应用和优势。文章详细说明了 ScaNN 如何通过优化算法实现对大规模数据集的快速近似最近邻搜索，提升了向量搜索的性能和精度。
52. https://oreil.ly/lKB61: （https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html）该资源是伯克利大学维护的 “Function Calling Leaderboard” 网页，展示了基于函数调用能力的智能代理或模型的评测结果和排名，旨在通过基准测试比较不同系统在函数调用任务上的表现。
53. https://oreil.ly/pbh3y: （https://ann-benchmarks.com/）该资源是一个名为 “ANN-Benchmarks” 的网站，提供不同近似最近邻（ANN）算法在多个数据集上的性能比较，使用包括索引和查询效率在内的四个主要指标，帮助用户评估和权衡各种算法的优劣。
54. https://oreil.ly/slO9x: （https://graphics.stanford.edu/courses/cs468-06-fall/Papers/06%20indyk%20motwani%20-%20stoc98.pdf）该资源是 Indyk 和 Motwani 于 1998 年在 STOC 会议上发表的论文，介绍了局部敏感哈希（Locality-Sensitive Hashing，LSH）的方法。这篇论文是 LSH 领域的开创性工作，提出了一种用于高维近似最近邻搜索的高效算法，极大地推动了相似性搜索和大规模数据处理技术的发展。
55. https://oreil.ly/v-T_4: （https://www.anthropic.com/news/contextual-retrieval）该 URL（https://www.anthropic.com/news/contextual-retrieval）指向 Anthropic 官网上的一篇关于 “上下文检索”（contextual retrieval）技术的新闻或介绍页面。页面内容可能涉及 Anthropic 推出的

Claude 模型在处理知识库时的应用建议，特别是关于在知识库规模较小（小于 20 万 tokens，大约 500 页材料）时，直接将全部知识库信息包含在提示中，而无需使用检索增强生成（RAG）等复杂方法的指导说明。该资源旨在帮助用户理解如何高效利用 Claude 模型进行信息检索和生成任务。

56. https://www.nltk.org: 该资源是“NLTK”（Natural Language Toolkit）的官方网站，提供一个用于自然语言处理（NLP）的经典开源工具包，包含分词、词性标注、语法分析、语料库接口等多种功能，广泛应用于文本处理和语言数据分析。
57. https://x.com/AravSrinivas/status/1737886080555446552: 该资源是一条由 Perplexity 公司 CEO Aravind Srinivas 在 X（原 Twitter）上发布的推文，内容涉及在信息检索领域中，相较于 BM25 或全文搜索，实现真正的改进具有较大难度。
58. https://x.com/ylecun/status/1702027572077326505: 该资源是一条来自 Meta 首席人工智能科学家 Yann LeCun 在 X（原 Twitter）上的状态更新（推文），内容明确表达了他对自回归大型语言模型（LLMs）规划能力的否定观点，指出“自回归 LLMs 不能进行规划”。这条推文被用作论据支持关于 LLMs 在推理和规划能力方面存在局限性的讨论。

第 7 章（117）

1. https://adapterhub.ml: 该资源是一个名为“AdapterHub”的平台，提供可共享和复用的 LoRA 适配器（LoRA adapters），用户可以在此获取经过微调的适配器模型，用于类似预训练模型的应用和集成，方便模型的模块化扩展和应用。
2. https://arxiv.org/abs/1511.00363: 该资源是 2015 年由 Courbariaux 等人发表在 arXiv 上的论文，题为“BinaryConnect”，介绍了一种将神经网络权重

二值化（即 1 位表示）的量化方法，以减少模型存储和计算复杂度，从而实现更高效的神经网络训练和推理。

3. https://arxiv.org/abs/1602.05629: 该资源是论文《Communication-Efficient Learning of Deep Networks from Decentralized Data》（作者：Brendan McMahan 等，2016 年），介绍了联邦学习（federated learning）的方法，特别是通过在多个设备上分别训练模型再合并的方法，实现分布式数据上的高效深度学习。论文提出了一种在保护数据隐私的前提下，利用多设备本地数据协同训练机器学习模型的框架。
4. https://arxiv.org/abs/1602.07360: 该资源是论文《SqueezeNet: AlexNet-level accuracy with 50x fewer parameters》，由 Iandola 等人在 2016 年发表，介绍了一种通过低秩分解和卷积层替换（如用 1×1 卷积代替 3×3 卷积）的轻量级卷积神经网络架构 SqueezeNet，能够在保持 AlexNet 级别图像分类准确率的同时，大幅减少模型参数量，提升计算效率。
5. https://arxiv.org/abs/1603.05279: 该资源是 Xnor-Net 论文，题为“Xnor-Net: ImageNet Classification Using Binary Convolutional Neural Networks”，由 Rastegari 等人于 2016 年发表在 arXiv 上。论文提出了一种基于二值卷积神经网络的模型压缩方法，实现了用 1-bit 表示权重和激活，从而大幅减少计算和存储需求，适用于高效的深度学习推理。该工作在模型量化领域具有开创性意义，推动了轻量级神经网络的发展，并促成了相关技术公司 Xnor.ai 的创立及后续被苹果公司收购。
6. https://arxiv.org/abs/1609.01596: 该资源是一篇由 Arild Nøkland 于 2016 年发布的学术论文，介绍了一种名为“直接反馈对齐”（Direct Feedback Alignment）的神经网络训练方法。该方法作为反向传播的替代方案，通过直接将误差信号反馈给隐藏层，而无需传统的链式反向传播，从而简化了权重更新过程，提高了训练效率。
7. https://arxiv.org/abs/1611.04558: 该资源是一篇由 Johnson 等人在 2016 年发表的论文，介绍了谷歌多语言翻译系统的研究成果。该系统通过迁移学习，实现了在没有直接训练数据的情况下，将葡萄牙语–英语和英语–西班牙语的翻译知识

迁移应用于葡萄牙语–西班牙语的直接翻译，体现了多语言神经机器翻译模型的跨语言通用能力和有效性。

8. https://arxiv.org/abs/1612.00796: 该资源是指向论文《Overcoming catastrophic forgetting in neural networks》（作者：Kirkpatrick 等人，2016 年），该论文提出了一种称为“弹性权重固化”（Elastic Weight Consolidation, EWC）的方法，用于缓解神经网络在顺序学习多个任务时出现的灾难性遗忘问题。论文详细分析了神经网络在顺序微调（sequential finetuning）过程中旧任务知识丢失的现象，并提出解决方案以保持模型在旧任务上的性能。
9. https://arxiv.org/abs/1709.00103: 该资源是 Zhong 等人在 2017 年发表的关于 WikiSQL 数据集的研究论文，主要涉及自然语言处理中的语义解析或文本理解任务。论文详细介绍了 WikiSQL 基准测试及其相关方法，是该领域的重要参考文献。
10. https://arxiv.org/abs/1712.05877: 该资源是论文《Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference》（Jacob et al., 2017），介绍了在神经网络训练过程中采用低精度（如 INT8）量化技术的方法，旨在提高模型训练和推理的计算效率，同时解决训练与推理阶段精度不匹配的问题。论文探讨了如何在保证模型性能的前提下，实现基于整数算术的高效神经网络训练和推理。
11. https://arxiv.org/abs/1804.07461: 该 URL（https://arxiv.org/abs/1804.07461）指向的是 Wang 等人于 2018 年发布的一篇关于 GLUE 基准测试的论文。GLUE（General Language Understanding Evaluation）是一个用于评估自然语言理解模型综合能力的基准集合，广泛用于衡量各种预训练语言模型在多种语言理解任务上的表现。
12. https://arxiv.org/abs/1804.08838: 该资源是 Li et al. (2018)发表在 arXiv 上的论文，主要研究了大型语言模型（LLMs）的参数规模与其内在维度之间的关系。论文提出，尽管 LLMs 拥有大量参数，但其内在维度实际上很低，且预训练过程隐式地最小化了模型的内在维度。这一发现表明，预训练过程相当于一种压缩机

制，使得训练更充分的模型可以通过较少的可训练参数和较少的数据进行高效微调。

13. https://arxiv.org/abs/1810.04805: 该资源是指向论文《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》（作者：Devlin et al., 2018），该论文提出了 BERT 模型，一种基于深度双向 Transformer 结构的预训练语言表示模型，显著提升了多项自然语言处理任务的性能。
14. https://arxiv.org/abs/1902.00751: 该资源是论文《Parameter-Efficient Transfer Learning for NLP》（Houlsby et al., 2019），介绍了一种通过在预训练语言模型（如 BERT large）中插入适配器（adapter）模块来实现参数高效微调的方法。该方法通过添加少量额外参数，显著减少了微调时需要更新的参数数量，同时保持了与全量微调相近的性能，提升了微调的效率和灵活性。论文在 GLUE 基准测试中展示了该方法的有效性，开创了基于适配器的“加法式”微调技术，成为后续诸如 LoRA 等参数高效微调方法的重要基础。
15. https://arxiv.org/abs/1910.05653: 该资源是论文《Model Fusion via Optimal Transport》（作者 Singh 和 Jaggi，2020），提出了一种通过最优传输（Optimal Transport）方法对模型参数进行对齐，从而确保函数相关的参数能够被正确地平均融合，旨在改进多模型融合的效果。
16. https://arxiv.org/abs/2009.03300: 该资源是关于“MMLU 基准测试”的论文或文档，MMLU（Massive Multitask Language Understanding）是一个用于评估语言模型多任务理解能力的基准测试集合。文中提到的研究通过该基准比较了 RAG 方法与微调（finetuning）技术在不同模型上的表现，表明 RAG 在大多数问题类别中表现更优。
17. https://arxiv.org/abs/2012.13255: 该资源是由 Aghajanyan 等人在 2020 年发表的一篇论文，讨论了大型语言模型（LLMs）在预训练过程中其“内在维度”（intrinsic dimension）的特性。论文提出预训练过程实际上隐式地最小化了模型的内在维度，体现出预训练对模型的某种压缩效果。研究发现，经过良好预训练的更大模型反而具有更低的内在维度，从而使得后续微调时只需较少的可训练

参数和较少的数据即可实现高效适应。该论文为理解预训练如何提升模型压缩与微调效率提供了理论依据，并对未来如何结合低秩训练方法提出了启发。

18. https://arxiv.org/abs/2101.00190: 该资源是 Li 和 Liang 于 2021 年发表的论文，介绍了“prefix-tuning”技术。这是一种软提示调优（soft prompt tuning）的方法，主要通过在每个 Transformer 层的输入前添加软提示标记（soft prompt tokens），以调整预训练语言模型的行为，从而实现高效的参数调优和下游任务适配。
19. https://arxiv.org/abs/2103.10385: 该资源是由 Liu 等人在 2021 年发布的论文，题为“P-Tuning”，介绍了一种软提示调优（soft prompt tuning）的方法。该方法通过在预训练语言模型的输入嵌入层插入可训练的软提示向量，实现对模型的高效微调，从而提升下游任务的性能。论文详细阐述了 P-Tuning 技术的设计理念及其与其他软提示调优方法（如 prefix-tuning 和 prompt tuning）的区别。
20. https://arxiv.org/abs/2104.08691: 该资源是 Lester 等人在 2021 年发表的一篇论文，题为“Prompt Tuning”，介绍了一种软提示调优（soft prompt tuning）的方法。该方法通过在模型输入的嵌入层前添加可训练的提示向量，实现对预训练语言模型的高效微调，从而在保持模型主体参数不变的情况下，适应下游任务需求。论文详细探讨了这种提示调优技术的设计和效果，是软提示调优领域的重要工作之一。
21. https://arxiv.org/abs/2106.09685: 该资源是由 Hu 等人在 2021 年发表的论文，介绍了一种名为 LoRA（Low-Rank Adaptation，低秩适配）的参数高效微调方法。该方法通过在模型中插入可合并回原始层的低秩矩阵模块，避免了增加额外推理延迟，同时显著降低了微调时需调整的参数数量。论文还探讨了大型语言模型（LLMs）存在低内在维度的现象，表明预训练过程隐式地减少了模型的内在维度，从而使得微调过程更加高效。LoRA 因其高效性和实用性，成为当前流行的适配器微调技术之一。
22. https://arxiv.org/abs/2106.10199: 该资源是论文《BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models》(Zaken et al., 2021)，介绍了一种名为 BitFit 的适配器

方法，用于在预训练 Transformer 语言模型上进行参数高效的微调。BitFit 方法通过只调整少量的偏置参数，显著减少了微调时的参数量，同时保持了较好的性能。该论文发表于 2021 年，与 LoRA 方法同期出现，是适配器微调领域的重要工作之一。

23. https://arxiv.org/abs/2203.05482: 该资源是一篇由 Wortsman 等人在 2022 年发表的论文，题为“Model soups”，介绍了一种通过线性组合多个微调后模型的整体权重来提升模型准确率的方法。这种方法无需增加推理时间，即可改善模型性能，展示了将多个模型权重简单平均的有效性。
24. https://arxiv.org/abs/2204.05862: 该资源是一篇由 Bai 等人于 2020 年发表的学术论文，讨论了“alignment tax”现象，即在人工智能模型对齐（alignment）过程中可能出现的性能或效率损失问题。论文地址为：https://arxiv.org/abs/2204.05862。
25. https://arxiv.org/abs/2205.05198: 该资源是一篇 2022 年发表在 arXiv 上的学术论文，题为《Reducing Activation Recomputation in Large Transformer Models》（减少大型 Transformer 模型中的激活重计算），作者为 Korthikanti 等人。论文主要研究如何降低大型 Transformer 模型在训练过程中激活值的重复计算，从而减少显存占用和提升计算效率。
26. https://arxiv.org/abs/2208.07339: 该资源是一篇由 Dettmers 等人在 2022 年发表的论文，介绍了将大型语言模型（LLMs）进行低精度量化的方法，特别是将模型量化到 8 位（int8）以减少计算和存储资源的消耗，从而提升推理效率。论文中提出的技术对深度学习模型的推理量化具有重要指导意义。
27. https://arxiv.org/abs/2209.04836: 该资源是论文《Git Re-Basin: Merging Models Modulo Permutation Symmetries》（作者：Ainsworth 等，2022 年），其内容聚焦于在模型融合过程中通过考虑参数的排列对称性来对齐模型参数，从而实现更有效的模型合并。论文提出了一种方法，使得功能相关的参数能够在合并前正确对齐，解决了直接线性组合时参数不匹配的问题。
28. https://arxiv.org/abs/2210.11416: 该资源是一篇由 Chung 等人于 2022 年发表在 arXiv 上的论文，介绍了 Flan-T5 模型的微调方法。论文展示了通过在约 82,000 条（指令，输出）对上进行微调，一个体积远小于 GPT-3 的 Flan-T5 模

型能够在多种写作辅助任务中实现更优的文本编辑性能，体现了小型专用微调模型在特定任务上的高效性和优势。

29. https://arxiv.org/abs/2212.04089: 该资源是 Ilharco 等人于 2022 年发表在 arXiv 上的一篇论文，介绍了“任务向量（task vectors）”的概念及其应用，提出了“任务算术（task arithmetic）”的方法。通过对任务向量进行加减运算，可以实现任务能力的组合或特定能力的去除，例如通过减去某些任务向量来减少模型中的不良行为，如去除面部识别能力或消除预训练期间获得的偏见。
30. https://arxiv.org/abs/2212.05055: 该资源是 Komatsuzaki 等人在 2022 年发表的一篇论文，题为《Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints》。论文介绍了一种通过层堆叠（layer stacking）方法训练混合专家模型（MoE）的技术。具体做法是基于预训练的密集模型，将某些层或模块复制多份，并添加一个路由器（router）将输入分配给最合适的专家副本，随后对合并后的模型及路由器进行进一步训练以提升性能。这种方法无需从头训练 MoE 模型，能够有效利用已有的密集模型参数，实现稀疏专家模型的训练与优化。
31. https://arxiv.org/abs/2303.17564: 该资源是一篇由 Wu 等人于 2023 年发表在 arXiv 上的学术论文，介绍了 BloombergGPT 这一专门面向金融领域的大型语言模型。论文详细描述了该模型的设计、训练过程及其在金融任务上的性能表现，并披露了训练所需的计算资源和成本估算。
32. https://arxiv.org/abs/2305.05862: 该资源是一篇由 Li 等人在 2023 年发表的学术论文，内容涉及对比评估 GPT-4-0314 与 BloombergGPT 在多个金融基准测试中的表现，结果显示 GPT-4-0314 在这些金融任务上具有显著优势。
33. https://arxiv.org/abs/2305.08252: 该资源是论文《Dutt et al., 2023》，介绍了将 LoRA（低秩适配）方法应用于卷积神经网络（CNN）架构的研究工作。论文探讨了 LoRA 在除 Transformer 模型之外的其他神经网络结构中的应用和效果，扩展了 LoRA 技术的适用范围。
34. https://arxiv.org/abs/2305.14314: 该资源是一篇由 Dettmers 等人于 2023 年发表的学术论文，介绍了一种名为 QLoRA 的低比特量化方法，专门用于大规模语言模型（LLMs）的高效微调。QLoRA 通过将模型权重以 4 位精度存储，并

在前向和反向传播时将其反量化为 BF16，从而显著减少了内存使用，同时保持了微调性能，解决了大模型微调中的内存瓶颈问题。

35. https://arxiv.org/abs/2305.17888: 该资源是一篇由 Liu 等人于 2023 年发布在 arXiv 上的学术论文，介绍了一种名为 LLM-QAT 的方法。该方法通过将神经网络中的权重和激活量量化到 4 位精度，同时保持词嵌入（embeddings）在 16 位精度，从而实现低精度训练中的权重和激活值有效量化，兼顾计算效率与模型性能。
36. https://arxiv.org/abs/2306.01708: 该资源是由 Yadav 等人于 2023 年发布在 arXiv 上的论文，题为“TIES-Merging: Resolving Interference When Merging Models”。论文研究了在合并机器学习模型时如何解决参数干扰问题，特别提出了一种方法，通过将大量任务向量（task vector）参数重置为其基础模型中的原始值（即将对应任务向量参数置零），可以在几乎不损失模型性能的情况下实现模型合并与干扰消除。
37. https://arxiv.org/abs/2307.05695: 该资源是论文《ReLoRA》（Lialin et al., 2023），介绍了一种用于训练低秩大语言模型（LLMs）的方法，适用于参数规模最高达 1.3 亿的基于 Transformer 的模型。
38. https://arxiv.org/abs/2307.09288: 该资源是由 Touvron 等人在 2023 年发布的论文，介绍了 Llama 2 模型，一种具有多达 39 亿参数的开源大型语言模型（LLM）。这篇论文详细描述了 Llama 2 在性能和推理效率方面的优化，是当前广泛使用的高效 Transformer 语言模型之一。
39. https://arxiv.org/abs/2308.12950: 该资源是论文《Code Llama: Open Foundation Models for Code》（作者 Rozière 等，2024 年），介绍了基于 Llama 2 的不同 Code Llama 模型的构建方法，重点探讨了多种微调技术，包括通过长上下文微调（long-context finetuning）将模型的最大上下文长度从 4,096 个标记扩展到 16,384 个标记，以支持更长的代码文件处理能力。论文还涉及指令微调（instruction finetuning）等监督微调方法。
40. https://arxiv.org/abs/2309.12307: 该资源是由 Chen 等人于 2023 年发表的一篇论文，题为“LongLoRA”，介绍了一种基于 LoRA 的变体方法。LongLoRA

通过引入注意力机制的修改技术，旨在扩展模型的上下文长度，从而提升在处理长文本时的性能。

41. https://arxiv.org/abs/2309.14717: 该资源是由 Xu 等人在 2023 年发表的一篇关于量化 LoRA（Low-Rank Adaptation）的研究论文，题目和内容围绕如何通过量化技术节省内存，提高 LoRA 在大模型微调中的效率和实用性。该论文属于量化 LoRA 领域的代表性工作之一。
42. https://arxiv.org/abs/2309.16119: 该资源为论文《ModuLoRA》（Yin et al., 2023），属于量化 LoRA（Low-Rank Adaptation）相关的研究工作，旨在通过模块化设计进一步提升 LoRA 在模型微调中的内存效率和性能表现，属于节省内存的量化 LoRA 技术领域的重要进展。
43. https://arxiv.org/abs/2310.11453: 该资源是由 Wang 等人在 2023 年发表的一篇关于 1-bit 神经网络（BitNet）的学术论文，介绍了极端量化技术在模型压缩中的应用，属于量化神经网络领域的重要研究成果。
44. https://arxiv.org/abs/2311.03099: 该资源是一篇由 Yu 等人于 2023 年发表在 arXiv 上的论文，介绍了一种名为 DARE（Drop And REscale）的模型合并技术。该方法通过先对任务向量中的冗余参数进行剪枝，从而在多个模型合并时有效减少参数间的干扰，显著提升合并后模型的性能。论文重点探讨了剪枝在提升最终合并模型质量中的重要作用，尤其是在需要合并大量模型的情况下。
45. https://arxiv.org/abs/2312.04339: 该资源是由 Tam 等人在 2023 年发表的一篇论文，标题为 “Merging by Matching Models in Task Parameter Subspaces”。论文研究了在模型融合过程中，通过在任务参数子空间中匹配模型参数的方法，以实现更有效的模型合并。该方法旨在解决直接线性组合模型参数时因参数未对齐而导致的性能下降问题，提出了一种在参数空间对齐模型的技术，从而提升模型融合效果。
46. https://arxiv.org/abs/2312.15166: 该资源是一篇由 Kim 等人于 2023 年发布的学术论文，介绍了一种称为“深度可分层扩展（depthwise scaling）”的模型扩展方法。该方法通过调整模型的层数，实现将一个具有 32 层、70 亿参数的模型扩展为 SOLAR 10.7B 模型，从而有效提升模型规模和性能。

