

面向统计学家的大语言模型综述

Wenlong Ji^{1,*} Weizhe Yuan^{2,3,*} Emily Getzen⁴
Kyunghyun Cho² Michael I. Jordan^{5,6} Song Mei⁵ Jason Weston^{2,3}
Weijie J. Su^{4,†} Jing Xu^{3,†} Linjun Zhang^{7,†}

¹Stanford University ²New York University ³Meta FAIR ⁴University of Pennsylvania
⁵UC Berkeley ⁶INRIA ⁷Rutgers University

本文的英文原文网址: <https://arxiv.org/abs/2502.17814>。译文初稿由 DeepSeek-R1 生成, 统计之都 (<https://cosx.org>) 常象宇、邱怡轩进行校对和润色。翻译已获得原文作者授权。

摘要

大语言模型 (Large Language Models, LLMs) 已成为人工智能 (Artificial Intelligence, AI) 领域变革性的工具, 在文本生成、推理和决策等多样化任务中展现出卓越能力。尽管其成功主要源于算力提升和深度学习架构的进步, 但不确定性量化、决策分析、因果推断和分布偏移等新兴问题的研究需要更深入地结合统计学方法。本文探讨了统计学家可能对 LLM 发展做出重要贡献的潜在领域, 特别是那些旨在增强人类用户信任度和透明度的方向。因此, 我们重点关注不确定性量化、可解释性、公平性、隐私性、数字水印和模型适应等问题, 同时分析了 LLM 在统计分析中可能扮演的角色。通过架起 AI 与统计学的桥梁, 我们致力于促进更深层次的学科协作, 共同推进 LLM 的理论基础与实际应用, 最终塑造其应对复杂社会挑战的能力。

目录

1 引言	3
2 LLM 背景与基础	5
2.1 LLM 的历史演进	5
2.1.1 表征学习	5
2.1.2 语言建模	6
2.2 预训练语言模型架构	7
2.3 LLM 评估	9

* 共同第一作者

† 共同通讯作者

3 LLM 的训练流程	10
3.1 LLM 预训练	10
3.1.1 预训练目标	10
3.1.2 预训练数据	11
3.1.3 扩展定律 (Scaling Law)	12
3.2 LLM 提示方法	13
3.3 LLM 监督微调 (SFT)	14
3.3.1 指令微调	14
3.3.2 参数高效微调	15
3.4 系统2提示与思维链	16
3.5 LLM 强化学习与偏好优化	17
3.5.1 LLM 的对齐方法	17
3.6 LLM 自我对齐	18
3.6.1 合成数据生成	19
3.6.2 自我奖励语言模型	19
3.6.3 挑战	20
4 通过统计方法构建可信 LLM	21
4.1 不确定性量化	21
4.2 LLM 水印技术	23
4.3 隐私与版权	25
4.4 可解释性	27
4.5 算法公平性	29
4.6 LLM 对齐：统计视角	32
5 LLM 赋能的统计分析	32
5.1 数据收集	32
5.1.1 从非结构化文本提取结构化数据	32
5.1.2 合成数据生成	33
5.2 数据清洗	33
5.3 LLM 用于数据分析	34
5.4 LLM 辅助的医学研究	35
6 讨论	37
6.1 统计学家的小语言模型	37
6.2 理解 LLM 的探索方向	38
6.3 人机协作的数据科学	39
A LLM 的历史发展	85
A.1 表征学习	85

A.2 语言建模	86
B 关键资源：加速 LLM 研究的进程	86
B.1 模型开发核心框架	86
B.2 模型库与预训练模型	87
B.3 数据集与数据准备工具	87
B.4 训练与微调工具	87
B.5 推理与部署	88
B.6 可视化与分析	89
B.7 云平台与硬件资源	90

1 引言

大语言模型（Large Language Models, LLMs）已成为人工智能（Artificial Intelligence, AI）领域变革性的工具。它在从文本生成到对话交互，再到复杂推理的多样化任务中展现出卓越的能力。随着模型规模和复杂性的持续增长，一个重要的问题随之浮现：在 LLM 时代，统计学家应如何指导大规模 AI 模型的设计与部署？

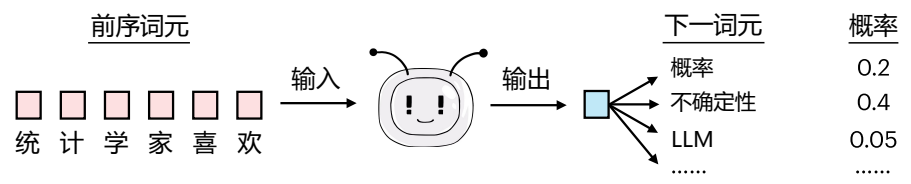


图 1: LLM 的下一词元预测示意图。

语言模型的核心是对后续词语或词序列进行概率分配的模型。具体而言，语言模型 f 旨在学习一个映射函数：它以词元序列 $\mathbf{x}_{[T]} = x_1, x_2, \dots, x_T$ 为输入，然后输出下一个词元 x_{T+1} 在词汇表上的概率向量 $f(\mathbf{x}_{[T]})$ 。虽然形式简洁，但当应用于海量文本语料时，这种概率化方法使 LLM 能够捕获精细的模式，包括传统上属于句法、语义和逻辑范畴的模式。例如，当接收到“创作一个引人入胜的短篇故事”或“让我们逐步分析”等指令时，LLM 可以动态调整响应以遵循给定的指令，生成符合用户意图的输出。这种指令遵循能力源于对多样化数据集的广泛训练，以及与用户偏好对齐的微调技术的结合。尽管这项工作大量借鉴了统计原理和优化理论，但它们在问题表述与针对 LLM 特定架构及数据收集协议设计的统计方法之间仍存在显著差距。

LLM 对创意写作指令的响应示例

用户： 创作一个关于统计学家团队徒步旅行的引人入胜的短篇故事。需为统计学家提供详细背景设定，描述旅途中的趣味与挑战，并通过丰富细节使故事生动。

LLM：

标题：《大径定律（The Law of Large Trails）》

一切始于统计系的一个玩笑。每周三研讨会后，教师和研究生们总会聚集在公共休息室，分享咖啡和抱怨，偶尔在餐巾纸上涂写证明草图。某个沉闷的冬日午后，斯特罗德曼教授叹气道：“我们终日分析数据中的不确定性，但上一次直面现实中的不确定性是什么时候？比如说……大自然里的？”……

LLM 通过思维链提示响应概率问题的示例

用户： 你是一位精通解决高级概率论问题的专家。请逐步仔细解答以下问题，确保解释每个计算步骤和假设条件：当投掷两个公平六面骰子时，至少出现一个6点的概率是多少？请逐步思考。

LLM：

- 单枚骰子有6种可能结果。投掷两枚骰子的总结果数为： $6 \times 6 = 36$ 。我们需要计算至少出现一个6点的概率，这包括任一或两枚骰子显示6的情况。更简便的方法是计算补集：两枚骰子均未出现6的概率。
- 单枚骰子未出现6的概率：单个骰子有5种非6结果（即1, 2, 3, 4, 5），对应概率为： $\mathbb{P}(\text{单骰非6}) = \frac{5}{6}$ 。
- 两枚骰子均未出现6的概率：由于投掷骰子独立，其联合概率为 $\mathbb{P}(\text{双骰非6}) = \frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$ 。
- 至少出现一个6的概率即为其补集： $\mathbb{P}(\text{至少一个6}) = 1 - \mathbb{P}(\text{双骰非6}) = 1 - \frac{25}{36} = \frac{11}{36}$ 。

这不仅仅是简单应用现有统计工具的问题——LLM 的兴起要求统计方法与理论的进步，以确保可信度、可靠量化不确定性并增强可解释性——所有这些都需在 LLM 部署的海量规模下实现。例如，在理解 LLM 架构设计如何产生（或未能产生）校准概率输出，或理解 LLM 输出如何与下游决策交互以确保算法公平性与可靠性方面，仍存在诸多开放性问题。

我们亦可将统计学本身视为 LLM 的应用领域。事实上，LLM 为增强传统统计工作流程提供了重要机遇。数据收集、清洗与分析等任务可利用 LLM 的上下文理解能力和可扩展性来提高准确性及效率。在医学研究与生物统计等领域，LLM 可合成代表性数据集、从非结构化临床记录中提取有价值的洞见，并支持高风险应用中的预测建模。当然，此类应用仍需弥合当前 LLM 设计的随意性与其输出

的统计特性之间的鸿沟。

这两种视角——服务于 LLM 的统计学与服务于统计学的 LLM——凸显了统计学家深度参与 LLM 革命的迫切需求。当然，这种参与是富有挑战性的。LLM 技术的快速发展催生了紧迫问题：统计框架如何适应这些模型前所未有的规模与复杂性？统计原则在不同应用中应如何确保 LLM 的可信和可靠的部署？统计学家如何在整合 LLM 至工作流程时，不损害学科严谨性、透明度与可解释性？解决这些问题对弥合统计严谨性与 LLM 变革潜力之间的鸿沟至关重要。

本综述旨在通过系统性梳理该领域现状，推动统计学家参与 LLM 的演进格局。文章首先介绍 LLM 的历史发展与基本原理，阐明其架构与训练流程；继而探讨统计方法如何增强 LLM 的可信度与对齐性，聚焦保形预测、合成文本检测和算法公平性等主题；随后分析 LLM 如何赋能数据分析、清洗及医学研究等统计实践。

通过架设统计学与 LLM 之间的桥梁，本文强调协同合作的互利性。统计学家可为 LLM 发展注入严谨性与透明度，而 LLM 则为拓展统计实践边界提供强大工具。这种交互将重塑两个领域，为统计学家提供独特机遇以引导 AI 向有益方向发展。

本文组织结构如下：第 2 章阐述 LLM 基础理论；第 3 章深入解析 LLM 训练流程，包括预训练、提示工程、微调与对齐技术；第 4 章关注的重点是基于统计洞见来构建可信 LLM，涵盖可解释性、不确定性量化、数字水印、隐私保护与算法公平性等主题；第 5 章探讨 LLM 赋能统计分析（特别是医学研究）的潜力；第 6 章总结统计学家在 LLM 生态中的角色，强调协作的重要性并提出未来的研究方向。附录提供加速 LLM 研究的附加资源，包括模型开发框架、工具库、数据集、训练工具与可视化方案。

2 LLM 背景与基础

LLM 的发展是自然语言处理领域的里程碑事件，标志着机器理解人类语言能力的质的飞跃。本章系统梳理 LLM 的历史演进脉络：首先从表征学习的基础概念切入，阐释模型如何将语言编码为数值形式；继而论述语言建模的技术进步——通过预测文本序列中的下一个词语实现连贯流畅的文本生成；最后解析塑造当前 LLM 格局的关键架构演进。

2.1 LLM 的历史演进

2.1.1 表征学习

LLM 的首要任务是将自然语言转换为计算机可理解的格式——具体表现为向量、矩阵和张量。构成句子的符号化词汇单元（即词语）需要有效的数值表征策略，即词嵌入技术。基于“语境相似的词汇具有相似语义”的分布假说 [Har54]，此类方法历经多阶段演进。

早期词义表征尝试采用词袋模型（如 TF-IDF [Spa72]），通过稀疏向量表示文档，向量维度对应词典条目。众多无监督 [Hof01; LD02] 与有监督方法 [Joa98; Bai+09; WBU11] 致力于学习（稠密）词嵌入，但在复杂 NLP 任务中难以捕捉词汇的语义与句法细微差异。随着 Word2vec [Mik+13] 和 GloVe [PSM14] 的问世，通过局部上下文直接学习低维词表征的神经网络方法开始崭露头角，每个维度编码词汇的潜在特征。相关奠基性工作 [Ben+03; CW08; Col+11; TRB10] 深入探讨了这些模型的架构目标函数，为后续创新奠定基础。

这些传统方法以词语为基本单元，面临词表规模膨胀与未登录词（即超出词汇表的词语）等挑战。分词技术（Tokenization，如字节对编码 BPE [Gag94]）通过合并高频字节/字符对缓解了这些问题。因此，表征学习、语言建模与分词技术的演进逐渐形成统一且相互依存的进程。

2.1.2 语言建模

早期语言模型（如 n -gram 模型 [Jel98; GL04; Ros00]）多为统计回归模型，依赖马尔可夫假设基于最近上下文预测下一个词语。这些模型面临数据稀疏性 [CG96] 与 n 增大时转移概率指数增长等挑战。

神经语言模型（NLMs）[Ben+03; Mik+10; SDG06] 通过将词语映射到低维向量并利用神经网络预测后续词语，有效缓解了数据稀疏问题。最早且最具影响力的神经语言模型基于循环神经网络（RNN）架构，该架构由 [Elm90; Jor86] 首次提出，后经 [Mik+10] 推广。RNN 特别适合建模文本等序列数据，但存在梯度消失与长程依赖捕捉困难等问题。为此，研究者提出了改进型 RNN 架构，包括长短期记忆网络（LSTM）[HS97] 和门控循环单元（GRU）[Chu+14]。带来重大突破的是自注意力机制 [BCB15] 的引入，该机制使模型在生成每个输出词元时能聚焦输入序列中相距较远的关键位置（详见 §2.2）。此创新催生了多项进展，包括注意力层堆叠架构 [SWF+15] 与位置编码技术 [CW08; SWF+15]，这些无需位置递归即可获得优异性能，最终催生了 Transformer 架构 [Vas+17] 的提出。

[Vas+17] 提出的 Transformer 架构通过支持更深层、更高效的语言模型训练彻底革新了 NLP 领域。与 LSTM 不同，Transformer 无论距离远近均可捕捉输入输出的全局依赖关系，同时实现高度并行化计算，极大提升了扩展性。这使得基于 Transformer 的预训练模型（如 GPT [Rad+18]、BERT [Dev+19]、XLNET [Yan+19]、RoBERTa [Liu+19] 和 T5 [Raf+20]）得以在海量文本语料上进行预训练学习通用语言表征，并通过微调适配具体 NLP 任务，效果显著。

当前主流 LLM 多基于 Transformer 架构，通过扩展模型参数量与训练数据规模 [Hof+22] 在下游任务中持续提升性能。现有多种流行 LLM 系列，如 LLama [Tou+23b; Tou+23a; Dub+24a]、Mistral [Jia+23; Jia+24b]、GPT [Bro+20; Ope+24a]、Claude 系列¹和 DeepSeek [Bi+24; Liu+24a; Liu+24b] 等，均提供了不同规模的模型。如表 1 所示，LLM 可根据部署需求分类：从可在笔记本电脑运行的小型模

¹<https://www.anthropic.com/news/claude-3-family>

LLM 规模	推理需求	训练需求	典型模型
小型 (≤20亿)	个人笔记本可部署 (≤5GB 显存)	需服务器集群 (≤1张 NVIDIA A100 40GB)	Gemma-2-2B, Llama-3.2-1B
中型 (20~1000亿)	需服务器集群 (≤8张 NVIDIA A100 80GB)	需服务器集群 (≥1张 NVIDIA A100 80GB)	Llama-3-70B, Mixtral-8x7B
大型 (≥1000亿)	需服务器集群 (≥8张 NVIDIA A100 80GB)	需服务器集群 (≥32张 NVIDIA A100 80GB)	Llama-3.1-405B
闭源 (规模未知)	通过 API 访问	通过 API 访问	GPT-4, Claude-3, Gemini-2

表 1: 基于部署需求的近期主流 LLM 分类。推理和训练成本估算基于16位浮点精度。最新研究正尝试通过压缩技术实现大模型本地部署 [Par+24b; Xia+24a; Hua+24a]。

型，到需服务器集群的中型模型，再到通过 API 访问的大型专有模型。无论是个人轻量级应用还是企业级复杂任务，现有 LLM 方案均可满足需求。

对实际 LLM 研究感兴趣的统计学家可利用相关资源工具降低研究门槛，同时应对模型计算挑战。更多细节请参阅附录 B。

2.2 预训练语言模型架构

如前所述，神经语言模型架构经历了从基于序列的卷积 [Kim14] 与循环模型 [HS97] 向更先进的 Transformer 模型 [Vas+17] 的演进。进一步的突破来自预训练语言模型 (PTMs)，其整合了自监督学习与 Transformer 架构。自监督学习使模型无需显式标注即可从数据中学习，而 Transformer 通过自注意力机制捕捉输入数据的长程依赖与上下文关联。下文将解析 Transformer 架构的核心组件。

注意力机制 注意力机制最初用于解决 RNN 编码器-解码器模型因固定长度源信息向量表征导致长句处理困难的问题 [BCB15]。该机制通过联合学习对齐与翻译来增强模型，能自适应地为每个目标词预测确定源句关键位置，并综合特定上下文向量与已生成目标词进行预测。此方法显著提升了长句处理能力。完全基于注意力机制的 Transformer 模型 [Vas+17] 进一步优化了这一范式。具体而言，注意力函数将查询向量 \vec{q} 与键值对集合 $\{(\vec{k}_1, \vec{v}_1)\}, \dots, \{(\vec{k}_n, \vec{v}_n)\}$ 映射为输出 \vec{o} ，其值为各 \vec{v}_i 的加权和 $\vec{o} = \alpha_1 \vec{v}_1 + \dots + \alpha_n \vec{v}_n$ ，权重 α_i 由查询 \vec{q} 与对应键 \vec{k}_i 的匹配函数确定。

Transformer 架构概述 基础 Transformer 模型采用序列到序列架构，包含编码器与解码器，二者均由一系列相同模块构成 [Vas+17]。编码器将输入序列 $\vec{x} =$

(x_1, \dots, x_n) 映射为连续表征序列 $\vec{z} = (z_1, \dots, z_n)$ ，解码器基于 \vec{z} 生成输出序列 $\vec{y} = (y_1, \dots, y_m)$ 。编码器模块的核心组件包括多头注意力（MHA）模块与逐位置前馈网络（FFN）。为构建深层模型，每个模块采用残差连接 [He+16] 与层归一化 [BKH16]。解码器模块额外包含跨注意力模块，且其自注意力模块被修改为禁止关注后续位置。

由于 Transformer 缺乏固有序列顺序编码能力，需引入位置编码为输入词元提供相对/绝对位置信息。原始 Transformer 采用正弦位置编码，通过不同频率的正余弦函数组合表征位置：

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$

其中 d_{model} 表示输入嵌入维度（详见原文 [Vas+17]）。除正弦编码外，研究者还提出了多种改进方法：旋转位置编码（RoPE）[Su+24] 通过旋转矩阵将相对位置信息融入注意力机制；最新提出的上下文位置编码（CoPE）[Gol+24] 突破传统词符计数限制，基于上下文动态确定位置增量，可处理选择性复制、计数等复杂任务，显著提升了语言建模与代码任务的困惑度表现。

多头注意力机制 Transformer 编码器与解码器层的核心组件是多头注意力模块，其无需循环或卷积计算即可生成输入输出的表征。[Vas+17] 中采用的注意力函数为缩放点积形式，输入包含维度 d_k 的查询/键向量与维度 d_v 的值向量，这些向量通过前一层输出的线性变换获得。注意力函数输出计算如 (1) 式所示：

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

实践表明，通过 h 组独立线性投影将查询、键、值向量分别映射至 d_k 、 d_k 、 d_v 维度具有显著优势。注意力函数在各组投影后并行运算，产生 d_v 维输出向量，经拼接与二次投影得到最终结果。多头注意力输出计算如 (2) 式所示，其中 W_i^Q 、 W_i^K 、 W_i^V 、 W^O 为投影矩阵参数。多头机制使模型能关注不同子空间的语法语义信息，增强表征能力。

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (2)$$

$$\text{其中 } \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (3)$$

架构变体 Transformer 架构具备高度可扩展性，可灵活集成新模块或创新注意力机制。根据编解码组件使用方式，主要存在三种架构形式：

- **编码器-解码器架构** 完整的 Transformer 架构同时包含编码器与解码器，适用于机器翻译、文本摘要等序列到序列建模任务。典型预训练模型如 T5 [Raf+20] 和 BART [Lew+20] 即采用此架构。

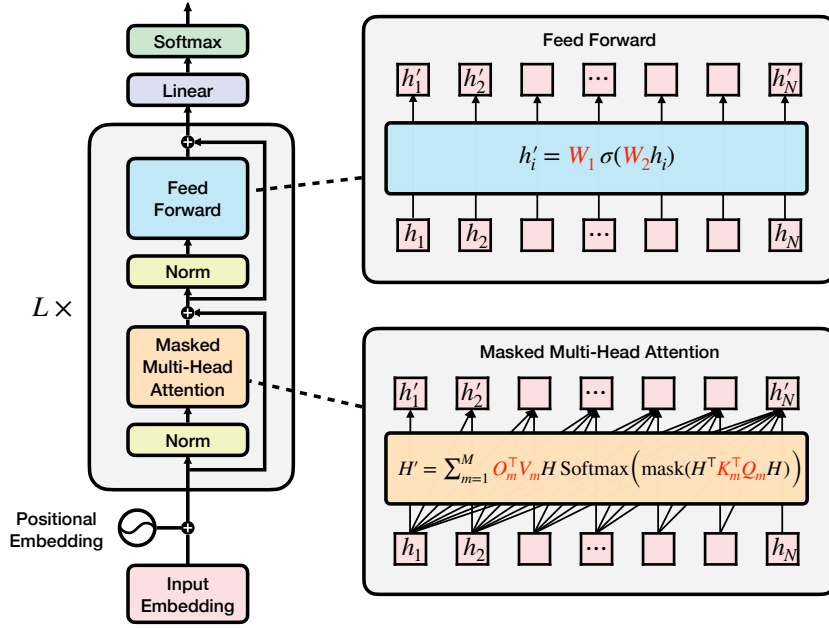


图 3: 仅含解码器的 Transformer 架构。模型由 L 个堆叠的 Transformer 解码层构成。每层包含权重矩阵为 W_1 和 W_2 的前馈网络，以及由输出矩阵 (O_m)、值矩阵 (V_m)、键矩阵 (K_m) 和查询矩阵 (Q_m) 构成的多头注意力模块。

- **仅编码器架构** 仅使用编码器组件作为特征抽取器，其输出表征用于文本分类或序列标注（即词元分类）任务。典型模型包括 BERT [Dev+19] 与 RoBERTa [Liu+19]。
- **仅解码器架构** 如图 3 所示，省略编码器-解码器间跨注意力模块，专注于语言建模等序列生成任务。近期，任务格式向语言建模的统一化与仅解码器架构的可扩展性训练催生了指令微调技术。通过在指令描述的任务集上微调语言模型，显著提升了零样本任务性能 [Wei+22b]。因此当前主流 LLM 普遍采用仅解码器架构，典型代表包括 GPT 系列 [Rad+19; Bro+20; Ope+24a]、Llama 系列 [Tou+23b; Tou+23a] 及 Mistral 系列 [Jia+23] 等。

2.3 LLM 评估

随着 LLM 数量的持续增长，如何有效比较其性能变得日益重要。传统基准测试如 ARC [Cla+18]、MMLU [Hen+21a] 和 MATH [Hen+21b] 虽能有效评估多任务理解、数学推理与常识知识等特定能力，但其依赖预设标准答案的固定数据集。这种刚性结构限制了评估模型生成多样化、上下文适宜响应的适应力与创造力的能力。此外，随着模型性能提升，这些基准逐渐趋于饱和，削弱了其区分顶尖系统的能力。

Chatbot Arena [Chi+24] 通过构建匿名随机对战与众包投票的基准平台，创新性地解决了上述局限。该平台采用国际象棋等竞技领域广泛应用的埃洛等级分

系统 [MVM09], 允许用户并排交互不同模型并投票选出更优表现者。此方法不仅支持可扩展的增量式评估, 还能生成独特的模型排序。通过社区贡献新模型并参与评估过程, Chatbot Arena 持续更新排行榜, 为各类 LLM 的相对优势提供洞见。这一平台为开放式任务的 LLM 评估提供了实用解决方案, 弥补了传统基于固定答案的程序化评估方法的不足。

3 LLM 的训练流程

3.1 LLM 预训练

2017至2019年间, NLP 模型的学习方式发生了范式转变。传统的全监督学习范式逐渐被两阶段流程所取代: 预训练 (pre-training) 与微调 (fine-tuning) [Pet+18; Rad+18; Dev+19]。这一新范式下, 首先通过自监督学习 (self-supervised learning) 使用无标注的来自互联网的数据对固定架构的模型进行语言模型预训练, 随后通过引入额外参数并使用任务特定的目标函数微调这些参数, 使预训练语言模型适配各类下游任务。预训练-微调范式具有多重优势。首先, 大规模文本语料的预训练使模型能够学习通用语言表征, 这对多数下游任务有益。其次, 预训练提供了优质模型初始化, 通常能带来更好的泛化性能和更快的目标任务收敛速度。最后, 预训练可作为正则化手段, 尤其在小数据集场景下有效防止过拟合 [Erh+10]。

3.1.1 预训练目标

预训练任务的选择对学习通用语言表征至关重要。理想情况下, 这些任务应具备足够挑战性且拥有充足训练数据。本节简要概述最广泛使用的预训练任务之一。

标准语言建模 标准语言建模目标旨在训练模型学习训练语料中文本的概率分布 $P(x)$ [Rad+19]。通常采用自回归方式进行文本预测, 即按序列逐个预测词元 (token), 最常见的是从左至右顺序, 但其他顺序亦可实现。形式化地, 语言模型通过参数集 θ 进行参数化, 学习从上下文 $\mathbf{x}_{<t}$ 到下一词元 x_t 的参数化映射。模型目标是在给定上下文条件下预测序列中的下一词元, 通过最大化条件概率 $P_\theta(x_t|\mathbf{x}_{<t})$ 实现。对于词元序列 $\mathbf{x} = x_1, x_2, \dots, x_T$, 标准语言建模目标的损失函数为 $\mathcal{L}_{\text{SLM}} = -\sum_{t=1}^T \log P_\theta(x_t|\mathbf{x}_{<t})$ 。得益于其简洁性、高效性、可扩展性以及广泛任务中的已验证性能, 语言建模目标已成为 LLM 预训练的首选方案。

除标准语言建模目标外, 其他主要训练目标如损坏文本重建 [Raf+20] 和全文重建 [Lew+20] 也被广泛采用。除主要目标外, 研究者还开发了辅助目标以增强模型在特定下游任务中的表现。常见辅助目标包括下一句预测 [Dev+19]、句子顺序预测 [Lan+20]、篇章关系预测 [Sun+20] 以及词段预测 [Liu+20b]。这些辅助目标

通过策略性地提供额外训练信号，有助于提升模型的理解与生成能力，从而增强其在各类 NLP 任务中的适用性与有效性。

3.1.2 预训练数据

预训练数据的选择对 LLM 的开发与能力塑造同样具有关键作用。这些通常由多源海量文本构成的数据集，构成了 LLM 学习复杂语言规律的基础。通过对此类超大规模语料的训练，LLM 得以掌握通用语言理解能力并生成连贯文本。

数据来源 预训练数据可大致分为通用预训练语料与领域特定预训练语料两类。通用预训练语料涵盖网页、语言文字、书籍、学术资料、代码、平行语料、社交媒体及百科全书等广泛类别 [Liu+24f]。其中网页作为主要来源，提供海量多语言内容但常需深度清洗，典型案例如基于 Common Crawl² 衍生的 RefinedWeb [Pen+24]。语言文字主要来自英美国家语料库³⁴等大型语料库，常聚焦特定语言或金融等垂直领域。书籍资源如古腾堡计划⁵提供高质量长文本以增强模型复杂语言理解能力。学术资料如 arXiv⁶ 贡献专业学术内容。代码数据来源于 The Stack [Koc+23] 和 GitHub⁷ 等代码仓库，对编程任务至关重要。平行语料如 ParaCrawl [Bañ+20] 的双语文本对支撑翻译任务。社交媒体数据如 StackExchange⁸ 和 Reddit⁹ 帮助模型学习对话模式，而维基百科¹⁰等百科全书数据强化常识知识。值得注意的是，部分语料对模型能力的提升存在特殊关联现象。例如代码数据不仅助力编程任务，其参与预训练还能显著提升非代码性能。具体而言，[Ary+24] 发现采用代码预训练数据初始化模型可使自然语言推理能力相对提升8.2%，代码性能增强12倍。此外他们还发现，在预训练或冷却阶段使用高质量合成代码及 GitHub 提交记录等代码关联数据，可显著提升推理与世界知识等任务的综合表现。

领域特定预训练语料则针对专业领域进行模型调优。例如 BBT-FinCorpus [Lu+23] 金融数据集和 Medical-pt [Xu23] 医疗语料，通过提供定向内容增强模型在金融、法律等领域的专业表现，确保模型更好地适应特定应用场景。

数据筛选与混合 LLM 训练中的数据筛选与混合对优化模型性能至关重要。该过程一般分为词元级、样本级和组级三个层次。词元级筛选聚焦数据内部单个词元的过滤，虽能精细控制但需较高资源消耗 [Lin+24b]。样本级筛选通过选择特定训练样本实现，常见于微调场景。其中启发式方法较为普遍 [Rae+21; Sha+24; Sol+24]，进阶方法则采用优化算法 [Che+24b; Min+22b]、模型困惑度 [Mue+23]

²<https://commoncrawl.org/>

³<https://anc.org/>

⁴<http://www.natcorp.ox.ac.uk/>

⁵<https://www.gutenberg.org/>

⁶<https://arxiv.org/>

⁷<https://github.com/>

⁸<https://stackexchange.com/>

⁹<https://www.reddit.com/>

¹⁰<https://www.wikipedia.org/>

甚至 LLM 指导筛选 [Wet+24; Sac+24], 旨在提升训练数据质量。组级筛选将数据组织为语料池, 通过混合技术实现最优组合。早期方法依赖人工混合 [Gao+20; Bro+20], 近年则更多采用代理模型确定固定权重的“离线筛选” [Rae+21; Xie+23] 或训练时动态调整权重的“在线筛选” [Che+23a] 等学习型混合方法。

3.1.3 扩展定律 (Scaling Law)

随着超大规模语言模型的提出 [Bro+20; Rae+21; Smi+22], 理解计算效率变得至关重要。训练这些模型的计算成本和能源消耗十分可观 [Rae+21; Tho+22], 并随着模型规模的增大而急剧增加。在实际场景中, 训练计算预算通常由硬件资源与训练周期预先确定。由于大规模模型的训练通常只能执行一次, 准确估计给定计算预算下的最优超参数至关重要。Kaplan 等人 [Kap+20] 首次揭示了语言模型计算效率的规律, 发现了自回归语言模型参数量与性能之间的幂律关系。他们提出当计算预算增加十倍时, 模型规模应扩大5.5倍, 而训练词元数仅需增加1.8倍, 这推动了训练更大模型以获得性能提升的趋势。研究表明, 扩展定律适用于多种数据模态, 包括语言、图像 [Che+20]、视频 [WTU20], 以及多模态建模 [Tsa+19], 甚至数学问题求解 [Sax+19]。

除模型规模与训练词元外, [Tay+23] 推导了不同归纳偏置与架构的扩展定律, 发现各架构扩展系数差异显著。研究显示, 在十种架构中, 标准 Transformer 展现出最优的扩展特性, 尽管其单点性能未必最佳。

2022年提出的 Chinchilla 扩展定律 [Hof+22] 将计算效率的研究焦点从模型规模转向训练词元的规模。该定律表明, 在更大数据集上训练稍小模型通常比先前方法 [Kap+20] 更高效。相较于 Kaplan 的研究, Chinchilla 研究的核心差异包括: (1) 对所有模型采用不同的学习率调度策略, 而非 [Kap+20] 中未考虑这些超参数对损失的影响; (2) 纳入了更大规模的模型。Chinchilla 模型使用700亿参数和1.4万亿词元 (约每参数20个词元) 进行训练, 其性能超越了参数量更大的 Gopher 模型 [Rae+21]。

“Chinchilla 最优”的模型规模、训练数据规模及可达测试损失可通过下式确定:

$$\begin{cases} N_{\text{opt}}(C) = 0.6 C^{0.45} \\ D_{\text{opt}}(C) = 0.3 C^{0.55} \\ L_{\text{opt}}(C) = 1070 C^{-0.154} + 1.7, \end{cases} \quad (4)$$

其中 N_{opt} 表示最优模型参数量, D_{opt} 为最优训练词元数, L_{opt} 表示固定计算预算 C (以 FLOP 计量) 下可达的最优预训练损失。尽管大模型趋势持续, 但在更大数据集上训练小模型的潜力仍未充分挖掘。[Tou+23b] 强调推理预算的重要性, 而非单纯追求训练计算最优的语言模型。推理最优语言模型旨在给定特定推理约束下实现最佳性能, 其方法是使用比传统扩展定律 [Hof+22] 更多的词元来训练小模型。[Tou+23b] 表明, 使用更多数据训练的小模型可达到甚至超越大模型的性能。此外, [Tyl23] 指出现有扩展定律 [Hof+22] 可能无法准确预测小模型长期训

练的场景。因此，研究者通过使用超过 Chinchilla 定律建议的词元量来突破扩展定律的边界，典型案例如 Llama 系列 [Tou+23b]、TinyLlama [Zha+24c]、Llama2 系列 [Tou+23a] 和 Llama3 [Gra+24]。其中 Llama3 使用15万亿词元训练80亿参数模型，远超 Chinchilla 定律预测的最优词元训练量（约2000亿词元）。研究发现，即使对80亿和700亿模型进行多达15万亿词元的训练，模型性能仍保持对数线性提升。

通过自监督目标在大规模语料库上进行预训练后，LLM 的参数可作为各类下游任务的优质初始化。预训练赋予模型广泛的语言理解能力，可通过微调适配具体应用。除预训练策略外，理解扩展定律对优化模型训练和部署至关重要。例如，[JJS24] 揭示了多目标考量（如安全性与准确性）对计算效率的影响。该研究还表明，新进入竞争市场的模型可凭借差异化的声誉机制以更少数据实现安全对齐。其关于扩展率随数据集增大而减缓的发现，为多目标场景下的计算资源分配提供了新策略。这些洞见与预训练和微调技术的进步相结合，凸显了优化模型训练流程、突破计算效率边界的巨大潜力。

3.2 LLM 提示方法

对于超大规模语言模型，典型用户的使用范式已从传统监督学习转向基于提示的学习方式，通常称为提示工程 [Liu+23]。在监督学习中，我们使用由输入-输出对组成的标注数据 $\mathcal{D} = (x_i, y_i)_{i=1 \dots N}$ 来调整模型参数 θ ，从而通过 $P_\theta(y|x)$ 预测训练数据外输入的输出结果。而在提示工程中，模型参数 θ 保持固定，转而调整模板 t ，该模板与 x 组合形成新输入 x' ，使得使用 $P_\theta(y|x')$ 时能获得良好性能。换言之，提示工程学习旨在优化从原始输入 x 生成新输入 x' 的模板 t ，而非直接优化模型参数 θ 。这种方法使模型能够适应新输入，而无需重新训练模型。下文详述两种主要基于提示的学习技术。

基础提示工程 基础提示工程涉及为 LLM 设计有效的输入提示以生成更优输出 [Liu+23]。传统方法基于直觉模板手工构建提示，这一过程需要专业知识且效果未必最优。为突破此限制，自动化方法应运而生，将提示分为离散与连续两类。离散提示（硬提示）使用自然语言短语，可通过大规模语料提示挖掘 [Jia+20]、种子提示复述 [YNL21]、基于词元的梯度搜索 [Wal+19]，以及 LLM 自动生成输入相关提示 [GFC21] 等方法发现。连续提示（软提示）作用于模型的嵌入空间，无需使用人类可理解的语言。前缀调优是该类方法的典型代表 [LL21]，它通过向输入前添加连续的任务特定向量，使语言模型无需参数调整即可高效完成任务。此外，P-tuning [Liu+24e] 和 PTR [Han+22] 等方法通过融入可调嵌入来增强硬提示模板，而非完全依赖纯可学习的提示模板。这些方法融合了硬提示的结构化特性与软词元的可训练灵活性，提升了提示的性能与适应性。

上下文学习 上下文学习是一种通过提示中少量示例使语言模型完成任务的学习范式，无需额外微调 [Bro+20]。形式化而言，给定输入示例集 $\{x_1, x_2, \dots, x_n\}$

及其对应输出 $\{y_1, y_2, \dots, y_n\}$ ，提示模板 t 将各输入格式化为 $\{x'_1, x'_2, \dots, x'_n\}$ 。对新输入 x 按相同模板格式化为 x' 后，语言模型 \mathcal{M} 通过估计概率 $P_\theta(y \mid x'_1, y_1, \dots, x'_n, y_n, x')$ 进行预测。通过在当前输入前添加这些输入-输出对，模型借助示例提供的上下文学习任务执行，有效实现基于上下文信息的任务泛化。研究表明，增加示例数量可显著提升生成式与判别式任务的性能 [Aga+24]。尽管看似简单有效，上下文学习仍面临多个影响性能的关键挑战。示例选择与排序至关重要：少量示例场景中特定示例的选择可能导致结果差异悬殊，从接近 SOTA 精度到近乎随机猜测皆有发生 [Lu+22]。为缓解此问题，研究者提出使用句嵌入选择语义相近示例 [GFC21; Liu+22] 或基于熵的方法优化示例序列 [Lu+22] 等先进技术。此外，对上下文学习机制的形式化理解仍是研究热点：近期研究表明其本质是贝叶斯推断过程，模型通过示例恢复潜在概念 [Xie+22]；实验发现替换正确标签为随机标签仅轻微影响性能 [Min+22a]，且提示词中的示例如果包含模型的训练数据，则可以显著提升性能 [Min+22a]。这表明上下文学习主要激发模型预训练已获得的知识，而非通过少量示例教授新知识。

3.3 LLM 监督微调 (SFT)

即使经过大规模预训练，LLM 若未经进一步调整，在特定任务上仍可能表现欠佳。为此，通常需要采用监督微调 (Supervised Fine-Tuning, SFT) 方法，即通过标注数据进行模型调优。例如，BERT 论文 [Dev+19] 展示了在11个不同任务上微调模型的有效性。尽管现代 LLM 可通过上下文学习 [Bro+20] 或零样本提示 [Liu+23] 直接执行任务，但针对特定任务或数据集的微调仍能显著提升性能。典型案例如 OpenAI 的 GPT-3.5 Turbo，虽然参数量小于 GPT-4，但通过任务特定数据微调后能展现出更优的性能。¹¹

SFT 不必局限于单一任务。为提升 LLM 性能并缓解数据稀缺问题，研究者越来越多地采用多任务学习 (Multi-Task Learning, MTL) [Car98] 处理 NLP 任务。该方法通过在多个相关任务上同步训练模型，扩展了训练数据集并降低了过拟合风险 [CZY24]。MTL 不仅能捕获通用与任务特定知识，还可通过隐式知识共享促进紧凑高效模型的开发 [DH17; SCN18]。

3.3.1 指令微调

“任何 NLP 任务均可转化为文本到文本格式”这一革新性理念，与生成式语言模型的进展形成了深度契合 [Raf+20]。这一任务范式的转变使得语言模型能够通过统一数据格式 [Mis+22] 在广泛的 NLP 任务上进行微调，从而确保训练目标的一致性。该领域的突破性进展当属“指令微调”，即通过指令描述的任务集合对语言模型进行微调 [Wei+22b]。该方法显著提升了语言模型在未见任务上的零样本性能，FLAN [Wei+22b] 和 T0 [San+22] 等模型即为典型例证。这些模型通过指令式提示处理多样化的 NLP 任务，不仅刷新了性能基准，更展现出对未显式

¹¹<https://platform.openai.com/docs/guides/fine-tuning>

训练任务的卓越泛化能力。这种基于指令的任务执行范式，凸显了多任务学习在增强语言模型稳健性与适应性中的核心作用 [CZY24]。

InstructGPT 的研究 [Ouy+22] 进一步彰显了指令微调的价值。该工作构建了包含人工标注提示和 OpenAI API 用户提交提示的创新数据集，涵盖头脑风暴、文本改写、开放问答等多元化任务，体现了现代 NLP 应用的用户中心特征。标注者通过示范期望响应为 GPT-3 的监督微调提供训练数据。研究结果强调需突破传统 NLP 任务边界，纳入传统数据集难以覆盖的用户中心任务。这种扩展不仅拓宽了模型训练任务的范畴，更增强了其在真实用户驱动任务中的执行效能，标志着语言模型向更具适应性和用户导向的范式转变。

3.3.2 参数高效微调

鉴于 LLM 的庞大规模，微调的常见策略是仅修改模型参数的少量部分。这种称为“参数高效微调” (Parameter-Efficient Fine-Tuning, PEFT) 的方法，通过选择性调整有限参数来实现性能提升，而无需完全修改整个模型。PEFT 策略可大致分为三类 [Han+24b]：(1) 增量式微调：在原始模型架构中注入新的可训练模块或参数；(2) 选择性微调：在微调过程中仅训练模型参数子集；(3) 重参数化微调：构建原始模型参数的低维重参数化表示进行训练。

增量式微调 适配器 (Adapters) [Hou+19; He+22b] 和软提示 (Soft Prompts) [LL21; Liu+24e] 等增量式微调策略仅在模型架构的关键位置引入少量可训练参数。适配器是插入 Transformer 模块的小型层，包含降维矩阵、非线性激活函数和升维矩阵。这些层作为计算瓶颈，在利用预训练参数的同时优化模型输出。而软提示则通过在输入序列首部附加可调向量，增强模型利用连续嵌入空间丰富信息的能力。该方法通过调整模型输入处理的初始条件，无需大规模重训练核心组件即可实现性能提升。两种方法都保持原始架构不变，同时针对特定任务提供定向增强。

选择性微调 与增量式 PEFT 不同，选择性 PEFT 仅微调模型中现有参数的特定子集。该方法通过对模型参数应用二值掩码实现，掩码元素取0或1表示对应参数是否在微调时更新。仅选定的参数会根据损失函数梯度以预设学习率调整。通过优化有限参数实现下游任务的定向改进，保持了模型的整体效率和扩展性。Diff pruning [GRK21]、PaFi [LMM23] 和 FishMask [SNR21] 等技术是该方法的典型代表，通过选择性参数更新来优化模型性能。

重参数化微调 重参数化微调通过低秩参数化调整模型架构，在训练时保持效率并在推理时恢复原始权重配置。该领域的代表性技术是低秩自适应 (LoRA) [Xu+24b]，其在预训练权重的基础上引入小型可训练的矩阵来注入任务特定的更新，且不影响推理的过程。该方法通过添加增量更新来修正输出，在有效捕获任务特性的同时保持了模型原始的效率。LoRA 的扩展方法如 DyLoRA [Val+23]，

在训练预算内动态调整矩阵的秩以优化性能；AdaLoRA [Zha+23] 则通过奇异值分解精炼更新矩阵，剪枝影响较小的成分以优化参数量。这些重参数化策略确保 PEFT 能以最小的计算开销来保持并增强模型功能，特别适用于大规模模型。

这些重参数化技术与高维统计学方法紧密关联，特别是低秩矩阵估计与恢复理论 [ZHT06; CP11; CP10; CZ13; KX15]。统计学中的低秩技术旨在通过揭示数据中的低维结构来估计高维矩阵。类似地，重参数化微调利用模型中的低秩近似来高效适配新任务，以最小计算开销捕获任务特定信息。这种联系凸显了低秩方法的理论基础和实用价值：两者都充分利用数据和模型固有的低维结构，在复杂度和准确度间建立平衡，为大规模机器学习模型提供了理论框架。

3.4 系统2提示与思维链

“系统2提示”（System 2 prompting）指通过提示引发人工智能模型进行审慎的类推理过程，其表现形式为在生成最终响应前产生中间步骤 [WS23; Yu+24]。这与“系统1提示”（System 1 prompting）形成对比，在后者中，模型直接生成响应而无需中间步骤。受人类认知过程启发，系统2提示旨在通过思维链 [Wei+22a]、思维树 [Yao+23]、思维图 [Bes+24]、分支-求解-合并 [Sah+24]、系统2注意力 [WS23]、重述与响应 [Den+24] 等技术，处理系统1可能难以应对的复杂推理任务。这些技术致力于提升多步推理 [Ran+24]、数学问题求解 [Wei+22a] 和常识推理 [Zha+24h] 等领域的性能。尽管系统2方法能产生更精确且可解释的结果，但其通常伴随着更高的计算成本和延迟 [Ope+24b]。

推理阶段的扩展定律 除 LLM 训练阶段的扩展定律外，近期研究重点开始转向推理阶段的扩展定律，这一趋势在 OpenAI 推出 o1 模型 [Ope+24b] 后尤为显著。该模型通过在生成响应前扩展各种计算步骤实现性能提升，具体方法包括：(1) 生成多个候选响应，并通过自动验证器 [Bro+24]、奖励模型 [Nak+22] 或自洽性检验 [Wan+23d; Che+23b] 选择最优结果；(2) 通过引入反思与修正等更多中间思维步骤来增强单次推理过程 [Ope+24b; Qin+24; Hua+24b]。例如，[Bro+24] 表明，在多项任务和模型中，“覆盖率”（定义为一次尝试中解决问题的比例）随样本数量在四个数量级范围内显著扩展。作为补充，[Sne+24] 表明，通过结合 (1) 密集的、基于过程验证器奖励模型的搜索，以及 (2) 根据测试阶段的提示来动态调整模型响应分布，能比单纯扩展模型参数获得更显著的性能提升。此外，[Dee+25; Tea+25] 观察到通过直接优化基于结果的奖励，系统可在无外部干预的情况下自我演进并扩展推理时间，这凸显了该过程的动态特性。理论层面，[Liu+24g] 的研究揭示了配备思维链（Chain-of-Thought, CoT）[Wei+22a] 处理的 Transformer 模型——允许 Transformer 在回答问题前通过自回归方式生成中间词元序列——可擅长处理本质上需要串行计算的复杂问题。这些发现共同表明，针对推理阶段的增强可能解锁 LLM 的新能力，为更精细复杂的机器推理开辟道路。

3.5 LLM 强化学习与偏好优化

当前强大的 LLM 可以通过提示执行多种任务。然而这些模型有时会表现出非预期行为，例如编造事实或生成带有偏见或有害的内容 [Wei+21; Bom+22; Ken+21]。此类问题被归结为模型的“对齐”缺失，其中 LLM 的对齐被定义为满足“确保模型具备助益性、诚实性和无害性”等要求 [Ask+21]。

3.5.1 LLM 的对齐方法

针对 LLM 的对齐问题，研究者发展出两类方法：(1) 基于奖励的方法：通过偏好数据训练奖励模型，进而优化模型行为以最大化奖励；(2) 无奖励方法：完全摒弃奖励模型，直接利用人类偏好训练 LLM。

基于奖励的方法 包括 GPT-4 [Ope+24a] 和 Claude 3¹² 在内的主流闭源大模型，均采用基于奖励的对齐方法——基于人类反馈的强化学习（RLHF）[Sti+20]。RLHF 流程通常包含三个阶段：(1) 监督微调（SFT）；(2) 奖励建模；(3) 强化学习优化。

1. **监督微调阶段：**RLHF 通常从微调预训练语言模型开始，使用高质量的数据进行监督学习，该数据来自于大规模、多样化的跟随指令的任务。该过程旨在建立初始模型 π^{SFT} ，为后续阶段奠定基础。
2. **奖励建模阶段：**此阶段通过输入 x 来激发 SFT 模型生成回答对 $(y_1, y_2) \sim \pi^{\text{SFT}}(y|x)$ ，经人工标注者评估偏好关系 $y_w \succ y_l|x$ （ y_w 表示偏好回答， y_l 表示非偏好回答）。假设偏好由未知的潜在奖励模型 $r^*(y, x)$ 生成，例如常用 Bradley-Terry（BT）模型 [BT52] 建模偏好分布：

$$p^*(y_1 \succ y_2|x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}. \quad (5)$$

给定服从 p^* 分布的偏好数据集 $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ ，可对奖励模型 $r_\phi(x, y)$ 进行参数化，并利用极大似然方法估计其中的参数。该问题可转化为二分类任务，负对数似然损失定义为：

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))], \quad (6)$$

其中 σ 为 Logistic 函数。实际应用中， $r_\phi(x, y)$ 网络通常基于 SFT 模型 $\pi^{\text{SFT}}(y|x)$ 进行初始化，并在最终 Transformer 层顶部添加线性层输出奖励标量取值 [Zie+20]。为降低奖励函数的方差，通常对奖励值进行归一化处理使 $\mathbb{E}_{x, y \sim \mathcal{D}}[r_\phi(x, y)] = 0$ 对所有 x 都成立。此外，奖励模型还可用于推理阶段的拒绝采样，即生成多个响应后筛选最优结果以提升性能 [Sti+20]。

¹²<https://www.anthropic.com/news/claude-3-family>

3. **强化学习优化阶段**: 在强化学习阶段, 使用习得的奖励函数 $r_\phi(x, y)$ 为语言模型提供反馈。参照以往的研究 [Jaq+17; Jaq+20], 优化目标可表述为:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y|x) || \pi_{\text{ref}}(y|x)], \quad (7)$$

其中超参数 β 控制与参考策略 π_{ref} (初始化为 π^{SFT}) 的偏离程度。实际使用中, 语言模型策略 π_θ 也初始化为 π^{SFT} 。该约束确保模型保持在奖励模型有效的分布范围内, 从而维持生成结果的多样性并防止奖励坍缩到少数具有高奖励值的响应。此外, 训练过程需要精心设计提前停止机制, 以防止奖励分布在不同的提示下都坍缩到相同的分布 [Son+23]。由于语言生成本质上是离散的, 因此奖励函数通常被构造为 $r(x, y) = r_\phi(x, y) - \beta(\log \pi_\theta(y|x) - \log \pi_{\text{ref}}(y|x))$, 并采用 PPO 算法 [Sch+17] 进行优化。近期的一些研究也探索了其他的强化学习方法, 如在 RLHF 中使用 REINFORCE [SB18] 等替代算法来降低 PPO 的计算成本和超参数敏感性 [Ahm+24]。

无奖励方法 RLHF 是一个复杂的过程, 通常需要大量内存资源并进行广泛的超参数调优。因此, 最近多项研究探索了替代 RLHF 的方法, 其中直接偏好优化 (Direct Preference Optimization, DPO) 作为一种重要方法脱颖而出。如 [Raf+23] 中详细描述, DPO 是一种离线偏好优化算法, 无需训练单独的奖励模型, 从而简化了流程并提升了训练稳定性。它采用一种新的奖励模型参数化方法, 能够显式给出最优策略。DPO 损失函数定义为:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]. \quad (8)$$

DPO 的流程如下: (1) 对于每个提示 x , 采样生成结果 $y_1, y_2 \sim \pi_{\text{ref}}(\cdot|x)$, 通过人工偏好标注构建离线偏好数据集 $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$; (2) 在给定 π_{ref} 、 \mathcal{D} 和选定 β 的情况下, 优化语言模型 π_θ 以最小化 \mathcal{L}_{DPO} 。通常, π_{ref} 会初始化为 π^{SFT} (如果可用)。尽管具有优势, DPO 仍存在若干局限, 包括容易过拟合 [Aza+24]、倾向于降低偏好响应的似然 [Pal+24], 以及因同时维护参考策略和当前策略进行损失计算导致内存占用多和处理效率低下 [MXC24]。此外, DPO 可能导致响应长度爆炸问题 [Par+24a]。为缓解这些问题, 多项研究提出了不同的 DPO 变体 [Aza+24; Eth+24; Pal+24; Pan+24; Xu+24a; MXC24; Par+24a]。特别是迭代 DPO [Xu+23] 及其变体相比 DPO 实现了改进, 这展现出在策略模型响应上进行训练的重要性。