47. https://arxiv.org/abs/2401.17868: 该资源是一篇由 Zhong 等人在 2024 年发表的学术论文，探讨了将低秩适配方法（LoRA）应用于卷积神经网络（CNN）架构的相关研究。论文内容主要涉及如何在非变换器（transformer）模型中有效利用 LoRA 技术，以扩展其在模型微调和参数高效调整方面的应用范围。
48. https://arxiv.org/abs/2402.04964: 该资源（https://arxiv.org/abs/2402.04964）是一篇由 Aleem 等人于 2024 年发布的学术论文，内容涉及将 LoRA（低秩适配技术）应用于卷积神经网络（CNN）架构的研究。论文展示了 LoRA 在非 Transformer 模型，尤其是卷积神经网络中的应用方法和效果，扩展了 LoRA 技术的适用范围。
49. https://arxiv.org/abs/2402.05445: 该 URL（https://arxiv.org/abs/2402.05445）指向的是由 Qin 等人于 2024 年发表的一篇关于“IR-QLoRA”的学术论文。论文内容聚焦于量化 LoRA（Low-Rank Adaptation）技术的研究，特别是在节省内存方面的改进与应用。该工作属于量化 LoRA 相关领域的重要研究成果之一，旨在提升模型微调效率和资源使用的优化。
50. https://arxiv.org/abs/2402.17764: 该资源是一篇由微软研究人员 Ma 等人在 2024 年发布的论文，介绍了一种名为 BitNet b1.58 的基于 Transformer 架构的大型语言模型（LLM）。该模型创新性地实现了每个参数仅使用 1.58 比特的极低位宽量化，同时在性能上能够与使用 16 位精度的 Llama 2 模型（最多 3.9 亿参数）相媲美，标志着进入了“1-bit LLM”时代。该论文详细阐述了模型设计、量化方法及其在推理效率和性能上的优势。
51. https://arxiv.org/abs/2403.03507: 该资源是由 Zhao 等人于 2024 年发表的一篇关于低秩（low-rank）大语言模型（LLMs）训练方法的论文，题为 GaLore。论文介绍了一种训练低秩大语言模型的新方法，GaLore 在 1 亿参数规模的模型上表现与全秩模型相当，并且在 7 亿参数规模的模型上也展现出有前景的性能。
52. https://arxiv.org/abs/2406.04692: 该资源是一篇由 Wang 等人于 2024 年发表的学术论文，介绍了 Together AI 通过层叠（layer stacking）方法将多个较弱的开源模型混合，构建出 Mixture-of-Agents 模型。该模型在某些基准测试中达到了与 OpenAI 的 GPT-4o 相当的性能，展示了通过组合多个模型提升整体表现的有效性。

53. https://arxiv.org/abs/2407.21783: 该资源是由 Dubey 等人在 2024 年发布的一篇论文，强调在模型训练后调整阶段（post-training）应侧重于使模型能够“知道自己知道什么”，而不是简单地向模型中添加新的知识。换言之，该论文探讨了模型对自身知识状态的校准和认知能力的提升，而非单纯增加知识量。
54. https://arxiv.org/html/2404.05086v1: 该资源是 Fomenko 等人在 2024 年发布于 arXiv 上的一篇论文，讨论了 LoRA（低秩适配）技术在深度学习模型中的应用，特别是比较和分析了基于前馈层和基于注意力层的 LoRA 方法，指出前馈层的 LoRA 可以作为注意力层 LoRA 的补充，尽管在内存限制下注意力层 LoRA 通常效果更佳。
55. https://en.wikipedia.org/wiki/Ensemble_learning: 该资源是维基百科关于“集成学习”（Ensemble Learning）的页面，介绍了集成学习的基本概念，即通过组合多个学习算法以获得比单一算法更优的预测性能，以及集成学习与模型合并的区别，强调集成通常是结合各模型的输出结果，而保持各个模型本身不变。
56. https://en.wikipedia.org/wiki/Floating-point_arithmetic: 该资源是维基百科中关于“浮点运算”（Floating-point arithmetic）的页面，介绍了浮点数的表示方法及其在计算机中的运算规则，特别是符合 IEEE 754 标准的浮点数格式，用于描述如何在计算机系统中有效且准确地表示和处理实数。
57. https://en.wikipedia.org/wiki/IEEE_754: 该 URL（https://en.wikipedia.org/wiki/IEEE_754）指向的是关于 IEEE 754 标准的维基百科页面。IEEE 754 是由电气和电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）制定的浮点数算术标准，定义了计算机中浮点数的表示和运算方式，是目前广泛采用的浮点数格式规范。该标准详细规范了浮点数的二进制编码格式、舍入规则、特殊值（如无穷大和 NaN）以及浮点运算的行为，广泛应用于科学计算、神经网络等需要高精度数值表示的领域。
58. https://en.wikipedia.org/wiki/Minifloat: 该资源介绍了“minifloat”格式，即一种用于表示浮点数的超小位宽数值格式，通常用于 8 位及以下的数值表示。文中提到的 FP8（8 位）和 FP4（4 位）即是这种格式的实例，4 位是遵循所有

IEEE 浮点数标准原则的最小浮点数表示位数。该格式主要用于在极低比特数下保持浮点数的表示能力，常用于需要节省存储空间和计算资源的场景。

59. https://github.com/Lightning-AI/litgpt: 该 URL 指向的资源是一个名为 LitGPT 的开源项目，托管在 GitHub 上。LitGPT 是一个支持多种微调（finetuning）方法的深度学习框架，尤其擅长实现基于适配器（adapter-based）技术的微调。它对流行的基础模型提供开箱即用的 LoRA（低秩适配）支持，方便用户在有限资源下高效地进行模型微调，适合有一定模型架构理解和编程能力的开发者使用。
60. https://github.com/axolotl-ai-cloud/axolotl: 该资源是一个名为“Axolotl”的开源项目，托管在 GitHub 上，属于模型微调（finetuning）框架。它支持多种微调方法，特别是基于适配器（adapter-based）的技术，并且能够方便地对流行基础模型（base models）进行低秩适配（LoRA）等高效微调操作。Axolotl 旨在简化和加速模型微调流程，适合有一定模型架构理解和编程能力的用户使用。
61. https://github.com/chiphuyen/aie-book: 该资源是一本关于机器学习和人工智能工程的书籍的 GitHub 代码仓库，提供了相关的学习资料、示例代码和实用指南，帮助读者理解模型微调、提示工程等技术，并包含针对书中内容的详细讲解和实现示例。
62. https://github.com/ggerganov/llama.cpp/pull/7150: 该资源是 llama.cpp 项目中一个关于 BF16 和 FP16 之间性能对比的性能测试（benchmark）相关的 Pull Request（PR）。
63. https://github.com/hiyouga/LLaMA-Factory: 该资源是一个名为“LLaMA-Factory”的 GitHub 开源项目，提供用于微调（finetuning）大型语言模型 LLaMA 的工具和框架，支持多种微调方法，尤其是基于适配器（adapter-based）的技术，方便用户使用自己的数据对模型进行定制训练。
64. https://github.com/huggingface/peft: 该资源是 Hugging Face 维护的一个名为 PEFT（Parameter-Efficient Fine-Tuning，参数高效微调）的开源 GitHub 代码仓库，提供多种高效的微调方法框架，尤其支持基于适配器（adapter-based）的技术。它旨在帮助用户在保持模型性能的同时，减少微调

时的计算资源和参数开销，方便快速地对大规模预训练模型进行定制化训练。该仓库广泛应用于包括 LoRA 等技术，并且支持流行基础模型的直接使用和扩展。

65. https://github.com/microsoft/DeepSpeed: 该资源是微软开源的分布式深度学习训练库 DeepSpeed，主要用于支持多机多卡环境下的高效模型微调和训练，帮助用户实现大规模模型的分布式训练加速和优化。
66. https://github.com/unslothai/unsloth: 该资源“https://github.com/unslothai/unsloth”是一个开源的模型微调框架（finetuning framework），支持多种微调方法，尤其是基于 adapter 的技术，能够方便地对流行基础模型进行 LoRA 等参数高效微调。它旨在简化模型微调过程，帮助用户在不同任务和数据上进行定制化训练。
67. https://huyenchip.com/llama-police: 该资源是一个名为“Llama Police”的网页，提供了一个更全面且最新的 LLaMA 模型微调（finetuning）框架和模型仓库的汇总列表，方便用户查找和使用各种微调工具及相关代码。
68. https://oreil.ly/0pZgw: （https://blogs.nvidia.com/blog/tensorfloat-32-precision-format/）该资源介绍了 NVIDIA 设计的 TensorFloat-32（TF32）浮点格式，重点讲解了 TF32 在 GPU 上的应用及其在人工智能计算中的优势，特别是在神经网络训练中兼顾了计算精度和效率，作为一种优化的浮点数表示格式。
69. https://oreil.ly/0vuze: （https://huggingface.co/TheBloke/Llama-2-70B-fp16）该 URL（https://huggingface.co/TheBloke/Llama-2-70B-fp16）指向的是 Hugging Face 平台上由用户 TheBloke 发布的 Llama 2 模型的一个版本，具体为 70 亿参数规模的 FP16（半精度浮点数）格式的模型权重文件。该资源通常用于支持机器学习和自然语言处理任务，尤其是在需要高效计算和节省显存的场景中使用 FP16 格式进行推理和训练。上下文中提到的 FP16 与 BF16 的讨论表明该模型版本涉及浮点数精度的技术细节和性能对比。
70. https://oreil.ly/5-5lw: （https://openai.com/index/instruction-following/）该资源是 OpenAI 官网上关于“instruction following”（指令跟随）技术的介绍页面，详细说明了通过微调（finetuning）使模型更好地理解

和执行用户指令，从而提升模型行为的能力，类似于 InstructGPT 论文中提出的思路，即微调帮助释放模型已有但难以通过普通提示访问的潜能。

71. https://oreil.ly/6lt6-:（https://llama.meta.com/llama3/）该资源是 Meta 公司发布的 Llama 3 系列大规模语言模型的官方网站，提供关于 Llama 3 模型（包括 70 亿参数版本）的介绍、技术细节及下载或使用说明，属于开源且支持自托管的高性能语言模型。
72. https://oreil.ly/7I6Ch:（https://docs.google.com/document/d/1rqj7dkuvl7Byd5KQPUJRxc19BJt8wo0yHNwK84KfU3Q/edit）该资源是 OpenAI 关于微调（finetuning）的最佳实践文档，介绍了两种模型开发路径：进阶路径（progression path）和蒸馏路径（distillation path）。
73. https://oreil.ly/82-jJ:（https://readme.fireworks.ai/docs/fine-tuning-models）该 URL 指向的资源是 Fireworks 平台关于微调（fine-tuning）模型的文档，内容可能涵盖如何在 Fireworks 环境下进行模型微调的具体方法、配置要求及相关限制说明。
74. https://oreil.ly/8ush1:（https://www.threads.net/@karpathy/post/CvlI8cqPPZq?hl=en）该 URL（https://www.threads.net/@karpathy/post/CvlI8cqPPZq?hl=en）指向的是 Karpathy 在 Threads 平台上的一篇帖子，内容很可能涉及 FP16 与 BF16 数值格式的技术讨论，尤其是在 Llama 3.1 模型中的应用和表现。这一帖子可能补充或延伸了上下文中提到的关于 FP16 和 BF16 混淆问题的讨论，提供了相关见解或实验结果。
75. https://oreil.ly/A-d5f:（https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms）该 URL（https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms）所指向的资源是一篇由 Raschka 撰写的关于大型语言模型（LLMs）微调的实用技巧文章。文章内容涉及微调参数设定的经验总结，例如 LoRA 方法中 rank 参数（r）的选择对模型性能的影响，讨论了不同 r 值对微调效果的权衡，以及在实际应用中如何合理配置以避免过拟合或受限于硬件资源。
76. https://oreil.ly/B59ci:（https://proceedings.mlr.press/v97/maheswaranathan19a.html）该资源链接指向 Maheswaranathan 等人在机

器学习领域的论文或研究文章，内容介绍了一种结合随机搜索与代理梯度（surrogate gradients）的方法，用于训练神经网络模型权重，作为传统反向传播以外的一种有效训练策略。文章发表在机器学习研究会议（MLR Press）上，具体探讨了进化策略在神经网络训练中的应用。

77. https://oreil.ly/BGXtn: （https://cloud.google.com/tpu/docs/bfloat16）该资源介绍了 Google 设计的浮点数格式 BFloat16（BF16），该格式旨在优化人工智能性能，特别是在 Google 的云端 TPU（张量处理单元）上使用。页面内容可能涵盖 BF16 的技术细节、优势及其在深度学习和神经网络计算中的应用。
78. https://oreil.ly/BR63I: （https://www.robots.ox.ac.uk/~vgg/publications/2014/Jaderberg14b/jaderberg14b.pdf）该资源是 Jaderberg 等人在 2014 年发表的论文，题为《Speeding up Convolutional Neural Networks with Low Rank Expansions》（通过低秩展开加速卷积神经网络），主要研究如何利用低秩矩阵分解技术提升卷积神经网络的计算效率与加速推理过程。
79. https://oreil.ly/D-wIG: （https://www.jmlr.org/papers/volume18/16-456/16-456.pdf）该资源是一篇发表在《Journal of Machine Learning Research》（JMLR）上的学术论文，题为“16-456”，内容涉及机器学习模型的训练方法，特别可能讨论了低精度（如 INT8）训练技术及其对训练效率和模型性能的影响。这篇论文详细分析了在训练过程中使用低精度表示带来的挑战，如反向传播对精度的敏感性及误差累积问题，为相关领域的研究和实践提供理论支持和实验验证。
80. https://oreil.ly/FIP9V: （https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing）该资源介绍了 NVIDIA 最新的 GPU 架构“Blackwell”平台，该平台支持以 4 位浮点数格式进行模型推理，旨在推动新一代计算的发展。
81. https://oreil.ly/Ftnxd: （https://huggingface.co/Sao10K/L3-70B-Euryale-v2.1）该资源是一个名为“L3-70B-Euryale-v2.1”的大型语言模型（可能基于 Llama 2-70B 架构）的版本，托管在 Hugging Face 平台上。它是“Euryale”

模型的一个版本，属于通过“frankenmerging”技术合并多个微调模型（如 Xwin 和 Euryale）以形成更强大模型的相关资源之一。

82. https://oreil.ly/GFeC7:（https://dl.acm.org/doi/10.1145/2987550.2987586）该资源是 2016 年发表在《Proceedings of the Seventh ACM Symposium on Cloud Computing》上的论文，题为《Ako: Decentralised Deep Learning with Partial Gradient Exchange》，由 Watcharapichat 等人撰写。论文介绍了一种去中心化的深度学习方法，提出通过部分梯度交换实现分布式训练中的梯度累积，从而优化模型更新过程。
83. https://oreil.ly/HZnfa:（https://qwenlm.github.io/blog/qwen2/）该资源是关于 Qwen2-72B-Instruct 模型的介绍页面，Qwen2-72B-Instruct 是一款参数规模为 72 亿、性能与 GPT-4 相当的开源大型语言模型，支持自托管使用。
84. https://oreil.ly/IM0Jc:（https://huggingface.co/alpindale/goliath-120b）该资源是一个名为 Goliath-120B 的大型语言模型，托管在 Hugging Face 平台上。Goliath-120B 是通过“frankenmerging”技术构建的，该技术将两个经过微调的 Llama 2-70B 模型（Xwin 和 Euryale）各自的 72 层（共 80 层）进行合并，形成一个新的 1200 亿参数规模的模型。
85. https://oreil.ly/J-soV:（https://www.anthropic.com/news/claude-3-5-sonnet）该资源介绍了 Anthropic 发布的 Claude 3.5 Sonnet 模型，这是一款拥有 700 亿参数的中型大语言模型，性能与 GPT-4 相当。
86. https://oreil.ly/J7kVB:（https://research.character.ai/optimizing-inference/）该资源是 Character.AI 发布的一篇关于“优化推理（inference）”的研究文章，内容聚焦于通过在低精度（如 INT8）下训练模型，以提升训练和推理阶段的效率，同时解决训练与推理精度不匹配的问题。文章探讨了低精度训练的技术挑战及其对模型性能的影响，旨在提供有效的优化策略以提高模型的计算效率和准确性。
87. https://oreil.ly/JZRsd:（https://developer.nvidia.com/automatic-mixed-precision）该资源是 NVIDIA 官方提供的关于自动混合精度（Automatic Mixed Precision, AMP）技术的介绍与工具页面，旨在帮助用户了解如何利用

AMP 功能在机器学习模型中自动设置部分计算为低精度，从而提升训练速度和效率，同时保持模型精度。

88. https://oreil.ly/LHLNz: （https://www.danielpovey.com/files/2018_interspeech_tdnnf.pdf）该资源是一篇由 Povey 等人在 2018 年发表的论文，题为“Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks”，介绍了一种用于深度神经网络训练的半正交低秩矩阵分解方法，旨在提高模型训练效率和性能。论文链接为：https://www.danielpovey.com/files/2018_interspeech_tdnnf.pdf。
89. https://oreil.ly/Np1Hn: （https://dl.acm.org/doi/10.1145/3394486.3406703）该资源是一篇由 Rasley 等人在 2020 年发表的学术论文，介绍了 DeepSpeed 框架及其在大规模深度学习模型训练中的优化技术，特别是通过 CPU 卸载（CPU offloading）来更高效地利用硬件内存，从而在有限 GPU 内存条件下实现更大模型的训练和推理。
90. https://oreil.ly/Ny1WI: （https://www.youtube.com/watch?v=ahnGLM-RC1Y&ab_channel=OpenAI）该 URL 指向的是 OpenAI 官方发布在 YouTube 上的视频资源，内容很可能是关于应用程序开发流程的示范或讲解，展示了应用开发过程中可能遵循的不同路径及下一步的建议操作，旨在帮助开发者理解和优化其开发工作流。
91. https://oreil.ly/QL2gL: （https://devblogs.nvidia.com/mixed-precision-nlp-speech-openseq2seq/）该资源是 NVIDIA 开发者技术博客上一篇关于混合精度训练在自然语言处理（NLP）和语音识别中的应用文章，具体介绍了使用 OpenSeq2Seq 框架进行混合精度训练的方法和实践经验，发布时间为 2018 年 10 月。
92. https://oreil.ly/RoPL4: （https://aclanthology.org/2022.aacl-short.38/）该资源是 2022 年 AACL 会议的短文论文，题目涉及通过微调语言模型以缓解性别和种族偏见，具体研究了在女性作者文本和非洲作者文本上的微调如何减少 BERT 类模型中的性别和种族偏见。
93. https://oreil.ly/T08JJ: （https://huggingface.co/models?library=peft&sort=downloads）该资源是 Hugging Face 上一个专门展示使用 PEFT

（Parameter-Efficient Fine-Tuning）库的模型集合页面，用户可以在此页面浏览和下载基于 LoRA 适配器等技术微调的预训练模型，以便进行高效的模型微调和复用。

94. https://oreil.ly/U8L4d: （https://x.com/HamelHusain/status/1695515530344689859）该 URL（https://x.com/HamelHusain/status/1695515530344689859）指向的是关于 Llama 3.1 模型中 FP16 与 BF16 数值格式混淆问题的讨论内容，属于围绕 Llama 3.1 和低精度浮点数（FP16、BF16）性能对比及相关技术细节的推文。
95. https://oreil.ly/URfbk: （https://huggingface.co/Xwin-LM/Xwin-LM-70B-V0.1）该资源是 Hugging Face 上的一个大型语言模型项目“Xwin-LM-70B-V0.1”，很可能是基于 Llama 2-70B 进行微调或改进的模型版本。根据上下文，该模型可能与“Goliath-120B”模型相关，后者是通过将两个微调后的 Llama 2-70B 模型（Xwin 和 Euryale）进行层级合并（frankenmerging）而成。因此，“Xwin-LM-70B-V0.1”应是“Goliath-120B”模型合并前的组成部分之一，代表 Xwin 团队发布的一个 70 亿参数级别的语言模型版本。
96. https://oreil.ly/Udw0Z: （https://www.informatica.si/index.php/informatica/article/viewFile/2828/1433）该资源是发表在《Informatica》期刊上的一篇学术文章或论文，内容涉及迁移学习（transfer learning）及其相关方法如微调（finetuning）。文章可能回顾了迁移学习的起源、基本概念及其在机器学习中的应用，引用了该领域的早期重要文献，如 Bozinovski 和 Fulgosi 在 1976 年提出的相关理论。该文档适合研究迁移学习理论及实践的学者和技术人员参考。
97. https://oreil.ly/V4pma: （https://www.geekwire.com/2020/exclusive-apple-acquires-xnor-ai-edge-ai-spin-paul-allens-ai2-price-200m-range/）该资源是一篇 GeekWire 的独家报道，内容涉及苹果公司于 2020 年初以约 2 亿美元收购了专注于模型压缩技术的边缘人工智能初创公司 Xnor.ai。
98. https://oreil.ly/WK_zT: （https://magazine.sebastianraschka.com/p/the-missing-bits-llama-2-weights）该 URL（https://magazine.sebastianraschka.com/p/the-missing-bits-llama-2-weights）指向的是 Sebastian

Raschka 撰写的一篇文章，内容聚焦于 Llama 2 模型权重中的“缺失位”（missing bits）问题，特别讨论了 FP16 和 BF16 数值格式在 Llama 模型中的使用及其性能表现差异。这篇文章 likely 提供了技术解析、实验对比以及对相关数值表示方法在大语言模型中的影响的深入探讨。