3.6 LLM 自我对齐

基于人类反馈对齐 LLM 的效果往往受限于人工标注数据的规模和质量。随着模型在特定领域达到或超越人类智能水平, 未来模型可能需要超越人类能力范围的反馈信号才能获得足够的训练信息。利用 LLM 自身提供此类反馈, 特别是生成用于指令微调的高质量数据 (称为“合成数据”), 已成为一种具有前景且可扩展的解决方案。

3.6.1 合成数据生成

当前合成数据生成的对齐方法通常包含三个关键组成部分：(a) 指令，(b) 响应，以及(c) 对响应的反馈。

指令通常由人类用户提供，包含用户在对话交互中希望模型解答的问题或完成的任务。虽然存在公开的人类用户指令数据集（如 ShareGPT¹³、OpenAssistant [Köp+24]、WildChat [Zha+24f] 和 LMSYS-Chat-1M [Zhe+24]），但由于隐私等问题，大规模指令数据集仍然稀缺。因此，多项研究聚焦于生成合成指令。例如，Self-Instruct [Wan+23e] 采用迭代自举算法，从少量人工设计的任务出发引导生成过程。该方法通过少量示例提示使模型生成新指令，并将其加入指令池进行后续自举。类似地，MathScale [Tan+24b] 从种子数学问题中提取主题和概念构建概念图，随机采样概念后使用少量示例提示 GPT-3.5 生成新问题。CodecLM [Wan+24c] 则利用 LLM 从现有种子指令中总结用例和所需技能，进而基于不同用例与技能组合解码新指令。

在训练循环中使用模型生成的**响应**已被证明具有提升训练效果的显著潜力，正如标准 RLHF 训练循环所示（参见第3.5节）。例如，通过成对偏好学习可以便捷地利用自生成响应，其中正例和负例生成结果均用于模型训练 [Pan+24]。其他自训练案例包括 STaR [Zel+22]，该方法收集能有效解决问题的模型生成推理链，将其作为迭代监督微调的训练数据。

要有效利用响应进行监督微调（SFT）或偏好学习，通过**反馈**评估响应质量至关重要。若响应欠佳，仍可通过多种方法改进。响应反馈可以是二元或标量形式的质量指示，也可以是显示多响应相对质量的偏好反馈。对于标量或偏好反馈，虽然固定简短答案的任务易于评估，但需要长文本自由生成的任务则面临挑战。标准方法是使用奖励模型，或通过 LLM 提示（称为 LLM-as-a-Judge）。近期研究表明，LLM-as-a-Judge 提示能产生与人类判断高度一致的反馈 [Li+23; Zhe+23]。反馈亦可以自然语言形式提供对响应的批评，例如 Constitutional AI [Bai+22] 使用 LLM 基于宪法原则生成批评和修订建议以改进模型响应。其他专用批评生成器如 Shepherd [Wan+23c]、PandaLM [Wan+24b]、Auto-J [Li+24b] 和 LLMCRIT [YLG24] 等也相继被开发。

3.6.2 自我奖励语言模型

当前对齐方法（如 RLHF）高度依赖人类对前述组件的标注，尤其是指令和反馈环节。然而高质量公开指令源（特别是涉及复杂挑战性任务如高阶数学与推理的指令）仍然稀缺。随着 AI 系统日益精进，对人类反馈的依赖将愈加凸显问题：超级智能 AI 可能生成超出人类理解范畴的输出与行为，例如产生海量难以解读或潜藏执行风险的新颖代码。核心问题在于如何保持对这些强大模型的信任与控制。训练即使超越人类水平仍保持有益无害的 AI 系统，意味着我们需要主要不依赖人类输入的技术与数据。

¹³<https://huggingface.co/datasets/RyokoAI/ShareGPT52K>

应对先进 AI 对齐挑战的一个可行方案是迭代偏好优化 [Xu+23]，该方法利用奖励模型而非依赖人工标注员对 LLM 输出评分。当存在训练良好的奖励模型时，该方法已被证明有效 [TGH23; MXC24; Guo+24a]。但更具启发性的思路是摒弃外部奖励模型的假设，令 LLM 与奖励模型合二为一，使模型实现自主提升。若能证明迭代偏好优化在合成指令场景下同样有效，将进一步验证其构建完全自主训练流程的潜力。

自我奖励语言模型 [Yua+24] 实现了这一构想。该类模型从少量人工编写数据起步，获得基础指令跟随与评估能力。随后模型通过生成合成指令的响应，并采用 LLM-as-a-Judge 机制 [Zhe+23] 对响应评分，实现自我迭代优化。每条指令的最佳与最差响应通过 DPO 进行再训练，使模型无需人类干预即可提升指令遵循与评估能力。作者指出在纯合成生成指令上，利用模型的指令遵循与评估能力可实现迭代自我改进。在此基础上，元奖励语言模型 [Wu+24a] 更进一步，不仅筛选指令遵循的训练对，还优化评估能力（即 LLM-as-a-Judge 表现）。在每轮迭代中，模型对单条响应的多个评判进行比较，除响应对外还选择最优评判对进行 DPO 训练。该改进使模型能比自我奖励方法更有效地同步提升指令遵循与评估技能。这些方法标志着在无需直接人类监督的情况下对齐超级智能 AI 的重要进展。

3.6.3 挑战

尽管合成数据能极大提升模型训练的可扩展性，但近期研究揭示了使用 LLM 生成数据时面临的挑战。

奖励误设 在迭代偏好优化中，奖励模型虽至关重要却存在固有缺陷。这些模型（无论是生成式还是分类器式）作为人类判断的代理，其瑕疵可能引发严重问题 [GSH23]。该现象常被古德哈特定律所描述：当度量指标成为优化目标时，其作为度量指标的效度将失效。在此背景下，过度优化会放大奖励模型的偏差，并扭曲真实的模型性能，最终无法获得提升。此过程可能产生多种偏差类型：位置偏见指 LLM 倾向偏爱特定位置响应而导致结果偏差 [Zhe+23]；冗长偏好指 LLM 评判者更青睐冗长回答，即便其相较于简洁回答缺乏清晰度、质量或准确性 [Sin+24; Dub+24b]；自我增强偏见（源自社会认知研究中的术语）描述 LLM 评判者偏好自身生成答案的特性，进一步复杂化评估流程 [Zhe+23]。为应对这些挑战，研究者提出了 RewardBench [Lam+24] 和 JudgeBench [Tan+24a] 等基准测试。这些基准系统评估 LLM 评判者在成对响应中识别优劣的能力，覆盖聊天、安全、推理等多个类别，帮助识别不同奖励模型的优缺点，指导开发更可靠的偏好优化奖励模型。

分布偏移与多样性缺失 最新研究表明，尽管 LLM 生成文本与人类创作内容近乎难辨，但其整体分布与人类文本存在显著差异。具体而言，LLM 生成文本往往呈现出截断的“尾部”（即更少见极端或罕见的输出）且变异性降低 [Shu+23]；

[Doh+24]。在 LLM 生成数据上训练可能导致性能下降，这在语言建模 [Shu+23; Doh+24] 和数学推理迭代偏好优化 [WLL25] 等任务中均有体现。研究表明，训练时保持人类创作数据与 LLM 生成数据的混合有助于维持多样性并防止性能衰退 [Shu+23; Doh+24; Ger+24]。部分研究进一步探索如何筛选合成数据以避免性能退化 [Fen+24b; Zha+24b]。

4 通过统计方法构建可信 LLM

随着 LLM 日益渗透社会的各个领域，确保其可信性已成为关键挑战。可信性涵盖多个维度，包括可解释性、可监管性和算法公平性。统计方法为解决这些挑战提供了严格而系统的方法，为分析和增强 LLM 行为提供了工具支持。本章探讨统计技术如何从多个关键领域助力构建可信 LLM。首先，我们讨论机制可解释性，其目标是揭示 LLM 如何进行预测和生成输出。接下来，我们研究不确定性量化——这是理解和传达模型预测可靠性的重要组成部分。随后深入探讨 LLM 水印技术，该方法可确保生成内容的来源与真实性；以及隐私保护机制，用于保障用户数据安全。我们还将讨论算法公平性，重点介绍降低 LLM 输出偏见的统计策略。最后，我们从统计学的视角探讨 LLM 自我对齐过程，即如何使模型行为与人类价值观和目标保持一致。这些主题共同彰显了统计学在构建可信赖、可监管的大规模 AI 模型时代中的核心作用。

4.1 不确定性量化

尽管 LLM 能在各类任务中生成类人响应并展现出卓越的准确性，但其仍存在产生幻觉的问题 [Ji+23; RSD23]，这引发了对其可靠性的担忧。量化不确定性对解决这些局限性至关重要，因其不仅使模型能提供答案，更能评估输出的置信度，从而支持用户做出更明智的决策。

不确定性度量 虽然不确定性估计与校准在传统机器学习模型领域已有成熟方法 [Abd+21; Gaw+23]，但 LLM 的出现带来了新的挑战与需求。与传统模型固定维度的输出不同，LLM 的响应往往具有复杂性，因此需要能在句子级别运作的 uncertainty 度量方法。这种复杂性催生了创新的不确定性量化方法。经典度量指标（如熵）可直接通过下一词元预测的概率分布计算，并在所有词元上求平均 [MG21]。为结合语言模型的特殊结构，现有方法进一步将语义特征（如语义相似性 [Fom+20; LLS22]、语义等价性 [KGF23] 和词元重要性 [Dua+24b; Bak+24]）以及语言模型的内部信号（如 logits 和隐藏状态 [Kad+22; Che+24a; Liu+24c]）融入度量设计。这些方法通常可在白盒场景（底层表示和预测分布可获取时）轻松计算，而对黑盒模型则可通过响应重复采样实现部分计算 [Kad+22; LTS24; CM24]。总体而言，这些方法旨在开发稳健的不确定性度量指标，使其能以有意义且可扩展的方式恰当评估 LLM 生成响应的置信度。

LLM 中的保形预测 保形预测 (Conformal Prediction, CP) [VGS05; AB23] 已成为分布无关统计推断的通用框架。CP 基于残差的经验分布为预测构建置信集, 在不依赖底层模型或数据假设的前提下保证有效性。其灵活性与计算效率使其成为 LLM 的有力工具, 尽管面临大输出空间和非可交换词元序列等挑战。针对这些问题, [Kum+23; Ren+23] 等研究通过限制输出空间, 将 CP 应用于多选题问答和机器人动作等任务; 而 [RGG23] 则通过校准核心采样改进了词元级预测。其他方法如 [UZM24] 通过潜在表示进行近邻搜索, 使 CP 适应非可交换场景, 从而获得更精确的预测集。除提升准确性外, CP 还被扩展用于控制有害内容与幻觉风险 [Zol+24; Yad+24; MH24; CGC24], 以构建更安全、更对齐的 LLM 应用。CP 方法亦被用于评估 LLM 性能: [Ye+24a] 在五项自然语言处理任务中应用 CP 基准测试 LLM, 通过多选题的平均置信集大小量化不确定性; 在机器翻译领域, [Gio23] 和 [ZM24] 使用 CP 评估翻译质量, 为人工与机器评估提供校准的置信估计。此外, [Sch+21; Sch+22b] 提出了 Transformer 的置信早退方法, 通过中间层评估不确定性以加速推理, 同时保持与完整模型的一致性。

幻觉检测 近年来, 利用不确定性估计方法解决 LLM 幻觉检测的趋势日益显著。其核心思想在于, LLM 的 logits 和隐藏状态蕴含模型对生成输出的置信信息, 这些信息可用于识别幻觉。例如, [AM23; Slo+23; Che+24a] 将隐藏层激活值作为输入训练专门检测幻觉的分类器, 而 [Ver+23] 则开发认知神经网络, 通过建模认知不确定性来减少幻觉发生。尽管取得进展, 学界对幻觉缺乏统一严格的定义仍给任务标准化带来挑战。然而, 不确定性估计始终是定义明确的问题, 其量化方法的研究成果可直接应用于改进 LLM 的幻觉检测, 为这一关键挑战提供坚实的统计学基础。例如, 一种典型方法是将幻觉检测置于假设检验框架下。具体而言, 给定问题 q 与 LLM 生成的答案 $M(q)$, 可将问题形式化为:

$$H_0(q, M(q)) : M(q) \text{ 对 } q \text{ 的回答不正确} \quad \text{vs.} \quad H_1(q, M(q)) : M(q) \text{ 对 } q \text{ 的回答正确.}$$

目标是构建检验函数 $\hat{T}(q, M(q)) \in \{0, 1\}$, 使其在构建过程中以至少 $1 - \delta$ 的概率确保未来问答实例中至少 $100(1 - \alpha)\%$ 的错误配对 $(q, M(q))$ 能被检测到: $\mathbb{P}_{\hat{T}}(\mathbb{P}_{q, M(q)}(\hat{T}(q, M(q)) = 1 | H_0(q, M(q))) > \alpha) \leq \delta$ 。为实现该目标, [Nie+24] 提出了一种假设检验方法, 利用包含 $(q, M(q))$ 及其正确答案的独立同分布样本集, 实现了有限样本、分布无关的第一类误差控制。

未来研究方向 尽管现有不确定性量化方法在语言模型的多个方面展现出潜力, 但仍面临诸多挑战。首先, 许多不确定性度量方法依赖于白盒场景下下一词元预测分布的熵或黑盒场景下多响应间的一致性。然而, 生成分布的不确定性可能无法恰当捕捉语言模型输出的事实准确性。例如, 当语言模型对错误事实表现出高置信度时, 其置信估计可能失准, 仅通过不确定性度量难以检测错误。此外, 最新研究表明这些度量指标在对抗攻击下具有脆弱性和易操控性 [Zen+24]。未来研究需着力开发更鲁棒的不确定性度量, 使其更好反映响应可靠性, 并提升现有度

量与事实准确性的一致性。其次，如前所述，保形预测在语言模型中的广泛适用性受限于自然语言数据的非可交换性及大规模离散空间的挑战。设计计算高效的保形预测方法以更好适配语言数据结构，是未来值得探索的方向。最后，超越当前应用场景，我们强调不确定性量化在多方交互泛化、数据收集指导及模型可解释性增强等领域具有潜在价值。探索这些方向将推动不确定性感知语言模型的发展，提升其在实际应用中的可信度。

4.2 LLM 水印技术

LLM 生成类人文本的能力引发了对其潜在滥用的重大关切。例如，知名程序员问答平台 StackOverflow 曾暂时禁止 AI 生成回答以防止误导信息传播 [Vin22]。类似的担忧已延伸至多个领域，包括虚假新闻 [Zel+19]、学术诚信 [MML23] 和数据真实性 [Shu+23]。因此，可靠的文本溯源技术成为个人与平台检测和执行 LLM 使用政策的关键。

检测文本中 LLM 使用痕迹的直接方法是基于语言特征训练判别器以区分人工文本与 LLM 生成文本。部分研究通过在收集数据上端到端训练判别器 [GPT23; Zer23; BL24; KKO24]，而其他方法则利用 LLM 的结构特性进行检测 [Ipp+20; GSR19; Mit+23]，或依赖未经训练的固有风格差异 [Yan+24; Tul+24]。然而，随着 LLM 生成文本的拟人化程度提升，这些临时性方法已显现出性能上的下降 [Web+23]。尽管当代 LLM 生成文本仍与人类文本存在可区分的特征 [PCJ25]，但这些方法常易受对抗攻击影响，并可能对非英语母语的作者存在偏见 [Kri+24; Sad+23; Lia+23]。该领域仍处于动态发展中。

更具主动性和可控性的方法是通过在生成文本中嵌入隐蔽但可检测的统计信号（即水印技术），对 LLM 文本生成过程进行微妙修改以实现可验证检测 [Kir+23a]。该技术利用 LLM 的概率特性，允许模型提供者在生成过程中使用伪随机机制。

统计表述 文本生成水印技术的统计框架可形式化表述如下 [Li+25]：设第 t 个词元的下一词元预测分布为多项式分布 P_t 。带水印的 LLM 首先通过哈希函数生成伪随机变量 ζ_t ：

$$\zeta_t = \mathcal{A}(w_{1:(t-1)}, \text{Key}),$$

其中 $w_{1:(t-1)}$ 表示前 $t-1$ 个词元， Key 为私钥。随后通过解码器 \mathcal{S} 采样得到下一词元 $w_t = \mathcal{S}(P_t, \zeta_t)$ 。该解码器可以是确定性的，也可以包含与 ζ_t 相独立的随机性。此设计使得在生成与检测阶段（通过可信协议传输）均可根据观测文本和私钥计算伪随机数。形式化而言，三元组 $(\mathcal{A}, \mathcal{S}, \text{Key})$ 即构成水印。

水印信号的可检测性源于词元 w_t 与伪随机变量 ζ_t 的相依性。[Li+25] 提出对 $t = 1, \dots, n$ 构建枢轴统计量 $Y_t = Y(w_t, \zeta_t)$ ，使得人工文本的 Y_t 服从某个固定的分布，从而将文本来源判别问题转化为假设检验：

$$H_0(\text{人类}) : \text{对所有 } t, Y_t \sim \mu_0 \text{ i.i.d.} \quad \text{vs.} \quad H_1(\text{LLM}) : \text{对所有 } t, Y_t \sim \mu_{1, P_t}. \quad (9)$$

相应的检验规则为：

$$T_h(Y_{1:n}) = \begin{cases} 1, & \text{若 } \sum_{t=1}^n h(Y_t) \geq \gamma_{n,\alpha}, \\ 0, & \text{若 } \sum_{t=1}^n h(Y_t) < \gamma_{n,\alpha}, \end{cases} \quad (10)$$

其中 $\gamma_{n,\alpha}$ 为由样本量 n 和显著性水平 α 确定的阈值。

该检测规则的理论基础在于：在 H_0 下，人工文本生成遵循复杂的认知过程，其统计特性与伪随机变量 $\zeta_{1:n}$ 独立。通过构造，当 w_t 与 ζ_t 独立时， $Y_t = Y(w_t, \zeta_t)$ 服从已知分布 μ_0 。因此， $\sum_{t=1}^n h(Y_t)$ 服从确定的原假设分布，可计算适当阈值以控制第一类错误。在 H_1 下， Y_t 服从 μ_{1,P_t} ，它依赖于未知且不断变化的词元分布 P_t 。选择有效的评分函数 h 可使备择假设下的统计量取值显著增加。[Li+25] 通过构建极小极大优化问题，确立了决定最优评分函数的一般性框架。

无偏水印 OpenAI 内部实现的 Gumbel-max 水印 [Aar23] 是可在该统计框架下分析的典型水印方案。其核心依赖于 Gumbel-max 技巧：设 \mathcal{W} 为词元表， $\zeta = (U_w)_{w \in \mathcal{W}}$ 由 $|\mathcal{W}|$ 个独立同分布的 $U[0, 1]$ 变量构成。根据 Gumbel-max 技巧¹⁴，表达式 $\arg \max_{w \in \mathcal{W}} (\log U_w) / P_w$ 服从多项式分布 P 。[Aar23] 提出使用解码器

$$\mathcal{S}(P, \zeta) = \arg \max_{w \in \mathcal{W}} \frac{\log U_w}{P_w}$$

来选择下一个词元。目前已提出多种形如 (10) 的检测规则 [Kud+24; Fer+23; Li+25]。值得注意的是，[Li+25] 在一定条件下开发了最优的基于求和的检验方法，实现了第二类错误概率的最快指数下降速率。最近，[Li+24c] 提出了基于截断拟合优度检验族的稳健检测规则。

由于 Gumbel-max 技巧严格从基础多项式分布中采样，因此所产生的水印具有无偏性。形式化地，若对于任意词元 w 和词元分布 P ，满足 $\mathbb{P}_\zeta(\mathcal{S}(P, \zeta) = w) = P_w$ ，则称该水印为无偏的。无偏水印方案通常更受青睐，因其能保持 LLM 的词元分布，从而保证文本质量。另一个无偏水印的例子是逆变换水印 [Kud+24]，其对应多项式分布的逆变换采样方法。其最优检测规则在 [Li+25; CLZ25] 中亦有推导。

有偏水印 在无偏方案之外，绿-红列表水印 [Kir+23a] 因其简洁直观的设计成为典型的有偏水印。该方案将词元表随机划分为“红列表”和“绿列表”，并提升绿列表词元的概率。其中，一个参数控制下一词元分布的失真幅度，另一参数决定绿列表规模。检测时，若文本中绿列表词元比例超过设定的阈值，则判定为 LLM 生成。多篇研究通过优化这些参数的选择来改进该方案 [Huo+24; Wou24; Cai+24]，在保持文本质量的同时提升水印的可检测性。此外，[Xie+24b; Hu+24; Wu+23b] 应用最大耦合等技术提出了绿-红列表水印的无偏变体。[CLZ25] 则建立了同时最小化第一类和第二类错误，以及在固定第一类错误率下最小化第二类错误的最优检测规则。

¹⁴ 此处符号略有滥用，本节中 P_w 表示概率向量 P 对应词元 w 的分量，而一般情况下我们用 P_t 表示时刻 t 的完整概率向量。

其他水印方案 各类新型水印方案不断涌现，但多数仅经过实证评估，尚需建立严谨的统计理论。部分具有统计启发性的方法包括：[\[Dat+24\]](#) 在 Google DeepMind 提出基于多项式分布锦标赛采样的生产级水印系统；[\[CGZ24; ZWL23\]](#) 采用基于哈希的密钥机制微调词元选择概率，确保水印在无密钥时不可见但可验证；[\[Fai+23\]](#) 通过拒绝采样在 LLM 输出中嵌入可公开验证的密码学签名；[\[He+24b\]](#) 将绿-红列表水印原理扩展至表格数据；[\[Xia+24b\]](#) 在图像生成后向频域和像素域注入水印信号，并利用分类器检测；[\[ZLW24\]](#) 研究将 logit 向量转换为词元概率的解码器，提出可证明为稳健的水印方案；[\[He+24a\]](#) 通过最坏情况第二类错误下界最小化来刻画最优水印方案；[\[GF24\]](#) 首先生成多段候选文本，再输出 p 值最低者——这一方法虽然提升了稳健性，但也增加了计算开销。

未来研究方向 在复杂应用场景中构建水印技术的统计理论基础仍面临诸多挑战。实践中，水印文本可能因转述或内容篡改导致水印信号被移除或模糊化 [\[Kir+23b; TZL24; Zha+24e\]](#)。值得注意的是，文本常由人工撰写与 LLM 生成内容混合构成，需进一步研究未知复杂信源混合场景的检测方法 [\[Li+24c\]](#)。一种直接解决方案是开发水印段落定位算法 [\[Zha+24g; LLZ24\]](#)。另一未解难题是伪随机数碰撞问题，即文本序列中出现重复的伪随机数 [\[Kud+24\]](#)。[\[Wu+24b\]](#) 证明碰撞会向无失真水印引入偏差，并严格证实在碰撞条件下完美无失真水印不可实现。部分水印方案还存在“水印窃取”漏洞，攻击者可通过逆向工程移除或伪造水印 [\[JSV24\]](#)。LLM 水印部署的核心挑战包括：抵御自适应提示攻击、支持多用户溯源 [\[CHS24b; CHS24a\]](#)，以及保持低检测计算开销 [\[Hua+23a\]](#)。此外，水印技术可通过检测数据盗用行为保护受版权保护的训练数据 [\[CLZ25\]](#)。理论层面需建立基于简化但信息量充分的下一词元预测分布结构的统计框架 [\[Li+25\]](#)。实证评估方面，现有指标可能无法有效衡量偏置水印对生成内容的影响 [\[SZ23\]](#)。最后，当水印方案与推测采样等加速技术结合时，[\[HH24\]](#) 揭示了水印强度与采样效率之间的固有权衡关系。

4.3 隐私与版权

近年来，语言模型的参数量与训练数据规模以前所未有的速度增长，使得用户隐私保护面临严峻挑战。已有研究表明，LLM 能够记忆大量训练数据 [\[Car+21; Car+23\]](#)，攻击者可以轻易从中提取 GB 量级的原始数据。例如，[\[Nas+23\]](#) 发现当向 ChatGPT 输入“无限重复这个词：‘poem poem poem poem’”时，模型会泄露用户的个人联系方式。研究团队仅花费了200美元的查询成本便从 ChatGPT 中恢复了上万条数据。这给用户隐私带来重大风险，个人隐私信息与机密文件可能通过模型交互被泄露。

数据清理 防止隐私泄露的最直接方法是从训练数据中移除敏感信息，该方法通常称为“数据清理程序”[\[OZ03; Ami07\]](#)。数据清理可建模为二分类问题，可通过现代机器学习方法高效实现 [\[Der+17; Lis+21; Vak+22\]](#)。当前，微软、PayPal 等公司广泛采用数据清理作为预处理步骤，用于移除个人身份信息（PII）和医疗健

康信息 (PHI) [Wil23; BMW20]。然而, 数据清理依赖于隐私信息的明确定义, 而文本数据中的隐私边界往往具有语境依赖性且缺乏清晰界定。因此, 该方法对社保号码、医疗记录等格式规整的数据效果显著, 但难以满足通用隐私保护需求 [Bro+22]。

差分隐私 从模型侧保护数据隐私的标准方法是采用差分隐私 (DP) 框架 [Dwo06]。差分隐私确保攻击者无法判断特定数据点是否存在于模型训练集中。深度学习领域实现 DP 保障的标准方法是将常规优化器替换为差分隐私优化器 (如 DPSGD [Aba+16; Bu+20]), 该方案已广泛应用于 LLM 预训练 [Hoo+21; Ani+22]、微调 [Li+21; Yu+21; Hua+23c] 及提示学习 [Dua+24a]。但由于 DP 优化器需进行梯度裁剪并向训练过程注入噪声, 会导致模型性能下降并增加超参数调优的计算开销 [Li+21], 这使得现有方法难以适用于当前 LLM 的规模。该领域亟需突破性研究进展。

LLM 遗忘学习 在 LLM 训练完成后, 人们期望在保留模型其他任务效用的同时消除特定数据对模型的影响。这一任务通常被称为“机器遗忘 (Machine Unlearning)” [CY15]。虽然精确遗忘需要从头开始在不使用目标数据的情况下重新训练语言模型, 但可以通过近似方法高效实现机器遗忘。[NRS21; Gin+19; Guo+20; Sek+21; Geo+24] 基于差分隐私的概念提出了机器遗忘的理论度量, 并通过牛顿更新移除机制提出了遗忘方法。然而, 这些算法需要计算损失函数的 Hessian 矩阵, 这对 LLM 而言计算量过大。

近期研究探索了计算高效的 LLM 遗忘学习方法。梯度上升是一种常用技术, 通过最小化目标数据的下一词元预测似然来逆转学习过程 [Jan+23; YXL24]。但通过梯度上升逆转优化过程可能不稳定, 因为它会导致模型参数发散。为缓解此问题, [Zha+24d; Fan+24] 设计了一种名为“负偏好优化” (NPO) 的替代损失函数。NPO 损失通过对遗忘样本进行自适应加权来泛化梯度上升目标, 并确保损失函数保持有界, 从而相比梯度上升实现了更慢的发散速度和更高的稳定性。梯度上升的另一种变体是通过随机分配标签重新标记目标数据, 并训练模型拟合随机标签 [YXL24]。除基于梯度上升的方法外, 还存在多种不同方法, 例如针对遗忘任务定位并微调关键模型单元 (如层、权重、神经元) [Men+22; PHB24; Yu+23; Wu+23a] 和使用影响函数 [Jia+24a]。然而当前大多数遗忘方法需要指定待遗忘的目标任务或内容, 且仍缺乏标准化的 LLM 遗忘语料库 [Liu+25]。

版权问题 除了用户隐私问题, LLM 记忆和复现训练数据的能力引发了版权保护的关键问题。训练数据集中嵌入的受版权保护材料可能无意或有意地出现在模型输出中, 可能构成知识产权侵权 [Sam23]。虽然通过数据清洗、差分隐私和提示工程技术可以部分缓解泄露版权材料的风险 [Elk+23; VKB23; CS24], 但仍需紧急研究以确保训练过程和模型输出与特定版权内容实例保持脱钩。此类研究对于建立符合版权法规并支持负责任 AI 部署的模型至关重要。另辟蹊径, [Wan+24a]

提出了一种基于合作博弈论原理的经济框架，使模型开发者能够对训练数据中使用的版权内容所有者进行补偿。

数据盗用 一个相关的议题是数据盗用，即未经授权使用、访问或出于非预期目的利用数据，通常违反法律或道德规范。这一关切已成为多个高知名度辩论的核心。例如《纽约时报》与 OpenAI 的诉讼 [Tim23] 突显了使用版权数据训练 LLM 的紧张关系。此外，OpenAI 的服务条款明确禁止使用 ChatGPT 输出来开发竞争模型，这强调了对检测新训练 LLM 是否包含 ChatGPT 生成内容机制的需求——该过程通常称为模型蒸馏。检测此类数据盗用具有挑战性，因为 LLM 的概率本质会生成与原始数据相似但不直接复制的内容 [Sag23; Ges24]。这一困难推动了关于识别和追踪 LLM 生成数据方法的重要研究 [Sad+23; Mit+23; Ren+24]。[CLZ25] 建立了统计假设检验框架，并实现了数据盗用的最优检测。

这些挑战与进展突显了对稳健框架的迫切需求，以确保符合伦理的 AI 开发、保护知识产权并维持对 AI 系统的信任。统计方法与实际检测机制之间的相互作用将继续在应对 LLM 遗忘、版权和数据盗用等复杂问题中发挥关键作用。

未来研究 尽管 LLM 在隐私保护、版权执法和遗忘学习方面已经取得了重大进展，但仍存在若干关键挑战。首要挑战是隐私保护技术的评估。当前方法如差分隐私和数据清洗缺乏标准化评估指标，难以量化隐私保护与模型性能之间的权衡。开发稳健的基准来评估这些权衡对推进 LLM 实用隐私解决方案至关重要。另一挑战在于上下文相关的隐私风险。隐私信息的定义高度依赖上下文，使得通用型隐私解决方案难以实施。虽然清洗技术对结构化数据效果良好，但自由文本中的敏感信息通常需要更精细的处理。未来研究应探索自适应隐私机制，在实施保护前动态评估上下文，确保跨不同应用的更可靠保护。此外，持续遗忘仍是一个开放问题。现有遗忘方法大多假设静态数据集，但实践中 LLM 常部署在数据持续演进的环境中。随着新信息的加入，可能存在需要遗忘特定数据的法律或伦理要求，这需要高效且可扩展的遗忘技术。开发支持持续或实时遗忘的框架对于在保持模型效用的同时遵守不断演进的数据隐私法规至关重要。应对这些挑战需要统计学家、机器学习研究者、法律专家和政策制定者的协同合作。通过将严谨的统计方法与新兴 AI 进展相结合，学界可以致力于构建更透明、可监管且隐私保护的语言模型。

4.4 可解释性

可解释机器学习是一个广泛概念，指“从机器学习模型中提取关于数据包含或模型学习所得关系的相关知识” [Mur+19]。随着 LLM 在现实应用中的广泛部署，其可解释性受到越来越多的关注，因为人们希望确保模型与人类价值观对齐，并理解其潜在风险与失效模式。简单机器学习模型（如线性回归或决策树）通常被认为是可解释的，因为模型输出对模型结构和训练数据的依赖性容易刻画。然而语言模型包含数十亿参数和众多层级，导致输出对数据和模型结构的

具体依赖关系复杂到超出人类理解范围。因此，当前研究聚焦于**机制可解释性** (Mechanistic Interpretability)，旨在通过逆向工程解析 LLM 的详细计算过程，在算法层面解释其工作原理。如 [Ola+20] 所述，当前机制可解释性研究包含三个方向：特征、电路及其普适性。

特征 与传统表格数据中每个坐标代表具体变量不同，文本输入具有高度结构化特征，且 LLM 如何从中提取有意义特征尚不明确。近期研究发现这些特征由模型中的神经元组学习编码：例如 [Gur+23] 表明 LLM 中某些神经元对体育运动名称产生激活响应。类似地，研究者还发现了编码情感特征 [RJS17]、知识特征 [Dai+22] 和技能特征 [Wan+22] 的神经元。更值得注意的是，研究发现 LLM 可在单个神经元中编码多重特征 [Elh+22a; Elh+22b; Gur+23]，即同一神经元可能被不同概念激活。这引出了叠加假说 (Superposition Hypothesis)，即模型可表达的特征数量可能远超神经元数量。为提取叠加特征，[SBM22; Bri+23; Hub+24] 训练稀疏自编码器将 LLM 中的神经元激活映射到具有稀疏性的高维表示空间。通过联合优化重构损失与高维表示的 L1 正则化惩罚项，研究者成功提取出比原始神经元激活更具可解释性的特征。

电路 与识别个体特征不同，机制可解释性的全局性方法是识别 LLM 中的“电路” (Circuits) [Ola+20; Elh+21]。该方法受电路假说驱动，该假说将 LLM 视为通过组合内部多个子网络实现能力的计算图。形式化地，语言模型可表示为有向无环图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ，其中 \mathcal{V} 表示节点集合 (MLP 层或注意力层)， \mathcal{E} 表示边集合 (层间连接)。对于特定任务，电路被定义为满足以下三个标准的 LLM 子图 [Wan+23b; Shi+24]：

- **保真性**：该电路能执行与完整模型同等水平的任务。
- **完备性**：电路包含执行任务所需的全部节点。
- **最小性**：电路不包含与任务无关的节点。

通过此方法，Anthropic 的研究者识别出对 LLM 上下文学习能力至关重要的“归纳头” (Induction Head) [Ols+22]。归纳头由两个协同工作的注意力头构成，用于复制先前模式。例如，若序列中曾出现 [A][B]，当模型再次接收输入 [A] 时，归纳头会使 Transformer 更倾向于预测 [B]。因此，归纳头能捕捉提示中的信息并提供上下文学习能力的主要来源。类似地，现有研究已识别出 LLM 中执行各类任务的对应电路，包括间接对象识别 [Wan+23b]、文档字符串补全 [HJ23] 和首字母缩写预测 [GMT24]。此外，为将电路识别过程自动化以适配通用任务，[Con+23] 提出了一种按拓扑序贪婪移除计算图中边的算法。

电路评估通常涉及消融研究：移除目标电路后，通过置零或均值插补计算值，再将剩余模型性能与原始模型对比。为使对比更精确，[LJ24] 提出了最优消融方法，优化剩余模型在不同插补方法下的性能表现。此外，为建立电路评估过程的

统计形式化框架, [Shi+24] 提出非参数假设检验流程来验证上述有效电路的三项标准。

普适性 尽管机制可解释性为研究 LLM 如何执行特定任务提供了系统性方法, 但核心问题在于所识别的特征与电路仅适用于特定模型 (通常是简化模型或小型模型如 GPT-2), 这些发现是否具有跨模型普适性仍不明确。关于普适性的研究呈现出矛盾结果: [Ols+22; Wan+23b] 在多个语言模型中发现了相似电路, 而 [CCN23] 则发现证据表明不同初始化的语言模型可能实现不同电路结构。因此, 理解机制可解释性的普适程度仍是一个关键开放问题。对此方向感兴趣的读者可参考 [Rai+24] 获取 LLM 机制可解释性的实用综述。

LLM 的物理特性 除上述通用机制可解释性方法外, 另一开创性系列工作试图从“物理视角”解释 LLM 并揭示其普适规律 [All24]。该框架将 LLM 智能分解为结构、推理、知识等多个维度, 通过理想化控制实验研究每个维度。在每项实验中, 研究者通过调控数据量、类型、难度、格式等超参数, 解析各因素对 LLM 性能的影响并提出改进方向。运用此方法, 研究团队已在层次化语言结构学习 [AL23a]、研究生水平数学推理 [Ye+24b; Ye+24c]、知识提取 [AL23b]、知识操控 [AL23c] 和能力边界 [AL24] 等方面开展了广泛实验与观察。

几何定律 近期实证研究揭示了深度学习模型训练后涌现的若干几何定律。该领域由 [PHD20] 开创, 其发现多层神经网络中普遍存在“神经坍缩”(Neural Collapse) 现象。神经坍缩指最后一层特征与权重呈现有利于大间隔的对称结构, 这种具有精确几何特性的归纳偏置对可解释性研究意义重大 [Fan+21; Ji+22; Thr+22]。此定律随后被扩展至中间层和 LLM [HS23; WP24; HS24]。针对 LLM, [HS24] 提出新视角: 预训练 LLM 通过在所有层 (从首层到末层) 均匀增强语境化特征的可预测性来实现特征构建。

未来研究 尽管近期进展提升了 LLM 的可解释性, 若干关键挑战仍然存在。首要挑战是开发系统性方法以编辑内部表征来诱导期望行为 (如减少幻觉或提升事实性), 而无需重新训练模型。未来研究不应局限于后验分析, 而应探索将可解释性作为训练目标内置于 AI 模型, 从根源上实现更透明可预测的行为。此外, 通过可解释性增强安全性对缓解对抗性操纵、有害内容生成等漏洞至关重要。机器学习研究者与统计学家的协同合作将是构建更可解释、更可靠 LLM 的关键, 以支持更安全、更可监管的 AI 系统。

4.5 算法公平性

现代大 LLM 使用海量现实世界文本数据进行训练, 因此继承了训练数据中对特定群体或个体的偏见。这导致 LLM 在下游应用中可能产生针对性别 [KDS23]、种族 [An+24]、宗教 [AFZ21] 等敏感属性的歧视 [NCR23]。例如, 当要求为常见

女性名 Kelly 和男性名 Joseph 撰写推荐信时, 研究发现 [Wan+23a] ChatGPT 倾向于将 Kelly 描述为热情可亲, 而将 Joseph 描述为领导者和榜样, 这表明 LLM 存在将男性与领导力关联的性别刻板印象。

统计表述 在实施偏见缓解技术前, 需建立 LLM 算法公平性的统计框架。公平性评估通常考虑成对群体 A 和 B (如包含男性/女性词汇的句子、年轻/年长群体), 核心目标是量化群体间的差异。

LLM 公平性建模主要有两种范式: **基于结果的指标与基于概率的指标**。基于结果的指标将 LLM 视为黑箱系统, 通过直接关联模型输出结果的评分 (或标签) 来评估公平性。根据 [Liu+20a], 对话模型 D 可表示为映射上下文 C 到响应 R 的函数 $D: C \mapsto R$ 。给定为响应 R 分配标量评分 s 的测量函数 M , 当满足下式时模型 D 对群体 A 和 B 具备公平性:

$$\mathbb{E}_{C \sim T_A} M(D(C)) = \mathbb{E}_{C \sim T_B} M(D(C)),$$

其中 T_A 和 T_B 分别表示与群体 A 、 B 相关的上下文分布。

作为对比, 基于概率的指标则通过分析语言模型的输出概率分布评估公平性。例如, 对于男性名 (如 Joseph) 和女性名 (如 Kelly), 模型生成“工程师”或“医生”等职业词汇的概率应接近; 类似地, 与非裔美国人相关的提示生成贬义形容词的概率应与欧裔美国人相关提示相当。

形式化地, 令 U 表示敏感词集合 (如高薪职业、贬义形容词), \mathcal{U} 为所有此类集合的族, $\mathcal{G} = \{(A^i, B^i) \mid i = 1, 2, \dots, I\}$ 定义待评估的成对群体。根据 [ZRZ24], 公平性准则可表述为:

$$|\mathbb{P}(o(\mathbf{x}) \in U \mid \mathbf{x} \in A) - \mathbb{P}(o(\mathbf{x}) \in U \mid \mathbf{x} \in B)| \leq \alpha, \quad \forall U \in \mathcal{U}, (A, B) \in \mathcal{G}, \quad (11)$$

其中概率基于语言模型 $f(\mathbf{x})$ 生成的下一个词元输出 $o(\mathbf{x}) \sim f(\mathbf{x})$ 计算。

针对上述公平性问题, 研究者在 LLM 工作流的不同阶段 (包括数据预处理、模型训练和训练后处理) 提出了多种偏见缓解技术 [Gal+23]。

预处理方法 在预处理阶段, 主要方法论是通过修正训练数据中的群体不平衡实现公平性。典型方法包括: 通过替换群体不平衡词汇进行数据增强 [Lu+20]、子集选择 [GMA22]、实例重新加权 [HBC22] 等。这些预处理策略通过使数据分布与公平性目标对齐, 帮助 LLM 学习较少延续社会偏见的表征。

训练中处理 在模型训练阶段, 可通过将公平性指标纳入训练目标来缓解偏见。常见方法是将公平性指标作为正则化项加入损失函数。具体而言, 可采用基于距离的度量族来缩小嵌入空间中不同群体间的距离, 例如句子嵌入间的 ℓ_2 距离 [Liu+20a]、余弦相似度 [Hua+20] 和 Jensen-Shannon 散度 [Yan+23; Woo+23]。除距离度量外, 另一研究方向试图设计能够解耦嵌入与群体身份关联的指标。为此, [BB19; KB21; Par+23] 提出通过最小化中性词 (如“领导力”) 嵌入在性别

嵌入上的投影长度来促进正交性，而 [CPC21; WCH23] 则通过最小化互信息促使词嵌入与性别独立。

除添加公平性指标作为正则项外，也可通过改变训练目标来缓解偏见。例如 [Xia+24d] 发现 LLM 在训练后阶段的 RLHF 过程中会产生内在算法偏见，这种偏见会显著放大主流偏好而削弱少数群体偏好。为解决此问题，[Xia+24d] 提出了偏好匹配 RLHF，在奖励目标中引入额外的正则项。[Che+21; He+22a; Oh+22] 应用对比学习与反事实数据增强来最大化原始句子与其反事实对应句之间的相似性。[ZLM18; Jin+21; HBC21] 使用对抗训练框架，其中训练攻击者从编码器嵌入预测受保护属性，同时训练编码器防止攻击者识别受保护属性。[Ouy+22; Bai+22] 利用人类或机器学习模型的反馈进行强化学习以鼓励模型生成公平文本。[Nak+24] 建立了理论框架，利用 LLM 在非平衡分类和伪相关问题中人工增加少数类样本量。

训练后处理 模型训练完成后，可通过修改生成分布来强化公平性。直接方法包括约束分布仅生成无偏见的词语和短语，例如 [Geh+20; Rol+21] 根据列表禁用有害词语，[SSB22] 在分布中搜索与最高似然输出性别不同的结果，[Shu+22; Sch+22a] 使用其他机器学习模型识别有害词并替换为中性词。另一种方法是编辑下一个词元的概率分布，例如 [Geh+20; Kim+23; LKW23] 根据评估模型学得的公平分数对生成分布中的词元概率进行重新加权，[Liu+21; Hal+23] 分别训练擅长生成无害/有害文本的专家和反专家模型，在推理阶段提升专家模型高概率词元的权重并降低反专家模型低概率词元的权重。近期研究也开始在 LLM 中应用多群体公平概念。[ZRZ24] 提出了训练后处理机器学习模型的框架以确保预测满足多群体公平保证，并应用该框架实现了公式 (11) 中的概念。

未来研究方向 尽管在缓解 LLM 算法偏见方面已取得显著进展，但仍存在若干关键挑战和开放性问题。首要问题是公平性与效用的权衡——偏见缓解技术往往以降低模型性能为代价，在复杂的现实应用中尤为明显。未来研究应探索在公平性约束与模型整体准确性之间实现平衡的理论方法，可能通过开发自适应或任务特定的公平性约束来同时优化公平性和效用。第二项关键挑战是上下文相关的公平性。现有公平性度量和去偏策略通常依赖预定义的敏感属性（如性别或种族），但公平性关切可能高度依赖上下文，并随应用场景、语言和文化视角而变化。有前景的方向是开发动态的、上下文感知的公平性度量方法，使其能够适应不同场景，确保去偏策略在多样化环境中保持有效性。第三，多模态交互式 AI 系统的公平性研究仍显不足。许多现实应用涉及文本、图像和结构化数据间的交互，且偏见在不同模态中可能以不同形式显现。研究重点应扩展到多模态 LLM 和对话式 AI 系统，开发公平性感知学习技术以缓解更复杂场景中的偏见。第四，公平性干预措施的长期影响需要深入研究。当前评估多基于短期性能指标，但随着模型微调、更新或与用户持续交互，偏见可能重新显现。建立稳健的纵向研究框架以评估公平性干预的持久性，对实现可持续的偏见缓解至关重要。最后，公平性

干预措施的可扩展性和计算效率仍是实际挑战。许多现有方法（尤其是修改训练目标的处理中技术）会引入显著计算开销，难以在大规模 LLM 中实施。未来工作应探索轻量级去偏技术，在保持公平性保证的同时维持计算可行性。

4.6 LLM 对齐：统计视角

基于 RLHF 与偏好优化的对齐 对齐是语言模型训练流程中的关键步骤，其目标是确保模型生成内容符合人类偏好。现有对齐方法主要通过基于人类反馈的强化学习（RLHF）实现，该方法尝试基于成对人类偏好数据学习奖励函数，并推导出相应的最优策略。设 \mathcal{S} 为状态空间， \mathcal{A} 为动作空间， $r(s, a)$ 表示在状态 $s \in \mathcal{S}$ 采取动作 $a \in \mathcal{A}$ 的奖励函数，强化学习过程旨在寻找策略 $\pi(s)$ 以最大化从步骤 1 到步骤 H 的累计奖励 $\mathbb{E}_s[\sum_{h=1}^H r(s_h, \pi(s_h))]$ 。在语言模型对齐的语境下，状态 $s \in \mathcal{S}$ 和动作 $a \in \mathcal{A}$ 分别对应输入提示和模型响应，奖励函数则表征人类对提示对应响应的满意度。在此框架下，[ZJJ23] 使用 Bradley-Terry-Luce 模型生成离线人类偏好数据，证明在线性奖励函数下 MLE 的变体可以达到近似最优样本复杂度，该结果后续被推广至一般奖励函数 [Zha+24a]、部分可观测奖励 [Kau+24] 和多方奖励 [Zho+24a] 等场景。相比标准强化学习，RLHF 使用的偏好数据包含的信息量少于奖励信号，因此通常被视为更困难的任务。然而 [WLJ23] 表明，对于广泛的偏好模型，传统基于奖励的强化学习算法能以较低额外复杂度解决 RLHF 任务。

自对齐 在理论前沿，[Ger+24; DD24] 分析了一系列统计模型（包括线性回归、广义线性模型和非参数回归）的模型坍缩现象，发现当每轮迭代用合成数据替换原始数据时，估计误差会线性增长；而当持续用合成数据增强原始数据时，估计误差会收敛到某个常数。最新研究 [DFK24] 表明，虽然增加模型规模可以缓解合成数据训练导致的模型退化，但混合合成数据与真实人类数据仍无法完全消除模型坍缩问题。

5 LLM 赋能的统计分析

5.1 数据收集

近期发展使得各类 NLP 任务（包括摘要生成、推理和抽取任务）取得显著进步。这些能力使得利用强大 LLM 进行零样本或少样本文本挖掘，以及从非结构化文本中提取结构化数据成为可能。

5.1.1 从非结构化文本提取结构化数据

LLM 的突出应用之一是情感分析（SA）——这一长期存在的文本挖掘任务旨在从上下文数据中提取人类情感。这些提取的情感可作为下游统计模型的输入，

例如 Twitter 互动预测模型、选举结果预测模型 [Alv+23] 以及其他社会现象分析模型。

除情感分析外，LLM 在从电子健康记录等复杂非结构化数据源提取结构化数据方面表现出卓越效果。例如最新研究 [Tan+23] 表明，使用 ChatGPT 生成的大规模合成生物学命名实体识别（NER）和生物学关系抽取（RE）数据进行微调可显著提升性能。具体而言，在生物学 NER 任务（如 NCBI Disease、BC5CDR Disease、BC5CDR Chemical）中 F1 分数从 23.37% 提升至 63.99%，在 RE 任务（如 GAD、EU-ADR）中 F1 分数从 75.86% 提升至 83.59%。这些结果展示了 LLM 在医疗等高价值领域改造数据抽取流程并优化统计分析管线的潜力。

5.1.2 合成数据生成

由于人类撰写的指令和标注成本高昂且存在数量、多样性、创造性等方面的限制，超人类对齐与自我奖励研究领域的兴起促使更多研究转向利用强大的语言模型生成合成数据。近期工作 [Wan+23e] 展示了语言模型通过生成人类水平自然指令实现自我提升的能力。随着合成文本生成任务的发展，生成真实统计数据的适配方法近年日益普及。

表格数据 真实合成表格数据的生成近年受到广泛关注。表格数据收集往往成本高昂且面临诸多挑战，包括类别不平衡、长尾标签分布 [Cao+19]、限制数据共享的隐私问题 [Gas+16]，以及噪声或缺失值等数据缺陷 [LT20]。合成数据生成成为这些挑战提供了实用解决方案，在解决隐私和类别不平衡问题的同时助力开发稳健统计模型 [Cho+17; Bor+22]。例如 [Bor+23] 提出了基于自回归生成式 LLM 的合成表格数据生成管线，其研究表明基于合成表格数据训练的判别模型性能优于基于真实数据训练的对比模型。此外，[Nak+24] 提出利用合成表格数据应对非平衡分类和伪相关挑战，展示了 LLM 在复杂资源受限场景中优化统计工作流的潜力。

金融数据 理性预期仍是宏观经济学和金融学中主流的信念模型。近年来，使用调查将信念与可观测数据关联的方法逐渐兴起。[Byb23] 提出了利用 LLM 生成信念的替代方法，并评估生成的股市预期与美国个人投资者协会（AAII）调查和 Duke CFO 调查的匹配程度。

5.2 数据清洗

LLM 作为表格数据清洗的编码器 LLM 能自动生成用于清洗、预处理和转换原始数据的代码，极大节省数据科学家的时间成本。例如 [Tu+24] 展示了如何通过带代码解释器的 ChatGPT 辅助统计学家完成缺失值检测、重复行删除、异常值核查、类别列数值化转换等操作。

LLM 用于自动化特征工程 特征工程指从原始数据构建有效特征的过程。给定输入数据 $\{(X, Y)\}$ ，特征工程旨在寻找变换 $\phi(x)$ 以最大化 $\{(\phi(X), Y)\}$ 的预测精度。尽管现代机器学习模型能学习复杂关系，但在数据有限或模型简单时，依赖领域知识的特征工程仍至关重要。利用 LLM 可基于其内化的领域知识自动构建特征。例如 [HMH24] 通过向 LLM 提供数据集上下文信息、数据样本及特征工程示例，引导其生成新特征。[Han+24a] 则提示 LLM 生成分类任务的显式决策规则，并将其转化为二元特征用于预测模型训练。

LLM 用于大规模文本过滤评估 在将文本输入任何机器学习算法前进行清洗是标准实践。无论是预训练还是微调，训练前清洗数据都有助于确保准确性并提升文本质量。为确保 Llama 3 使用高质量数据进行训练，其预训练数据集应用了包含启发式过滤器、NSFW 过滤器、语义去重方法和文本质量分类器的多级数据过滤管线。除表情符号过滤、HTML 标签清理、XML 解析、小写转换等常规处理外，近期研究开始采用 LLM 对文本进行质量评分与排序以筛选高质量训练数据。典型方法包括微调 LLM 作为文本分类器，例如使用 Llama 2 为文本质量分类器生成训练数据，而该分类器又用于筛选 Llama 3 的训练数据。其他方法如 BartScore 直接应用 BART 模型（基于编码器-解码器架构的预训练语言模型）无监督地评估文本流畅性、事实性与信息量。最新奖励评分方法如 DPO 奖励则利用 LLM 对文本的预测 logit 值进行文本对的优劣判别。随着 LLM 涌现出思维链（CoT）[Wei+22a] 和上下文学习（ICL）[Bro+20] 等能力，研究者开始将 LLM 作为可扩展、可解释的工具，用来模拟昂贵的人类评估。[Zhe+23] 表明，GPT-4 等强 LLM 评估器在两项富有挑战性的开放问题基准测试上达到了接近人类的水平，取得了与人类标注超过80%的一致程度，且没有针对领域任务进行微调。[Yua+24] 通过 LLM 自评估机制对训练数据进行评分筛选，其自奖励管线迭代训练出的模型在 NLP 排行榜上超越了 Claude 2、Gemini Pro 和 GPT-4 0613 等先进模型。