99. https://oreil.ly/Xe7h6: （https://blog.eleuther.ai/transformer-math/）该资源是 EleutherAI 发布的名为“Transformer Math 101”的文章或教程，内容介绍了与 Transformer 模型相关的训练内存计算方法和原理，帮助读者理解 Transformer 模型在训练过程中所需的内存资源估算。
100. https://oreil.ly/ZKRPR: （https://openreview.net/forum?id=BkluqlSFDS）该 URL “https://openreview.net/forum?id=BkluqlSFDS” 指向的是一个学术论文或会议投稿页面，内容可能与机器学习中的线性组合方法相关，特别是在模型融合或联邦学习等领域的研究。这类资源通常包含该方法的理论分析、应用案例及实验结果，探讨如何通过简单的线性组合提升模型性能。
101. https://oreil.ly/atIgi: （https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus）该资源是一篇由 Google 发布的博客文章，介绍了 BFloat16 数据格式及其在 Cloud TPU（谷歌云张量处理单元）上的高性能应用，详细阐述了 BFloat16 如何提升机器学习模型的计算效率和性能。
102. https://oreil.ly/avDPk: （https://arxiv.org/pdf/2205.05638）该资源（https://arxiv.org/pdf/2205.05638）是一篇关于 IA3 方法的学术论文。IA3 是一种新型的适配器（adapter）方法，采用高效的混合任务批处理策略，特别适合多任务微调（multi-task finetuning）。研究表明，IA3 在某些情况下的性能优于 LoRA 方法，甚至超过了全模型微调。
103. https://oreil.ly/eXC02: （https://dl.acm.org/doi/abs/10.5555/193284）该 URL（https://dl.acm.org/doi/abs/10.5555/193284）指向的是一篇由 Perrone 于 1993 年发表的学术论文，内容涉及“线性组合”在机器学习中的应用。论文探讨了如何通过线性组合多个模型来提升整体性能的思想，这一方法在早期就被提出并研究，为后续如联邦学习等领域中模型融合技术的发展奠定了理论基础。

104. https://oreil.ly/hRV9P: （https://huggingface.co/open-llm-leaderboard）该资源是 Hugging Face 提供的一个公开的大型语言模型（LLM）排行榜，展示了当前在各种任务和指标上表现领先的开源语言模型信息。
105. https://oreil.ly/hxUAk: （https://pytorch.org/tutorials/distributed/home.html）该资源是 PyTorch 官方提供的分布式训练教程页面，介绍如何使用 PyTorch 的分布式功能进行多机多卡模型微调与训练。
106. https://oreil.ly/iPwB_: （https://openaccess.thecvf.com/content/ICCV2023/papers/Wang_Overwriting_Pretrained_Bias_with_Finetuning_Data_ICCV_2023_paper.pdf）该资源是一篇发表于 ICCV 2023 会议的论文，题为《Overwriting Pretrained Bias with Finetuning Data》。论文探讨了通过微调（finetuning）技术来减轻预训练模型中固有偏见的方法，具体内容包括如何利用精心设计的微调数据集来抵消模型在预训练阶段学到的性别、种族等偏见，从而提升模型的公平性和公正性。
107. https://oreil.ly/lqLfv: （https://machinelearning.apple.com/research/introducing-apple-foundation-models）该资源介绍了苹果公司推出的基础模型（Apple Foundation Models），重点讲述了苹果在设备端高效部署机器学习模型的技术细节，包括混合精度（mixed precision）量化方案，采用 2 位和 4 位混合格式来降低模型权重的存储和计算需求，从而提升模型推理效率和性能。文章还可能涵盖苹果在基础模型研发及优化方面的最新进展与应用场景。
108. https://oreil.ly/mqHMU: （https://cims.nyu.edu/~sbowman/multinli/）该资源是纽约大学（NYU）Samuel Bowman 教授团队维护的 MultiNLI（Multi-Genre Natural Language Inference，多领域自然语言推理）数据集的主页，提供关于该数据集的详细介绍、下载链接及相关研究资料。
109. https://oreil.ly/pBaQM: （https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html）该资源是 NVIDIA 官方文档，介绍混合精度训练（mixed precision training）技术在深度学习中的应用与优化方法。文档详细讲解了如何在训练过程中同时使用高精度和低精度的数据表示（如权重用高精度，梯度和激活值用低精度），以提高训练效率和减少

计算资源消耗，同时保持模型性能。该资源适合希望了解或实现混合精度训练以加速深度学习模型训练的开发者和研究人员。

110. https://oreil.ly/t9HTH: （https://arxiv.org/pdf/2312.05934）该资源是 2023 年 12 月发布在 arXiv 上的一篇学术论文，标题可能与“Fine-Tuning or Retrieval?”相关，作者包括 Ovadia 等人。论文研究了在需要最新信息的任务中（例如时事问答），基于检索增强生成（RAG）的方法相较于传统微调模型的表现，发现 RAG 不仅优于单纯的微调模型，而且使用基础模型的 RAG 表现优于基于微调模型的 RAG。此外，论文还探讨了 RAG 与微调技术的结合使用，指出两者结合在某些情况下能提升模型性能，但也存在性能提升有限的情况。该论文为如何结合微调和检索技术以优化模型性能提供了实证分析和见解。
111. https://oreil.ly/u7wYx: （https://kipp.ly/transformer-inference-arithmetic/）该资源是 kipply 博客中由 Carol Chen 撰写的一篇文章，题为《Transformer Inference Arithmetic》，内容主要讲解 Transformer 模型推理阶段的内存计算方法。
112. https://oreil.ly/u_cVP: （https://oreillymedia.pxf.io/eKD9M1）该资源是一本由 O’Reilly 出版的书籍，标题为《Designing Machine Learning Systems》（设计机器学习系统），内容包括关于“云端和边缘计算中的机器学习”的章节。URL “https://oreillymedia.pxf.io/eKD9M1” 很可能指向这本书的购买页面或详细介绍页面。
113. https://oreil.ly/vfXqE: （https://machinelearning.apple.com/research/introducing-apple-foundation-models）该资源介绍了苹果公司推出的“Apple Foundation Models”，即基础大模型，重点讲解了其在多任务适配和高效部署方面的创新技术，如通过多 LoRA（低秩适配器）方法实现对同一基础模型的多任务适配，以及利用量化技术显著降低模型和适配器的内存占用，从而支持在 iPhone 等设备上的本地高效运行。
114. https://oreil.ly/xzdiG: （https://ieeexplore.ieee.org/document/6638949）该资源是 IEEE Xplore 上发表的一篇关于低秩矩阵分解和低秩神经网络训练的学术论文，可能涉及通过低秩方法加速神经网络的训练和推理，提升模型在高维输出目标上的效率和性能。

115. https://oreil.ly/zzREV: （https://www.databricks.com/blog/efficient-fine-tuning-lora-guide-llms）该资源为 Databricks 官方博客文章，内容聚焦于高效微调大型语言模型（LLMs）的方法指南，特别介绍了基于 LoRA（Low-Rank Adaptation）技术的微调策略，强调在模型的前馈层（feedforward layers）应用 LoRA 可以显著提升性能，提供了理论和实践层面的详尽分析与经验总结。
116. https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/: 该资源是一本名为《Designing Machine Learning Systems》的书籍，内容主要围绕设计和构建机器学习系统，可能涵盖包括集成方法（ensemble methods）在内的机器学习相关技术和实践。
117. https://x.com/abacaj/status/1695334296792264792?s=20: 该 URL（https://x.com/abacaj/status/1695334296792264792?s=20）指向的是社交平台 X（原 Twitter）上用户 abacaj 发布的一条状态（推文），内容涉及关于 Llama 3.1 模型在 FP16 与 BF16 数值格式上的混淆问题的讨论。这条推文作为对该技术话题的补充资料或观点引用，被用作 Llama 3.1 相关数值格式讨论的参考来源之一。

第 8 章（100）

1. https://arxiv.org/abs/1302.4389: 该资源是 2013 年由 Goodfellow 等人发表在 arXiv 上的论文，介绍了通过对训练数据进行扰动（perturbation）来提升机器学习模型性能和增强其对攻击的鲁棒性的方法。这篇论文在对抗性训练和模型稳健性研究中具有重要影响。
2. https://arxiv.org/abs/1503.02531: 该资源是由 Hinton 等人在 2015 年发表的一篇论文，介绍了“模型蒸馏”（Model Distillation，也称为知识蒸馏，Knowledge Distillation）的方法。该方法通过训练一个小型模型（学生模型）

来模仿大型模型（教师模型）的行为，将大型模型的知识“蒸馏”到小型模型中，从而实现模型压缩和加速。

3. https://arxiv.org/abs/1511.04599: 该资源是 Moosavi-Dezfooli 等人在 2015 年发表的一篇论文，介绍了通过对输入数据进行扰动来提升模型性能和增强模型对攻击的鲁棒性的方法，具体内容涉及对抗样本或对抗扰动的生成与防御技术。
4. https://arxiv.org/abs/1710.08864: 该资源是论文《One Pixel Attack for Fooling Deep Neural Networks》（Su et al., 2017），该论文提出了一种针对深度神经网络的对抗攻击方法——仅通过修改输入图像中的一个像素，即可使模型产生错误分类。论文中展示了该方法在 CIFAR-10 和 ImageNet 等数据集上的有效性，揭示了深度学习模型在面对微小扰动时的脆弱性，强调了此类攻击在安全性方面的潜在风险。
5. https://arxiv.org/abs/1711.03938: 该资源是论文《CARLA: An Open Urban Driving Simulator》（Dosovitskiy et al., 2017），介绍了 CARLA，一个用于自动驾驶汽车研究的开源城市驾驶模拟器。CARLA 提供了高保真度的虚拟环境，支持在安全、可控的条件下模拟复杂的城市驾驶场景，用于测试和验证自动驾驶系统的性能。
6. https://arxiv.org/abs/1810.04805: 该资源是指向论文《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》（作者：Jacob Devlin 等人，2018 年），该论文介绍了 BERT 模型及其预训练方法，其中包括通过随机替换部分词汇（约 1.5%）的扰动技术来提升模型性能。
7. https://arxiv.org/abs/1903.12261: 该资源是 Hendrycks 和 Dietterich 于 2019 年发布的论文，介绍了 ImageNet-C 和 ImageNet-P 两个数据集，这两个数据集通过对 ImageNet 图像施加 15 种常见的视觉扰动（如亮度变化、降雪、对比度变化和噪声添加）来评估模型在图像分类任务中的鲁棒性和泛化能力。
8. https://arxiv.org/abs/1910.01108: 该资源是论文《DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter》（作者 Sanh 等，2019 年），介绍了一种通过模型蒸馏技术将 BERT 模型压缩为更小、更快的 DistilBERT 模型的方法。该方法在保持接近原始 BERT 语言理解能力的同时，显

著减少了模型大小（约减少 40%）并提升了推理速度（约快 60%），以便于在资源受限环境中的部署和应用。

9. https://arxiv.org/abs/2102.01293: 该资源是论文《The Mythos of Model Pre-Training》(Hernandez et al., 2021)，讨论了预训练模型在微调过程中的“骨化”现象，即预训练权重可能限制模型对新数据的适应能力，特别是在训练数据量很大时，微调效果可能不如从头训练。文章指出较小的模型更容易出现这种骨化问题。
10. https://arxiv.org/abs/2103.00020: 该资源是 Radford 等人在 2021 年发布的一篇论文，介绍了 CLIP（Contrastive Language–Image Pre-training）模型的设计与训练方法。CLIP 是一种结合图像和文本的多模态模型，通过大规模的图文对进行对比学习，实现了在多种视觉理解任务上的零样本学习能力。
11. https://arxiv.org/abs/2107.06499: 该资源是 Lee 等人（2021 年）发表在 arXiv 上的一篇论文，研究了训练数据重复对模型性能的负面影响，指出训练数据中的重复样本会导致模型性能下降。
12. https://arxiv.org/abs/2205.10487: 该资源是一篇由 Hernandez 等人在 2022 年发表的论文，研究了训练数据中重复样本对大规模语言模型性能的影响。文中指出，即使只有 0.1% 的数据被重复多次（例如重复 100 次），也会显著降低模型表现，导致一个拥有 8 亿参数的模型性能退化到只有 4 亿参数模型的水平，强调了训练数据去重的重要性。
13. https://arxiv.org/abs/2206.14486: 该资源是由 Sorscher 等人于 2022 年发表的一篇关于数据剪枝（data pruning）的学术论文，论文探讨了如何通过发现有效的数据剪枝指标来筛选训练数据，从而显著降低现代深度学习模型的计算资源开销。
14. https://arxiv.org/abs/2210.11416: 该资源为论文《Scaling Instruction-Finetuned Language Models》(Chung et al., 2022)，研究了指令微调语言模型时，微调任务数量和多样性对模型性能的影响。论文发现，随着微调任务从 9 个增加到 282 个，模型性能显著提升；超过 282 个任务后，性能提升趋于平缓，但直到 1836 个任务仍有小幅提升，表明多样化的微调任务对模型能力提升具有重要作用。

15. https://arxiv.org/abs/2212.09251: 该资源是一篇题为《Discovering Language Model Behaviors with Model-Written Evaluations》的学术论文，由 Perez 等人在 2022 年发表，介绍了利用语言模型生成的数据合成技术来构建特定测试集，从而评估语言模型的 154 种不同行为特征（如个性、政治观点、伦理立场和社会偏见）。论文发现，语言模型生成的数据集在质量上接近甚至有时超过了人工撰写的数据集。
16. https://arxiv.org/abs/2212.10560: 该资源是 Wang 等人在 2022 年发布于 arXiv 的论文，题为“Self-Instruct: Aligning Language Models with Self-Generated Instructions”。论文介绍了一种基于自我生成指令的方式，构建多样且高质量的指令-响应数据集，用于指导和微调大型语言模型，使其更好地理解和执行人类指令。该方法被后续工作如 Alpaca 采用，作为生成训练数据的基础。
17. https://arxiv.org/abs/2304.03277: 该资源是一篇由微软研究人员于 2023 年发布的论文，内容涉及通过统计分析（如动词-直接宾语-名词三元组的分布及回答长度）来比较 GPT-3 与 GPT-4 在相同指令下生成文本的差异，旨在帮助评估生成的数据质量及模型性能。
18. https://arxiv.org/abs/2304.08460: 该资源（https://arxiv.org/abs/2304.08460）是一篇由 Köksal 等人在 2023 年发表的学术论文，介绍了一种“反向指令”（reverse instruction）的方法。该方法通过利用已有的长篇高质量文本（如故事、书籍和维基百科文章），并使用人工智能生成能够引出这些文本内容的提示语，从而合成高质量的指令数据。这种方法能够有效避免 AI 生成内容时出现的虚假信息（hallucination），提升生成指令数据的质量。
19. https://arxiv.org/abs/2304.14108: 该资源是一篇由 Gadre 等人于 2023 年发布在 arXiv 上的论文，介绍了 DataComp 项目及其举办的数据集竞赛。该竞赛旨在创建用于训练 CLIP 模型的最佳数据集，采用统一的训练脚本在提交的数据集上训练 CLIP 模型，并通过模型在 38 个下游任务上的表现来评估数据集质量。
20. https://arxiv.org/abs/2305.11171: 该资源是由 Gekhman 等人于 2022 年发表在 arXiv 上的一篇论文，题目与利用大型语言模型（LLMs）生成具有特定属性的数据有关。具体来说，论文介绍了如何使用 LLMs 生成事实不一致的摘要，并将

这些合成数据用于训练模型，从而提升模型检测事实不一致性的能力。这种方法为改进模型在事实核查和文本生成质量控制方面提供了有效手段。

21. https://arxiv.org/abs/2305.11206: 该资源是由 Zhou 等人在 2023 年发布于 arXiv 上的一篇论文，标题为《LIMA: Less Is More for Alignment》。论文展示了通过对 65 亿参数的 Llama 模型使用仅 1000 条精心筛选的提示和回答进行微调，模型在生成回答时能在 43% 的情况下达到或超过 GPT-4 的水平（根据人工评估）。此外，论文还探讨了数据多样性和质量对模型性能的影响，进行了使用不同质量和多样性数据集训练 7 亿参数语言模型的实验，分析了数据特性对生成质量的具体影响。总体而言，该论文强调了高质量少量数据在模型对齐和性能优化中的重要作用。
22. https://arxiv.org/abs/2305.14233: 该资源为论文《UltraChat: A Large-scale Multi-turn Dialogue Dataset for Chatbot Training》（Ding et al., 2023），介绍了通过利用 ChatGPT 生成多主题、多子话题的多轮对话数据集 UltraChat 的方法。该数据集旨在提升聊天语言模型的性能，强调数据的质量和多样性，包括丰富的话题覆盖和多样的对话指令设计，适用于训练更通用和高效的对话系统。
23. https://arxiv.org/abs/2305.15717: 该资源是由 Gudibande 等人于 2023 年发布在 arXiv 上的一篇学术论文，题为《The False Promise of Imitating Proprietary LLMs》。论文指出，通过模仿专有大型语言模型（LLMs）所获得的性能表现可能是表面的，模仿模型虽然能够很好地复制教师模型的风格，但在事实准确性和对训练数据之外任务的泛化能力方面存在不足。
24. https://arxiv.org/abs/2305.17493: 该资源是论文《The Curse of Recursion: Training on Generated Data Makes Models Forget》（作者：Shumailov 等，2023 年），发表于 arXiv，研究了在模型训练中递归使用 AI 生成的数据所导致的“模型崩溃”（model collapse）现象。论文通过实验证明，这种现象会使得包括变分自编码器（Variational Autoencoders）、高斯混合模型（Gaussian mixture models）和大型语言模型（LLMs）在内的多种模型性能随着训练过程不断恶化，且该问题既存在于预训练阶段，也存在于后训练阶段。
25. https://arxiv.org/abs/2308.06259: 该资源是 2023 年 Li 等人发表在 arXiv 上的一篇论文，介绍了一种“反向指令”（reverse instruction）的方法。该方法通

过利用已有的长篇高质量文本（如故事、书籍、维基百科文章），由 AI 生成能够引出这些内容的提示语，从而合成高质量的指令数据。此方法能够有效避免 AI 生成回答时出现的幻觉问题，提升生成内容的质量和可靠性。论文还讨论了基于该方法进行模型自举（model bootstrapping）的应用实例。

26. https://arxiv.org/abs/2308.12284: 该资源是 Tirumala 等人于 2023 年发表在 arXiv 上的一篇研究论文，内容涉及训练数据重复对模型性能的负面影响。论文分析了数据重复如何导致模型性能下降，进一步支持了多个研究结果，强调了在训练过程中避免数据重复的重要性。
27. https://arxiv.org/abs/2309.05447: 该资源是由 Chen 等人（2023 年）发表的一篇论文，提出了一种“反向指令”（reverse instruction）的方法：通过利用现有的高质量长文本内容（如故事、书籍和维基百科文章），并使用 AI 生成能够引出这些内容的提示语，从而合成高质量的指令数据。这种方法有助于提升生成指令数据的质量，减少 AI 在生成长篇回答时出现的幻觉问题。
28. https://arxiv.org/abs/2309.12284: 该资源是一篇由 Yu 等人于 2023 年发表的论文，介绍了他们通过对 MATH 和 GSM-8K 数据集中的 1.5 万个样本进行多样化重写，构建了一个名为 MetaMath 的新数学题库，包含近 40 万个样本。论文展示了在该数据集上训练的模型在相关数学基准测试中，性能优于更大规模的模型。
29. https://arxiv.org/abs/2310.00429: 该资源是 Bertrand 等人于 2023 年发表在 arXiv 上的论文，研究了在训练数据中混合使用合成数据与真实数据以避免模型崩溃（model collapse）的问题，展示了通过合理混合数据类型可以减缓或避免模型崩溃的现象，但未给出合成数据与真实数据的具体比例建议。
30. https://arxiv.org/abs/2402.07043: 该资源是 Dohmatob 等人于 2024 年发布的一篇论文（预印本），研究了在训练机器学习模型时，使用合成数据与真实数据混合的方法以避免模型崩溃（model collapse）的问题。论文展示了通过合理混合合成数据和真实数据，可以改善训练效果，但尚未给出合成数据与真实数据具体比例的明确建议。
31. https://arxiv.org/abs/2403.04652: 该资源是 Young 等人在 2024 年发表的一篇论文，讨论了数据质量对模型性能的影响，具体指出少量高质量的精心设计指令

（约 1 万条）优于大量噪声较多的指令数据（数十万条），强调了数据策划和数据质量在模型训练中的重要性。

32. https://arxiv.org/abs/2403.04706: 该资源是由 Li 等人于 2024 年发布在 arXiv 上的一篇论文，题为“Common 7B Language Models Already Possess Strong Math Capabilities”。论文展示了在数学问题微调 Llama 2-7B 模型时，使用合成数据几乎能达到与真实数据相当的效果。实验结果表明，合成数据的效果在规模扩展到约一百万样本时并未出现明显的性能饱和，说明大量合成数据在提升模型数学能力方面具有较大潜力。
33. https://arxiv.org/abs/2403.07714: 该资源是一篇由 Guo 等人于 2024 年发表的论文，题为“StableToolBench”，介绍了利用强大的人工智能模型来模拟 API 调用的技术。该方法无需实际执行 API 请求，就能通过 AI 模拟 API 的预期结果，从而在训练模型与 API 交互时节省成本和时间，提高效率。
34. https://arxiv.org/abs/2404.01413: 该资源是一篇由 Gerstgrasser 等人于 2024 年发表的学术论文，题为《Is Model Collapse Inevitable?》。论文探讨了模型崩溃（model collapse）的问题，指出当训练数据集完全由合成数据构成时，模型崩溃是不可避免的，但通过将合成数据与真实数据混合使用，可以避免模型崩溃。论文与 Bertrand 等人（2023）和 Dohmatob 等人（2024）的相关研究结果类似，均未对合成数据与真实数据的具体比例给出明确建议。
35. https://arxiv.org/abs/2406.11704: 该资源是 2024 年由 Adler 等人发布在 arXiv 上的一篇论文，介绍了 NVIDIA 研究团队开发的 Nemotron 模型。论文重点讨论了为了提升通用聊天语言模型的性能，构建一个包含任务多样性、话题多样性和指令多样性的高质量训练数据集的重要性。该数据集涵盖了不同的输出格式、输出长度以及开放式和是非型回答的指令，旨在通过丰富多样的训练数据改进模型的表现。
36. https://arxiv.org/abs/2406.11794: 该 URL（https://arxiv.org/abs/2406.11794）对应的资源是一篇由 Li 等人在 2024 年发布的论文，内容介绍了 DataComp 在 2024 年举办的一场类似于 2023 年 CLIP 模型数据集竞赛的新比赛，旨在评估用于训练规模从 4.12 亿到 70 亿参数的语言模型的数据集质量。论

文中采用了标准化的方法，通过模型在多个下游任务上的表现来衡量数据集的优劣，推动数据集构建和评估的规范化和标准化。

37. https://arxiv.org/abs/2407.21783: 该 URL（https://arxiv.org/abs/2407.21783）指向的是由 Dubey 等人于 2024 年发布的有关 Llama 3 模型的学术论文。该论文详细介绍了 Llama 3 在模型架构上与之前版本的延续性，以及其性能提升主要依赖于数据质量、多样性和训练规模的提升。论文还重点阐述了训练数据的覆盖范围，涵盖预训练、监督微调和偏好微调三个阶段。

此外，论文提出并设计了一种支持工具调用的多消息聊天格式，用于解决 AI 在工具使用场景中需要向不同目标发送多条消息的问题；探讨了人工生成数据易出现错误，因而开发了 AI 辅助标注工具以保证数据质量；并介绍了 Llama 3 大量依赖合成数据的训练方法，包括代码翻译、代码反向翻译以及从零合成编程指令数据等创新技术。总之，该论文为理解 Llama 3 的训练数据策略、工具使用支持、多样化数据合成方法以及模型性能提升机制提供了详尽的技术细节和实践案例。

38. https://arxiv.org/html/2311.14212v2: 该资源是由 Kern 等人于 2024 年发表的一篇学术论文，内容涉及数据标注质量的研究。论文发现，在标注任务中，标注员在标注会话的后半段所做的标注质量较低，可能是由于疲劳或厌倦导致的。这一发现可作为检测低质量数据的启发式依据，强调了对数据进行人工检查的重要性。
39. https://chancejs.com: 该资源是一个名为 Chance 的 JavaScript 库，提供生成各种假数据（如姓名、地址、电话号码、电子邮件地址等）的功能，主要用于软件开发中的测试和数据模拟，帮助开发者生成多样化的测试数据以验证程序的健壮性。
40. https://data.gov: 该资源是美国政府的开放数据门户网站，提供数十万个公开数据集，供公众免费访问和使用。
41. https://data.gov.in: 该资源是印度政府的开放数据门户网站，提供数以万计的公开数据集，旨在促进数据的透明共享和利用。
42. https://en.wikipedia.org/wiki/Bloom_filter: 该资源是维基百科中关于“Bloom filter”（布隆过滤器）的条目页面，介绍了一种基于哈希技术的空间高

效概率型数据结构，用于测试一个元素是否属于一个集合，常用于去重和快速查找操作。

43. https://en.wikipedia.org/wiki/Data_quality: 该资源是维基百科中关于“数据质量”（Data Quality）的条目，介绍了数据质量的多维度定义和相关概念，包括数据的完整性、唯一性、有效性、及时性、准确性、一致性、适用性，以及额外的可访问性、可比性、可信度、灵活性和合理性等方面。文章旨在全面阐述数据质量的内涵及其在不同应用场景中的重要性。
44. https://en.wikipedia.org/wiki/MinHash: 该资源是维基百科中关于“MinHash”的页面，介绍了一种基于哈希的去重（deduplication）方法。MinHash 通过将数据映射到多个哈希桶中，只在同一桶内的样本之间进行相似性比较，从而高效地估计集合之间的相似度，常用于大规模数据的近似重复检测和相似性搜索。
45. https://github.com/EleutherAI/lm-evaluation-harness: 该资源是 Eleuther AI 提供的一个名为 “lm-evaluation-harness” 的评估工具库，包含超过 400 个基准测试数据集，每个数据集平均约有 2000 多个样本，适用于语言模型的性能评估和 PEFT（参数高效微调）训练。
46. https://github.com/arsenetar/dupeguru: 该资源是一个名为 dupeGuru 的开源项目，提供文件或数据去重（重复数据检测与删除）功能的工具库，帮助用户快速查找和处理重复内容。
47. https://github.com/chiphuyen/lazynlp: 该资源是一个名为 “lazyNLP” 的开源库，提供自然语言处理相关功能，其中包括利用布隆过滤器（Bloom filter）进行文本重叠估计和去重的工具。
48. https://github.com/dedupeio/dedupe: 该资源“https://github.com/dedupeio/dedupe”是一个用于数据去重（deduplication）的开源库，旨在帮助用户识别和合并重复或相似的数据记录，从而提高数据清洗和整理的效率。
49. https://github.com/ekzhu/datasketch: 该资源是一个名为 “datasketch” 的开源库，托管在 GitHub 上，主要用于数据去重和相似性检测，提供高效的近似集合和重叠估计算法，帮助开发者进行大规模数据的去重和相似性分析。

50. https://github.com/google-research/deduplicate-text-datasets: 该资源是谷歌研究团队开发的一个开源库，名为“deduplicate-text-datasets”，用于文本数据集的去重处理，帮助用户高效识别和移除文本数据中的重复内容。
51. https://github.com/joke2k/faker: 该资源是一个名为 Faker 的开源库，托管在 GitHub 上，主要用于生成各种格式的伪造数据，如姓名、地址、电话号码和电子邮件地址，方便开发者在软件测试过程中模拟和验证不同类型的数据输入。
52. https://github.com/life4/textdistance: 该资源是一个名为 TextDistance 的开源库，托管在 GitHub 上，主要用于计算字符串之间的距离和相似度，常用于文本去重和重复数据检测等任务。
53. https://github.com/seatgeek/thefuzz: 该资源是一个名为“TheFuzz”的开源库，主要用于文本去重（deduplication）和模糊匹配，帮助开发者实现相似文本的快速识别与去重处理。
54. https://github.com/tatsu-lab/stanford_alpaca: 该资源是一个名为“Alpaca”的开源项目，托管在 GitHub 上，由斯坦福大学团队开发。Alpaca 通过对大型语言模型 Llama-7B 进行微调，利用由强大模型（如 text-davinci-003）生成的示例数据，训练出一个体积更小（约为原模型 4% 大小）但性能相似的语言模型。该项目旨在提供一个轻量级、高效的语言模型微调方案。
55. https://learning.oreilly.com/library/view/designing-machine-learning/9781098107956/ch04.html#perturbation: 该资源是《Designing Machine Learning Systems》一书的在线章节，具体为第 4 章的内容，重点讨论了数据增强（data augmentation）技术中的扰动（perturbation）方法。
56. https://oreil.ly/-R3Wd: （https://docs.google.com/document/d/1rqj7dkuvl7Byd5KQPUJRxc19BJt8wo0yHNwK84KfU3Q/edit）该资源是一个 Google 文档链接，内容很可能与 OpenAI 的微调（finetuning）相关，具体介绍了不同模型在少量（约 100 个）和大量（约 55 万）示例上的微调效果比较，说明更先进的模型在少量数据下表现更优，而在大量数据下各模型表现趋于一致。

57. https://oreil.ly/-Vd_q:（https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1）该资源是一个名为“Mixtral-8x7B-Instruct-v0.1”的大型语言模型，具有 560 亿参数，采用专家混合（mixture-of-experts）架构。它由 Jiang 等人于 2024 年开发，主要用于生成指令和偏好数据，作为其他更大规模模型（如 NVIDIA 的 Nemotron-4-340B-Instruct）微调的“教师模型”。尽管参数规模较小，但该模型在指导学生模型学习过程中起到了关键作用。
58. https://oreil.ly/0ymnI:（https://huggingface.co/datasets/HuggingFaceTB/cosmopedia）该资源是一个名为“Cosmopedia”的大型合成数据集，包含约 250 亿个标记（tokens），由 Mixtral-8x7B-Instruct-v0.1 模型生成，涵盖合成的教科书、博客文章、故事、帖子和 WikiHow 文章等内容，主要用于训练和评估基于合成文本的大规模语言模型。
59. https://oreil.ly/1YFbA:（https://eng.snap.com/synthetic-data-for-machine-learning）该资源是 Snap 公司关于利用合成数据增强机器学习视觉模型训练的案例研究，详细介绍了如何通过生成具有不同肤色、体型、发型、服装乃至面部表情的合成角色，来扩充数据集，覆盖未充分代表的边缘情况，减少数据偏差，从而提升 AI 模型的表现和公平性。
60. https://oreil.ly/2JlmX:（https://https-deeplearning-ai.github.io/data-centric-comp/）该资源是一个与数据驱动型人工智能（data-centric AI）相关的项目或竞赛主页，介绍了 Andrew Ng 于 2021 年发起的数据中心 AI 竞赛，参与者需通过修正数据标签、增加边缘案例、数据增强等方法改进同一基础数据集，以提升 AI 模型性能。
61. https://oreil.ly/3d_EG:（https://www.ibm.com/topics/data-quality）该资源是 IBM 官网关于“数据质量”（Data Quality）主题的介绍页面，详细阐述了数据质量的定义及其七个关键维度，包括完整性、唯一性、有效性、时效性、准确性、一致性和适用性，帮助用户理解数据质量的核心要素及其在实际应用中的重要性。
62. https://oreil.ly/BHEh1:（https://www.datacentricai.cc/benchmark/）该 URL（https://www.datacentricai.cc/benchmark/）指向的是一个与数据驱动型人工智能评测相关的资源页面，主要介绍和提供用于衡量和比较不同数据集

质量的基准测试平台。该平台可能涵盖类似于 DataComp 竞赛的标准化评测脚本和任务，帮助研究者和开发者通过统一的下游任务性能指标来评估训练数据集对模型效果的影响，从而推动数据集构建和优化的研究与实践。

63. https://oreil.ly/DZ5uV: （https://registry.opendata.aws/）该资源是亚马逊云服务（AWS）提供的一个开放数据注册中心，汇集并托管了多个公开开放的数据集，供用户免费访问和使用，支持数据驱动的研究与开发。
64. https://oreil.ly/F9Fyc: （https://arxiv.org/pdf/2303.08774）该资源是指向一篇发表于 arXiv 的学术论文，文档编号为 2303.08774。根据上下文，该论文很可能涉及大规模语言模型（如 GPT-3 和 GPT-4）训练过程中数据收集、数据处理和数据标注的重要性和方法，特别强调了从 GPT-3 到 GPT-4 在数据工作投入上的显著增加，以及数据在模型训练中的关键作用。
65. https://oreil.ly/FyHwn: （https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1）该资源是一个名为“Mixtral-8x7B-Instruct-v0.1”的大型语言模型，托管在 Hugging Face 平台上。根据上下文，该模型由 Jiang 等人于 2024 年发布，具备生成高质量合成文本的能力，已用于创建包含 250 亿词元的合成语料库“Cosmopedia”，涵盖教科书、博客文章、故事和 WikiHow 指南等多种文本类型，适合用于自然语言处理和生成任务的研究与应用。
66. https://oreil.ly/Gbu2T: （https://www.databricks.com/blog/announcing-mlflow-28-llm-judge-metrics-and-best-practices-llm-evaluation-rag-applications-part）该 URL 指向的资源是一篇由 Databricks 发布的博客文章，内容介绍了 MLflow 2.8 版本的发布情况，重点涵盖了大型语言模型（LLM）的评估指标、判定方法以及最佳实践，特别是在 LLM 评测和基于检索增强生成（RAG）应用中的应用。文章可能还包括如何通过数据清理（如去除多余的 HTML 标签）来提升模型性能的相关经验和技术细节。
67. https://oreil.ly/HMJX_: （https://www.tensorflow.org/datasets/catalog/overview#all_datasets）该 URL 指向的是 TensorFlow 官方文档中关于数据集的概览页面，介绍了 TensorFlow 中可用的预构建数据集集合，用户可以方便地加载和使用这些数据集进行机器学习模型的训练和测试。

68. https://oreil.ly/IK-1c: （https://www.dataperf.org/home）该 URL（https://www.dataperf.org/home）指向的是 DataPerf 项目的官方网站。DataPerf 是一个数据集性能评估基准平台，旨在通过标准化的评测方法，衡量和比较用于训练机器学习模型（如语言模型）的数据集质量和效果，推动数据驱动的模型训练和优化。
69. https://oreil.ly/IUA3j: （https://developer.nvidia.com/blog/leverage-our-latest-open-models-for-synthetic-data-generation-with-nvidia-nemotron-4-340b/）该资源是 NVIDIA 关于其最新开源大规模语言模型 Nemotron-4 340B 的博客文章，介绍了如何利用该模型生成合成数据以进行模型微调和优化。文中强调 Nemotron-4 340B 在指令微调和偏好微调阶段使用了 98% 的合成数据，展示了合成数据在提升模型性能方面的重要作用。
70. https://oreil.ly/OZxiz: （https://proceedings.mlr.press/v202/taori23a.html）该资源是 2023 年 Taori 和 Hashimoto 发表的一篇学术论文，题为《Data Feedback Loops: Model-driven Amplification of Dataset Biases》。论文探讨了当模型训练数据中包含先前模型输出时，模型中的偏差如何被放大，以及模型输出越忠实于原始训练分布，反馈循环越稳定，从而减少偏差放大的风险。链接指向该论文在机器学习领域重要会议或期刊的正式发表页面。
71. https://oreil.ly/R4-VI: （https://arxiv.org/pdf/2005.14165）该 URL “https://arxiv.org/pdf/2005.14165” 指向的是一篇学术论文，内容大概率涉及人工智能或自然语言处理领域，特别是与大型语言模型训练数据相关的研究。这推断基于上下文中提到的数据收集、过滤、去重和训练数据重叠分析等内容，以及 GPT-3 和 GPT-4 训练数据处理人员数量的变化。该论文可能详细介绍了用于训练大型语言模型的数据处理方法、数据集构建策略或数据质量控制技术，是理解和改进大型语言模型训练数据管理的重要参考资料。
72. https://oreil.ly/Tb4-W: （https://proceedings.neurips.cc/paper_files/paper/2023/file/6b9aa8f418bde2840d5f4ab7a02f663b-Paper-Conference.pdf）该资源是 2023 年 NeurIPS 会议论文集中的一篇论文 PDF，论文主题与“数据筛选”或“数据剪枝”（data pruning）相关，探讨如何通过选择对模型训练最有价值的数据样本来提升训练效率，从而降低深度学习的资源消耗。

73. https://oreil.ly/TgOaR: （https://datasetsearch.research.google.com/）该资源是谷歌提供的一个数据集搜索引擎，帮助用户方便地查找和访问各种公开的数据集。
74. https://oreil.ly/U7gfm: （https://tech.buzzfeed.com/lessons-learned-building-products-powered-by-generative-ai-7f6c23bff376）该资源是 BuzzFeed 分享的关于构建由生成式人工智能驱动的产品的经验教训，内容可能涉及他们如何使用合成指令数据和 LoRA 技术对 Flan-T5 模型进行微调，从而显著降低推理成本的实践案例与技术细节。
75. https://oreil.ly/VhVzp: （https://www.icpsr.umich.edu/web/pages/ICPSR/index.html）该资源是密歇根大学社会研究所（Institute for Social Research）下的 ICPSR（Inter-university Consortium for Political and Social Research）官方网站，提供来自成千上万社会科学研究的数据存储、管理和共享服务。
76. https://oreil.ly/Xe50R: （https://www.datacomp.ai/index.html）该 URL（https://www.datacomp.ai/index.html）指向的是 DataComp 竞赛的官方网站。DataComp 是一个旨在通过公开竞赛推动高质量训练数据集建设的项目，最初在 2023 年举办，目标是创建用于训练 CLIP 模型的最佳数据集；随后在 2024 年举办了类似的竞赛，评估不同规模语言模型的数据集质量。该网站可能提供竞赛介绍、规则、提交入口、结果展示及相关论文链接等信息。
77. https://oreil.ly/YnbiK: （https://electrek.co/2022/09/20/tesla-creating-simulation-san-francisco-unreal-engine/）该资源是一篇来自 Electrek 的文章，介绍了特斯拉利用虚幻引擎（Unreal Engine）创建旧金山的虚拟仿真环境，用于自动驾驶汽车的测试与验证。通过这种高精度的虚拟仿真，特斯拉能够在安全、低成本的环境中模拟复杂的交通场景，提升自动驾驶系统的性能和安全性。
78. https://oreil.ly/_tW6P: （https://www.opendatanetwork.com/）该资源是一个名为“Open Data Network”的平台，提供对成千上万数据集的搜索服务，方便用户查找和访问各种公开数据。

79. https://oreil.ly/d-Yty: （https://www.openml.org/search?type=data&sort=runs&status=active）该 URL 链接指向 OpenML 网站上的数据集搜索页面，用户可以按照运行次数（runs）排序，查看当前活跃状态的机器学习数据集资源。
80. https://oreil.ly/eb_Bn: （https://snap.stanford.edu/data/）该资源是由斯坦福大学提供的大型网络数据集集合，汇集了丰富的图数据集，广泛用于图网络分析和相关研究。
81. https://oreil.ly/ez6Iw: （https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf）该资源是 2012 年由 Krizhevsky 等人发表的著名论文《ImageNet Classification with Deep Convolutional Neural Networks》（即 AlexNet 论文）的 PDF 全文，详细介绍了使用深度卷积神经网络在 ImageNet 图像分类任务中的突破性成果，以及通过数据增强技术提升模型性能的方法。
82. https://oreil.ly/f8LPj: （https://d1qx31qr3h6wln.cloudfront.net/publications/Nemotron_4_340B_8T_0.pdf）该资源是由 NVIDIA 发布的关于 AI 模型后训练（post-training）数据合成和评估方法的技术文档或研究报告，详细介绍了利用 AI 判定多个响应中更优答案的策略，以及如何克服模型偏见（如首位偏见）的方法。
83. https://oreil.ly/g8A4a: （https://www.kaggle.com/datasets?fileType=csv）该 URL 指向 Kaggle 网站上专门筛选为 CSV 文件格式的数据集页面，用户可以在此浏览和下载大量以 CSV 格式提供的公开数据集。
84. https://oreil.ly/h-lAl: （https://www.reddit.com/r/MachineLearning/comments/404r9m/comment/cysfexz/?utm_source=share&utm_medium=web3x&utm_name=web3xcss&utm_term=1&utm_content=share_button）该 URL 指向 Reddit 上 r/MachineLearning 版块中一条评论，内容很可能是关于 OpenAI 在 GPT-3 到 GPT-4 期间，数据处理工作量和人员投入显著增加的讨论，具体涉及数据收集、筛选、去重、标注等环节，以及 OpenAI 团队的人员构成和研究数据来源的相关话题。

85. https://oreil.ly/hJhTF: （https://www.nature.com/articles/s41586-024-07566-y）该 URL 指向 2024 年 7 月发表于《Nature》期刊的一篇学术文章，题为“AI Models Collapse When Trained on Recursively Generated Data”，由相关作者撰写，内容探讨了人工智能模型在使用递归生成数据训练时可能出现的崩溃现象。
86. https://oreil.ly/i7hpS: （https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf）该资源是 ImageNet 项目的核心论文，题为《ImageNet: A Large-Scale Hierarchical Image Database》，由 Deng 等人于 2009 年发表，详细介绍了 ImageNet 数据集的构建方法、层次结构及其在计算机视觉领域的重要性。
87. https://oreil.ly/iGToR: （https://d1qx31qr3h6wln.cloudfront.net/publications/Nemotron_4_340B_8T_0.pdf）该资源是 NVIDIA 发布的关于 Nemotron-4 模型的技术文档或研究报告，具体介绍了 Nemotron-4-340B-Instruct 模型的预训练和微调过程。文档详细描述了基于 3400 亿参数的 Nemotron-4 基础模型，如何利用由 56 亿参数混合专家模型 Mixtral-8x7B-Instruct-v0.1 生成的指令和偏好数据进行微调，从而使得学生模型 Nemotron-4-340B-Instruct 在多项任务上超越了教师模型的性能。
88. https://oreil.ly/imdhy: （https://www.jmlr.org/papers/v25/23-0870.html）该资源是发表在《Journal of Machine Learning Research》（JMLR）上的一篇论文，标题为“Scaling Instruction-Finetuned Language Models”。论文主要研究了在语言模型的微调过程中，加入包含逐步推理（chain-of-thought, CoT）响应的数据，如何显著提升模型在 CoT 任务上的表现，尤其是在不同规模模型上的效果提升，准确率在某些任务中几乎翻倍。
89. https://oreil.ly/jAR9e: （https://archive.ics.uci.edu/datasets）该资源是加州大学欧文分校（UC Irvine）提供的机器学习数据集存储库，俗称“UCI 机器学习库”。它是一个历史悠久且广泛使用的数据集集合，包含数千个用于机器学习和数据挖掘研究的公开数据集。研究人员和开发者可以从中获取各种领域的数据，支持算法开发、模型训练和性能评估。

90. https://oreil.ly/mUEJO: （https://www.fast.ai/posts/2023-09-04-learning-jumps/）该资源是一篇由 fast.ai 发布的文章，标题为“Learning jumps”，发表于 2023 年 9 月 4 日。文章探讨了机器学习中关于数据量需求的主题，讨论了从极少量数据（如单个示例）到海量数据（数百万示例）对模型训练效果的影响，结合实际实验案例，帮助读者理解不同情境下的数据需求差异及其对模型学习的影响。
91. https://oreil.ly/prIw9: （https://www.nature.com/articles/nature16961）该资源是 DeepMind 团队发表在《Nature》期刊上的论文，介绍了 AlphaGo 项目的研究成果，详细阐述了利用自我对弈（self-play）训练人工智能的方法，使 AI 通过模拟大量围棋对局数据，最终实现超越人类顶尖围棋选手的突破。
92. https://oreil.ly/rX6oc: （https://openai.com/index/openai-five-defeats-dota-2-world-champions/）该资源介绍了 OpenAI 开发的 Dota 2 人工智能系统“OpenAI Five”，重点讲述了该 AI 通过大量自我对战训练，最终击败了 Dota 2 世界冠军的过程与成果。
93. https://oreil.ly/skn8z: （https://www.nature.com/articles/s41586-023-06747-5）该资源为一篇发表于《自然》（Nature）杂志的学术文章，标题为“s41586-023-06747-5”，内容可能涉及人工智能领域，尤其是使用模板生成数学方程或训练高水平几何推理模型（如 AlphaGeometry）的方法与研究，展示了利用大量合成数据提升 AI 模型解决奥林匹克级几何问题能力的最新进展。
94. https://oreil.ly/tlt5h: （https://huggingface.co/datasets）该资源是 Hugging Face 提供的一个数据集平台，汇集了海量的公开数据集，供机器学习和数据科学社区用于模型训练、评估和研究。
95. https://oreil.ly/u9ghd: （https://crfm.stanford.edu/2023/03/13/alpaca.html）该资源是斯坦福大学发布的关于 Alpaca 模型的介绍页面，详细说明了 Alpaca 模型的训练方法。具体来说，Alpaca 基于 Self-Instruct 数据集的少量种子示例，利用 OpenAI 的 GPT-3（text-davinci-003）模型自动生成了大量指令-响应对，从而构建了一个高质量的指令微调数据集，用于训练轻量级的指令跟随语言模型。