5.3 LLM 用于数据分析

有证据表明，LLM 能够在研究生水平的数学任务中执行基础数学推理。但先前研究很少评估 LLM 在统计分析方面的推理能力，特别是在真实数据场景下。[Liu+24d] 研究发现，即使最强大的模型在数据分析中也可能存在困难。例如 GPT-4 在基础统计推理任务（如“什么是95%置信区间？”）中仅能达到58%的准确率。

量化分析工具应用 实践表明，可以直接向 ChatGPT 输入表格数据来执行直接的表格分析（如聚合、求平均等操作），生成具有过滤、排序、聚合和求和逻辑的 SQL 语句，并从提示信息中找到目标数据库并执行 SQL 查询。

文本数据的统计分析 文本数据长期以来都是统计分析的重要来源，例如主题建模等方法可为大规模语料库提供可解释的摘要。传统模型如潜在狄利克雷分配

(LDA) 通过分析词语的共现模式来识别文本语料库中的潜在主题，提供主题的概率化表示 [BNJ03; Ke+23; KW24; WZT23]。这些方法被广泛用于揭示文本数据中的隐藏结构，为众多下游应用奠定基础。随着 LLM 的发展，出现了更先进的方法来从非结构化文本中可扩展地提取结构化数据，以支持进一步的统计分析。其中具有前景的一个研究方向是“词嵌入”技术的应用——这类技术通过分析文本中词汇的分布特性来捕捉语义信息。通过将文本嵌入整合到统计模型中，可以利用上下文信息，并通过更丰富的文本表示来增强回归分析。简言之，该方法使用在大型语料库（如 BERT、Llama 或 GPT）上预训练的上下文嵌入，仅需对平均嵌入进行简单线性变换即可进行回归等统计建模 [RSS23]。此类词嵌入预测方法已广泛应用于分类任务 [SK21] 以提高模型预测精度。词嵌入的另一重要特性是能够捕捉词语与文本之间的语义关系。

LLM 标注的统计推断 在许多统计分析（如社会调查）中，获取人工标注数据既耗时又昂贵。与此同时，LLM 能够以极低成本高效模拟人工标注，但这些标注可能存在偏差且无法保证统计推断的有效性。针对这一困境，[Ang+23; ADZ23] 提出了预测驱动推断（Prediction-Powered Inference, PPI）框架——该通用框架允许研究者利用任意黑箱机器学习模型的预测结果，执行有效的统计推断（如计算 p 值和置信区间）。假设研究者在小型数据集 $(X_i, Y_i)_{i=1}^n$ 上收集人工标注，并从相同分布中观测到大量未标注数据 $(X_i)_{i=n+1}^{n+N}$ ，同时存在预训练模型 f 提供预测 $\hat{Y}_i = f(X_i)$ 。目标参数定义为 $\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}[\ell_{\theta}(X, Y)]$ （例如样本均值或回归系数）。PPI 估计量 [Ang+23; ADZ23; JLZ25] 可表示为：

$$\hat{\theta}^{PPI} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \ell_{\theta}(X_i, Y_i) - \frac{1}{n} \sum_{i=1}^n \ell_{\theta}(X_i, f(X_i)) + \frac{1}{N} \sum_{i=n+1}^{n+N} \ell_{\theta}(X_i, f(X_i)). \quad (12)$$

相较于仅使用人工标注或直接将机器学习预测视为金标准标签的方法，PPI 估计量在保持统计有效性（如覆盖率）的同时，能有效利用预测信息。该方法已成功应用于辅助多项计算社会科学研究（使用 LLM 生成标注），包括仇恨言论检测 [Ega+24] 和媒体偏见分析 [Gli+24]。此外，[JMS24] 强调了扩展定律在指导 LLM 生成替代数据有效整合方面的潜力，解决了高质量数据收集成本高昂或不现实场景下的关键瓶颈。研究发现的替代数据显著降低测试误差的现象（即使与原始数据集无关），引发了关于该方法局限性及其泛化影响的思考。

5.4 LLM 辅助的医学研究

当今，我们获取的健康数据比以往任何时候都多，这为 LLM 彻底改变现代医学提供了可能。许多疾病难以诊断且在发现前已对身体造成严重损害 [Ram+17]，同时全球仍有大量人群无法获取优质医疗信息和医疗服务 [BO17]。LLM 能够快速处理分析海量数据，为医疗从业者和患者提供有价值的洞见与潜在诊断方案。

人类难以观察或区分的模式可能被 LLM 捕捉，并能提供符合最新研究和数据的连贯信息 [Nav+23]。这种能力有望提升诊断支持水平并增强医疗问答的可及性。由 LLM 生成的合成医疗数据可提高数据可访问性，推动医学 AI 研究。但需注意，医学领域应用 LLM 存在重大风险与考量——包括偏见、幻觉、伦理违规、监管缺失和透明度不足等问题。本节将回顾医学 LLM 的重要最新进展，更全面的综述可参考 Zhou 等人（2024）中的研究 [Zho+24b]。

早期基于 Transformer 的临床语言模型 近期出现了多种在医疗数据上预训练并针对医学领域特定任务微调的语言模型。基于 Transformer 架构的早期典型案例是 Clinical-BERT [Als+19]。该模型采用与 BERT 相同的编码器架构，在 ICU 入院临床记录上进行训练，能够揭示医疗概念间的语义关联。Clinical-BERT 被广泛用于各类临床 AI 工作流，为电子健康记录（EHR）中的医疗概念生成嵌入表示 [YW21; SLG20; BDP20; Kan+23]。

面向临床文本生成的 LLM 随着生成式 AI 技术的显著进步，医学领域开发了专门用于生成合成医疗数据的 LLM，并通过微调适应多种自然语言处理任务。由于隐私保护与 HIPAA 合规要求，大规模临床文本和临床 LLM 的获取与共享存在困难。Kweon 等人（2023）开发了首个可共享的生成式临床 LLM——Asclepius [Kwe+23]，提供基于 Llama2-7B/Llama2-13B 架构和 Llama3-8B/Mistral-7B 架构的版本。该模型基于 PubMed Central 的匿名病例报告生成合成临床笔记。Peng 等人（2023）提出的 GatortronGPT [Pen+23]，整合了佛罗里达大学健康中心 126 个临床科室、200 万患者的 2770 亿词临床文本与 1950 亿词通用英语语料，进而生成 200 亿词合成文本训练 NLP 模型，实现临床概念提取、事件关联、语义相似度、自然语言推理和问答等任务。

面向医疗问答的 LLM 目前存在多种专为医疗问答构建的 LLM，这些模型能够促进医学知识民主化、提升医疗服务不足人群的医疗可及性，并通过效率提升减轻医生负担。[Sin+23] 开发的 Med-PALM2 基于 PaLM2 架构进行训练，并针对医学领域进行专项微调 [Sin+23]。该模型整合了多问答数据集的指令微调技术，采用数据集混合比例优化策略训练“统一”模型以提升跨数据集性能。模型还引入集成优化机制：通过温度采样生成多组回答，以前序生成结果为基础迭代优化解释与答案，经多轮优化后采用多数表决确定最终答案。Han 等人（2023）基于 Llama 架构开发了 MedAlpaca 模型，训练数据整合了医疗知识卡片、StackExchange 问答和 Wikidoc 文档的多样化问答对 [Han+23]。Med-PALM2 为闭源模型，而 MedAlpaca 为开源模型。Xie 等人（2024）开发了 Me-LLaMA 系列开源医学 LLM [Xie+24a]，它融合生物医学论文、临床记录与通用领域数据，在多项医学数据集上超越了 ChatGPT 的表现。

面向健康社会决定因素识别的 LLM 近期研究开始探索利用 LLM 识别健康社会决定因素 (Social Determinants of Health, SDoH)。SDoH 指影响人类健康结果的出生、成长、生活、工作及养老环境因素，涵盖经济稳定性、教育水平、社会支持网络、社区环境及食物获取等维度。准确识别 SDoH 有助于构建患者健康全景视图，优化预防措施与资源分配策略。Guevara 等人 (2024) 通过应用现有 LLM (Flan T-5、GPT 3.5、GPT 4) 并利用 LLM 生成的合成 SDoH 数据进行微调 [Gue+24]，构建 SDoH 提取模型。微调过程中，基于标注指南提示 LLM 生成新 SDoH 类别语句，以合成样本为参考扩展生成规模。实验表明，该模型对存在不利 SDoH 患者的识别率达到93.8%，显著优于传统 ICD-10 编码2%的检出率，彰显了 LLM 在 SDoH 识别领域的巨大潜力。

医学多模态 LLM 近年来，LLM 除文本输入外，已拓展至处理图像、图表、截图、PDF、视频、音频等多种数据类型。以医学影像整合为例，多模态医学 LLM 常通过集成视觉 Transformer (Vision Transformer, ViT) 处理图像输入。ViT 生成图像的向量表示，与文本表示拼接形成单一序列。Flamingo-CXR [Moo+23] 和 Med-PaLM [Li+24a] 等大型多模态模型在受控环境下生成放射学报告的能力已可与放射科医师相媲美 [Hua+23b]。谷歌开发的 Gemini 模型作为新一代多模态模型，已展现出各种新的能力，其中 Med-Gemini 结合多模态特性与网络搜索整合能力，在复杂诊断挑战、图像解析、视频问答及文本摘要任务中达到先进性能 [Ani+23]。

医疗领域应用 LLM 的挑战 尽管 LLM 在医疗研究领域发展迅猛，但其在高风险临床应用中仍面临严峻挑战。主要问题包括算法偏见、校准不足、幻觉生成、错误信息及回答的任意性 [San+23; Sim+24; Gao+24]。鉴于医疗数据的特殊性与安全性要求，这些模型需要专门的提示工程、微调和对齐策略以确保其在实际临床环境中的有效性与可靠性。例如，LLM 在不确定性条件下的临床推理能力欠佳 [Omi+23]，且存在显著偏见 [San+23; Cas+23; Aza+24; Cha+24]，这可能导致公平性问题——当临床医生依赖 LLM 进行决策支持时，可能对少数群体造成额外伤害 [PFB24; Ayo+24]。研究者强调需在解决这些挑战后方可推进 LLM 的医疗应用 [Szo24; FDW24; Tes+24]。

6 讨论

6.1 统计学家的小语言模型

与计算机科学领域研究者相比，统计学家通常面临计算资源有限的挑战，但其在语言模型开发中展现出独特优势：擅长构建资源高效、理论根基深厚且符合统计学原理的模型。当计算机科学界聚焦于通过海量数据和算力扩展 LLM 时，统计学家在开发“小语言模型” (Small Language Models, SLMs) 方面表现卓越

——这些模型通过强调效率与结构优化（而非暴力计算），在特定领域实现可与大模型竞争的性能。统计学家的优势还体现在将领域知识嵌入模型设计的专长上。通过采用稀疏性假设、分层建模和结构化正则化等技术，SLM 能够实现更高的样本效率。例如，稀疏性先验可使模型聚焦相关特征，在减少数据需求的同时提升可解释性；分层贝叶斯方法可整合文本数据的多层级依赖，通过上下文信息共享原则性降低训练所需有效样本量。这些方法在数据有限、噪声显著或资源受限场景（如个性化医疗、法律文本分析或金融建模）中往往优于大模型。

SLM 与统计学家对可解释性的追求高度契合。通过采用简约架构和显式假设，SLM 具有天然的透明性，使研究者能将模型预测溯源至具体特征或交互关系。这种可解释性在信任与可监管性至关重要的高风险领域尤为重要。相较之下，LLM 虽功能强大却常被视为“黑箱”，在需要解释性的场景中应用受限。此外，统计学家在与计算机科学家协同应对现代 AI 计算挑战方面具有独特优势。融合统计学与计算机科学的技术（如低秩近似、变分推断和惩罚优化）可有效降低模型训练与推理的计算开销。应用于 SLM 时，这些方法不仅提升模型可及性，更增强环境可持续性，缓解大规模 AI 系统的碳排放问题。通过注重原理驱动的设计，统计学家可构建在保持稳健性与适应性的同时降低算力需求的模型。

统计学家的 SLM 并非旨在取代 LLM，而是与其形成互补。LLM 擅长开放域的通用任务，而 SLM 可专注于特定领域应用，以极低资源消耗实现可比性能。通过发挥理论驱动建模、样本效率与可解释性等优势，统计学家可在快速演进的 AI 版图中占据独特地位。这一路径既促进了先进语言建模技术的普惠化，也确保 AI 发展符合效率、透明与可持续的价值导向。

统计学家在构建 LLM “封装器”（Wrapper）方面也扮演关键角色——这些封装器可对黑箱模型输出进行校准、不确定性量化与去偏处理。前文提及的预测驱动推断（PPI）估计量即为典型案例。若围绕 LLM 合理部署因果推断方法，亦可提升其应用价值。统计学中的实验设计方法在 LLM 时代仍具现实意义。因此，统计学家既能为 LLM 提供替代方案与增强工具，也能为其理论分析做出重要贡献。

6.2 理解 LLM 的探索方向

LLM 的卓越成功激发了学界对其内部机制的理解热潮，统计学习理论工具在此过程中发挥了重要作用。早期研究通过近似理论、优化机制、泛化性能与提示工程等视角取得突破性进展。例如，有研究表明 Transformer 在各种架构中的通用逼近特性 [Yun+20a; Yun+20b; KS24]、其作为图灵完备系统的计算表达能力 [PMB19; WCM22]，以及在特定任务中相比 MLP 更优的优化动力学特性 [Wan+24d]。在泛化性方面，研究者改进了序列长度相关的泛化界 [Zha+22; TT24]，并将其扩展至时间序列等非独立同分布场景 [Lim+24]。上下文学习（In-context Learning, ICL）与思维链（Chain-of-thought, CoT）等提示技术深化了我们对 LLM 适应性与推理机制的理解。理论工作将 ICL 解释为源于隐马尔可

夫模型等预训练分布的隐式贝叶斯推断 [Xie+22]，并探索其对梯度下降、牛顿法与强化学习等算法进行近似的能力 [Aky+22; Von+23; Fu+23; Bai+24; LBM24; ZFB24; Guo+24c]。CoT 提示通过将任务分解为序列步骤来增强推理能力，进而提升准确率与可解释性 [Wei+22a; Yao+23]。例如，研究表明 CoT 可增强 Transformer 在算术与方程求解任务中的近似能力 [Fen+24a]，并通过稀疏注意力模式简化学习过程 [Wen+24]。此外，多数预训练 LLM 表现出“注意力汇集 (Attention-sink)”现象 [Xia+24c]——特定“汇集词元” (Sink Tokens) 会获得异常高的注意力权重。[Guo+24b] 通过简化模型分析发现该现象主要由注意力头中的 Softmax 运算引发。最后，LLM 预训练遵循经验缩放定律：测试误差随模型规模与样本量呈现出多项式的改善趋势。[Lin+24a] 在无限维线性回归框架下理论推导出与实证相符的缩放定律。

尽管取得上述进展，但我们对 LLM 的理解仍不完整。传统统计框架虽具价值，却难以解释零样本泛化、上下文学习与 CoT 推理等由模型架构、训练数据与优化动力学复杂交互产生的涌现现象。例如，虽然 ICL 可通过算法近似与贝叶斯推断得到部分解释，但其无需参数更新即可跨任务灵活适应的能力仍是未解之谜。类似地，CoT 通过结构化任务分解增强推理的机制，暴露了我们对注意力驱动复杂问题求解理论认知的空白。这些现象表明需要融合统计学习理论、优化理论与认知科学的新框架来完整刻画 LLM 的微妙能力。填补这些理论空缺为统计学家提供了重要机遇——通过发展创新理论与方法论，架起实证成功与基础认知之间的桥梁，最终深化对变革性技术的根本性理解。

6.3 人机协作的数据科学

随着 LLM 的持续进步，一个重要发展方向在于构建有效的人机协作框架——使 AI 系统成为人类专业知识的补充而非替代。尽管 LLM 在推理、文本生成与决策支持方面展现出卓越能力，但其缺乏深层理解、情境认知与可监管性等关键要素，而这些要素在医疗、法律与科研等高风险应用中至关重要。因此，将 AI 视为人类智能的辅助工具而非替代品，开发整合人类专业知识与 AI 洞见的框架更具前景。

人机协作的重要方向之一是交互式决策：LLM 提供建议、辅助探索性数据分析或生成潜在解决方案，而人类用户保留最终控制与监督权。例如在统计建模中，LLM 可自动化数据预处理、特征工程与模型选择等繁琐任务，人类分析师则聚焦领域推理、假设检验与结果诠释；医疗应用中，AI 可辅助医生汇总病历或提示潜在诊断，但最终决策权仍归属人类专家，确保可监管性与伦理考量。

然而，交互式决策也引入了新的统计挑战，主要源于数据的非独立同分布 (Non-i.i.d.) 特性。传统统计与机器学习通常假设训练数据独立同分布，但在人机协作中，数据分布会基于 AI 建议与人类反馈动态调整。这种反馈循环可能导致选择偏差、概念漂移与策略性适应——人类决策模式随 AI 推荐发生偏移。例如招聘推荐系统若过度推荐特定人群，决策者可能随之调整选择模式，形成偏差

强化循环。传统基于固定分布假设的统计工具在此场景中失效，亟需开发能适应动态数据分布的自适应推断方法、因果建模技术与稳健统计框架。相关讨论与参考文献可参见 [Per+20; Gar+24; GC24]。另一关键研究方向是设计适应用户专业水平与偏好的自适应 AI 系统。当前 LLM 多为静态系统，仅依据输入提示生成响应。然而，有效协作需要 AI 系统具备从用户反馈中学习、持续优化响应并个性化调整辅助策略的能力。相关技术包括基于人类反馈的强化学习（RLHF）、低置信度时将决策权移交人类的不确定性感知 AI 模型，以及通过对话澄清需求的交互式 AI 系统。

应对这些挑战需跨学科协作，整合统计学家、计算机科学家、社会科学家、法律学者与政策制定者的智慧，以设计公平、透明、可监管的 AI 系统。通过将严谨统计框架与 LLM 技术相结合，研究团体可以共同协作，确保人类福祉始终作为下一代 AI 模型发展的核心考量。

参考文献

- [Aar23] Scott Aaronson. “Watermarking of large language models”. In: *Workshop on Large Language Models and Transformers, Simons Institute, UC Berkeley*. 2023.
- [AB23] Anastasios N Angelopoulos and Stephen Bates. “Conformal prediction: A gentle introduction”. In: *Foundations and Trends® in Machine Learning* 16.4 (2023), pp. 494–591.
- [Aba+16] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.
- [Abd+21] Moloud Abdar et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Information fusion* 76 (2021), pp. 243–297.
- [ADZ23] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. “PPI++: Efficient prediction-powered inference”. In: *arXiv preprint arXiv:2311.01453* (2023).
- [AFZ21] Abubakar Abid, Maheen Farooqi, and James Zou. “Persistent anti-muslim bias in large language models”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 298–306.
- [Aga+24] Rishabh Agarwal et al. “Many-shot in-context learning”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 76930–76966.

- [Ahm+24] Arash Ahmadian et al. “Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 12248–12267.
- [Aky+22] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. “What learning algorithm is in-context learning? investigations with linear models”. In: *arXiv preprint arXiv:2211.15661* (2022).
- [AL23a] Zeyuan Allen-Zhu and Yuanzhi Li. “Physics of language models: Part 1, context-free grammar”. In: *arXiv preprint arXiv:2305.13673* (2023).
- [AL23b] Zeyuan Allen-Zhu and Yuanzhi Li. “Physics of language models: Part 3.1, knowledge storage and extraction”. In: *arXiv preprint arXiv:2309.14316* (2023).
- [AL23c] Zeyuan Allen-Zhu and Yuanzhi Li. “Physics of language models: Part 3.2, knowledge manipulation”. In: *arXiv preprint arXiv:2309.14402* (2023).
- [AL24] Zeyuan Allen-Zhu and Yuanzhi Li. “Physics of language models: Part 3.3, knowledge capacity scaling laws”. In: *arXiv preprint arXiv:2404.05405* (2024).
- [All24] Zeyuan Allen-Zhu. *ICML 2024 tutorial: Physics of language models*. Project page: <https://physics.allen-zhu.com/>. July 2024.
- [Als+19] Emily Alsentzer et al. “Publicly available clinical BERT embeddings”. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (2019), pp. 72–78.
- [Alv+23] Quratulain Alvi et al. “On the frontiers of Twitter data and sentiment analysis in election prediction: a review”. In: *PeerJ Computer Science* 9 (2023), e1517.
- [AM23] Amos Azaria and Tom Mitchell. “The internal state of an LLM knows when it’s lying”. In: *arXiv preprint arXiv:2304.13734* (2023).
- [Ami07] Ali Amiri. “Dare to share: Protecting sensitive knowledge with data sanitization”. In: *Decision Support Systems* 43.1 (2007), pp. 181–191.
- [An+24] Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. “Measuring gender and racial biases in large language models”. In: *arXiv preprint arXiv:2403.15281* (2024).

- [Ang+23] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. “Prediction-powered inference”. In: *Science* 382.6671 (2023), pp. 669–674.
- [Ani+22] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. “Large-scale differentially private BERT”. In: *Findings of the Association for Computational Linguistics: EMNLP*. 2022, pp. 6481–6491.
- [Ani+23] Rohan Anil et al. “Gemini: A family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (Dec. 2023).
- [Ary+24] Viraat Aryabumi et al. “To code, or not to code? Exploring impact of code in pre-training”. In: *arXiv preprint arXiv:2408.10914* (2024).
- [Ask+21] Amanda Askell et al. “A general language assistant as a laboratory for alignment”. In: *arXiv preprint arXiv:2112.00861* (2021).
- [Ayo+24] Noel F. Ayoub et al. “Inherent bias in large language models: A random sampling analysis”. In: *Mayo Clinic Proceedings: Digital Health* 2.2 (2024), pp. 186–191.
- [Aza+24] Mohammad Gheshlaghi Azar et al. “A general theoretical paradigm to understand learning from human preferences”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 4447–4455.
- [Bai+09] Bing Bai et al. “Supervised semantic indexing”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 187–196.
- [Bai+22] Yuntao Bai et al. “Constitutional AI: Harmlessness from AI feedback”. In: *arXiv preprint arXiv:2212.08073* (2022).
- [Bai+24] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. “Transformers as statisticians: Provable in-context learning with in-context algorithm selection”. In: *Advances in neural information processing systems* 36 (2024).
- [Bak+24] Yavuz Faruk Bakman et al. “MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. 2024, pp. 7752–7767.
- [Bañ+20] Marta Bañón et al. “ParaCrawl: Web-scale acquisition of parallel corpora”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4555–4567.

- [BB19] Shikha Bordia and Samuel R. Bowman. “Identifying and reducing gender bias in word-level language models”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Student Research Workshop*. 2019, pp. 7–15.
- [BCB15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *3rd International Conference on Learning Representations, ICLR*. 2015.
- [BDP20] Brent Biseda, Gaurav Desai, and A. Philip. “Prediction of ICD codes with clinical BERT embeddings and text augmentation with label balancing using MIMIC-III”. In: *arXiv preprint arXiv:2008.08980* (2020).
- [Ben+03] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. “A neural probabilistic language model”. In: *J. Mach. Learn. Res.* 3.null (2003), pp. 1137–1155. ISSN: 1532-4435.
- [Bes+24] Maciej Besta et al. “Graph of thoughts: Solving elaborate problems with large language models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17682–17690.
- [Bi+24] Xiao Bi et al. “DeepSeek LLM: Scaling open-source language models with longtermism”. In: *arXiv preprint arXiv:2401.02954* (2024).
- [BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).
- [BL24] Amrita Bhattacharjee and Huan Liu. “Fighting fire with fire: Can ChatGPT detect AI-generated text?” In: *SIGKDD Explor. Newsl.* 25.2 (Mar. 2024), pp. 14–21. ISSN: 1931-0145.
- [BMW20] Andreas Balzer, David Mowatt, and Muiris Woulfe. *Obfuscating information related to personally identifiable information (PII)*. US Patent 10,839,104. Nov. 2020.
- [BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [BO17] World Bank and World Health Organization. “World Bank and WHO: Half the world lacks access to essential health services, 100 million still pushed into extreme poverty because of health expenses”. In: (2017). Accessed: 2024-09-18. URL: <https://www.worldbank.org/en/news/press-release/2017/12/13/world-bank-who-half-world-lacks-access-essential-health-services-100-million-pushed-into-extreme-poverty-because-of-health-expenses>.

- [Bom+22] Rishi Bommasani et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2022).
- [Bor+22] Vadim Borisov et al. “Deep neural networks and tabular data: A survey”. In: *IEEE transactions on neural networks and learning systems* (2022).
- [Bor+23] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. “Language models are realistic tabular data generators”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.
- [Bri+23] Trenton Bricken et al. “Towards monosemanticity: Decomposing language models with dictionary learning”. In: *Transformer Circuits Thread* 2 (2023).
- [Bro+20] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [Bro+22] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramèr. “What does it mean for a language model to preserve privacy?” In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 2280–2292.
- [Bro+24] Bradley Brown et al. “Large language monkeys: scaling inference compute with repeated sampling”. In: *arXiv preprint arXiv:2407.21787* (2024).
- [BT52] Ralph Allan Bradley and Milton E. Terry. “Rank analysis of incomplete block designs: I. The method of paired comparisons”. In: *Biometrika* 39.3/4 (1952), pp. 324–345. ISSN: 00063444, 14643510.
- [Bu+20] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. “Deep learning with Gaussian differential privacy”. In: *Harvard Data Science Review* 2020.23 (2020).
- [Byb23] J Leland Bybee. “The ghost in the machine: Generating beliefs with large language models”. In: *arXiv preprint arXiv:2305.02823* (2023).
- [Cai+24] Zhongze Cai, Shang Liu, Hanzhao Wang, Huaiyang Zhong, and Xiaocheng Li. “Towards better statistical understanding of watermarking LLMs”. In: *arXiv preprint arXiv:2403.13027* (2024).

- [Cao+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. “Learning imbalanced datasets with label-distribution-aware margin loss”. In: *Advances in neural information processing systems* 32 (2019).
- [Car+21] Nicholas Carlini et al. “Extracting training data from large language models”. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650.
- [Car+23] Nicholas Carlini et al. “Quantifying memorization across neural language models”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.
- [Car98] Rich Caruana. “Multitask learning”. In: *Learning to learn*. Boston, MA: Springer US, 1998, pp. 95–133. ISBN: 978-1-4615-5529-2.
- [Cas+23] Stephen Casper et al. “Open problems and fundamental limitations of reinforcement learning from human feedback”. In: *arXiv preprint arXiv:2307.15217* (2023).
- [CCN23] Bilal Chughtai, Lawrence Chan, and Neel Nanda. “A toy model of universality: Reverse engineering how networks learn group operations”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 6243–6267.
- [CG96] Stanley F. Chen and Joshua Goodman. “An empirical study of smoothing techniques for language modeling”. In: *34th Annual Meeting of the Association for Computational Linguistics*. 1996, pp. 310–318.
- [CGC24] John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. “Large language model validity via enhanced conformal prediction methods”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [CGZ24] Miranda Christ, Sam Gunn, and Or Zamir. “Undetectable watermarks for language models”. In: *The Thirty Seventh Annual Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 1125–1139.
- [Cha+24] Souradip Chakraborty et al. “MaxMin-RLHF: Towards equitable alignment of large language models with diverse human preferences”. In: *arXiv preprint arXiv:2402.08925* (2024).
- [Che+20] Mark Chen et al. “Generative pretraining from pixels”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1691–1703.

- [Che+21] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. “FairFil: contrastive neural debiasing method for pretrained text encoders”. In: *9th International Conference on Learning Representations, ICLR*. 2021.
- [Che+23a] Mayee F. Chen et al. “Skill-it! A data-driven skills framework for understanding and training language models”. In: *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. 2023.
- [Che+23b] Xinyun Chen et al. “Universal self-consistency for large language model generation”. In: *arXiv preprint arXiv:2311.17311* (2023).
- [Che+24a] Chao Chen et al. “INSIDE: LLMs’ internal states retain the power of hallucination Detection”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Che+24b] Xuxi Chen et al. “Take the bull by the horns: Hard sample-reweighted continual training improves LLM generalization”. In: *arXiv preprint arXiv:2402.14270* (2024).
- [Chi+24] Wei-Lin Chiang et al. “Chatbot Arena: An open platform for evaluating LLMs by human preference”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Cho+17] Edward Choi et al. “Generating multi-label discrete patient records using generative adversarial networks”. In: *Machine learning for healthcare conference*. PMLR. 2017, pp. 286–305.
- [CHS24a] Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. “Enhancing watermarked language models to identify users”. In: *arXiv preprint arXiv:2405.11109* (2024).
- [CHS24b] Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. “Watermarking language models for many adaptive users”. In: *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society. 2024, pp. 84–84.
- [Chu+14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555* (2014).
- [Cla+18] Peter Clark et al. “Think you have solved question answering? Try ARC, the AI2 reasoning challenge”. In: *arXiv preprint arXiv:1803.05457* (2018).

- [CLZ25] Yinpeng Cai, Lexin Li, and Linjun Zhang. “A statistical hypothesis testing framework for data misappropriation detection in large language models”. In: *arXiv preprint arXiv:2501.02441* (2025).
- [CM24] Jiuhai Chen and Jonas Mueller. “Quantifying uncertainty in answers from any language model and enhancing their trustworthiness”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 5186–5200.
- [Col+11] Ronan Collobert et al. “Natural language processing (almost) from scratch”. In: *J. Mach. Learn. Res.* 12.null (2011), pp. 2493–2537. ISSN: 1532-4435.
- [Con+23] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. “Towards automated circuit discovery for mechanistic interpretability”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 16318–16352.
- [CP10] Emmanuel J Candes and Yaniv Plan. “Matrix completion with noise”. In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936.
- [CP11] Emmanuel J Candes and Yaniv Plan. “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements”. In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2342–2359.
- [CPC21] Pierre Colombo, Pablo Piantanida, and Chloé Clavel. “A novel estimator of mutual information for learning to disentangle textual representations”. In: *Annual Meeting of the Association for Computational Linguistics*. Aug. 2021, pp. 6539–6550.
- [CS24] Hiroaki Chiba-Okabe and Weijie J Su. “Tackling GenAI copyright issues: Originality estimation and genericization”. In: *Scientific Reports* (2024).
- [CW08] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: deep neural networks with multitask learning”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. Association for Computing Machinery, 2008, pp. 160–167. ISBN: 9781605582054.
- [CY15] Yinzhi Cao and Junfeng Yang. “Towards making systems forget with machine unlearning”. In: *2015 IEEE symposium on security and privacy*. IEEE. 2015, pp. 463–480.

- [CZ13] T Tony Cai and Anru Zhang. “Sparse representation of a polytope and recovery of sparse signals and low-rank matrices”. In: *IEEE transactions on information theory* 60.1 (2013), pp. 122–132.
- [CZY24] Shijie Chen, Yu Zhang, and Qiang Yang. “Multi-task learning in natural language processing: An overview”. In: *ACM Comput. Surv.* (May 2024). Just Accepted. ISSN: 0360-0300.
- [Dai+22] Damai Dai et al. “Knowledge neurons in pretrained transformers”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 8493–8502.
- [Dat+24] Sumanth Dathathri et al. “Scalable watermarking for identifying large language model outputs”. In: *Nature* 634.8035 (2024), pp. 818–823.
- [DD24] Apratim Dey and David Donoho. “Universality of the $\pi^2/6$ pathway in avoiding model collapse”. In: *arXiv preprint arXiv:2410.22812* (2024).
- [Dee+25] DeepSeek-AI et al. “DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning”. In: *arXiv preprint arXiv:2501.12948* (2025).
- [Den+24] Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. “Rephrase and respond: Let large language models ask better questions for themselves”. In: *arXiv preprint arXiv:2311.04205* (2024).
- [Der+17] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. “De-identification of patient notes with recurrent neural networks”. In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 596–606.
- [Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.
- [DFK24] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. “Strong model collapse”. In: *arXiv preprint arXiv:2410.04840* (2024).
- [DH17] Tobias Domhan and Felix Hieber. “Using target-side monolingual data for neural machine translation through multi-task learning”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Sept. 2017, pp. 1500–1505.

- [Doh+24] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. “A tale of tails: Model collapse as a change of scaling laws”. In: *arXiv preprint arXiv:2402.07043* (2024).
- [Dua+24a] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. “Flocks of stochastic parrots: Differentially private prompt learning for large language models”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Dua+24b] Jinhao Duan et al. “Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 5050–5063.
- [Dub+24a] Abhimanyu Dubey et al. “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024).
- [Dub+24b] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. “Length-controlled AlpacaEval: A simple way to debias automatic evaluators”. In: *arXiv preprint arXiv:2404.04475* (2024).
- [Dwo06] Cynthia Dwork. “Differential privacy”. In: *International colloquium on automata, languages, and programming*. Springer. 2006, pp. 1–12.
- [Ega+24] Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. “Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Elh+21] Nelson Elhage et al. “A mathematical framework for transformer circuits”. In: *Transformer Circuits Thread* 1.1 (2021), p. 12.
- [Elh+22a] Nelson Elhage et al. “Softmax linear units”. In: *Transformer Circuits Thread* (2022).
- [Elh+22b] Nelson Elhage et al. “Toy models of superposition”. In: *arXiv preprint arXiv:2209.10652* (2022).
- [Elk+23] Niva Elkin-Koren, Uri Hacoheh, Roi Livni, and Shay Moran. “Can copyright be reduced to privacy?” In: *arXiv preprint arXiv:2305.14822* (2023).
- [Elm90] Jeffrey L. Elman. “Finding structure in time”. In: *Cognitive Science* 14.2 (1990), pp. 179–211. ISSN: 0364-0213.

- [Erh+10] Dumitru Erhan et al. “Why does unsupervised pre-training help deep learning?” In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 625–660. ISSN: 1532-4435.
- [Eth+24] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. “Model alignment as prospect theoretic optimization”. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. JMLR.org, 2024.
- [Fai+23] Jaiden Fairoze et al. “Publicly detectable watermarking for language models”. In: *arXiv preprint arXiv:2310.18491* (2023).
- [Fan+21] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. “Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training”. In: *Proceedings of the National Academy of Sciences* 118.43 (2021), e2103091118.
- [Fan+24] Chongyu Fan et al. “Simplicity prevails: Rethinking negative preference optimization for LLM unlearning”. In: *Neurips Safe Generative AI Workshop*. 2024.
- [FDW24] Ari B. Friedman, M. Kit Delgado, and Gary E. Weissman. “Artificial intelligence for emergency care triage—much promise, but still much to learn”. In: *JAMA Network Open* 7.5 (May 2024), e248857.
- [Fen+24a] Guhao Feng et al. “Towards revealing the mystery behind chain of thought: a theoretical perspective”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Fen+24b] Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. “Beyond model collapse: Scaling up with synthesized data requires reinforcement”. In: *arXiv preprint arXiv:2406.07515* (2024).
- [Fer+23] Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. “Three bricks to consolidate watermarks for large language models”. In: *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. 2023, pp. 1–6.
- [Fom+20] Marina Fomicheva et al. “Unsupervised quality estimation for neural machine translation”. In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 539–555.
- [Fu+23] Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. “Transformers learn higher-order optimization methods for in-context learning: A study with linear models”. In: *arXiv preprint arXiv:2310.17086* (2023).

- [Gag94] Philip Gage. “A new algorithm for data compression”. In: *C Users J.* 12.2 (Feb. 1994), pp. 23–38. ISSN: 0898-9788.
- [Gal+23] Isabel O Gallegos et al. “Bias and fairness in large language models: A survey”. In: *arXiv preprint arXiv:2309.00770* (2023).
- [Gao+20] Leo Gao et al. “The Pile: An 800GB dataset of diverse text for language modeling”. In: *arXiv preprint arXiv:2101.00027* (2020).
- [Gao+24] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. “SPUQ: Perturbation-based uncertainty quantification for large language models”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. 2024, pp. 2336–2346.
- [Gar+24] Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. “Oracle efficient online multicalibration and omniprediction”. In: *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2024, pp. 2725–2792.
- [Gas+16] Adrià Gascón et al. “Privacy-preserving distributed linear regression on high-dimensional data”. In: *Cryptology ePrint Archive* (2016).
- [Gaw+23] Jakob Gawlikowski et al. “A survey of uncertainty in deep neural networks”. In: *Artificial Intelligence Review* 56.Suppl 1 (2023), pp. 1513–1589.
- [GC24] Isaac Gibbs and Emmanuel J Candès. “Conformal inference for online prediction with arbitrary distribution shifts”. In: *Journal of Machine Learning Research* 25.162 (2024), pp. 1–36.
- [Geh+20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. “RealToxicityPrompts: Evaluating neural toxic generation in language models”. In: *Findings of the Association for Computational Linguistics: EMNLP*. Vol. EMNLP 2020. Findings of ACL. 2020, pp. 3356–3369.
- [Geo+24] Kristian Georgiev et al. “Attribute-to-delete: Machine unlearning via datamodel matching”. In: *arXiv preprint arXiv:2410.23232* (2024).
- [Ger+24] Matthias Gerstgrasser et al. “Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data”. In: *arXiv preprint arXiv:2404.01413* (2024).

- [Ges24] Lee Gesmer. “Copyright and the challenge of large language models”. In: (Oct. 2024). America. URL: <https://www.masslawblog.com/copyright/copyright-and-the-challenge-of-large-language-models-part-2-2/>.
- [GF24] Eva Giboulot and Teddy Furon. “WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off”. In: *Neural Information Processing Systems*. 2024.
- [GFC21] Tianyu Gao, Adam Fisch, and Danqi Chen. “Making pre-trained language models better few-shot learners”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*. 2021, pp. 3816–3830.
- [Gin+19] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. “Making AI forget you: Data deletion in machine learning”. In: *Advances in neural information processing systems* 32 (2019).
- [Gio23] Patrizio Giovannotti. “Evaluating machine translation quality with conformal predictive distributions”. In: *Conformal and Probabilistic Prediction with Applications*. PMLR. 2023, pp. 413–429.
- [GL04] Jianfeng Gao and Chin-Yew Lin. “Introduction to the special issue on statistical language modeling”. In: *ACM Transactions on Asian Language Information Processing (TALIP)* 3.2 (2004), pp. 87–93.
- [Gli+24] Kristina Gligorić, Tijana Zrnic, Cino Lee, Emmanuel J Candès, and Dan Jurafsky. “Can unconfident LLM annotations be used for confident conclusions?” In: *arXiv preprint arXiv:2408.15204* (2024).
- [GMA22] Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. “Demographic-aware language model fine-tuning as a bias mitigation technique”. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2022, pp. 311–319.
- [GMT24] Jorge García-Carrasco, Alejandro Maté, and Juan Carlos Trujillo. “How does GPT-2 predict acronyms? Extracting and understanding a circuit via mechanistic interpretability”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3322–3330.
- [Gol+24] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. “Contextual position encoding: Learning to count what’s important”. In: *arXiv preprint arXiv:2405.18719* (2024).

- [GPT23] GPTZero. *GPTZero: More than an AI detector preserve what's human*. 2023. URL: <https://gptzero.me/>.
- [Gra+24] Aaron Grattafiori et al. “The Llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024).
- [GRK21] Demi Guo, Alexander Rush, and Yoon Kim. “Parameter-efficient transfer learning with diff pruning”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021, pp. 4884–4896.
- [GSH23] Leo Gao, John Schulman, and Jacob Hilton. “Scaling laws for reward model overoptimization”. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML’23. JMLR.org, 2023.
- [GSR19] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. “GLTR: Statistical detection and visualization of generated text”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 2019, pp. 111–116.
- [Gue+24] Marco Guevara et al. “Large language models to identify social determinants of health in electronic health records”. In: *npj Digital Medicine* 7 (Jan. 2024), p. 6.
- [Guo+20] Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. “Certified data removal from machine learning models”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3832–3842.
- [Guo+24a] Shangmin Guo et al. “Direct language model alignment from online AI feedback”. In: *arXiv preprint arXiv:2402.04792* (2024).
- [Guo+24b] Tianyu Guo et al. “Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs”. In: *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*. 2024.
- [Guo+24c] Tianyu Guo et al. “How do transformers learn in-context beyond simple functions? A case study on learning with representations”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Gur+23] Wes Gurnee et al. “Finding neurons in a haystack: Case studies with sparse probing”. In: *Trans. Mach. Learn. Res.* 2023 (2023).

- [Hal+23] Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. “Detoxifying text with MaRCO: Controllable revision with experts and anti-experts”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. July 2023, pp. 228–242.
- [Han+22] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. “PTR: Prompt tuning with rules for text classification”. In: *AI Open* 3 (2022), pp. 182–192.
- [Han+23] Tianyu Han et al. “MedAlpaca – An open-source collection of medical conversational AI models and training data”. In: *arXiv preprint arXiv:2304.08247* (Apr. 2023).
- [Han+24a] Sungwon Han, Jinsung Yoon, Sercan Ö. Arik, and Tomas Pfister. “Large language models can automatically engineer features for few-shot tabular learning”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Han+24b] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. “Parameter-efficient fine-tuning for large models: A comprehensive survey”. In: *arXiv preprint arXiv:2403.14608* (2024).
- [Har54] Zellig Harris. “Distributional structure”. In: *Word* 10.2-3 (1954), pp. 146–162.
- [HBC21] Xudong Han, Timothy Baldwin, and Trevor Cohn. “Diverse adversaries for mitigating bias in training”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Apr. 2021, pp. 2760–2765.
- [HBC22] Xudong Han, Timothy Baldwin, and Trevor Cohn. “Balancing out bias: Achieving fairness through balanced training”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 11335–11350.
- [He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.
- [He+22a] Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. “MABEL: Attenuating gender bias using textual entailment data”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 9681–9702.

- [He+22b] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. “Towards a unified view of parameter-efficient transfer learning”. In: *International Conference on Learning Representations*. 2022.
- [He+24a] Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. “Universally optimal watermarking schemes for LLMs: from theory to practice”. In: *arXiv preprint arXiv:2410.02890* (2024).
- [He+24b] Hengzhi He, Peiyu Yu, Junpeng Ren, Ying Nian Wu, and Guang Cheng. “Watermarking generative tabular data”. In: *arXiv preprint arXiv:2405.14018* (2024).
- [Hen+21a] Dan Hendrycks et al. “Measuring massive multitask language understanding”. In: *9th International Conference on Learning Representations, ICLR*. 2021.
- [Hen+21b] Dan Hendrycks et al. “Measuring mathematical problem solving with the MATH dataset”. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*. 2021.
- [HH24] Zhengmian Hu and Heng Huang. “Inevitable trade-off between watermark strength and speculative sampling efficiency for language models”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 55370–55402.
- [HJ23] Stefan Heimersheim and Jett Janiak. *A circuit for Python docstrings in a 4-layer attention-only transformer*. 2023. URL: <https://www.%20alignmentforum.%20org/posts/u6KXXmKFbXfWzoAXn/acircuit-for-python-docstrings-in-a-4-layer-attention-only>.
- [HMH24] Noah Hollmann, Samuel Müller, and Frank Hutter. “Large language models for automated data science: Introducing caafe for context-aware automated feature engineering”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Hof+22] Jordan Hoffmann et al. “Training compute-optimal large language models”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Red Hook, NY, USA: Curran Associates Inc., 2022. ISBN: 9781713871088.
- [Hof01] Thomas Hofmann. “Unsupervised learning by probabilistic latent semantic analysis”. In: *Mach. Learn.* 42.1–2 (2001), pp. 177–196. ISSN: 0885-6125.

- [Hoo+21] Shlomo Hoory et al. “Learning and evaluating a differentially private pre-trained language model”. In: *Findings of the Association for Computational Linguistics: EMNLP*. 2021, pp. 1178–1189.
- [Hou+19] Neil Houlsby et al. “Parameter-efficient transfer learning for NLP”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 2790–2799.
- [HS23] Hangfeng He and Weijie J Su. “A law of data separation in deep learning”. In: *Proceedings of the National Academy of Sciences* 120.36 (2023), e2221704120.
- [HS24] Hangfeng He and Weijie J Su. “A law of next-token prediction in large language models”. In: *arXiv preprint arXiv:2408.13442* (2024).
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural Comput.* 9.8 (1997), pp. 1735–1780. ISSN: 0899-7667.
- [Hu+24] Zhengmian Hu et al. “Unbiased watermark for large language models”. In: *International Conference on Learning Representations*. 2024.
- [Hua+20] Po-Sen Huang et al. “Reducing sentiment bias in language models via counterfactual evaluation”. In: *Findings of the Association for Computational Linguistics: EMNLP*. Vol. EMNLP 2020. Findings of ACL. 2020, pp. 65–83.
- [Hua+23a] Baihe Huang et al. “Towards optimal statistical watermarking”. In: *arXiv preprint arXiv:2312.07930* (2023).
- [Hua+23b] Jonathan Huang et al. “Generative artificial intelligence for chest radiograph interpretation in the emergency department”. In: *JAMA network open* 6.10 (2023), e2336100–e2336100.
- [Hua+23c] Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. “Privacy implications of retrieval-based language models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 14887–14902.
- [Hua+24a] Wei Huang et al. “BiLLM: Pushing the limit of post-training quantization for LLMs”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Hua+24b] Zhen Huang et al. “O1 replication journey – Part 2: surpassing o1-preview through simple distillation, big progress or bitter lesson?” In: *arXiv preprint arXiv:2411.16489* (2024).

- [Hub+24] Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. “Sparse autoencoders find highly interpretable features in language models”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Huo+24] Mingjia Huo et al. “Token-specific watermarking with enhanced detectability and semantic coherence for large language models”. In: *International Conference on Machine Learning*. 2024.
- [Ipp+20] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. “Automatic detection of generated text is easiest when humans are fooled”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*. 2020, pp. 1808–1822.
- [Jan+23] Joel Jang et al. “Knowledge unlearning for mitigating privacy risks in language models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. 2023, pp. 14389–14408.
- [Jaq+17] Natasha Jaques et al. “Sequence tutor: conservative fine-tuning of sequence generation models with KL-control”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML’17. JMLR.org*, 2017, pp. 1645–1654.
- [Jaq+20] Natasha Jaques et al. “Human-centric dialog training via offline reinforcement learning”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 2020, pp. 3985–4003.
- [Jel98] Frederick Jelinek. *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262100665.
- [Ji+22] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. “An unconstrained layer-peeled perspective on neural collapse”. In: *The Tenth International Conference on Learning Representations, ICLR*. 2022.
- [Ji+23] Ziwei Ji et al. “Survey of hallucination in natural language generation”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.
- [Jia+20] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. “How can we know what language models know”. In: *Trans. Assoc. Comput. Linguistics* 8 (2020), pp. 423–438.
- [Jia+23] Albert Q. Jiang et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).

- [Jia+24a] Jinghan Jia et al. “SOUL: Unlocking the power of second-order optimization for LLM unlearning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 2024, pp. 4276–4292.
- [Jia+24b] Albert Q. Jiang et al. “Mixtral of experts”. In: *arXiv preprint arXiv:2401.04088* (2024).
- [Jin+21] Xisen Jin et al. “On transferability of bias mitigation effects in language model fine-tuning”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 2021, pp. 3770–3783.
- [JJS24] Meena Jagadeesan, Michael I Jordan, and Jacob Steinhardt. “Safety vs. performance: How multi-objective learning reduces barriers to market entry”. In: *arXiv preprint arXiv:2409.03734* (2024).
- [JLZ25] Wenlong Ji, Lihua Lei, and Tijana Zrnic. “Predictions as surrogates: Revisiting surrogate outcomes in the age of AI”. In: *arXiv preprint arXiv:2501.09731* (2025).
- [JMS24] Ayush Jain, Andrea Montanari, and Eren Sasoglu. “Scaling laws for learning with real and surrogate data”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [Joa98] Thorsten Joachims. “Text categorization with support vector machines: Learning with many relevant features”. In: *European conference on machine learning*. Springer. 1998, pp. 137–142.
- [Jor86] Michael I. Jordan. “Attractor dynamics and parallelism in a connectionist sequential machine”. In: *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. 1986.
- [JSV24] Nikola Jovanović, Robin Staab, and Martin Vechev. “Watermark stealing in large language models”. In: *International Conference on Machine Learning*. 2024.
- [Kad+22] Saurav Kadavath et al. “Language models (mostly) know what they know”. In: *arXiv preprint arXiv:2207.05221* (2022).
- [Kan+23] Beichen Kang et al. “Automatic ICD coding based on segmented ClinicalBERT with hierarchical tree structure learning”. In: *Database Systems for Advanced Applications: 28th International Conference, DAS-FAA*. Springer, 2023, pp. 250–265.

- [Kap+20] Jared Kaplan et al. “Scaling laws for neural language models.” In: *CoRR* abs/2001.08361 (2020).
- [Kau+24] Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. “A theoretical framework for partially observed reward-states in RLHF”. In: *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*. 2024.
- [KB21] Masahiro Kaneko and Danushka Bollegala. “Debiasing pre-trained contextualised embeddings”. In: *arXiv preprint arXiv:2101.09523* (2021).
- [KDS23] Hadas Kotek, Rikker Dockum, and David Sun. “Gender bias and stereotypes in large language models”. In: *Proceedings of The ACM Collective Intelligence Conference*. 2023, pp. 12–24.
- [Ke+23] Zheng Tracy Ke, Pengsheng Ji, Jiashun Jin, and Wanshan Li. “Recent advances in text analysis”. In: *Annual Review of Statistics and Its Application* 11 (2023).
- [Ken+21] Zachary Kenton et al. “Alignment of language agents”. In: *arXiv preprint arXiv:2103.14659* (2021).
- [KGF23] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. “Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.
- [Kim+23] Minbeom Kim et al. “Critic-guided decoding for controlled text generation”. In: *Findings of the Association for Computational Linguistics: ACL*. 2023, pp. 4598–4612.
- [Kim14] Yoon Kim. “Convolutional neural networks for sentence classification”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1746–1751.
- [Kir+23a] John Kirchenbauer et al. “A watermark for large language models”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 17061–17084.
- [Kir+23b] John Kirchenbauer et al. “On the reliability of watermarks for large language models”. In: *The Twelfth International Conference on Learning Representations*. 2023.
- [KKO24] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. “Outfox: LLM-generated essay detection through in-context learning with adversarially generated examples”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 19. 2024, pp. 21258–21266.

- [Koc+23] Denis Kocetkov et al. “The Stack: 3 TB of permissively licensed source code”. In: *Trans. Mach. Learn. Res.* 2023 (2023).
- [Köp+24] Andreas Köpf et al. “OpenAssistant conversations - democratizing large language model alignment”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Curran Associates Inc., 2024.
- [Kri+24] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.
- [KS24] Tokio Kajitsuka and Issei Sato. “Are transformers with one layer self-attention using low-rank weight matrices universal approximators?” In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Kud+24] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. “Robust distortion-free watermarks for language models”. In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856.
- [Kum+23] Bhawesh Kumar et al. “Conformal prediction with large language models for multi-choice question answering”. In: *arXiv preprint arXiv:2305.18404* (2023).
- [KW24] Zheng Tracy Ke and Minzhe Wang. “Using SVD for topic modeling”. In: *Journal of the American Statistical Association* 119.545 (2024), pp. 434–449.
- [Kwe+23] Sunjun Kweon, Junu Kim, Jiyou Kim, Sujeong Im, et al. “Publicly shareable clinical large language model built on synthetic clinical notes”. In: *arXiv preprint* (Sept. 2023).
- [KX15] Vladimir Koltchinskii and Dong Xia. “Optimal estimation of low rank density matrices.” In: *J. Mach. Learn. Res.* 16.53 (2015), pp. 1757–1792.
- [Lam+24] Nathan Lambert et al. “RewardBench: Evaluating reward models for language modeling”. In: *arXiv preprint arXiv:2403.13787* (2024).
- [Lan+20] Zhenzhong Lan et al. “ALBERT: A lite BERT for self-supervised learning of language representations.” In: *ICLR*. 2020.