96. https://oreil.ly/xbyXd: （https://waymo.com/blog/2021/07/simulation-city/）该资源是 Waymo 公司发布的一个关于其“SimulationCity”自驾车仿真平台的介绍或博客文章。该平台用于在虚拟环境中模拟各种驾驶场景，以安全、低成本地测试和验证自动驾驶车辆的反应和性能，避免在现实中进行可能危险或昂贵的实验。
97. https://www.arxiv.org/abs/2408.04154: 该资源是 Shen 等人在 2024 年发表的一篇论文，题为《The Data Addition Dilemma》。论文探讨了在训练通用聊天语言模型时，虽然通常认为增加更多多样化的数据能提升模型性能，但实际上在某些情况下，加入更多异质数据反而可能导致模型性能下降的问题。
98. https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product?_l=en_US: 该资源是一篇由 LinkedIn 工程团队发布的博客文章，标题大致为“关于构建生成式人工智能产品的思考”（Musings on Building a Generative AI Product）。文章主要探讨了在生成式 AI 产品开发过程中遇到的挑战，特别强调了标注指南（annotation guidelines）制定的复杂性和困难，以及在数据标注阶段因耗时耗力导致放弃细致标注、寄希望于模型自动学习的风险。内容旨在分享 LinkedIn 团队在生成式 AI 产品工程实践中的经验和见解。
99. https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/: 该资源是 O’Reilly 出版的《Designing Machine Learning Systems》一书，内容涉及设计机器学习系统的方法和技术，特别包括减少对标注数据需求的策略，如弱监督学习、半监督学习和主动学习等。
100. https://x.com/gdb/status/1622683988736479232: 该资源是 OpenAI 联合创始人 Greg Brockman 在社交平台 X（原 Twitter）上的一条推文，内容强调了“手动检查数据”在机器学习中的重要性，指出这种人工数据观察通常能带来高效且宝贵的洞见，且性价比极高。

第 9 章（91）

1. https://arxiv.org/abs/1710.01878: 该资源是 Zhu et al. 于 2017 年发表在 arXiv 上的一篇论文，主要讨论了在模型剪枝（pruning）后，剪枝得到的较小模型架构可以从头开始训练（train from scratch），而不必依赖于预训练的大模型。论文探讨了剪枝技术对模型结构和训练流程的影响，提出了剪枝后模型可独立训练的观点。
2. https://arxiv.org/abs/1802.04799: 该资源是题为《Scalable MatMul-free Language Modeling》的学术论文，由 Zhu 等人于 2024 年 6 月发布在 arXiv 上，探讨了一种可扩展的、无需矩阵乘法（MatMul）的语言建模方法，旨在优化神经网络中的计算效率，减少传统矩阵乘法带来的计算负担。
3. https://arxiv.org/abs/1810.05270: 该资源是论文《Rethinking the Value of Network Pruning》（作者：Liu et al., 2018），讨论了网络剪枝（pruning）技术在深度学习模型中的应用。论文重点探讨了剪枝模型不仅可以直接使用或微调以恢复性能，还能帮助发现更优的、更小的模型结构，这些剪枝后的架构甚至可以从头开始训练，表现出与原始大模型相当的效果。
4. https://arxiv.org/abs/1811.03115: 该资源是一篇由 Stern 等人于 2018 年发表在 arXiv 上的论文，题为《Blockwise Parallel Decoding for Deep Autoregressive Models》。论文提出了一种针对深度自回归模型的区块并行解码方法，旨在通过使模型在生成序列时能并行地预测多个标记，从而提高解码速度和效率。
5. https://arxiv.org/abs/1911.02150?ref=research.character.ai: 该资源是一篇由 Shazeer 于 2019 年发表在 arXiv 上的论文，介绍了“多查询注意力”（multi-query attention）机制。该机制通过在注意力模型中跨查询头共享键（key）和值（value）向量，有效减少了 KV 缓存的内存占用，从而提升了模型的计算效率和资源利用率。
6. https://arxiv.org/abs/2004.05150v2: 该资源是一篇由 Beltagy 等人在 2020 年发表的学术论文，介绍了“局部窗口注意力”（local windowed attention）机制。该机制通过只关注固定大小的邻近令牌窗口，而不是全部之前的令牌，显著

减少了注意力计算的复杂度和键值缓存的大小，适用于处理超长序列的自然语言处理任务。

7. https://arxiv.org/abs/2007.00072: 该资源是 Ivanov 等人于 2021 年发布在 arXiv 上的一篇论文，题为《Data Movement Is All You Need: A Case Study on Optimizing Transformers》。论文聚焦于优化 Transformer 模型中的数据移动问题，探讨了矩阵乘法（matmul）在神经网络中占据大量浮点运算比例的现象，并提出相应的优化方法。
8. https://arxiv.org/abs/2204.02311: 该资源是一篇由 Chowdhery 等人在 2022 年发表的论文，首次在 PaLM 论文中提出并介绍了“MFU”（模型浮点运算利用率，Model FLOP Utilization）这一概念，讨论了在深度学习模型训练过程中浮点运算效率的衡量与优化。
9. https://arxiv.org/abs/2211.05102: 该资源是 2022 年由 Pope 等人发表在 arXiv 上的一篇论文，内容涉及大型神经网络模型（如 5000 亿参数以上模型）中的多头注意力机制及其键值（KV）缓存的存储需求。论文中分析了随着批次大小和上下文长度增加，KV 缓存占用的显著内存规模（例如 3TB），并指出其缓存大小可达到模型权重的数倍，强调了高效管理和优化 KV 缓存对于大规模模型推理的重要性。
10. https://arxiv.org/abs/2211.17192: 该资源是一篇由 Laviathan 等人在 2022 年发表的论文，介绍了通过训练较小的草稿模型（draft model）来加速大型语言模型（如 T5-XXL）解码过程的方法，从而显著降低响应延迟，同时保持输出质量。
11. https://arxiv.org/abs/2302.01318: 该资源是一篇由 Chen 等人于 2023 年发表在 arXiv 上的论文，介绍了通过训练一个参数规模较小（约 4 亿参数）的草稿模型（draft model），以加速大型语言模型 Chinchilla-70B 的解码过程。该草稿模型能够以比目标模型快约 8 倍的速度生成文本，从而在不降低响应质量的前提下，大幅减少整体响应延迟。
12. https://arxiv.org/abs/2304.04487: 该资源是一篇由杨等人于 2023 年发表的学术论文，题为《Inference with Reference: Lossless Acceleration of Large Language Models》。论文介绍了一种名为“带参考的推理”的方法，该方法无

需额外模型即可加速大型语言模型的生成过程，特别适用于上下文与输出高度重合的场景，如检索系统、代码生成和多轮对话，能够实现生成速度提升约两倍。

13. https://arxiv.org/abs/2305.13245: 该资源是一篇由 Ainslie 等人于 2023 年发表的论文，介绍了一种名为“Grouped-query attention”的注意力机制。这种机制是多查询注意力（multi-query attention）的推广，通过将查询头分成较小的组，并在组内共享键值对，从而在查询头数量和键值对数量之间实现更灵活的平衡。
14. https://arxiv.org/abs/2309.06180: 该资源是一篇由 Kwon 等人在 2023 年发布的学术论文，介绍了 PagedAttention 技术。该技术通过将 KV 缓存划分为非连续块，优化了内存管理，减少了内存碎片，并实现了灵活的内存共享，从而提升了大规模语言模型（LLM）推理框架的服务效率。
15. https://arxiv.org/abs/2310.01801: 该资源是一篇由 Ge 等人于 2023 年发表的学术论文，主题涉及“自适应键值（KV）缓存压缩”技术，属于提升缓存效率和模型性能的研究方向。文章详细讨论了如何通过自适应方法对 KV 缓存进行压缩，以优化计算资源的使用。
16. https://arxiv.org/abs/2401.10774: 该资源是由 Cai 等人在 2024 年发表的论文，介绍了一种名为 Medusa 的多头解码方法。Medusa 通过在原始模型基础上增加多个解码头（小型神经网络层），使每个解码头分别预测未来不同位置的令牌，从而实现并行令牌生成。该方法在训练时保持原始模型参数冻结，仅训练这些解码头，显著提升了生成速度。NVIDIA 报告称，Medusa 在其 HGX H200 GPU 上将 Llama 3.1 的令牌生成速度提升了近 1.9 倍。
17. https://arxiv.org/abs/2401.11181: 该资源是由 Hu 等人在 2024 年发表的一篇论文，标题为“Inference Without Interference”。论文研究了大规模语言模型（LLM）推理服务中的优化技术，提出将预填充（prefill）和解码（decode）操作分配给不同计算实例（如不同 GPU），以显著提升处理请求的吞吐量，同时满足延迟要求。尽管这种解耦需要在实例间传输中间状态，但论文证明在配备高速互联（如 NVLink）的现代 GPU 集群中，通信开销并不显著，从而验证了该方法的实用性和效率。

18. https://arxiv.org/abs/2401.18079: 该资源为一篇由 Hooper 等人在 2024 年发布的学术论文，内容涉及 KV 缓存（Key-Value Cache）的量化技术，属于提升缓存效率和模型推理性能的研究领域。
19. https://arxiv.org/abs/2402.02057: 该资源是论文《Lookahead Decoding》（Fu et al., 2024），介绍了一种在解码过程中通过同一个解码器并行生成多个未来 tokens 的方法。这种方法旨在加速语言模型的生成过程，提高推理效率。论文提出的“Lookahead 解码”技术通过提前预测后续词元，实现了更快速的序列生成，属于提升生成速度的一种创新解码策略。
20. https://arxiv.org/abs/2403.05527: 该资源是一篇由 Kang 等人在 2024 年发表的论文，内容涉及“KV 缓存量化”（KV cache quantization）技术，属于提升缓存效率和模型推理性能的相关研究。
21. https://arxiv.org/abs/2405.12981?ref=research.character.ai: 该资源是一篇由 Brandon 等人在 2024 年发表的学术论文，介绍了“cross-layer attention”技术。该技术通过在相邻的多个神经网络层之间共享键（key）和值（value）向量，有效减少了键值缓存（KV cache）的内存占用，从而提高了模型的计算效率和资源利用率。具体来说，三个层共享同一组键值向量意味着将 KV 缓存的大小减少了三倍。该论文详细探讨了这种跨层注意力机制的设计和优势。
22. https://arxiv.org/abs/2407.08608: 该资源是由 Shah 等人在 2024 年发布于 arXiv 的一篇论文，介绍了针对最新硬件架构（如 NVIDIA H100 GPU）优化的 FlashAttention-3 算法，旨在提升注意力机制计算的效率和性能。
23. https://arxiv.org/html/2401.09670v1: 该资源是一篇由 Zhong 等人于 2024 年发表的学术论文，题为“DistServe”，讨论了推理服务器中预填充（prefill）和解码（decode）操作的解耦优化技术。论文展示了将预填充和解码任务分配到不同计算实例（例如不同 GPU）上，能够在保持延迟要求的前提下显著提升处理请求的吞吐量。同时，虽然该方法需要在实例间传输中间状态，但在具备高速互联（如 NVLink）的现代 GPU 集群中，通信开销并不显著。
24. https://en.wikipedia.org/wiki/Accelerated_Linear_Algebra: 该资源介绍了“Accelerated Linear Algebra”（XLA），这是一个最初由 TensorFlow 开发的加速线性代数计算的编译器。XLA 旨在通过对机器学习模型的计算图进行优化和

编译，提高模型的执行效率。该技术后来开源，形成了 OpenXLA 项目，支持更广泛的硬件和框架。XLA 通常集成在机器学习和推理框架中，用于提升计算性能，尤其是在深度学习任务中。

25. https://en.wikipedia.org/wiki/CUDA: 该资源是关于 CUDA（Compute Unified Device Architecture）的维基百科页面，介绍了 CUDA 作为 NVIDIA 开发的用于 GPU 编程的架构和平台，广泛应用于高性能计算和深度学习领域，帮助开发者更高效地利用 GPU 的并行计算能力。
26. https://en.wikipedia.org/wiki/DDR_SDRAM: 该资源是维基百科页面，介绍了 DDR SDRAM（双倍数据率同步动态随机存取存储器），这是一种广泛用于 CPU 的内存类型，具有双倍数据传输速率和二维结构的特点。
27. https://en.wikipedia.org/wiki/GDDR_SDRAM: 该资源是维基百科中关于 GDDR SDRAM（图形双倍数据速率同步动态随机存取存储器）的页面，介绍了这种专门用于图形处理单元（GPU）的高速显存类型，通常应用于中低端到中端的显卡中。
28. https://en.wikipedia.org/wiki/Goodput: 该资源是维基百科中关于“Goodput”的条目，介绍了“Goodput”这一网络性能指标。Goodput 衡量的是网络传输中有效载荷数据的传输速率，具体指成功传输且满足软件级服务目标（SLO）的请求数量。文中提到该指标被应用于大规模语言模型（LLM）推理服务中，用以评估满足性能要求的请求处理能力，区别于单纯关注吞吐量和成本的指标。
29. https://en.wikipedia.org/wiki/Graphics_processing_unit: 该资源介绍的是“图形处理器”（Graphics Processing Unit，简称 GPU），这是一种最初用于图形渲染和图像处理的专用处理器。随着技术发展，GPU 因其高度并行计算能力，被广泛应用于深度学习和神经网络训练等计算密集型任务，极大推动了人工智能领域的研究和发展。该页面详细说明了 GPU 的定义、功能演变及其在现代计算中的重要作用。
30. https://en.wikipedia.org/wiki/Green_data_center: 该资源介绍了“绿色数据中心”（green data centers）的相关内容，重点讨论了数据中心在能源消耗和环境影响方面的挑战，以及采用节能环保技术以减少碳足迹的重要性。文章可能

涵盖了绿色数据中心的设计理念、能源效率优化措施、可再生能源的利用，以及在建设和运营大规模数据中心时如何平衡电力供应、成本和性能需求等方面的信息。

31. https://en.wikipedia.org/wiki/High_Bandwidth_Memory: 该资源是关于“高带宽内存”（High Bandwidth Memory，简称 HBM）的维基百科页面，介绍了一种用于高性能图形处理器（GPU）等设备的先进内存技术。HBM 采用三维堆叠结构设计，能够提供比传统 DDR SDRAM 更高的数据传输带宽，适用于高端 GPU 以满足其对高速数据访问的需求。
32. https://en.wikipedia.org/wiki/Hopper_(microarchitecture): 该资源介绍了 NVIDIA Hopper 微架构，这是一种用于 GPU 加速器的先进微处理器架构，具有约 800 亿个晶体管，支持高效的计算性能和能耗管理，广泛应用于人工智能和高性能计算领域。
33. https://en.wikipedia.org/wiki/Jacobi_method: 该资源介绍了“Jacobi 方法”，这是一种迭代算法，允许在计算过程中多个部分的解同时且独立地更新。文中提到该方法被用于并行解码中的验证步骤，以确保非顺序生成的令牌能够正确组合。
34. https://en.wikipedia.org/wiki/List_of_AMD_graphics_processing_units: 该资源是维基百科页面，列出了 AMD（Advanced Micro Devices）公司生产的各种图形处理单元（GPU）的详细列表，涵盖了其不同世代和型号的 GPU 产品。
35. https://en.wikipedia.org/wiki/List_of_quantum_processors: 该 URL（https://en.wikipedia.org/wiki/List_of_quantum_processors）指向的资源是一个维基百科页面，内容是“量子处理器列表”。该页面列举并介绍了各种量子处理器（Quantum Processing Units，QPU），包括不同厂商和研究机构开发的量子计算硬件，展示了当前量子计算领域中各种量子处理器的型号、架构和技术特点。
36. https://en.wikipedia.org/wiki/NVLink: 该资源介绍了 NVLink，这是一种由英伟达（NVIDIA）开发的高速互连技术，主要用于连接 GPU 与 GPU 之间或 GPU 与 CPU 之间，以实现高带宽、低延迟的数据传输。NVLink 广泛应用于现代 GPU 集群，支持在同一计算节点内多个 GPU 之间高效共享数据，从而显著提

升多 GPU 系统的整体性能，特别适合需要大规模并行计算和数据交换的深度学习推理等场景。

37. https://en.wikipedia.org/wiki/Neural_Engine: 该资源介绍了苹果公司设计的用于机器学习推理的专用芯片——Neural Engine。Neural Engine 是一种针对低精度计算和高速内存访问进行优化的处理单元，旨在提升设备上人工智能任务（如图像识别和自然语言处理）的效率和性能。
38. https://en.wikipedia.org/wiki/Perceptrons_(book): 该资源是一本由人工智能先驱 Marvin Minsky 和 Seymour Papert 于 1969 年出版的书籍，书名为《Perceptrons: An Introduction to Computational Geometry》。书中论述了当时一层神经网络（感知器）的局限性，并质疑多层神经网络的计算能力，认为其能力不会显著优于低阶感知器。这一观点在当时由于计算能力不足无法被反驳，导致了 1970 年代人工智能研究资金的减少。
39. https://en.wikipedia.org/wiki/Tensor_Processing_Unit: 该资源介绍了谷歌开发的张量处理单元（Tensor Processing Unit，TPU），这是一种专门用于加速人工智能和机器学习工作负载的定制加速器硬件。TPU 旨在提升深度学习模型的计算效率和性能，是继 NVIDIA GPU 之后的重要 AI 计算硬件之一。
40. https://github.com/Dao-AILab/flash-attention: 该资源是一个名为 FlashAttention 的开源项目，托管在 GitHub 上（https://github.com/Dao-AILab/flash-attention）。FlashAttention 是一种针对注意力机制计算进行高度优化的核函数实现，旨在通过融合变换器模型中的多个常用操作，提高计算速度和效率。最初该项目主要针对 NVIDIA A100 GPU 进行优化，后续版本（如 FlashAttention-3）扩展支持了更先进的硬件如 NVIDIA H100 GPU。该项目由 Dao 等人在 2022 年开发，广泛应用于加速基于变换器的深度学习模型的训练和推理。
41. https://github.com/NVIDIA/TensorRT: 该资源是 NVIDIA 开源的 TensorRT 项目，提供了一个针对 NVIDIA GPU 优化的高性能深度学习推理编译器和加速库，旨在加速神经网络模型的推理过程，提高推理效率和吞吐量。
42. https://github.com/NVIDIA/TensorRT-LLM: 该资源是 NVIDIA 发布的开源项目“TensorRT-LLM”，它是一个基于 TensorRT 的高效大型语言模型推理框

架，旨在通过优化推理性能和资源利用率，方便用户快速部署和运行大规模语言模型。

43. https://github.com/ROCm/ROCm: 该资源是 AMD 推出的开源 GPU 计算平台“ROCm”（Radeon Open Compute）的官方代码库，提供了用于高性能计算和 GPU 编程的工具和框架，是 NVIDIA 专有 CUDA 的开源替代方案。
44. https://github.com/apache/tvm: 该资源是 Apache TVM 的官方代码仓库，Apache TVM 是一个开源的深度学习编译器框架，用于将深度学习模型高效地编译和优化以部署在多种硬件平台上，从而提升模型的运行性能和兼容性。
45. https://github.com/ggerganov/llama.cpp/blob/master/examples/main/README.md#prompt-caching: 该 URL 指向的是 llama.cpp 项目中 examples/main 目录下 README.md 文件的“prompt caching”章节。该章节介绍了 llama.cpp 实现的提示缓存机制，内容包括其缓存的是整个提示（prompt），仅在同一聊天会话中生效，可能通过缓存之前的对话内容，仅处理最新消息，以提升长对话中的处理效率。文档内容较为简略，主要通过代码推测其工作原理。
46. https://github.com/ggerganov/llama.cpp/pull/2926: 该 URL 指向的是 GitHub 上开源项目 llama.cpp 的一个 Pull Request（PR 2926），该 PR 涉及将某种易于实现且不影响模型质量的方法集成到 llama.cpp 推理框架中。具体来说，这表明该 PR 是在 llama.cpp 中引入了类似于 PyTorch 中 50 行代码实现的优化技术，目的是提升模型推理效率或功能，类似于 vLLM 和 TensorRT-LLM 中所做的改进。
47. https://github.com/openxla/xla: 该资源是 OpenXLA 项目的官方代码仓库，OpenXLA 是 XLA（Accelerated Linear Algebra）的开源实现，XLA 最初由 TensorFlow 开发，用于加速线性代数计算并优化机器学习模型的编译和执行效率。
48. https://github.com/triton-lang/triton: 该资源是 OpenAI 开发的开源 GPU 编程语言 Triton 的 GitHub 代码仓库，旨在为 AI 研究人员和工程师提供对 GPU 内存访问的细粒度控制，从而优化 GPU 计算性能。