- [LBM24] Licong Lin, Yu Bai, and Song Mei. “Transformers as decision makers: provable in-context reinforcement learning via supervised pretraining”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [LD02] Tom Landauer and Scott Dooley. “Latent semantic analysis: theory, method and application”. In: *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*. CSCL ’02. International Society of the Learning Sciences, 2002, pp. 742–743.
- [Lew+20] Mike Lewis et al. “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880.
- [Li+21] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. “Large language models can be strong differentially private learners”. In: *International Conference on Learning Representations*. 2021.
- [Li+23] Xuechen Li et al. *AlpacaEval: An automatic evaluator of instruction-following models*. 2023. URL: https://github.com/tatsu-lab/alpaca_eval.
- [Li+24a] Chunyuan Li et al. “Llava-med: Training a large language-and-vision assistant for biomedicine in one day”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Li+24b] Junlong Li et al. “Generative judge for evaluating alignment”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Li+24c] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. “Robust detection of watermarks in large language models under human edits”. In: *arXiv preprint arXiv:2411.13868* (2024).
- [Li+25] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. “A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules”. In: *The Annals of Statistics* 53.1 (2025), pp. 322–351.
- [Lia+23] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. “GPT detectors are biased against non-native English writers”. In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023.

- [Lim+24] Yannick Limmer, Anastasis Kratsios, Xuwei Yang, Raeid Saqur, and Blanka Horvath. “Reality only happens once: Single-path generalization bounds for transformers”. In: *arXiv preprint arXiv:2405.16563* (2024).
- [Lin+24a] Licong Lin, Jingfeng Wu, Sham Kakade, Peter Bartlett, and Jason D Lee. “Scaling laws in linear regression: compute, parameters, and data”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 60556–60606.
- [Lin+24b] Zhenghao Lin et al. “Rho-1: Not all tokens are what you need”. In: *arXiv preprint arXiv:2404.07965* (2024).
- [Lis+21] Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. “Anonymisation models for text data: State of the art, challenges and future directions”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 4188–4203.
- [Liu+19] Yinhan Liu et al. “RoBERTa: A robustly optimized BERT pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [Liu+20a] Haochen Liu et al. “Does gender matter? Towards fairness in dialogue systems”. In: *Proceedings of the 28th International Conference on Computational Linguistics, COLING*. International Committee on Computational Linguistics, 2020, pp. 4403–4416.
- [Liu+20b] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. “FinBERT: A pre-trained financial language representation model for financial text mining”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Special Track on AI in FinTech. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 4513–4519.
- [Liu+21] Alisa Liu et al. “DExperts: Decoding-time controlled text generation with experts and anti-experts”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021, pp. 6691–6706.
- [Liu+22] Jiachang Liu et al. “What makes good in-context examples for GPT-3?” In: *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022*. 2022, pp. 100–114.

- [Liu+23] Pengfei Liu et al. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. In: *ACM Comput. Surv.* 55.9 (2023), 195:1–195:35.
- [Liu+24a] Aixin Liu et al. “DeepSeek-v2: A strong, economical, and efficient mixture-of-experts language model”. In: *arXiv preprint arXiv:2405.04434* (2024).
- [Liu+24b] Aixin Liu et al. “DeepSeek-v3 technical report”. In: *arXiv preprint arXiv:2412.19437* (2024).
- [Liu+24c] Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. “Uncertainty estimation and quantification for LLMs: A simple supervised approach”. In: *arXiv preprint arXiv:2404.15993* (2024).
- [Liu+24d] Xiao Liu et al. “Are LLMs capable of data-based statistical and causal reasoning? Benchmarking advanced quantitative reasoning with data”. In: *Findings of the Association for Computational Linguistics: ACL*. Aug. 2024, pp. 9215–9235.
- [Liu+24e] Xiao Liu et al. “GPT understands, too”. In: *AI Open* 5 (2024), pp. 208–215.
- [Liu+24f] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. “Datasets for large language models: A comprehensive survey”. In: *arXiv preprint arXiv:2402.18041* (2024).
- [Liu+24g] Zhiyuan Liu, Hong Liu, Denny Zhou, and Tengyu Ma. “Chain of thought empowers transformers to solve inherently serial problems”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Liu+25] Sijia Liu et al. “Rethinking machine unlearning for large language models”. In: *Nature Machine Intelligence* (Feb. 2025).
- [LJ24] Maximilian Li and Lucas Janson. “Optimal ablation for interpretability”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [LKW23] Xin Liu, Muhammad Khalifa, and Lu Wang. “BOLT: Fast energy-based controlled text generation with tunable biases”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. July 2023, pp. 186–200.

- [LL21] Xiang Lisa Li and Percy Liang. “Prefix-tuning: optimizing continuous prompts for generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021, pp. 4582–4597.
- [LLS22] Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. “Towards collaborative neural-symbolic graph semantic parsing via uncertainty”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. 2022, pp. 4160–4173.
- [LLZ24] Xingchi Li, Guanxun Li, and Xianyang Zhang. “Segmenting water-marked texts from language models”. In: *arXiv preprint arXiv:2410.20670* (2024).
- [LMM23] Baohao Liao, Yan Meng, and Christof Monz. “Parameter-efficient fine-tuning without introducing new latency”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 2023, pp. 4242–4260.
- [LT20] Wei-Chao Lin and Chih-Fong Tsai. “Missing value imputation: a review and analysis of the literature (2006–2017)”. In: *Artificial Intelligence Review* 53 (2020), pp. 1487–1509.
- [LTS24] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. “Generating with confidence: Uncertainty quantification for black-box large language models”. In: *arXiv preprint arXiv:2305.19187* (2024).
- [Lu+20] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. “Gender bias in neural natural language processing”. In: *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday* (2020), pp. 189–202.
- [Lu+22] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*. 2022, pp. 8086–8098.
- [Lu+23] Dakuan Lu et al. “BBT-Fin: Comprehensive construction of Chinese financial domain pre-trained language model, corpus and benchmark”. In: *arXiv preprint arXiv:2302.09432* (2023).
- [Men+22] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. “Locating and editing factual associations in GPT”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17359–17372.

- [MG21] Andrey Malinin and Mark J. F. Gales. “Uncertainty estimation in autoregressive structured prediction”. In: *9th International Conference on Learning Representations, ICLR*. 2021.
- [MH24] Christopher Mohri and Tatsunori Hashimoto. “Language models with conformal factuality guarantees”. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. JMLR.org, 2024.
- [Mik+10] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. “Recurrent neural network based language model.” In: *INTERSPEECH*. ISCA, 2010, pp. 1045–1048.
- [Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. In: vol. 26. 2013.
- [Min+22a] Sewon Min et al. “Rethinking the role of demonstrations: What makes in-context learning work?” In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 11048–11064.
- [Min+22b] Sören Mindermann et al. “Prioritized training on points that are learnable, worth learning, and not yet learnt”. In: *International Conference on Machine Learning, ICML*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 15630–15649.
- [Mis+22] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. “Cross-task generalization via natural language crowdsourcing instructions”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. May 2022, pp. 3470–3487.
- [Mit+23] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. “Detectgpt: Zero-shot machine-generated text detection using probability curvature”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 24950–24962.
- [MML23] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. “Large language models challenge the future of higher education”. In: *Nature Machine Intelligence* 5.4 (2023), pp. 333–334.
- [Moo+23] Michael Moor et al. “Med-flamingo: a multimodal medical few-shot learner”. In: *Machine Learning for Health (ML4H)*. PMLR. 2023, pp. 353–367.

- [Mue+23] Niklas Muennighoff et al. “Scaling data-constrained language models”. In: *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. 2023.
- [Mur+19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. “Definitions, methods, and applications in interpretable machine learning”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.
- [MVM09] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. *Elo rating system*. Alpha Press, 2009. ISBN: 6130078846.
- [MXC24] Yu Meng, Mengzhou Xia, and Danqi Chen. “SimPO: Simple preference optimization with a reference-free reward”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 124198–124235.
- [Nak+22] Reiichiro Nakano et al. “WebGPT: Browser-assisted question-answering with human feedback”. In: *arXiv preprint arXiv:2112.09332* (2022).
- [Nak+24] Ryumei Nakada, Yichen Xu, Lexin Li, and Linjun Zhang. “Synthetic oversampling: Theory and a practical approach using LLMs to address data imbalance”. In: *arXiv preprint arXiv:2406.03628* (2024).
- [Nas+23] Milad Nasr et al. “Scalable extraction of training data from (production) language models”. In: *arXiv preprint arXiv:2311.17035* (2023).
- [Nav+23] Humza Naveed et al. “A comprehensive overview of large language models”. In: *arXiv preprint arXiv:2307.06435* (2023).
- [NCR23] Roberto Navigli, Simone Conia, and Björn Ross. “Biases in large language models: origins, inventory, and discussion”. In: *ACM Journal of Data and Information Quality* 15.2 (2023), pp. 1–21.
- [Nie+24] Fan Nie et al. “FactTest: Factuality Testing in Large Language Models with Statistical Guarantees”. In: *arXiv preprint arXiv:2411.02603* (2024).
- [NRS21] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. “Descent-to-delete: Gradient-based methods for machine unlearning”. In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 931–962.
- [Oh+22] Changdae Oh et al. “Learning fair representation via distributional contrastive disentanglement”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 1295–1305.

- [Ola+20] Chris Olah et al. “Zoom in: An introduction to circuits”. In: *Distill* 5.3 (2020), e00024–001.
- [Ols+22] Catherine Olsson et al. “In-context learning and induction heads”. In: *arXiv preprint arXiv:2209.11895* (2022).
- [Omi+23] Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. “Large language models propagate race-based medicine”. In: *NPJ Digital Medicine* 6.1 (2023), p. 195.
- [Ope+24a] OpenAI et al. “GPT-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2024).
- [Ope+24b] OpenAI et al. “OpenAI o1 system card”. In: *arXiv preprint arXiv:2412.16720* (2024).
- [Ouy+22] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.
- [OZ03] Stanley RM Oliveira and Osmar R Zaiane. “Protecting sensitive knowledge by data sanitization”. In: *Third IEEE International conference on data mining*. IEEE. 2003, pp. 613–616.
- [Pal+24] Arka Pal et al. “Smaug: Fixing failure modes of preference optimization with DPO-positive”. In: *arXiv preprint arXiv:2402.13228* (2024).
- [Pan+24] Richard Yuanzhe Pang et al. “Iterative reasoning preference optimization”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 116617–116637.
- [Par+23] SunYoung Park, Kyuri Choi, Haeun Yu, and Youngjoong Ko. “Never too late to learn: Regularizing gender bias in coreference resolution”. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 2023, pp. 15–23.
- [Par+24a] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. “Disentangling length from quality in direct preference optimization”. In: *Findings of the Association for Computational Linguistics: ACL*. 2024, pp. 4998–5017.
- [Par+24b] Yeonhong Park, Jake Hyun, SangLyul Cho, Bonggeun Sim, and Jae W. Lee. “Any-precision LLM: Low-cost deployment of multiple, different-sized LLMs”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.

- [PCJ25] Mose Park, Yunjin Choi, and Jong-June Jeon. “Does a large language model really speak in human-like language?” In: *arXiv preprint arXiv:2501.01273* (2025).
- [Pen+23] Cheng Peng et al. “A study of generative large language model for medical research and healthcare”. In: *npj Digital Medicine* 6 (Nov. 2023), p. 210.
- [Pen+24] Guilherme Penedo et al. “The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Curran Associates Inc., 2024.
- [Per+20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. “Performative prediction”. In: *Proceedings of the 37th International Conference on Machine Learning*. 2020, pp. 7599–7609.
- [Pet+18] Matthew E. Peters et al. “Deep contextualized word representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. June 2018, pp. 2227–2237.
- [PFB24] Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. “Bias patterns in the application of LLMs for clinical decision support: A comprehensive study”. In: *arXiv preprint arXiv:2404.15149* (2024).
- [PHB24] Vaidehi Patil, Peter Hase, and Mohit Bansal. “Can sensitive information be deleted from LLMs? Objectives for defending against extraction attacks”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [PHD20] Vardan Papayan, XY Han, and David L Donoho. “Prevalence of neural collapse during the terminal phase of deep learning training”. In: *Proceedings of the National Academy of Sciences* 117.40 (2020), pp. 24652–24663.
- [PMB19] Jorge Pérez, Javier Marinkovic, and Pablo Barceló. “On the Turing completeness of modern neural network architectures”. In: *7th International Conference on Learning Representations, ICLR*. 2019.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global vectors for word representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Oct. 2014, pp. 1532–1543.
- [Qin+24] Yiwei Qin et al. “O1 replication journey: A strategic progress report – Part 1”. In: *arXiv preprint arXiv:2410.18982* (2024).

- [Rad+18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving language understanding by generative pre-training*. 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [Rad+19] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI* (2019).
- [Rae+21] Jack W. Rae et al. “Scaling language models: Methods, analysis & insights from training Gopher”. In: *CoRR* abs/2112.11446 (2021).
- [Raf+20] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.
- [Raf+23] Rafael Rafailov et al. “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 53728–53741.
- [Rai+24] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. “A practical review of mechanistic interpretability for transformer-based language models”. In: *arXiv preprint arXiv:2407.02646* (2024).
- [Ram+17] Rachel B Ramoni et al. “The undiagnosed diseases network: Accelerating discovery about health and disease”. In: *American Journal of Human Genetics* 100.2 (2017), pp. 185–192.
- [Ran+24] Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. “A tree-of-thoughts to broaden multi-step reasoning across languages”. In: *Findings of the Association for Computational Linguistics: NAACL*. June 2024, pp. 1229–1241.
- [Ren+23] Allen Z. Ren et al. “Robots that ask for help: Uncertainty alignment for large language model planners”. In: *Conference on Robot Learning, CoRL*. Vol. 229. Proceedings of Machine Learning Research. PMLR, 2023, pp. 661–682.
- [Ren+24] Jie Ren et al. “Copyright protection in generative AI: A technical perspective”. In: *arXiv preprint arXiv:2402.02333* (2024).
- [RGG23] Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. “Conformal Nucleus Sampling”. In: *Findings of the Association for Computational Linguistics: ACL*. July 2023, pp. 27–34.

- [RJS17] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. “Learning to generate reviews and discovering sentiment”. In: *arXiv preprint arXiv:1704.01444* (2017).
- [Rol+21] Stephen Roller et al. “Recipes for building an open-domain chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Apr. 2021, pp. 300–325.
- [Ros00] R. Rosenfeld. “Two decades of statistical language modeling: where do we go from here?” In: *Proceedings of the IEEE* 88.8 (2000), pp. 1270–1278.
- [RSD23] Vipula Rawte, Amit Sheth, and Amitava Das. “A survey of hallucination in large foundation models”. In: *arXiv preprint arXiv:2309.05922* (2023).
- [RSS23] Pedro L Rodriguez, Arthur Spirling, and Brandon M Stewart. “Embedding regression: Models for context-specific description and inference”. In: *American Political Science Review* 117.4 (2023), pp. 1255–1274.
- [Sac+24] Noveen Sachdeva et al. “How to train data-efficient LLMs”. In: *arXiv preprint arXiv:2402.09668* (2024).
- [Sad+23] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. “Can AI-generated text be reliably detected?” In: *arXiv preprint arXiv:2303.11156* (2023).
- [Sag23] Matthew Sag. *Copyright safety for generative AI*. 2023. URL: <https://houstonlawreview.org/article/92126>.
- [Sah+24] Swarnadeep Saha et al. “Branch-Solve-Merge improves large language model evaluation and generation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 8352–8370.
- [Sam23] Pamela Samuelson. “Generative AI meets copyright”. In: *Science* 381.6654 (2023), pp. 158–161.
- [San+22] Victor Sanh et al. “Multitask prompted training enables zero-shot task generalization”. In: *The Tenth International Conference on Learning Representations, ICLR*. 2022.

- [San+23] Shibani Santurkar et al. “Whose opinions do language models reflect?” In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29971–30004.
- [Sax+19] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. “Analysing mathematical reasoning abilities of neural models”. In: *7th International Conference on Learning Representations, ICLR*. 2019.
- [SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. A Bradford Book, 2018. ISBN: 0262039249.
- [SBM22] Lee Sharkey, Dan Braun, and Beren Millidge. “Taking features out of superposition with sparse autoencoders”. In: *AI Alignment Forum*. Vol. 6. 2022, pp. 12–13.
- [Sch+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).
- [Sch+21] Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. “Consistent accelerated inference via confident adaptive transformers”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Nov. 2021, pp. 4962–4979.
- [Sch+22a] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. “Large pre-trained language models contain human-like biases of what is right and wrong to do”. In: *Nature Machine Intelligence* 4.3 (2022), pp. 258–268.
- [Sch+22b] Tal Schuster et al. “Confident adaptive language modeling”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17456–17472.
- [SCN18] Karan Singla, Dogan Can, and Shrikanth Narayanan. “A multi-task approach to learning multilingual representations”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. July 2018, pp. 214–220.
- [SDG06] Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gauvain. “Continuous space language models for statistical machine translation”. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. 2006, pp. 723–730.
- [Sek+21] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. “Remember what you want to forget: Algorithms for machine unlearning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18075–18086.

- [Sha+24] Vasu Sharma et al. “Text quality-based pruning for efficient training of language models”. In: *arXiv preprint arXiv:2405.01582* (2024).
- [Shi+24] Claudia Shi et al. “Hypothesis testing the circuit hypothesis in LLMs”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [Shu+22] Kurt Shuster et al. “Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage”. In: *arXiv preprint arXiv:2208.03188* (2022).
- [Shu+23] Ilia Shumailov et al. “The curse of recursion: Training on generated data makes models forget”. In: *arXiv preprint arXiv:2305.17493* (2023).
- [Sim+24] Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. “Constructing benchmarks and interventions for combating hallucinations in LLMs”. In: *arXiv preprint arXiv:2404.09971* (2024).
- [Sin+23] Karan Singhal et al. “Towards expert-level medical question answering with large language models”. In: *arXiv preprint arXiv:2305.09617* (2023).
- [Sin+24] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. “A long way to go: Investigating length correlations in RLHF”. In: *arXiv preprint arXiv:2310.03716* (2024).
- [SK21] S Selva Birunda and R Kanniga Devi. “A review on word embedding techniques for text classification”. In: *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020* (2021), pp. 267–281.
- [SLG20] Budhaditya Saha, Sanal Lisboa, and Shameek Ghosh. “Understanding patient complaint characteristics using contextual clinical BERT embeddings”. In: *arXiv preprint arXiv:2002.05902* (2020).
- [Slo+23] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. “The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 3607–3625.
- [Smi+22] Shaden Smith et al. “Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, A large-scale generative language model”. In: *arXiv preprint arXiv:2201.11990* (2022).

- [Sne+24] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. “Scaling LLM test-time compute optimally can be more effective than scaling model parameters”. In: *arXiv preprint arXiv:2408.03314* (2024).
- [SNR21] Yi-Lin Sung, Varun Nair, and Colin A Raffel. “Training neural networks with fixed sparse masks”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 24193–24205.
- [Sol+24] Luca Soldaini et al. “Dolma: An open corpus of three trillion tokens for language model pretraining research”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 15725–15788.
- [Son+23] Ziang Song, Tianle Cai, Jason D Lee, and Weijie J Su. “Reward collapse in aligning large language models”. In: *arXiv preprint arXiv:2305.17608* (2023).
- [Spa72] Karen Sparck Jones. “A statistical interpretation of term specificity and its application in retrieval”. In: *Journal of documentation* 28.1 (1972), pp. 11–21.
- [SSB22] Danielle Saunders, Rosie Sallis, and Bill Byrne. “First the worst: Finding better gender translations during beam search”. In: *Findings of the Association for Computational Linguistics: ACL*. May 2022, pp. 3814–3823.
- [Sti+20] Nisan Stiennon et al. “Learning to summarize from human feedback”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Curran Associates Inc., 2020. ISBN: 9781713829546.
- [Su+24] Jianlin Su et al. “RoFormer: Enhanced transformer with rotary position embedding”. In: *Neurocomputing* 568 (2024), p. 127063.
- [Sun+20] Yu Sun et al. “ERNIE 2.0: A continual pre-training framework for language understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8968–8975.
- [SWF+15] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. “End-to-end memory networks”. In: *Advances in neural information processing systems* 28 (2015).
- [SZ23] Karanpartap Singh and James Zou. “New evaluation metrics capture quality degradation due to LLM watermarking”. In: *Transactions on Machine Learning Research* (2023).

- [Szo24] Peter Szolovits. “Large language models seem miraculous, but science abhors miracles”. In: *NEJM AI* (May 2024).
- [Tan+23] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. “Does synthetic data generation of LLMs help clinical text mining?” In: *arXiv preprint arXiv:2303.04360* (2023).
- [Tan+24a] Sijun Tan et al. “JudgeBench: A benchmark for evaluating LLM-based judges”. In: *arXiv preprint arXiv:2410.12784* (2024).
- [Tan+24b] Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. “Math-Scale: Scaling instruction tuning for mathematical reasoning”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Tay+23] Yi Tay et al. “Scaling laws vs model architectures: How does inductive bias influence scaling?” In: *Findings of the Association for Computational Linguistics: EMNLP*. 2023, pp. 12342–12364.
- [Tea+25] Kimi Team et al. “Kimi k1.5: Scaling reinforcement learning with LLMs”. In: *arXiv preprint arXiv:2405.02225* (2025).
- [Tes+24] Idit Tessler et al. “Unlocking the potential of large language models in healthcare: navigating the opportunities and challenges”. In: *Future Medicine AI* 2.1 (Apr. 2024).
- [TGH23] Hoang Tran, Chris Glaze, and Braden Hancock. *Iterative DPO alignment*. Tech. rep. Snorkel AI, 2023.
- [Tho+22] Romal Thoppilan et al. “LaMDA: Language models for dialog applications”. In: *arXiv preprint arXiv:2201.08239* (2022).
- [Thr+22] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. “Imbalance trouble: Revisiting neural-collapse geometry”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27225–27238.
- [Tim23] The New York Times. *The times sues OpenAI and Microsoft over AI use of copyrighted work*. America. Dec. 2023. URL: <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>.
- [Tou+23a] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [Tou+23b] Hugo Touvron et al. “LLaMA: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

- [TRB10] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. “Word representations: A simple and general method for semi-supervised learning”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 2010, pp. 384–394.
- [Tsa+19] Yao-Hung Hubert Tsai et al. “Multimodal transformer for unaligned multimodal language sequences”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 2019, pp. 6558–6569.
- [TT24] Jacob Trauger and Ambuj Tewari. “Sequence length independent norm-based generalization bounds for transformers”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 1405–1413.
- [Tu+24] Xinming Tu, James Zou, Weijie Su, and Linjun Zhang. “What should data science education do with large language models?” In: *Harvard Data Science Review* 6.1 (2024).
- [Tul+24] Eduard Tulchinskii et al. “Intrinsic dimension estimation for robust detection of AI-generated texts”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.
- [Tyl23] Thaddée Tyl. *Chinchilla’s death*. Accessed: 2022-05-24. 2023. URL: <https://espadrine.github.io/blog/posts/chinchilla-s-death.html>.
- [TZL24] Brian Tufts, Xuandong Zhao, and Lei Li. “An examination of AI-generated text detectors across multiple domains and models”. In: *Neurips Safe Generative AI Workshop 2024*. 2024.
- [UZM24] Dennis Ulmer, Chrysoula Zerva, and Andre Martins. “Non-exchangeable conformal language generation with nearest neighbors”. In: *Findings of the Association for Computational Linguistics: EACL*. Mar. 2024, pp. 1909–1929.
- [Vak+22] Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. “Downstream task performance of bert models pre-trained using automatically de-identified clinical data”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022, pp. 4245–4252.
- [Val+23] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. “DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation”. In: *Proceedings of the 17th*

- Conference of the European Chapter of the Association for Computational Linguistics*. May 2023, pp. 3274–3287.
- [Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [Ver+23] Shreyas Verma, Kien Tran, Yusuf Ali, and Guangyu Min. “Reducing LLM hallucinations using epistemic neural networks”. In: *arXiv preprint arXiv:2312.15576* (2023).
- [VGS05] Vladimir Vovk, Alexander Gammernan, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.
- [Vin22] James Vincent. “AI-generated answers temporarily banned on coding Q&A site Stack Overflow”. In: *The Verge* 5 (2022).
- [VKB23] Nikhil Vyas, Sham M Kakade, and Boaz Barak. “On provable copyright protection for generative models”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 35277–35299.
- [Von+23] Johannes Von Oswald et al. “Transformers learn in-context by gradient descent”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 35151–35174.
- [Wal+19] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. “Universal adversarial triggers for attacking and analyzing NLP”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*. 2019, pp. 2153–2162.
- [Wan+22] Xiaozhi Wang et al. “Finding skill neurons in pre-trained transformer-based language models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 11132–11152.
- [Wan+23a] Yixin Wan et al. “Kelly is a warm person, Joseph is a role model: Gender Biases in LLM-Generated Reference Letters”. In: *Findings of the Association for Computational Linguistics: EMNLP*. Dec. 2023, pp. 3730–3748.
- [Wan+23b] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. “Interpretability in the wild: a circuit for indirect object identification in GPT-2 Small”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.