49. https://github.com/vllm-project/vllm: 该资源是一个名为 vLLM 的推理框架项目的 GitHub 仓库。vLLM 以其提出的 PagedAttention 技术著称，该技术通过将键值（KV）缓存划分为非连续块，优化内存管理，减少内存碎片，并支持灵活的内存共享，从而提升大语言模型（LLM）推理的效率。
50. https://mlir.llvm.org: 该资源是 MLIR（Multi-Level Intermediate Representation）的官方网站，MLIR 是一种多层次中间表示框架，用于构建可扩展且模块化的编译器基础设施，广泛应用于机器学习和编译器优化领域。
51. https://oreil.ly/42LSB: （https://aws.amazon.com/machine-learning/inferentia/）该资源是 AWS Inferentia，一种由亚马逊网络服务（AWS）设计的专门用于机器学习推理的芯片，旨在通过优化低精度计算和更快的内存访问，提高推理性能和效率。
52. https://oreil.ly/5hFSF: （https://cerebras.ai/blog/llama3.1-model-quality-evaluation-cerebras-groq-together-and-fireworks）该资源是一篇由 Cerebras 于 2024 年发布的博客文章，内容涉及对 LLaMA 3.1 模型质量的评估，重点探讨了在不同推理服务提供商通过优化技术可能导致模型行为和质量出现细微差异的现象，并介绍了 Cerebras 与 Groq 合作进行的相关实验及结果分析。
53. https://oreil.ly/5vRsP: （https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf）该资源是 NVIDIA 关于 Ampere 架构的数据中心解决方案白皮书（PDF 格式），详细介绍了 NVIDIA Ampere GPU 架构的设计理念、技术特点及其在数据中心中的应用，重点涵盖了芯片的计算能力、能效优化及功耗管理等内容。
54. https://oreil.ly/6bjVM: （https://pytorch.org/tutorials/intermediate/torch_compile_tutorial.html）该 URL（https://pytorch.org/tutorials/intermediate/torch_compile_tutorial.html）指向的是 PyTorch 官方教程中关于 torch.compile 功能的中级教学资源。该教程介绍了 PyTorch 中集成的编译器工具 torch.compile，它能够将 PyTorch 模型进行编译优化，从而提升模型的执行效率。教程内容通常涵盖如何使用 torch.compile 来加速深度学习模型的训练和推理，以及相关的实现原理和性能提升示例。

55. https://oreil.ly/6ySTY: （https://www.graphcore.ai/products/ipu）该 URL “https://www.graphcore.ai/products/ipu” 所指向的资源是 Graphcore 公司推出的智能处理单元（Intelligent Processing Unit，简称 IPU）产品页面。IPU 是一种专门设计用于加速人工智能和机器学习工作负载的处理器，旨在提升 AI 计算效率和性能，支持复杂的神经网络模型训练与推理。
56. https://oreil.ly/8rtsF: （https://www.anthropic.com/news/prompt-caching）该资源是 Anthropic 公司关于“prompt caching”（提示缓存）技术的介绍或新闻页面，内容聚焦于通过缓存输入提示实现显著降低模型调用成本（最高可节省 90%）和减少响应延迟（最高可减少 75%）的方案，展示了该技术在不同应用场景中对成本和延迟的具体影响。
57. https://oreil.ly/ACIty: （https://www.cerebras.net/product-chip/）该 URL（https://www.cerebras.net/product-chip/）所指向的资源介绍了 Cerebras 公司推出的 Wafer-Scale Quant Processing Unit（QPU）芯片，这是一种专为加速人工智能计算设计的大规模片上系统，具有极高的并行处理能力和计算性能，用于提升 AI 工作负载的处理效率。
58. https://oreil.ly/DlIPs: （https://nvidia.github.io/TensorRT-LLM/advanced/batch-manager.html）该 URL（https://nvidia.github.io/TensorRT-LLM/advanced/batch-manager.html）指向的是 NVIDIA TensorRT-LLM 项目中关于“批处理管理器”（Batch Manager）的高级文档。该资源详细介绍了连续批处理（Continuous batching）技术的实现原理与应用，讲解如何通过选择性地批处理不互相阻塞的操作，实现响应的快速返回和批处理的动态更新，从而提升大规模语言模型推理的效率和资源利用率。文档内容还结合了相关论文（如 Orca）和可视化示例，帮助读者理解和使用该批处理管理机制。
59. https://oreil.ly/FWYf5: （https://developer.nvidia.com/blog/low-latency-inference-chapter-1-up-to-1-9x-higher-llama-3-1-performance-with-medusa-on-nvidia-hgx-h200-with-nvlink-switch/）该 URL 指向的资源是一篇来自 NVIDIA 开发者博客的技术文章，介绍了通过 Medusa 技术在 NVIDIA HGX H200 平台上利用 NVLink Switch 实现低延迟推理，从而使 Llama 3.1 模

型的生成性能提升最高达 1.9 倍的相关内容。文章详细阐述了 Medusa 如何通过增加多个解码头并行预测未来 token，进而加速大语言模型的推理过程。

60. https://oreil.ly/IaPOB: （https://pytorch.org/blog/accelerating-generative-ai-2/）该资源是一篇来自 PyTorch 官方博客的文章，介绍了加速生成式人工智能（Generative AI）的方法和技术，重点讲解了如何通过简洁的代码实现高效推理，同时保持模型质量不变。文章还提及了该方法在多个流行推理框架（如 vLLM、TensorRT-LLM 和 llama.cpp）中的应用与集成。
61. https://oreil.ly/K3j6t: （https://crd.lbl.gov/divisions/amcr/computer-science-amcr/par/research/roofline/introduction/）该资源是劳伦斯伯克利国家实验室（LBNL）计算科学部门关于“Roofline 模型”的介绍页面。该页面详细解释了 Roofline 模型的基本概念，包括计算受限（compute-bound）和内存带宽受限（memory bandwidth-bound）的定义，以及如何通过算术强度（arithmetic intensity）来区分不同类型的操作。页面还介绍了 Roofline 图表的用途，这类图表常用于硬件性能分析，帮助开发者理解和优化程序的性能瓶颈。
62. https://oreil.ly/M_aGR: （https://dl.acm.org/doi/pdf/10.1145/1498765.1498785）该资源是指向 Williams 等人在 2009 年发表的一篇论文，标题为“Roofline”，该论文首次提出并系统阐述了“compute-bound”（计算瓶颈）与“memory bandwidth-bound”（内存带宽瓶颈）这两个概念。论文中通过引入“算术强度”（arithmetic intensity，指每字节内存访问所执行的算术操作数量）这一指标，建立了 Roofline 模型，用以分析和评估计算机系统在不同工作负载下的性能瓶颈。Roofline 模型通过图形化的 Roofline 图表，直观展示了计算性能与内存带宽之间的关系，成为硬件性能分析和优化的重要工具。
63. https://oreil.ly/On2-B: （https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html）该 URL 指向的资源是 NVIDIA Megatron-Core 开发者指南中关于“context parallelism”（上下文并行）技术的 API 文档。该文档详细介绍了如何将输入序列在多设备间拆分处理，以实现并行计算，例如将输入序列的前半部分在机器 1 上处理，后半部分在机器 2 上处理，从而提高大规模模型的训练效率。

64. https://oreil.ly/PRZSQ: （https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-series/）该资源介绍了 NVIDIA Jetson Xavier 系列嵌入式系统，这是一款面向边缘计算和推理任务优化的芯片及平台，专为加速人工智能推理和嵌入式机器自动化而设计。
65. https://oreil.ly/Pd6Pk: （https://arxiv.org/pdf/2311.04934）该 URL（https://arxiv.org/pdf/2311.04934）指向的是一篇学术论文，内容可能涉及“prompt caching”（提示缓存）技术的研究与应用。根据上下文，该论文很可能是由 Gim 等人在 2023 年 11 月首次提出 prompt 缓存概念的相关工作，介绍了 prompt 缓存如何显著降低大语言模型调用的计算成本和延迟，并探讨了其在实际模型 API（如 Google Gemini 和 Anthropic）中的应用和效果。
66. https://oreil.ly/QYdG8: （https://www.anyscale.com/blog/reproducible-performance-metrics-for-llm-inference）该资源是一篇由 Anyscale 发布的博客文章，标题为《可复现的大型语言模型（LLM）推理性能指标》。文章讨论了在使用自回归语言模型进行推理时，性能测量的挑战与瓶颈，特别是单个输出 token 生成过程中的延迟和计算资源消耗问题。通过深入分析推理中的带宽瓶颈和计算效率问题，该博客旨在提供一套可复现的性能指标方法，帮助开发者优化大型语言模型的推理速度和成本，从而提升用户体验。
67. https://oreil.ly/R7gXn: （https://wow.groq.com/why-groq/）该 URL “https://wow.groq.com/why-groq/” 指向的是 Groq 公司官网上的一个页面，内容主要介绍 Groq 的加速器产品——语言处理单元（Language Processing Unit, LPU），以及其设计理念、技术优势和为何选择 Groq 作为 AI 计算硬件解决方案的理由。页面旨在说明 Groq LPU 在 AI 工作负载加速方面与其他竞争产品（如 NVIDIA GPU、Google TPU、Intel Habana Gaudi 等）的区别和优势。
68. https://oreil.ly/RqY-3: （https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/）该资源是一篇来自 MIT 出版社《The Reader》的文章，探讨了计算设备和云计算对环境造成的巨大生态影响，特别关注芯片及数据中心的高能耗问题及其带来的环境压力，强调了建设绿色数据中心和解决电力供应瓶颈的必要性。

69. https://oreil.ly/SJ7Mb: （https://www.usenix.org/conference/osdi22/presentation/yu）该 URL（https://www.usenix.org/conference/osdi22/presentation/yu）指向的是 2022 年 USENIX 操作系统设计与实现大会（OSDI 2022）上 Yu 等人关于“Orca”系统的演讲或报告页面。该资源介绍了“Continuous batching”（连续批处理）技术，这是一种允许批处理中已完成的响应可以立即返回用户的机制，通过选择性地批处理不会相互阻塞响应生成的操作，实现批处理请求的持续更新与高效利用。该技术旨在提升系统吞吐量和响应效率，类似于公交车在乘客下车后立即接载下一位乘客以最大化载客率。
70. https://oreil.ly/XH2bh: （https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/）该资源介绍了 Meta 开发的下一代训练与推理加速芯片——Meta Training and Inference Accelerator（MTIA），重点展示其在人工智能模型推理和训练中的优化设计与性能提升。
71. https://oreil.ly/Xpwco: （https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html）该资源是一篇发表于《纽约时报》的报道，标题大致为“在庞大的计算机网络中，机器学习的证据”，内容介绍了机器学习特别是深度学习领域的发展历史，重点探讨了 2012 年前后深度学习复兴的关键因素之一——利用图形处理单元（GPU）进行神经网络训练的突破。这篇文章讲述了 AlexNet 等开创性模型如何借助 GPU 加速训练，从而推动了深度学习研究的快速发展，以及相关技术和产业背景的演变。
72. https://oreil.ly/Yv4V7: （https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf）该资源是 2012 年 NeurIPS 会议论文《ImageNet Classification with Deep Convolutional Neural Networks》（通常称为 AlexNet）的官方 PDF 文档。该论文由 Krizhevsky 等人撰写，开创性地展示了利用 GPU 加速训练深度卷积神经网络在大规模图像分类任务（ImageNet）中的成功应用，标志着深度学习研究的重大突破和复兴。
73. https://oreil.ly/_5Nqa: （https://pytorch.org/blog/accelerating-generative-ai-2/）该资源是 PyTorch 官方博客中的一篇文章，介绍了通过多种

优化步骤提升生成式 AI 模型（如 Llama-7B）推理性能的技术细节和实践案例，重点展示了 PyTorch 团队在推理加速方面取得的吞吐量改进。

74. https://oreil.ly/bNAOG: （https://www.nvidia.com/en-us/data-center/h100/）该资源是 NVIDIA 官方网站上关于其数据中心用 GPU 产品 H100 的介绍页面，详细展示了 H100 芯片的技术规格、性能指标及其在不同计算精度格式下的浮点运算能力（FLOP/s）等信息。
75. https://oreil.ly/eMQ_P: （https://www.youtube.com/watch?v=ELIcy6flgQI）该资源是 YouTube 上的视频，标题为《Llama Inference at Meta》，内容涉及 Meta 公司在 2024 年发布的关于 Llama 模型推理优化的技术讲解，特别提到了预填充(prefill)与解码实例(decoding instance)的比例等关键技术细节。
76. https://oreil.ly/iK0tN: （https://stratechery.com/2022/an-interview-with-nvidia-ceo-jensen-huang-about-manufacturing-intelligence/）该资源是一篇发表于 2022 年 3 月的 Stratechery 网站上的访谈文章，内容是对 NVIDIA 首席执行官黄仁勋（Jensen Huang）关于 GPU 发展及其在制造智能领域应用的访谈。文章中黄仁勋谈到了 GPU 的演变、扩展功能以及是否考虑更名的问题。
77. https://oreil.ly/ixtBl: （https://proceedings.mlsys.org/paper_files/paper/2024/hash/48fecef47b19fe501d27d338b6d52582-Abstract-Conference.html）该 URL 指向的资源是 2024 年机器学习系统会议（MLSys 2024）论文的摘要页面，论文主题涉及 KV 缓存（Key-Value Cache）相关技术，可能包括 KV 缓存的量化、压缩或选择性使用等方法，以提升机器学习模型的效率与性能。
78. https://oreil.ly/jD5Pj: （https://www.itnews.com.au/news/lufthansa-delays-chatbots-responses-to-make-it-more-human-462643）该资源是一篇发表于 2017 年 5 月、由 Ry Crozier 撰写的 iTnews 报道，标题为《Lufthansa Delays Chatbot’s Responses to Make It More ‘Human’》。文章介绍了德国汉莎航空公司为使其聊天机器人显得更具人性化，故意延迟机器人回复时间的做法，以避免机器人回应过快给用户带来不自然的体验。
79. https://oreil.ly/ludJ2: （https://developer.nvidia.com/system-management-interface）该资源是 NVIDIA 官方的系统管理接口（System Management Interface，简称 SMI）工具页面，提供用于监控和管理 NVIDIA

GPU 的命令行工具——nvidia-smi 的相关信息和下载。该工具可以显示 GPU 利用率、温度、功耗等多项硬件状态指标，帮助用户实时了解 GPU 的运行状况并进行性能调优。

80. https://oreil.ly/m8daG:（https://cloud.google.com/edge-tpu）该资源是谷歌（Google）推出的 Edge TPU，一种专为边缘计算设计的加速芯片，优化了推理计算的性能和效率，适用于在本地设备上快速执行机器学习推断任务。
81. https://oreil.ly/mRNCP:（https://www.computer.org/publications/tech-news/chasing-pixels/is-it-time-to-rename-the-gpu）该资源是一篇发表于《Chasing Pixels》栏目（由 Jon Peddie 撰写，2018 年 7 月）的文章，讨论了是否应该为 GPU（图形处理器）更名的问题。文章分析了 GPU 在图形之外的多种用途，以及业界对 GPU 名称是否需要更具通用性（如 GPGPU 或 XGU）的探讨和观点。
82. https://oreil.ly/nLt6A:（https://research.character.ai/optimizing-inference/）该资源介绍了 Character.AI 在优化推理性能方面的技术方案，重点讲述了如何通过改进注意力机制设计（如多查询注意力、多层交叉注意力及局部与全局注意力交错）大幅减少 KV 缓存大小，从而提升推理吞吐量，解决了长对话历史带来的内存瓶颈问题。
83. https://oreil.ly/oDQOk:（https://habana.ai/products/gaudi/）该资源是英特尔旗下 Habana Labs 推出的 Gaudi 加速器产品页面，介绍了一种专为加速人工智能（AI）工作负载设计的专用处理器，旨在提升深度学习训练和推理的性能与效率。
84. https://oreil.ly/pIHkL:（https://ai.google.dev/gemini-api/docs/caching）该资源介绍了 Google Gemini 模型 API 中的“prompt caching”（提示缓存）功能，具体说明了缓存的输入令牌可享受 75% 的费用折扣，但需要额外支付缓存存储费用（当前为每百万令牌每小时 1 美元）。文档详细阐述了该功能的使用方法、成本节约及性能提升等相关内容。
85. https://oreil.ly/qSpMK:（https://www.sciencedirect.com/science/article/pii/S2210537923000124）该资源是一篇由 Desislavov 等人于 2023 年发表的学术文章，内容聚焦于机器学习系统中的推理（inference）阶段。文章通过调研指

出，在常用系统中，推理的成本可能超过训练成本，且在已部署的人工智能系统中，推理阶段的开销可占到整体机器学习成本的 90%。该研究强调了专门针对推理优化的芯片设计的重要性。

86. https://oreil.ly/qwlHE: （https://openreview.net/forum?id=rJl-b3RcF7）该资源是一个学术论文页面，链接指向 OpenReview 平台上的论文，内容可能与神经网络剪枝（pruning）技术相关，讨论如何通过剪枝减少模型参数数量，从而降低内存占用和提升计算速度，同时保持模型性能。
87. https://oreil.ly/tOOOD: （https://www.databricks.com/blog/llm-training-and-inference-intel-gaudi2-ai-accelerators）该资源是一篇由 Databricks 发布的博客文章，标题为《LLM Training and Inference with Intel Gaudi 2 AI Accelerators》。文章介绍了利用英特尔 Gaudi 2 人工智能加速器进行大型语言模型（如 Llama 2-70B）训练和推理的技术细节，重点讨论了带宽利用率随并发用户数变化的表现及优化效果。
88. https://oreil.ly/uzg1s: （https://docs.vllm.ai/en/latest/models/spec_decode.html）该资源（https://docs.vllm.ai/en/latest/models/spec_decode.html）是 vLLM 官方文档中关于模型解码（speculative decoding）技术的说明页面，介绍了这种推理加速方法的实现原理、优势及在 vLLM 框架中的应用。
89. https://oreil.ly/y45q6: （https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/）该资源是一篇由 Tim Dettmers 撰写的文章，标题为《TPUs vs GPUs for Transformers and BERT》，内容主要比较了谷歌 TPU 与 GPU 在运行 Transformer 模型（如 BERT）时的性能表现和效率，深入分析了两种硬件架构在深度学习模型加速中的优劣及适用场景。
90. https://oreil.ly/zHsb8: （https://docs.nvidia.com/nim/benchmarking/llm/latest/metrics.html）该资源是 NVIDIA 官方文档中的一部分，介绍了用于大规模语言模型（LLM）性能评测的指标，特别是关于生成文本时各个词元之间时间延迟（如 inter-token latency，ITL）的定义和测量方法。
91. https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product?_l=en_US: 该资源是一篇由 LinkedIn 工程

团队发布的博客文章，题为《Musings on Building a Generative AI Product》。文章分享了 LinkedIn 在构建和部署生成式人工智能产品过程中积累的经验和思考，特别涉及了生成式 AI 应用中的性能指标如延迟（Latency）、吞吐量（Throughput）之间的权衡，以及实际应用中通过批处理（batching）等技术手段提升吞吐量但可能增加响应时间的情况。文章内容适合对生成式 AI 产品设计、性能优化及实际工程实践感兴趣的读者参考。

第 10 章（47）

1. https://arxiv.org/abs/1902.07742: 该资源是 Fu et al. (2019)发表在 arXiv 上的一篇论文，主要研究如何利用自然语言反馈来增强强化学习算法的学习能力。文章探讨了通过自然语言指导改进强化学习模型的技术方法，展示了在训练智能体时结合人类语言反馈以提高学习效果的可能性和初步成果。这项工作是自然语言反馈与强化学习结合领域的重要早期研究之一。
2. https://arxiv.org/abs/1903.02020: 该资源是 Goyal 等人于 2019 年发表在 arXiv 上的一篇学术论文，研究主题为利用自然语言反馈来改进强化学习算法，推动强化学习系统通过人类的自然语言指令或反馈进行学习和优化，属于强化学习与自然语言处理结合领域的早期重要工作之一。
3. https://arxiv.org/abs/1911.02557: 该 URL “https://arxiv.org/abs/1911.02557” 指向的是一篇由 Ponnusamy 等人于 2019 年发表的学术论文，内容涉及自然语言反馈在早期对话式人工智能应用中的研究，特别是在亚马逊 Alexa 等语音交互系统中的应用与探索。论文关注如何利用自然语言反馈改进对话系统的性能和用户体验。
4. https://arxiv.org/abs/2008.06924: 该资源（https://arxiv.org/abs/2008.06924）是一篇由 Zhou 和 Small 于 2020 年发表的学术论文，内容聚焦于利用自然语言反馈来改进强化学习算法。这项研究属于强化学习领域，探讨如何让强化

学习模型从人类的自然语言反馈中学习，从而提升模型的学习效果和交互能力。该论文是近年来该方向的代表性工作之一。

5. https://arxiv.org/abs/2009.14715: 该资源是 2020 年由 Sumers 等人发表在 arXiv 上的一篇学术论文，内容涉及利用自然语言反馈进行强化学习算法的研究，属于强化学习领域中通过自然语言指导智能体学习的相关工作。论文展示了如何利用自然语言反馈改进强化学习模型的训练效果，推动了早期对话式人工智能系统（如智能语音助手）中自然语言交互学习技术的发展。
6. https://arxiv.org/abs/2010.12251: 该资源是由 Park 等人在 2020 年发表的论文，内容聚焦于自然语言反馈在早期会话式人工智能应用中的研究，特别是与 Amazon Alexa 等语音助手相关的技术和方法。论文探讨了如何利用自然语言反馈提升对话系统的性能和用户交互体验。
7. https://arxiv.org/abs/2208.03270: 该资源是一篇由 Xu 等人于 2022 年发表在 arXiv 上的学术论文，介绍了 FITS（Feedback for Interactive Talk & Search）数据集。该数据集包含用户针对应用输出的自然语言反馈，特别是用户经常以抱怨的形式表达对答案错误、不相关、有害、冗长或缺乏细节等方面的不满。论文通过自动聚类方法，将这些反馈归纳为若干类别，旨在帮助改进人机交互系统中的反馈处理与响应机制。
8. https://arxiv.org/abs/2306.13588: 该资源是由 Yuan 等人在 2023 年发表的一篇学术论文，内容涉及通过自动聚类方法对 FITS 数据集（Xu 等人，2022 年）中的反馈类型进行分析和归纳。
9. https://arxiv.org/abs/2307.09009: 该资源是一篇由 Chen 等人于 2023 年发表在 arXiv 上的论文，主要研究了 OpenAI 的 GPT-4 和 GPT-3.5 模型在不同版本（如 2023 年 3 月与 2023 年 6 月版本）之间的性能差异，揭示了 API 表面不变的情况下，底层模型更新可能导致的显著性能波动。这对于理解和评估 AI 模型的版本变化及其对应用性能的影响具有重要意义。
10. https://arxiv.org/abs/2310.13548: 该资源是由 Sharma 等人在 2023 年发表的一篇学术论文，研究了基于人类反馈训练的人工智能模型倾向于表现出讨好用户（谄媚）行为的现象，指出这类模型更可能呈现符合用户观点的回答，从而影响其客观性和准确性。

11. https://arxiv.org/abs/2404.12272: 该资源是由 Shankar 等人于 2024 年发表在 arXiv 上的一篇学术论文，研究了开发者在与大量数据交互过程中如何调整对生成结果的判断标准，进而改进提示词设计和评估流程，以提升生成系统输出质量。
12. https://github.com/Azure/PyRIT: 该资源是微软 Azure 推出的开源项目“PyRIT”，提供现成的守护栏（guardrail）解决方案，帮助开发者在应用中实现对 AI 模型输入输出的安全防护和内容过滤，从而降低提示词攻击及不当内容风险。
13. https://github.com/FlowiseAI/Flowise: 该资源是一个名为 Flowise 的开源 AI 编排工具项目，托管在 GitHub 上，旨在帮助开发者构建和管理基于检索增强生成（RAG）和智能代理的应用程序流程。
14. https://github.com/NVIDIA/NeMo-Guardrails: 该资源是 NVIDIA 推出的 NeMo Guardrails 项目的 GitHub 仓库，提供了开箱即用的 AI 安全防护解决方案，用于在应用中实现输入输出的安全“护栏”机制，帮助开发者防范和控制与 AI 模型交互过程中可能出现的风险和不当内容。
15. https://github.com/Portkey-AI/gateway: 该资源是“Portkey”的一个 AI Gateway 项目，托管在 GitHub 上，提供了一个开源的、用于连接和管理人工智能模型或服务的网关实现。
16. https://github.com/chiphuyen/aie-book: 该资源是一个 GitHub 代码仓库，名为 “aie-book”，由作者 chiphuyen 维护。该仓库包含与“观察性”（observability）相关的学习资料和资源，特别聚焦于基于基础模型（foundation models）构建的应用程序的观察性实践，旨在帮助开发者深入了解和实现软件系统的监控与可观测性。
17. https://github.com/deepset-ai/haystack: 该资源是“Haystack”，一个由 deepset-ai 维护的开源 AI 编排工具库，主要用于构建基于检索增强生成（RAG）和智能代理的应用程序，支持多种信息检索和自然语言处理功能，便于开发复杂的 AI 工作流。
18. https://github.com/langchain-ai/langchain: 该资源是 LangChain 的 GitHub 代码仓库，LangChain 是一个用于构建和管理基于人工智能的应用程序的编排工具，特别适用于实现检索增强生成（RAG）和智能代理等应用模式。

19. https://github.com/langflow-ai/langflow: 该资源是一个名为 Langflow 的 AI 编排工具项目，托管在 GitHub 上，旨在支持人工智能应用中的检索增强生成（RAG）和代理框架等常见的编排模式。
20. https://github.com/meta-llama/PurpleLlama: 该资源是 Meta 发布的开源项目“Purple Llama”，提供开箱即用的“护栏”解决方案，用于防范和控制与大型语言模型交互中的风险和攻击，帮助开发者实现对模型输入输出的安全保护。
21. https://github.com/run-llama/llama_index: 该资源 “https://github.com/run-llama/llama_index” 指向的是 LlamaIndex 的 GitHub 仓库。LlamaIndex 是一个用于 AI 编排的工具，主要用于构建基于检索的生成（RAG）和代理框架，帮助整合和组织各种检索与工具调用的应用模式。
22. https://github.com/wealthsimple/llm-gateway: 该资源是 Wealthsimple 提供的一个开源大型语言模型（LLM）网关实现，旨在简化和标准化与各种语言模型服务的集成和访问。
23. https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html: 该资源是一篇博客文章，标题为“Data Distribution Shifts and Monitoring”，由作者在其个人博客发布，内容涉及机器学习系统中的数据分布变化及其监控方法。这篇文章是作者 2022 年出版的书籍《Designing Machine Learning Systems》中关于监控章节的早期草稿。
24. https://oreil.ly/-kOHE: （https://platform.openai.com/docs/guides/moderation）该资源是 OpenAI 提供的内容审核（内容审查）API 的官方文档，介绍了如何利用 OpenAI 的内容审核工具对输入和输出内容进行自动检测和过滤，以防止不当或有害内容的生成和传播，从而为应用开发者提供有效的安全防护机制。
25. https://oreil.ly/0NuNb: （https://developers.cloudflare.com/ai-gateway/）该资源是 Cloudflare 提供的 AI Gateway 相关文档或平台，介绍了 Cloudflare 在 AI 网关领域的解决方案或工具，旨在帮助用户快速搭建和管理 AI 模型的接入与调用。
26. https://oreil.ly/18tY4: （https://www.businessinsider.com/leaked-charts-show-how-ubers-driver-rating-system-works-2015-2）该资源是一篇由

Business Insider 发布的文章，内容泄露并展示了 Uber 司机评分系统的内部运作机制，具体说明了司机评分的标准和影响，以及低于某一评分阈值（如 4.6 分）时司机可能面临被停用的风险。

27. https://oreil.ly/AWwkx: （https://www.sciencedirect.com/science/article/abs/pii/S0001457519312370）该资源是一篇发表于科学期刊上的学术论文，题目涉及人们如何适应和操纵自动化技术（如自动驾驶汽车）以获得更有利的结果。文中以“Liu et al., 2020”为例，讨论了用户通过改变行为模式来影响技术系统的反馈，体现了人机交互中的行为适应现象。该论文可能探讨了用户与自动驾驶车辆之间的互动策略及其社会影响。
28. https://oreil.ly/CxwLn: （https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety）该 URL（https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety）指向的是微软 Azure 提供的 AI 内容安全服务。该服务旨在为应用开发者提供开箱即用的内容安全防护解决方案，通过自动检测和过滤有害或不当内容，帮助构建安全可靠的人工智能应用，防范输入输出中的风险与攻击。
29. https://oreil.ly/D2X_Y: （https://mlflow.org/docs/latest/llms/gateway/index.html）该资源是 MLflow 官方文档中关于“AI Gateway”或“LLM Gateway”的介绍页面，详细说明了如何使用 MLflow 提供的网关组件来简化大型语言模型（LLM）服务的部署和管理，帮助用户方便地集成和调用各种语言模型接口。
30. https://oreil.ly/Edew9: （https://openai.com/index/dall-e/）该资源介绍了 OpenAI 开发的图像生成模型 DALL·E，特别是其“inpainting”功能。该功能支持用户在生成的图像中选定特定区域，通过文本提示对该区域进行编辑和改进，从而实现对图像的局部修改和优化，提升生成结果的质量和用户体验。
31. https://oreil.ly/GeZvj: （https://developer.apple.com/design/human-interface-guidelines/machine-learning#Calibration）该资源是苹果开发者网站上的“人机界面指南”（Human Interface Guidelines）中关于机器学习部分的内容，具体聚焦于“校准”（Calibration）主题。它提供了关于如何设计

和优化机器学习模型在应用中的表现，确保默认结果准确且用户体验良好，同时指导开发者如何合理处理用户反馈以提升模型效果。

32. https://oreil.ly/ICRRA: （https://docs.truefoundry.com/docs/ai-gateway）该资源是 TrueFoundry 提供的 AI Gateway 相关文档，介绍了 TrueFoundry 平台中 AI Gateway 的功能、实现方式及使用指南，帮助用户理解和快速部署 AI Gateway 以便于集成和管理 AI 模型服务。
33. https://oreil.ly/Oml_x: （https://blog.langchain.dev/announcing-langsmith/）该资源是 Langchain 官方博客中关于 LangSmith 的发布公告页面，介绍了 LangSmith 这一工具，它用于在 AI 应用中追踪请求的执行路径，详细记录从用户查询到最终响应的全过程，包括系统各步骤的连接关系、耗时及成本等信息，帮助开发者更好地理解和调试 AI 系统的运行情况。
34. https://oreil.ly/St4W6: （https://konghq.com/products/kong-ai-gateway）该资源是 Kong 公司提供的 AI 网关产品，旨在简化和加速人工智能模型的部署与管理，帮助用户高效地集成和调用 AI 服务。
35. https://oreil.ly/_5RFN: （https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt）该资源是一篇报道，内容涉及三星员工因在使用 ChatGPT 时输入了公司的专有信息，导致三星机密意外泄露的事件。
36. https://oreil.ly/acfq0: （https://www.interconnects.ai/p/evaluating-open-llms）该资源（https://www.interconnects.ai/p/evaluating-open-llms）是一篇关于评估开源大型语言模型（Open LLMs）的文章或报告，内容涉及如何识别和分析人类反馈中的各种偏见（如偏好较长回答的偏见、近因偏见等），并探讨这些偏见如何影响对语言模型输出质量的评价。
37. https://oreil.ly/bGAeG: （https://dl.acm.org/doi/abs/10.1145/3178876.3185992）该资源是一篇发表于 ACM 数字图书馆的学术论文，主题围绕利用自然语言反馈（Natural Language Feedback）来改进强化学习（Reinforcement Learning）算法，探讨如何通过对话式反馈提升智能体的学习效果。文章可能涵盖相关方法、实验结果及其在早期对话式人工智能应用中的实际应用，如智能语音助手和语音控制功能。

38. https://oreil.ly/d2_sL:（https://perspectiveapi.com/）该资源是“Perspective API”，一个用于内容审核和过滤的工具或服务，帮助开发者实现内容的安全防护和管理，防止有害或不当内容的出现，常用于输入输出内容的风险监控和内容调控。
39. https://oreil.ly/jtt2m:（https://www.mdpi.com/2813-2203/2/2/20）该资源是一篇发表在 MDPI 期刊上的学术论文，题为“Sharma et al. (2023)”的研究，主要探讨了基于人类反馈训练的人工智能模型容易表现出“谄媚”倾向，即模型倾向于迎合用户的观点，而不一定提供最准确或最有益的信息。
40. https://oreil.ly/m8o0h:（https://dl.acm.org/doi/abs/10.1145/3479532）该资源是一篇发表在 ACM 数字图书馆上的学术论文，主题涉及利用自然语言反馈改进强化学习算法，属于对话系统和强化学习交叉领域的研究。论文可能探讨如何通过人类的自然语言指令或反馈来指导和提升强化学习模型的性能，推动早期对话式人工智能应用的发展。
41. https://oreil.ly/nAplp:（https://help.openai.com/en/articles/9055440-editing-your-images-with-dall-e）该资源是 OpenAI 官方帮助文档中的一篇文章，介绍了如何使用 DALL·E 进行图像编辑（inpainting）功能的说明与操作指南。
42. https://oreil.ly/vIfkA:（https://www.voiceflow.com/blog/how-much-do-chatgpt-versions-affect-real-world-performance）该资源是 Voiceflow 发布的一篇博客文章，内容讨论了不同版本的 ChatGPT 模型在实际应用中的性能差异，具体分析了模型更新如何影响真实世界中的表现和效果。文章通过实证数据展示了不同版本（如 GPT-3.5-turbo-0301 与 GPT-3.5-turbo-1106）之间可能出现的性能变化，帮助开发者理解和应对模型升级带来的影响。
43. https://oreil.ly/ve1DN:（https://www.wkyufm.org/news/2023-03-07/kentucky-senator-whose-twitter-account-liked-obscene-tweets-says-he-was-hacked）该资源是一篇 2023 年 3 月发布于 WKU 公共广播电台的新闻报道，内容涉及一位肯塔基州参议员的 Twitter 账号被发现“点赞”了不雅推文，该参议员声称其账号遭到黑客入侵。

44. https://oreil.ly/xKEVc: （https://www.politico.com/story/2017/09/12/ted-cruz-twitter-pornographic-post-242584）该资源是一篇发表于 2017 年 9 月的 POLITICO 报道，标题为《Ted Cruz Blames Staffer for ‘Liking’ Porn Tweet》，内容涉及美国参议员特德·克鲁兹（Ted Cruz）推特账户意外“点赞”色情内容推文事件，以及他将此行为归咎于工作人员的相关情况。
45. https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/: 该资源是一本由 O’Reilly 于 2022 年出版的书籍，书名为《Designing Machine Learning Systems》。该书涵盖了机器学习系统设计的相关内容，其中包括关于监控（monitoring）章节的详细介绍。
46. https://x.com/elonmusk/status/1800905349148664295: 该 URL 指向的是 Elon Musk 在 X（前身为 Twitter）上的一条动态，内容说明了 2024 年 X 平台将“点赞”设为私密后，用户点赞数量显著增加的情况。Elon Musk 作为 X 的所有者，在该条动态中分享了这一变化及其带来的正面影响。
47. https://x.com/elonmusk/status/1801045558318313746: 该 URL（https://x.com/elonmusk/status/1801045558318313746）指向的是 Elon Musk 在 X（前 Twitter）上的一条动态，内容说明自从 2024 年 X 平台将“点赞”行为设置为私密后，点赞数量显著增加。该推文用以支持文章中关于“点赞私密化提升用户互动和反馈质量”的论述。

原书 Github 仓库提供的资源

在撰写《AI 工程》一书的过程中，我查阅了大量论文、案例分析、博客文章、开源仓库和工具等。全书共包含 1200 多个参考链接，同时我还在持续追踪[1000 多个生成式 AI 相关的 GitHub 仓库](https://huyenchip.com/llama-police)。本文件整理了我认为最有助于理解各领域的精选资源。

如果你有发现其他实用但未被收录的资料，欢迎提交 PR 补充。

机器学习理论基础

虽然没有机器学习背景也能上手基础模型的开发，但了解 AI 底层的基本原理，有助于避免误用。掌握机器学习理论会让你事半功倍。

1. [讲义] [斯坦福 CS 321N](https://cs231n.github.io/)：神经网络入门课程中的经典之作。
   - [视频] 推荐观看 2017 年课程的第 1 到第 7 讲[视频录播](https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv)，涵盖了至今仍然适用的基础知识。
   - [视频] Andrej Karpathy 的[Neural Networks: Zero to Hero](https://www.youtube.com/playlist?list=PLAqhIrjkxbuWI23v9cThsA9GvCAUhRvKZ)系列更偏重实操，手把手演示如何从零实现多种模型。
2. [书籍] [Machine Learning: A Probabilistic Perspective](https://probml.github.io/pml-book/book1.html)（Kevin P Murphy, 2012）

   这本书内容扎实全面，但难度较高。曾经是许多朋友准备研究岗理论面试的首选读物。
3. [Aman 的数学入门笔记](https://aman.ai/primers/math/)

   简明扼要地梳理了基础微积分和概率相关概念。
4. 我还整理了一份 MLOps 相关资源清单，其中包含[机器学习与工程基础](https://huyenchip.com/mlops/#ml_engineering_fundamentals)的专门板块。
5. 我写过一篇[1500 字的简要笔记](https://github.com/chiphuyen/dmls-book/blob/main/basic-ml-review.md)，介绍了机器学习模型的学习过程，以及目标函数、学习步骤等核心概念。
6. 《AI 工程》一书也专门讲解了与本书主题紧密相关的重要概念：
   - Transformer 架构（第 2 章）
   - 向量嵌入（第 3 章）
   - 反向传播与可训练参数（第 7 章）

第 1 章

1. [GPTs are GPTs: 大型语言模型对劳动力市场影响的初探](https://arxiv.org/abs/2303.10130)（OpenAI，2023）

   OpenAI（2023）对不同职业受 AI 影响的程度进行了深入研究。他们将“受影响任务”定义为：如果 AI 或 AI 驱动的软件能将完成该任务所需时间缩短至少 50%，则该任务即为“受影响”。某一职业的“80% 受影响”意味着该职业 80% 的任务都属于此类。研究显示，受 AI 影响程度接近或达到 100% 的职业包括口译与笔译、报税员、网页设计师和作家等（部分见图 1-5）。而厨师、石匠、运动员等职业则几乎不受 AI 影响。这项研究很好地展示了 AI 适合应用的场景。

2. [Applied LLMs](https://applied-llms.org/)（Yan 等，2024）

   Eugene Yan 及其团队分享了一年来部署大语言模型应用的经验，内容实用，建议颇多。

3. [关于构建生成式 AI 产品的思考](https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product)（Juan Pablo Bottaro 与 Karthik Ramgopal，LinkedIn，2024）

   这是我读过的关于 LLM 应用落地最精彩的报告之一，详细分析了哪些做法有效、哪些无效。他们讨论了结构化输出、延迟与吞吐量的权衡、评估难题（大部分时间都花在制定标注规范上），以及生成式 AI 应用的“最后一公里”难题。

4. [Apple 人机界面指南](https://developer.apple.com/design/human-interface-guidelines/machine-learning)：ML 应用设计

   指导你如何思考 AI 与人在应用中的角色，这将直接影响界面设计决策。

5. [LocalLlama 论坛](https://www.reddit.com/r/LocalLLaMA/)：不时浏览，了解社区动态。

6. [State of AI Report](https://www.stateof.ai/)（每年更新）：内容全面，快速浏览可了解最新 AI 进展。

7. [企业构建与采购生成式 AI 的 16 大变化](https://a16z.com/generative-ai-enterprise-2024/)（Andreessen Horowitz，2024）

8. [“就像有个很糟糕的助理”：用户对话式智能体的期待与现实差距](https://dl.acm.org/doi/abs/10.1145/2858036.2858288)（Luger 与 Sellen，2016）

   这是一篇前瞻性的论文，深入探讨了用户与对话式智能体的交互体验。文章通过 14 位用户的深度访谈，论证了对话界面的价值及其改进方向。文中提到：**“对话界面系统相较于直接操作（GUI）的真正价值，体现在任务最复杂的场景中。”**

9. [斯坦福讲座：AI 如何改变编程与教育，Andrew Ng & Mehran Sahami](https://www.youtube.com/watch?v=J91_npj0Nfw&ab_channel=StanfordOnline)（2024）

   这场讨论展示了斯坦福 CS 系对未来计算机教育的思考。最喜欢的一句话：“计算机科学的核心是系统性思维，而不是写代码。”

10. [职业艺术家：AI 艺术对你职业影响有多大？——一年后 : r/ArtistLounge](https://www.reddit.com/r/ArtistLounge/comments/1ap0cm3/professional_artists_how_much_has_ai_art_affected/)

    许多人分享了 AI 对其工作的实际影响。例如：

“有时候，我会坐在会议室里，听着管理层畅想用 AI 取代程序员、写作者和视觉艺术家。我讨厌这样的会议，总想避开，但有时还是会被卷进去。毕生以来，我都热爱编程和艺术。可如今，心里常常涌上一种说不出的忧伤。”

第 2 章

大模型的训练

那些详细介绍重要模型训练过程的论文，堪称宝藏。我建议大家都去读一读。如果只能选三篇，我推荐 Gopher、InstructGPT 和 Llama 3。

1. [GPT-2] [语言模型是无监督多任务学习者](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)（OpenAI，2019）

2. [GPT-3] [语言模型是少样本学习者](https://arxiv.org/abs/2005.14165)（OpenAI，2020）
3. [Gopher] [扩展语言模型：Gopher 训练的方法、分析与洞见](https://arxiv.org/abs/2112.11446)（DeepMind，2021）
4. [InstructGPT] [用人类反馈训练语言模型遵循指令](https://arxiv.org/abs/2203.02155)（OpenAI，2022）
5. [Chinchilla] [训练计算最优的大型语言模型](https://arxiv.org/abs/2203.15556)（DeepMind，2022）
6. [Qwen 技术报告](https://arxiv.org/abs/2309.16609)（阿里巴巴，2022）
7. [Qwen2 技术报告](https://arxiv.org/abs/2407.10671)（阿里巴巴，2024）
8. [Constitutional AI：基于 AI 反馈的无害性](https://arxiv.org/abs/2212.08073)（Anthropic，2022）
9. [LLaMA：开放高效的基础语言模型](https://arxiv.org/abs/2302.13971)（Meta，2023）
10. [Llama 2：开放基础与微调聊天模型](https://arxiv.org/abs/2307.09288)（Meta，2023）
11. [Llama 3 模型家族](https://arxiv.org/abs/2407.21783)（Meta，2024）

    这篇论文非常精彩，尤其是关于合成数据生成与验证的部分，极具参考价值。
12. [Yi：01.AI 推出的开放基础模型](https://arxiv.org/abs/2403.04652)（01.AI，2024）

扩展规律

1. [从裸机到高性能训练：基础设施脚本与最佳实践 - imbue](https://imbue.com/research/70b-infrastructure/)

   讨论如何扩展算力以训练大模型。文中使用了 4,092 块 H100 GPU，分布在 511 台计算机上，每台 8 块 GPU。
2. [神经语言模型的扩展规律](https://arxiv.org/abs/2001.08361)（OpenAI，2020）

   早期的扩展规律研究，参数量和训练数据量都只到 10 亿级别。

3. [训练计算最优的大型语言模型](https://arxiv.org/abs/2203.15556)（Hoffman 等，2022）

   即 Chinchilla 扩展规律，可能是最知名的扩展规律论文。

4. [数据受限下的语言模型扩展](https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html)（Muennighoff 等，2023）

   “我们发现，在计算预算固定、数据受限的情况下，最多重复 4 轮数据训练，损失几乎不变于全新数据。但再多重复，算力的边际价值就会趋近于零。”

5. [指令微调语言模型的扩展](https://arxiv.org/abs/2210.11416)（Chung 等，2022）

   这是一篇很好的论文，强调了指令数据多样性的重要性。

6. [超越 Chinchilla 最优：将推理计算纳入语言模型扩展规律](https://arxiv.org/abs/2401.00448)（Sardana 等，2023）