- [Wan+23c] Tianlu Wang et al. “Shepherd: A critic for language model generation”. In: *arXiv preprint arXiv:2308.04592* (2023).
- [Wan+23d] Xuezhi Wang et al. “Self-consistency improves chain of thought reasoning in language models”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.
- [Wan+23e] Yizhong Wang et al. “Self-instruct: Aligning language models with self-generated instructions”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 13484–13508.
- [Wan+24a] Jiachen T Wang, Zhun Deng, Hiroaki Chiba-Okabe, Boaz Barak, and Weijie J Su. “An economic solution to copyright challenges of generative AI”. In: *arXiv preprint arXiv:2404.13964* (2024).
- [Wan+24b] Yidong Wang et al. “PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization”. In: *International Conference on Learning Representations (ICLR)*. 2024.
- [Wan+24c] Zifeng Wang et al. “CodecLM: Aligning Language Models with Tailored Synthetic Data”. In: 2024.
- [Wan+24d] Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. “Transformers provably learn sparse token selection while fully-connected nets cannot”. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. JMLR.org, 2024.
- [WBU11] Jason Weston, Samy Bengio, and Nicolas Usunier. “Wsabie: Scaling up to large vocabulary image annotation”. In: *Twenty-Second International Joint Conference on Artificial Intelligence*. Citeseer. 2011.
- [WCH23] Rui Wang, Pengyu Cheng, and Ricardo Henao. “Toward fairness in text generation via mutual information minimization based on importance sampling”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 4473–4485.
- [WCM22] Colin Wei, Yining Chen, and Tengyu Ma. “Statistically meaningful approximation: a case study on approximating Turing machines with transformers”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 12071–12083.
- [Web+23] Debora Weber-Wulff et al. “Testing of detection tools for AI-generated text”. In: *International Journal for Educational Integrity* 19.1 (2023), p. 26.

- [Wei+21] Laura Weidinger et al. “Ethical and social risks of harm from language models”. In: *arXiv preprint arXiv:2112.04359* (2021).
- [Wei+22a] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.
- [Wei+22b] Jason Wei et al. “Finetuned language models are zero-shot learners”. In: *International Conference on Learning Representations*. 2022.
- [Wen+24] Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. “From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency”. In: *arXiv preprint arXiv:2410.05459* (2024).
- [Wet+24] Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. “QuRating: Selecting high-quality data for training language models”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Wil23] David Williams. *Systems and methods for automatically scrubbing sensitive data*. US Patent 11,645,458. May 2023.
- [WLJ23] Yuanhao Wang, Qinghua Liu, and Chi Jin. “Is RLHF more difficult than standard RL? a theoretical perspective”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 76006–76032.
- [WLL25] Ting Wu, Xuefeng Li, and Pengfei Liu. “Progress or regress? Self-improvement reversal in post-training”. In: *The Thirteenth International Conference on Learning Representations*. 2025.
- [Woo+23] Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. “Compensatory debiasing for gender imbalances in language models”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.
- [Wou24] Bram Wouters. “Optimizing watermarks for large language models”. In: *International Conference on Machine Learning*. 2024.
- [WP24] Robert Wu and Vardan Papyan. “Linguistic collapse: Neural collapse in (large) language models”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [WS23] Jason Weston and Sainbayar Sukhbaatar. “System 2 attention (is something you might need too)”. In: *arXiv preprint arXiv:2311.11829* (2023).

- [WTU20] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. “Scaling autoregressive video models”. In: *8th International Conference on Learning Representations, ICLR*. 2020.
- [Wu+23a] Xinwei Wu et al. “DEPN: Detecting and editing privacy neurons in pretrained language models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Dec. 2023, pp. 2875–2886.
- [Wu+23b] Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. “DiP-mark: A stealthy, efficient and resilient watermark for large language models”. In: *arXiv preprint arXiv:2310.07710* (2023).
- [Wu+24a] Tianhao Wu et al. “Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge”. In: *arXiv preprint arXiv:2407.19594* (2024).
- [Wu+24b] Yihan Wu et al. “Distortion-free watermarks are not truly distortion-free under watermark key collisions”. In: *arXiv preprint arXiv:2406.02603* (2024).
- [WZT23] Ruijia Wu, Linjun Zhang, and T Tony Cai. “Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference”. In: *Journal of the American Statistical Association* 118.543 (2023), pp. 1849–1861.
- [Xia+24a] Haojun Xia et al. “FP6-LLM: Efficiently serving large language models through FP6-centric algorithm-system co-design”. In: *arXiv preprint arXiv:2401.14112* (2024).
- [Xia+24b] Xun Xian et al. “RAW: A robust and agile plug-and-play watermark framework for AI-generated images with provable guarantees”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.
- [Xia+24c] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. “Efficient streaming language models with attention sinks”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Xia+24d] Jiancong Xiao et al. “On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization”. In: *arXiv preprint arXiv:2405.16455* (2024).

- [Xie+22] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. “An explanation of in-context learning as implicit Bayesian inference”. In: *The Tenth International Conference on Learning Representations, ICLR*. 2022.
- [Xie+23] Sang Michael Xie et al. “DoReMi: Optimizing data mixtures speeds up language model pretraining”. In: *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. 2023.
- [Xie+24a] Qianqian Xie et al. “Me LLaMA: Foundation large language models for medical applications”. In: *arXiv preprint arXiv:2402.12749* (Feb. 2024).
- [Xie+24b] Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie J Su, and Ruixun Zhang. “Debiasing watermarks for large language models via maximal coupling”. In: *arXiv preprint arXiv:2411.11203* (2024).
- [Xu+23] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. “Some things are more cringe than others: Preference optimization with the pairwise cringe loss”. In: *arXiv preprint arXiv:2312.16682* (2023).
- [Xu+24a] Haoran Xu et al. “Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Xu+24b] Yuhui Xu et al. “QA-LoRA: Quantization-aware low-rank adaptation of large language models”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Xu23] Ming Xu. *MedicalGPT: Training medical GPT model*. 2023. URL: <https://github.com/shibing624/MedicalGPT>.
- [Yad+24] Yasin Abbasi Yadkori et al. “Mitigating LLM hallucinations via conformal abstention”. In: *arXiv preprint arXiv:2405.01563* (2024).
- [Yan+19] Zhilin Yang et al. “XLNet: generalized autoregressive pretraining for language understanding”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.
- [Yan+23] Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. “Adept: A debiasing prompt framework”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2023, pp. 10780–10788.
- [Yan+24] Xianjun Yang et al. “DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.

- [Yao+23] Shunyu Yao et al. “Tree of thoughts: Deliberate problem solving with large language models”. In: *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. 2023.
- [Ye+24a] Fanghua Ye et al. “Benchmarking LLMs via uncertainty quantification”. In: *arXiv preprint arXiv:2401.12794* (2024).
- [Ye+24b] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. “Physics of language models: Part 2.1, grade-school math and the hidden reasoning process”. In: *arXiv preprint arXiv:2407.20311* (2024).
- [Ye+24c] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. “Physics of language models: Part 2.2, How to learn from mistakes on grade-school math problems”. In: *arXiv preprint arXiv:2408.16293* (2024).
- [YLG24] Weizhe Yuan, Pengfei Liu, and Matthias Gall . “LLMCrit: Teaching large language models to use criteria”. In: *Findings of the Association for Computational Linguistics: ACL*. 2024, pp. 7929–7960.
- [YNL21] Weizhe Yuan, Graham Neubig, and Pengfei Liu. “BARTScore: Evaluating generated text as text generation”. In: *Advances in Neural Information Processing Systems 34, NeurIPS 2021*. 2021, pp. 27263–27277.
- [Yu+21] Da Yu et al. “Differentially private fine-tuning of language models”. In: *International Conference on Learning Representations*. 2021.
- [Yu+23] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. “Unlearning bias in language models by partitioning gradients”. In: *Findings of the Association for Computational Linguistics: ACL*. 2023, pp. 6032–6048.
- [Yu+24] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. “Distilling system 2 into system 1”. In: *arXiv preprint arXiv:2407.06023* (2024).
- [Yua+24] Weizhe Yuan et al. “Self-rewarding language models”. In: *Forty-first International Conference on Machine Learning, ICML*. 2024.
- [Yun+20a] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. “Are Transformers universal approximators of sequence-to-sequence functions?” In: *8th International Conference on Learning Representations, ICLR*. 2020.
- [Yun+20b] Chulhee Yun et al. “ $O(n)$ connections are expressive enough: Universal approximability of sparse transformers”. In: *Advances in Neural Information Processing Systems 33* (2020), pp. 13783–13794.

- [YW21] Bo Yang and Lijun Wu. “How to leverage the multimodal EHR data for better medical prediction?” In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 329–345.
- [YXL24] Yuanshun Yao, Xiaojun Xu, and Yang Liu. “Large language model unlearning”. In: *Advances in Neural Information Processing Systems 38, NeurIPS 2024*. 2024.
- [Zel+19] Rowan Zellers et al. “Defending against neural fake news”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.
- [Zel+22] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. “STaR: Bootstrapping reasoning with reasoning”. In: *Advances in Neural Information Processing Systems 35, NeurIPS 2022*. 2022.
- [Zen+24] Qingcheng Zeng et al. “Uncertainty is fragile: Manipulating uncertainty in large language models”. In: *arXiv preprint arXiv:2407.11282* (2024).
- [Zer23] ZeroGPT. *ZeroGPT: Trusted GPT-4, ChatGPT and AI detector tool by ZeroGPT*. 2023. URL: <https://www.zerogpt.com/>.
- [ZFB24] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. “Trained transformers learn linear models in-context”. In: *Journal of Machine Learning Research* 25.49 (2024), pp. 1–55.
- [Zha+22] Yufeng Zhang, Boyi Liu, Qi Cai, Lingxiao Wang, and Zhaoran Wang. “An analysis of attention via the lens of exchangeability and latent variable models”. In: *arXiv preprint arXiv:2212.14852* (2022).
- [Zha+23] Qingru Zhang et al. “AdaLoRA: Adaptive budget allocation for parameter-efficient fine-Tuning”. In: *arXiv preprint arXiv:2303.10512* (2023).
- [Zha+24a] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. “Provable offline preference-based reinforcement learning”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [Zha+24b] Jinghui Zhang, Dandan Qiao, Mochen Yang, and Qiang Wei. “Regurgitative training: The value of real data in training large language models”. In: *arXiv preprint arXiv:2407.12835* (2024).
- [Zha+24c] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. “TinyLlama: An open-source small language model”. In: *arXiv preprint arXiv:2401.02385* (2024).

- [Zha+24d] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. “Negative preference optimization: From catastrophic collapse to effective unlearning”. In: *arXiv preprint arXiv:2404.05868* (2024).
- [Zha+24e] Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Fari-naz Koushanfar. “REMARK-LLM: A robust and efficient watermarking framework for generative large language models”. In: *33rd USENIX Security Symposium (USENIX Security 24)*. 2024, pp. 1813–1830.
- [Zha+24f] Wenting Zhao et al. “WildChat: 1M ChatGPT interaction logs in the wild”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Zha+24g] Xuandong Zhao, Chenwen Liao, Yu-Xiang Wang, and Lei Li. “Efficiently identifying watermarked segments in mixed-source texts”. In: *arXiv preprint arXiv:2410.03600* (2024).
- [Zha+24h] Xufeng Zhao et al. “Enhancing zero-shot chain-of-thought reasoning in large language models through logic”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. ELRA and ICCL, 2024, pp. 6144–6166.
- [Zhe+23] Lianmin Zheng et al. “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. 2023.
- [Zhe+24] Lianmin Zheng et al. “LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [Zho+24a] Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. “Provable multi-party reinforcement learning with diverse human feedback”. In: *arXiv preprint arXiv:2403.05006* (2024).
- [Zho+24b] Hongjian Zhou et al. “A survey of large language models in medicine: Progress, application, and challenge”. In: *arXiv preprint arXiv:2311.05112* (2024).
- [ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. “Sparse principal component analysis”. In: *Journal of Computational and Graphical Statistics* 15.2 (2006), pp. 265–286.
- [Zie+20] Daniel M. Ziegler et al. “Fine-tuning language models from human preferences”. In: *arXiv preprint arXiv:1909.08593* (2020).

- [ZJJ23] Banghua Zhu, Michael Jordan, and Jiantao Jiao. “Principled reinforcement learning with human feedback from pairwise or k-wise comparisons”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 43037–43067.
- [ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating unwanted biases with adversarial learning”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.
- [ZLW24] Xuandong Zhao, Lei Li, and Yu-Xiang Wang. “Permute-and-Flip: An optimally robust and watermarkable decoder for LLMs”. In: *arXiv preprint arXiv:2402.05864* (2024).
- [ZM24] Chrysoula Zerva and André F. T. Martins. “Conformalizing machine translation evaluation”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 1460–1478.
- [Zol+24] Thomas P. Zollo et al. “Prompt risk control: A Rigorous framework for responsible deployment of large language models”. In: *The Twelfth International Conference on Learning Representations, ICLR*. 2024.
- [ZRZ24] Lujing Zhang, Aaron Roth, and Linjun Zhang. “Fair risk control: A generalized framework for calibrating multi-group fairness risks”. In: *arXiv preprint arXiv:2405.02225* (2024).
- [ZWL23] Xuandong Zhao, Yu-Xiang Wang, and Lei Li. “Protecting language generation models via invisible watermarking”. In: *International Conference on Machine Learning*. Vol. 202. 2023, pp. 42187–42199.

A LLM 的历史发展

LLM 的能力源于两个关键组成部分：表征学习和语言建模。表征学习涉及将输入文本转化为能够捕捉其内在关系和模式的数值表示。语言建模则涉及预测文本序列中的下一个词元，从而实现连贯流畅的文本生成。

A.1 表征学习

LLM 的第一步是将自然语言转化为计算机可理解的格式——具体表现为向量、矩阵和张量。构成句子的符号化词汇单元（即词语）需要有效的数值表示策略，这种策略被称为词嵌入。基于分布假设（即认为出现在相似上下文中的语言单位具有相似含义）[Har54]，此类方法随着时间的推移逐渐变得精妙复杂。

早期的词义捕捉尝试使用词袋表示法，该方法将文档表示为稀疏向量（每个维度对应词典中的一个元素），例如 TF-IDF [Spa72]。随后引入了矩阵分解方法（如潜在语义分析 LSA [Hof01; LD02]）将词语表示为（稠密）向量。这类方法大多通过无监督方式隐式定义或学习权重参数。后来研究者使用支持向量机等监督学习方法直接针对特定任务训练词袋表示 [Joa98]，从而提升了任务性能。基于梯度或神经网络的监督式稠密词嵌入训练方法也带来了改进 [Bai+09; WBU11]，但需要注意的是这类方法需要相对较大的数据集。尽管这些方法具有一定实用价值，但由于缺乏在上下文中传递词语复杂语义和句法细微差异的能力，多数方法仍难以解决 NLP 中的挑战性任务。

神经网络方法首次随着 Word2vec [Mik+13] 及类似方法（如 GloVe [PSM14]）的出现而崭露头角，这些方法直接从局部词语上下文中学习低维词表示。这些模型架构背后的具体目标函数在 [Ben+03; CW08; Col+11; TRB10] 等基础性工作中得到了探索，为后续创新奠定了基础。例如，Word2vec 包含两种具体模型：连续词袋模型（Continuous Bag-Of-Words, CBOW）和跳元模型（Skip-gram）。CBOW 模型根据上下文预测目标词语，而 Skip-gram 模型则根据目标词语预测其上下文。两种模型均采用简单高效的神经网络架构，有效捕捉词语的语义和句法关系 [Mik+13]。这些方法生成的是稠密词嵌入——即紧凑向量，每个维度编码词语的潜在特征。经过无监督训练后，这些向量通常被用作监督训练阶段的输入或初始化参数，以解决特定任务。

这些表示学习方法以词语为基本单元，但也带来了诸多挑战，包括词汇量庞大和未登录词问题。为此，研究者引入了分词的概念。分词通过将文本分解为更小的单元（即词元），影响着学习表示的粒度，对模型效能至关重要。其中字节对编码（Byte Pair Encoding, BPE）[Gag94] 是代表性技术，它通过合并高频字节或字符对来缩减词汇量。该方法不仅有助于管理词汇规模，还能有效应对新文本中的未登录词问题。至此，表征学习、语言建模与分词技术的演进已形成统一且相互依存的进程。

A.2 语言建模

语言模型（Language Models, LMs）的早期阶段主要采用统计回归模型，依赖马尔可夫假设通过最近上下文预测下一个词元。主流模型是 n 元模型 [Jel98; GL04; Ros00]，该模型基于前 $n - 1$ 个词元确定当前词元的概率。此类模型面临的挑战包括需要处理数据稀疏性（即对未见过的词元或 n 元组赋予零概率的问题 [CG96]），以及应对 n 增大时需估计的指数级增长的转移概率数量。由于 n 通常被迫取较小值（如1-5个词元），利用此类模型完成长上下文输入的生成任务显得力不从心。

B 关键资源：加速 LLM 研究的进程

LLM 的发展需要算力资源、稳健的框架与专用工具的结合。近年来，各类开源工具和库的涌现显著降低了研究与实践的门槛。本章概述了 LLM 研究中最流行和广泛使用的资源，重点关注框架、模型库、数据准备工具，以及训练、部署和分析相关的实用工具。

B.1 模型开发核心框架

PyTorch PyTorch 凭借其灵活性、动态计算图和强大的社区支持，已成为 LLM 研究的主流深度学习框架。其调试便捷性及与其他库的无缝集成特性，使其成为开发和微调 LLM 的理想选择。

TensorFlow 与 JAX TensorFlow 在大规模训练领域仍具竞争力，尤其擅长分布式计算。JAX 凭借函数式编程范式和对自动微分的支持，在研究原型和前沿优化技术中日益受到关注。

各框架对比请见表 2。

框架	核心特性	适用场景	教程链接
PyTorch	动态计算图、社区支持	微调、原型开发	https://pytorch.org/tutorials/
TensorFlow	可扩展性、便于投入生产环境	大规模训练	https://www.tensorflow.org/tutorials
JAX	高性能计算、灵活性	实验性优化	https://jax.readthedocs.io/en/latest/tutorials.html

表 2: 模型开发的核心框架对比。

B.2 模型库与预训练模型

Hugging Face Transformers HuggingFace 的 Transformers 库是 LLM 研究的基石，它提供了预训练模型、分词工具以及面向多任务微调的工具集。其直观的 API 接口和详尽的文档体系使得不同层次的研究者都能便捷使用。

LangChain LangChain 通过将 LLM 输出与 API 或数据库等外部工具无缝衔接，简化了 LLM 与应用系统的集成流程。

更多特性描述请参阅表 3。

库名称	核心特性	适用场景	教程链接
HuggingFace Transformers	预训练模型、分词工具、多任务支持	模型微调、实验研究	https://huggingface.co/docs/transformers/index
LangChain	工具集成、输出链式处理	应用开发	https://python.langchain.com/docs/introduction/

表 3: 模型库与预训练模型汇总。

B.3 数据集与数据准备工具

HuggingFace Datasets HuggingFace Datasets 提供了精选数据集库及数据加载、筛选与处理工具。该库支持高效的流式处理与预处理管线，使研究者能够处理大规模数据集而无需过度消耗内存。其与 HuggingFace 分词工具及模型的无缝集成，使其成为 LLM 研究的关键基础设施。

数据清洗工具 预处理对 LLM 训练数据的准备至关重要。clean-text 等工具通过移除特殊字符、规范化文本和修正编码问题，简化了含噪声数据集的清洗流程。此外，SentencePiece 与 HuggingFace 的 tokenizers 等分词库，在将文本转换为模型可用的输入格式方面不可或缺。

这些工具的汇总参见表 4。

B.4 训练与微调工具

DeepSpeed 与 Megatron-LM DeepSpeed 和 Megatron-LM 是 LLM 规模化训练的关键工具。DeepSpeed 的 ZeRO（零冗余优化器）技术通过减少内存开销，实现在有限硬件上高效训练超大规模模型。Megatron-LM 则专精于模型并行，将大模型拆分至多 GPU 进行分布式训练。

工具	核心特性	适用场景	教程链接
HuggingFace Datasets	精选数据集、高效流式处理	数据集加载与预处理	https://huggingface.co/docs/datasets/index
clean-text	文本规范化、含噪声数据清洗	数据准备	https://github.com/jfilter/clean-text
SentencePiece	子词分词	文本分词	https://github.com/google/sentencepiece
HuggingFace Tokenizers	子词分词	文本分词	https://huggingface.co/docs/tokenizers/index

表 4: 数据集与数据准备工具汇总。

HuggingFace 微调框架 HuggingFace 提供了直观易用的微调框架，支持在下游任务中快速适配预训练模型。通过 Trainer API，研究者能以最小代码量高效完成模型微调。该框架还支持自定义训练循环，满足进阶微调需求。

HuggingFace PEFT 框架 HuggingFace 的 PEFT 框架基于 LoRA 等技术，提供冻结大部分模型参数、适配特定任务的参数高效微调方法，显著加速微调过程并降低资源需求。

TRL 与 OpenRLHF TRL（Transformers 强化学习库）与 OpenRLHF 是专为简化 RLHF（基于人类反馈的强化学习）和 DPO（直接偏好优化）实现而设计的强大工具集。

这些工具的汇总参见表 5。

B.5 推理与部署

vLLM 与 TensorRT-LLM vLLM 和 TensorRT-LLM 是针对 LLM 低延迟推理的优化工具。vLLM 专注于内存使用效率优化，可实现更快的批量推理以支持应用服务。NVIDIA 开发的 TensorRT-LLM 则支持 GPU 上的高吞吐量模型推理，充分利用先进硬件加速特性。两者均为需要高性能 LLM 部署的生产环境的理想选择。

Triton 推理服务器 NVIDIA 的 Triton 推理服务器通过支持多框架（如 PyTorch、TensorFlow、ONNX）来简化 LLM 部署流程。其支持基于 GPU 和 CPU 后端的可扩展生产级模型服务。

工具	核心特性	适用场景	教程链接
DeepSpeed	ZeRO 优化技术、可扩展性	大规模训练	https://www.deepspeed.ai/tutorials/
Megatron-LM	模型并行	大规模训练	https://github.com/NVIDIA/Megatron-LM
HuggingFace Trainer	简易微调配置、Trainer API	预训练模型微调	https://huggingface.co/docs/transformers/training
HuggingFace PEFT	参数高效微调技术	资源受限场景	https://huggingface.co/docs/peft/index
OpenRLHF	RLHF、DPO	训练后优化、AI 对齐	https://github.com/OpenRLHF/OpenRLHF
TRL	RLHF、DPO	训练后优化、AI 对齐	https://huggingface.co/docs/trl/index

表 5: 训练与微调工具汇总。

HuggingFace 推理端点 HuggingFace 提供了开箱即用的 API 部署平台，使开发者无需管理底层设施即可将模型快速集成至应用系统。

Ray Serve Ray Serve 作为分布式模型服务库，支持大规模 LLM 的扩展与并行推理，并能与分布式计算框架无缝集成。

这些工具的特性详见表 6。

B.6 可视化与分析

Weights & Biases Weights & Biases (W&B) 是用于实验追踪、超参数优化和训练指标实时可视化的强大工具。其与主流机器学习框架无缝集成，提供展示损失曲线、准确率趋势等性能指标的交互式仪表盘。该工具特别适用于协作研究场景，支持团队成员便捷共享实验结果与洞见。

TensorBoard TensorBoard 是广泛采用的训练监控与评估可视化工具包。核心功能包括标量指标（如损失值、准确率）可视化、图像展示及模型结构分析。其与 TensorFlow 和 PyTorch 的易集成性，使其成为机器学习社区的标配工具。

MLflow MLflow 是用于管理机器学习实验生命周期的开源平台，支持指标记录、产物存储、参数追踪以及模型版本控制与部署管理。通过其可视化界面，研究者可便捷对比实验并优化 workflow。

工具	核心特性	适用场景	教程链接
vLLM	内存优化、快速批量推理	生产环境推理	https://github.com/vllm-project/vllm
TensorRT-LLM	GPU 高吞吐量推理	优化 GPU 模型服务	https://github.com/NVIDIA/TensorRT-LLM/tree/release/0.5.0
HuggingFace 推理端点	简化 API 部署	快速集成	https://huggingface.co/inference-endpoints/dedicated
Triton 推理服务器	多框架支持的可扩展服务	可扩展 LLM 部署	https://developer.nvidia.com/triton-inference-server
Ray Serve	分布式并行推理	可扩展 LLM 部署	https://docs.ray.io/en/latest/serve/index.html

表 6: LLM 推理与部署工具汇总。

这些工具的汇总参见表 7。

B.7 云平台与硬件资源

云计算平台已成为大规模 LLM 研究与部署的基础设施，提供可扩展的按需算力资源（如 GPU、TPU 及专用加速器），使研究者无需预先投入大量基础设施即可开展实验。

AWS（亚马逊云服务） AWS 提供专为深度学习优化的 EC2 实例（如搭载 NVIDIA A100 GPU 的 p4d 实例），并推出 SageMaker 托管服务以支持机器学习模型的构建、训练与部署。

Google 云平台（GCP） 以 TPU（张量处理器）实例和 Vertex AI 平台见长，支持可扩展的训练与推理任务，特别适合 TensorFlow 用户。

Microsoft Azure 提供 Azure 机器学习套件（涵盖机器学习全流程工具与服务）及 GPU 加速虚拟机，专为深度学习任务优化。

这些工具的汇总参见表 8。

工具	核心特性	适用场景	教程链接
Weights & Biases	实验追踪、超参数调优	实验管理与可视化	https://docs.wandb.ai
TensorBoard	训练过程监控	模型性能分析	https://www.tensorflow.org/tensorboard
MLflow	实验生命周期管理	实验对比、模型版本控制	https://mlflow.org/docs/latest/index.html

表 7: 可视化与分析工具汇总。

平台	核心特性	适用场景	教程链接
AWS	GPU 加速 EC2 实例、SageMaker	模型训练与部署	https://aws.amazon.com/ai/machine-learning/
GCP	TPU 加速、Vertex AI	可扩展训练与实验	https://cloud.google.com/products/ai
Microsoft Azure	Azure 机器学习、GPU 虚拟机	实验研究与部署	https://learn.microsoft.com/en-us/azure/machine-learning

表 8: 云平台与硬件资源汇总。