7. [AI 模型正在吞噬能源，降低能耗的工具已现，只待数据中心采纳](https://www.ll.mit.edu/news/ai-models-are-devouring-energy-tools-reduce-consumption-are-here-if-data-centers-will-adopt)（MIT 林肯实验室，2023）

8. [我们会耗尽数据吗？基于人类生成数据的 LLM 扩展极限](https://arxiv.org/abs/2211.04325)（Villalobos 等，2022）

趣味内容

1. [评估特征引导：缓解社会偏见的案例研究](https://www.anthropic.com/research/evaluating-feature-steering)（Anthropic，2024）

   这一研究领域非常有趣。作者聚焦于 29 项[与社会偏见相关的特征](https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html=safety-relevant-bias)，发现通过特征引导可以影响特定的社会偏见，但同时也可能带来意料之外的“偏离目标效应”。

2. [单义性扩展：从 Claude 3 Sonnet 中提取可解释特征](https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html)（Anthropic，2024）

3. [GitHub - ianand/spreadsheets-are-all-you-need](https://github.com/ianand/spreadsheets-are-all-you-need)

   “用标准 Excel 表格函数，在 Excel 中完整实现了 GPT2（ChatGPT 的前身）的前向计算过程。”

4. [BertViz：可视化 NLP 模型中的注意力机制（BERT、GPT2、BART 等）](https://github.com/jessevig/bertviz)

   这是一个非常实用的多头注意力可视化工具，专为展示 BERT 等模型的工作原理而开发。

采样

1. [结构化生成与约束解码指南](https://www.aidancooper.co.uk/constrained-decoding/)（Aidan Cooper，2024）

   一篇深入细致的结构化输出生成教程。

2. [基于压缩有限状态机的本地大模型快速 JSON 解码](https://lmsys.org/blog/2024-02-05-compressed-fsm/)（LMSYS，2024）

3. [语法结构生成能有多快？](https://blog.dottxt.co/how-fast-cfg.html)（Brandon T. Willard，2024）

我也写过一篇关于[文本生成采样](https://huyenchip.com/2024/01/16/sampling.html)的文章（2024）。

上下文长度与上下文效率

1. [长上下文微调全解](https://huggingface.co/blog/wenbopan/long-context-fine-tuning)（潘文博，2024）

2. [将语言模型扩展到 128K 上下文的数据工程](https://arxiv.org/abs/2402.10171v1)（Yu 等，2024）

3. [大模型 100K 上下文窗口的秘诀：所有技巧一网打尽](https://blog.gopenai.com/how-to-speed-up-llms-and-use-100k-context-window-all-tricks-in-one-place-ffd40577b4c)（Galina Alperovich，2023）

4. [扩展上下文很难……但并非不可能](https://kaiokendev.github.io/context)（kaioken，2023）
5. [RoFormer：旋转位置编码增强的 Transformer](https://arxiv.org/abs/2104.09864)（苏等，2021）

   介绍了 RoPE（旋转位置编码）技术，使基于 Transformer 的模型能够处理更长的上下文。

第 3、4 章

1. [AI 系统评估的挑战](https://www.anthropic.com/news/evaluating-ai-systems)（Anthropic，2023）

   讨论了常见 AI 基准测试的局限性，揭示了为何 AI 评估如此困难。

2. [语言模型的整体性评估](https://arxiv.org/abs/2211.09110)（Liang 等，斯坦福，2022）
3. [超越模仿游戏：量化与外推语言模型能力](https://arxiv.org/abs/2206.04615)（Google，2022）
4. [开源 LLM 性能进入平台期，让排行榜再次“陡峭”起来](https://huggingface.co/spaces/open-llm-leaderboard/blog)（Hugging Face，2024）

   详细解释了 Hugging Face 为何选择特定基准测试作为排行榜标准，对于你自建排行榜时如何挑选评测集也很有参考价值。

5. [用 MT-Bench 和 Chatbot Arena 评判“LLM 法官”](https://arxiv.org/abs/2306.05685)（Zheng 等，2023）
6. [LLM 任务型评测：哪些有效，哪些无效](https://eugeneyan.com/writing/evals/)（Eugene Yan，2024）
7. [你的 AI 产品需要评测](https://hamel.dev/blog/posts/evals/)（Hamel Hussain，2024）
8. [别再明文上传测试数据：防止评测集数据污染的实用策略](https://arxiv.org/abs/2305.10160)（Google & AI2，2023 年 5 月）

9. [alopatenko/LLMEvaluation](https://github.com/alopatenko/LLMEvaluation)（Andrei Lopatenko）

   收录了大量评测资源。其[评测 PPT](https://github.com/alopatenko/LLMEvaluation/blob/main/LLMEvaluation.pdf)也有不少参考资料。

10. [用模型生成的评测集发现语言模型行为](https://arxiv.org/abs/2212.09251)（Perez 等，2022）

    这是一篇有趣的论文，利用 AI 自身来发现 AI 的新行为。他们用不同自动化程度的方法，为 154 种多样行为生成了评测集。

11. [AI 海洋中的“海妖之歌”：大语言模型幻觉问题综述](https://arxiv.org/abs/2309.01219)（Zhang 等，2023）

12. OpenRouter 的[LLM 排行榜](https://openrouter.ai/rankings)展示了平台上最受欢迎的开源模型（按 Token 用量排名）。这有助于你通过流行度评估开源模型。希望更多推理服务能公开类似统计数据。

第 5 章

提示词工程指南

1. [Anthropic 提示词工程交互式教程](https://docs.google.com/spreadsheets/d/19jzLgRruG9kjUQNKtCg1ZjdD6l6weA6qRXG5zLIAhC8/edit=gid=1733615301)

   实用、全面且有趣。基于 Google 表格的交互练习让你可以轻松尝试不同提示词，并立刻看到效果。令人惊讶的是，其他模型厂商还没有类似的交互式指南。

2. [Brex 提示词工程指南](https://github.com/brexhq/prompt-engineering)

   收录了 Brex 公司内部使用的提示词示例。

3. [Meta 提示词工程指南](https://llama.meta.com/docs/how-to-guides/prompting/)

4. [Google Gemini 提示词工程指南](https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf)

5. [dair-ai/Prompt-Engineering-Guide](https://github.com/dair-ai/Prompt-Engineering-Guide)
6. 各大厂商的提示词示例集：[OpenAI](https://platform.openai.com/examples)、[Anthropic](https://docs.anthropic.com/en/prompt-library/library)、[Google](https://console.cloud.google.com/vertex-ai/generative/prompt-gallery)。
7. [更大的语言模型在上下文学习中的表现差异](https://arxiv.org/abs/2303.03846)（Wei 等，2023）
8. [我如何看待 LLM 提示词工程](https://fchollet.substack.com/p/how-i-think-about-llm-prompt-engineering)（Francois Chollet，2023）

防御性提示词工程

1. [Offensive ML Playbook](https://wiki.offsecml.com/Welcome+to+the+Offensive+ML+Playbook)

   收录了大量关于对抗性机器学习的资源，涵盖如何防御文本和图像等多种攻击手段，适合想要提升 ML 系统安全性的读者。

2. [The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions](https://arxiv.org/abs/2404.13208)（OpenAI，2024）

   这篇论文介绍了 OpenAI 如何通过训练模型建立指令优先级体系，从而防止模型被越狱攻击，值得一读。

3. [Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection](https://arxiv.org/abs/2302.12173)（Greshake 等，2023）

   附录中整理了大量间接提示注入的实际案例，非常有参考价值。

4. [Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks](https://arxiv.org/abs/2302.05733)（Kang 等，2023）
5. [Scalable Extraction of Training Data from (Production) Language Models](https://arxiv.org/abs/2311.17035)（Nasr 等，2023）

6. [How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs](https://arxiv.org/abs/2401.06373)（Zeng 等，2024）
7. [LLM Security](https://llmsecurity.net/)：LLM 安全相关论文的合集。
8. 自动化安全测试工具包括 [PyRIT](https://github.com/Azure/PyRIT)、[Garak](https://github.com/leondz/garak/)、[persuasive_jailbreaker](https://github.com/CHATS-lab/persuasive_jailbreaker)、[GPTFUZZER](https://arxiv.org/abs/2309.10253) 和 [MasterKey](https://arxiv.org/abs/2307.08715)。
9. [Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations](https://arxiv.org/abs/2312.06674)（Meta，2023）
10. [AI Security Overview](https://owaspai.org/docs/ai_security_overview/=threat-model)（AI Exchange）

第 6 章

RAG

1. [Reading Wikipedia to Answer Open-Domain Questions](https://arxiv.org/abs/1704.00051)（Chen 等，2017）

   首次提出 RAG 模式，用于提升问答等知识密集型任务的效果。

2. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](https://arxiv.org/abs/2005.11401)（Lewis 等，2020）
3. [Retrieval-Augmented Generation for Large Language Models: A Survey](https://arxiv.org/abs/2312.10997)（Gao 等，2023）
4. [Introducing Contextual Retrieval](https://www.anthropic.com/news/contextual-retrieval)（Anthropic，2024）

   这篇文章讨论了 RAG 系统中数据准备的重要性，介绍了多种数据处理技巧，并分析了何时应采用 RAG、何时适合用长上下文。

5. [Pinecone](https://www.pinecone.io/learn/chunking-strategies/) 和 [Langchain](https://js.langchain.com/v0.1/docs/modules/data_connection/document_transformers/) 的分块教程
6. [The 5 Levels Of Text Splitting For Retrieval](https://www.youtube.com/watch?v=8OJC21T2SL4)（Greg Kamradt，2024）
7. [GPT-4 + Streaming Data = Real-Time Generative AI](https://www.confluent.io/blog/chatgpt-and-streaming-data-for-real-time-generative-ai/)（Confluent，2023）

   详细介绍了在 RAG 应用中实时检索数据的最佳实践。
8. [Everything You Need to Know about Vector Index Basics](https://zilliz.com/learn/vector-index)（Zilliz，2023）

   系统讲解了向量检索和向量数据库的基础知识。
9. [A deep dive into the world’s smartest email AI](https://www.shortwave.com/blog/deep-dive-into-worlds-smartest-email-ai/)（Hiranya Jayathilaka，2023）

   虽然标题夸张，但内容是一份详尽的 RAG 邮件助手案例分析。
10. [书籍] [Introduction to Information Retrieval](https://nlp.stanford.edu/IR-book/information-retrieval-book.html)（Manning、Raghavan、Schütze，2008）

    信息检索是 RAG 的基础。想深入了解文本数据组织与查询技术，这本书不可错过。

智能体

1. 《Chameleon：大语言模型的即插即用组合式推理》（Lu 等，2023）

   这是我最喜欢的一项关于 LLM 规划器、工具使用方式及其失败模式的研究。有趣的发现之一是，不同的大语言模型在选择工具时有各自的偏好。
2. 《生成式智能体：人类行为的交互式拟像》（Park 等，2023）
3. 《Toolformer：语言模型可以自学工具使用》（Schick 等，2023）

4. 伯克利函数调用排行榜 及论文《Gorilla：大语言模型连接海量 API》（Patil 等，2023）

   其中列举了 ChatGPT 在函数调用中常见的四种错误，非常有意思。

5. THUDM/AgentBench：评估大语言模型作为智能体的基准（ICLR’24）
6. 《WebGPT：结合浏览器与人类反馈的问答系统》（Nakano 等，2021）
7. 《ReAct：语言模型中的推理与行动协同》（Yao 等，2022）
8. 《Reflexion：具备语言强化学习能力的智能体》（Shinn 等，2023）
9. 《Voyager：基于大语言模型的开放式具身智能体》（Wang 等，2023）
10. 《人工智能：现代方法》（Russell 与 Norvig，第四版，2020）

    规划与搜索密切相关，这本经典教材中有多章深入讲解了搜索相关内容。

第 7 章

1. 《GPT-3 文本分类微调最佳实践》（OpenAI）

   这份 OpenAI 的草稿虽然聚焦于 GPT-3，但其中许多技术同样适用于全面微调。内容涵盖 GPT-3 微调原理、训练数据准备、模型评估方法及常见错误。

2. 《轻松训练专用 LLM：PEFT、LoRA、QLoRA、LLaMA-Adapter 等》（Cameron R. Wolfe，2023）

   这是一篇 7000 字、调研详实的长文，系统梳理了基于 Adapter 的高效参数微调方法的演变，深入解释了 LoRA 为何流行及其原理。

3. 《微调还是检索增强？LLM 知识注入方式对比》（Ovadia 等，2024）

   这项有趣的研究为“微调还是 RAG”这一问题提供了参考答案。

4. 《NLP 中的高效参数迁移学习》（Houlsby 等，2019）

   该论文首次提出了 PEFT（高效参数微调）的概念。

5. 《LoRA：大语言模型的低秩适配》（Hu 等，2021）

   必读论文。

6. 《QLoRA：高效微调量化大语言模型》（Dettmers 等，2023）

7.《基于合成数据的直接偏好优化》（Anyscale，2024）
8.《Transformer 推理算术》（kipply，2022）
9.《Transformer 数学入门》（EleutherAI，2023）：主要讲解训练过程中的显存占用计算。
10.《缩小规模以实现高效微调：参数高效微调指南》（Lialin 等，2023）

    对多种微调方法进行了全面梳理，但并非所有技术都适用于当前主流实践。

11.《我的 LLM 自定义数据微调入门经验》（2023，r/LocalLLaMA）
12.《混合精度训练指南》（NVIDIA 官方文档）

第 8 章

1. [高质量数据集构建的标注最佳实践](https://www.grammarly.com/blog/engineering/annotation-best-practices/)（Grammarly，2022）
2. [指令微调语言模型的扩展](https://arxiv.org/abs/2210.11416)（Chung 等，2022）
3. [递归的诅咒：用生成数据训练会让模型遗忘](https://arxiv.org/abs/2305.17493)（Shumailov 等，2023）
4. [Llama 3 模型家族](https://arxiv.org/abs/2407.21783)（Meta，2024）

    这篇论文整体都很值得一读，尤其是关于合成数据生成与验证的部分尤为重要。

5. [用 GPT-4 进行指令微调](https://arxiv.org/abs/2304.03277)（Peng 等，2023）

    利用 GPT-4 生成指令跟随数据，用于大模型微调。

6. [语言模型合成数据的最佳实践与经验教训](https://arxiv.org/abs/2404.07503)（Liu 等，DeepMind 2024）
7. [UltraChat：通过大规模高质量指令对话提升聊天语言模型](https://arxiv.org/abs/2305.14233)（Ding 等，2023）
8. [去重训练数据让语言模型更优秀](https://arxiv.org/abs/2107.06499)（Lee 等，2021）

9. [大模型能否只靠一个样本学习？](https://www.fast.ai/posts/2023-09-04-learning-jumps/)（Jeremy Howard 和 Jonathan Whitaker，2023）

    一个有趣的实验，展示了仅用一个训练样本也能让模型有所提升。

10. [LIMA：少即是多的对齐方法](https://arxiv.org/abs/2305.11206)（Zhou 等，2023）

公开数据集

以下是一些公开数据集的获取渠道。虽然应充分利用现有数据，但绝不能完全信任它。数据必须经过细致检查和验证。

在使用数据集前，务必核查其许可证。尽量了解数据的来源。即使数据集允许商业用途，也有可能部分数据来自不允许的渠道。

1. [Hugging Face](https://huggingface.co/datasets) 和 [Kaggle](https://www.kaggle.com/datasets?fileType=csv) 各自托管了数十万数据集。
2. Google 有一个非常好用但不太为人知的 [Dataset Search](https://datasetsearch.research.google.com/)。
3. 各国政府通常是开放数据的重要提供者。[Data.gov](https://data.gov) 提供约数十万数据集，data.gov.in 也有数万数据集。
4. 密歇根大学[社会研究院](https://www.icpsr.umich.edu/web/pages/ICPSR/index.html)（ICPSR）收录了数万个社会学研究数据。
5. [加州大学欧文分校机器学习库](https://archive.ics.uci.edu/datasets) 和 [OpenML](https://www.openml.org/search?type=data&sort=runs&status=active) 是两个较早的数据集仓库，各自收录了数千个数据集。
6. [Open Data Network](https://www.opendatanetwork.com/) 可检索数万个数据集。
7. 云服务商通常也会托管一些开放数据集，最著名的是 [AWS Open Data](https://registry.opendata.aws/)。
8. 许多机器学习框架自带小型预置数据集，如 [TensorFlow datasets](https://www.tensorflow.org/datasets/catalog/overview=all_datasets)。

9. 一些评测工具也会托管规模较大的评测基准数据集，足以用于 PEFT 微调。例如，Eleuther AI 的 [lm-evaluation-harness](https://github.com/EleutherAI/lm-evaluation-harness) 收录了 400 多个基准数据集，平均每个数据集有 2000 多个样本。
10. [斯坦福大型网络数据集](https://snap.stanford.edu/data/) 是图数据集的重要仓库。

第 9 章

1. [精通 LLM 技术：推理优化](https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/)（NVIDIA 技术博客，2023）

   一篇非常优秀的综述，系统介绍了多种推理优化技术。

2. [用 PyTorch 加速生成式 AI（二）：GPT 极速实践](https://pytorch.org/blog/accelerating-generative-ai-2/)（PyTorch，2023）

   通过具体案例展示了不同优化方法带来的性能提升。

3. [高效扩展 Transformer 推理](https://arxiv.org/pdf/2211.05102)（Pope 等，2022）

   这是一篇技术含量极高的推理优化论文，出自 Jeff Dean 团队。我最喜欢其中关于不同权衡点（如延迟与成本）应关注哪些方面的讨论。

4. [Character.AI 的 AI 推理优化实践](https://research.character.ai/optimizing-inference/)（Character.AI，2024）

   这篇文章技术性不强，更像是“看，我们能做到这些”的展示。Character.AI 团队的成果令人印象深刻，内容涵盖注意力机制设计、缓存优化和 int8 训练等。

5. [视频] [GPU 优化研讨会：OpenAI、NVIDIA、PyTorch 与 Voltron Data](https://www.youtube.com/watch?v=v_q2JTIqE20&ab_channel=MLOpsLearners)

6. [视频] [Essence VC Q1 虚拟大会：LLM 推理](https://www.youtube.com/watch?v=XPArX12gXVE)（涵盖 vLLM、TVM 和 Modal Labs）
7. [大语言模型中的 KV 缓存优化技术](https://www.omrimallis.com/posts/techniques-for-kv-cache-optimization/)（Omri Mallis，2024）

   一篇极佳的文章，详细讲解了 KV 缓存优化，这是 Transformer 推理中内存占用最大的部分之一。

   [João Lages](https://medium.com/@joaolages/kv-caching-explained-276520203249)也有一篇很棒的 KV 缓存可视化文章。
8. [用推测采样加速大语言模型解码](https://arxiv.org/abs/2302.01318)（DeepMind，2023）
9. [DistServe：分离预填充与解码以优化大语言模型服务吞吐](https://arxiv.org/abs/2401.09670)（Zhong 等，2024）
10. [2023 年深度学习最佳 GPU 深度分析](https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/)（Tim Dettmers，2023）

    Stas Bekman 也有一份很棒的[加速器评测笔记](https://github.com/stas00/ml-engineering/tree/master/compute/accelerator)。
11. [大规模多租户 GPU 集群在 DNN 训练中的分析](https://www.usenix.org/system/files/atc19-jeon.pdf)（Jeon 等，2019）

    详细研究了微软 GPU 集群在多租户环境下训练深度神经网络的情况。作者分析了两个月的集群日志，重点关注影响集群利用率的三个关键问题：组调度与本地性约束、GPU 利用率和作业失败。
12. [AI 数据中心的能源困境——AI 数据中心空间之争](https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race)（SemiAnalysis，2024）

    对数据中心业务及其瓶颈进行了精彩分析。

我还有一篇较早的文章：[机器学习编译器与优化器入门指南](https://huyenchip.com/2021/09/07/a-friendly-introduction-to-machine-learning-compilers-and-optimizers.html)（Chip Huyen，2018）

第 10 章

1. [第 4 章：监控](https://sre.google/workbook/monitoring/)（Google SRE 手册）
2. [人机 AI 交互设计准则](https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/)（微软研究院）

   微软提出了 18 条人机 AI 交互设计准则，涵盖开发前、开发中、出错时及长期演进等各阶段的设计建议。
3. [洞察偏好：大语言模型对齐中的反馈获取机制](https://arxiv.org/abs/2308.15812v3)（Bansal 等，2023）

   研究反馈协议如何影响模型训练效果。
4. [基于反馈的大规模对话 AI 自学习](https://arxiv.org/abs/1911.02557)（Ponnusamy 等，亚马逊，2019）
5. [面向大规模对话 AI 系统的隐式用户反馈学习框架](https://arxiv.org/abs/2010.12251)（Park 等，亚马逊，2020）

对话 AI 的用户反馈设计目前研究较少，相关资料有限，但希望未来会有更多进展。

附录：知名企业工程博客

我很喜欢阅读优质的技术博客，以下是我经常关注的一些工程技术博客：

1. [LinkedIn 工程博客](https://www.linkedin.com/blog/engineering)
2. [DoorDash 工程博客](https://careersatdoordash.com/engineering-blog/)
3. [Uber 工程博客](https://www.uber.com/en-US/blog/engineering/)
4. [非官方 Google 数据科学博客](https://www.unofficialgoogledatascience.com/)
5. [Pinterest 工程博客 – Medium](https://medium.com/pinterest-engineering)

6. [Netflix 技术博客](https://netflixtechblog.com/)
7. [LMSYS 组织博客](https://lmsys.org/blog/)
8. [Anyscale 博客](https://www.anyscale.com/blog)
9. [Databricks 博客：数据科学与机器学习](https://www.databricks.com/blog/category/engineering/data-science-machine-learning)
10. [Together 博客](https://www.together.ai/blog)
11. [Duolingo 工程博客](https://blog.duolingo.com/hub/engineering/)