Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(1), 194-208
Published Online January 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.141019

Hans汉斯

基于大语言模型的文本生成综述

刘津睿，都云程

北京信息科技大学计算机学院，北京

收稿日期：2024年10月30日；录用日期：2025年1月17日；发布日期：2025年1月29日

摘　要

文本生成(Text Generation)是自然语言处理(NLP)领域的一项核心技术。由于自然语言自身的复杂性，在内容创作、人机对话、机器翻译等领域的实际应用需求驱动下，文本生成技术长期以来一直是NLP研究的重点、难点和热点。随着深度学习、预训练语言模型等技术的产生和发展，文本生成技术得到长足发展，而基于Transformer的大语言模型(LLM)的产生，则彻底使文本生成技术取得革命性突破。本文旨在对文本生成的技术、模型、范式等方面的历史和现状进行总结，特别侧重于大语言模型对文本生成在框架模型、技术方案、评估基准等方面所带来的变革，以及大语言模型在文本生成领域的典型应用场景，并对文本生成在大语言模型背景下的技术发展趋势进行展望。

关键词

文本生成，大语言模型，自然语言处理

A Review of Text Generation Based on Large Language Model

Jinrui Liu, Yuncheng Du

Computer School, Beijing Information Science and Technology University, Beijing

Received: Oct. 30th, 2024; accepted: Jan. 17th, 2025; published: Jan. 29th, 2025

Abstract

Text generation is a fundamental technology in the field of Natural Language Processing (NLP). Due to the intrinsic complexity of natural language and the practical demands in applications such as content creation, human-computer interaction, and machine translation, text generation has long been a focal point of NLP research, characterized by its challenges and significant research interest. With the development of deep learning and pre-trained language models, text generation technology has made considerable advancements. The emergence of large language model (LLM) based on

文章引用: 刘津睿, 都云程. 基于大语言模型的文本生成综述[J]. 人工智能与机器人研究, 2025, 14(1): 194-208.
DOI: 10.12677/airr.2025.141019

the Transformer architecture has brought about a paradigm shift, leading to groundbreaking progress in the field. This paper seeks to provide a comprehensive review of the evolution and current state of text generation techniques, models, and paradigms, with a particular emphasis on the transformative impact of LLM on the design frameworks, technical approaches, and evaluation benchmarks in text generation. Furthermore, this paper explores the representative application scenarios of LLM in text generation and discusses future research directions and technological trends in this domain within the context of LLM.

Keywords

Text Generation, Large Language Model, Natural Language Processing

1. 引言

文本生成，又称自然语言生成，是自然语言处理中最重要的子领域之一。文本生成任务的目标是根据各种输入数据(如文本、图像、表格和知识库等)生成合理且可读的自然语言文本。过去几十年中，文本生成技术广泛应用于各类场景，如对话系统、机器翻译以及文本摘要。

文本生成的主要目标是从数据中自动学习从输入到输出的映射，以最少的人为干预构建端到端的解决方案。这种映射函数使生成系统能够在不同的领域中进行泛化，并根据给定的条件生成自然语言文本。早期的文本生成方法通常使用统计语言模型来建模 n-gram 上下文中的单词条件概率[1] [2]，但因数据稀疏性问题而受限。尽管已有多种平滑技术用于估计未登录词的概率[3] [4]，然而这些方法基于单词级表示，难以捕捉词汇之间的相似性。

随着深度学习的兴起，文本生成逐渐转向神经网络模型，并取得了显著进展，成为文本生成技术的主导方法。常见的文本生成模型采用基于编码器-解码器的序列到序列框架[5]，编码器将输入序列映射为低维嵌入，解码器再基于该嵌入生成目标文本。与统计方法相比，嵌入表示更能捕捉输入与输出之间的潜在关系。基于神经网络的文本生成系统随着图神经网络(GNN) [6]、递归神经网络(RNN) [7]等多种架构的相继提出，同时结合注意力机制[8]和复制机制[9]进一步提升了文本生成的性能。在文本生成领域中，神经网络模型的优势在于通过端到端的学习实现语义映射，避免了繁重的特征工程，并通过低维表示[10]有效缓解了数据稀疏性问题。

随后，文本生成技术在 PLM 的推动下取得了重大进展[11]，其核心范式是在大规模无监督语料库上进行预训练，然后在下游任务中进行微调。这一“预训练 + 微调”范式取得了当时最优性能。随着 Transformer [12]和算力的提升，基于 PLM 的文本生成模型从浅层逐渐发展为深层架构，如 BERT [13]和 GPT [14]。大量研究表明，这些模型能够通过预训练目标(如掩码语言建模)学习到丰富的上下文表示，极大地减少了从头训练模型的需求。在实际应用中，文本生成得益于 PLM 的语言理解能力，通过有效的数据获取、模型选择、输入处理等步骤，所生成的文本展现出更强的流畅性和语义一致性。PLM 的引入使文本生成不仅具备了更广泛的领域适应性，还能够生成符合语境的高质量自然语言文本[15]-[17]。

随着 LLM 的出现，文本生成技术在近几年取得了突破性进展。通过增加 PLM 模型的参数和扩大训练数据规模，文本生成系统在多个下游任务中表现出卓越的性能，这种现象被称为扩展法则[18]。如 GPT-3 [19]和 PaLM [20]等大型模型的成功，展示了模型规模的扩大对文本生成能力的提升。尽管这些模型架

构与较小模型类似，但是随着参数规模与训练数据规模的扩大，基于 LLM 的文本生成模型在处理复杂语言任务时展现出更为出色的“涌现能力”，不仅能够生成语法准确且语义丰富的自然语言文本，而且在处理长距离依赖和保持上下文一致性方面表现优异，有效解决了数据稀疏性问题，生成的文本更加流畅、自然，且具备较强的适应性[21]。

基于 LLM 的文本生成技术如今已经广泛应用于对话生成、机器翻译、技术文档撰写等领域，显著提升了文本生成的质量和效率，并推动了广告、新闻自动撰写、自动摘要等应用的发展[22]。随着 LLM 模型规模、训练数据规模的扩大及算法的不断进步，文本生成技术将在更广泛的领域拓展应用，并继续推动 NLP 和自动生成内容(AIGC)的创新和突破性发展[22]。然而，现有文献[10] [23] [24]对基于 LLM 文本生成技术的系统性综述仍较为欠缺，特别是 LLM 如何支持文本生成的框架及其技术贡献方面。因此，本文旨在填补这一空白，深入探讨基于 LLM 的文本生成技术及其带来的重要突破。

本文的组织如下：首先简要地介绍了文本生成的概念以及 LLM 核心架构 Transformer、扩展法则及其涌现能力(第二节)。其次，介绍了基于 LLM 的文本生成技术如何通过预训练、指令微调和提示工程等关键基础技术得到推动以及基于人类反馈的强化学习(RLHF)和检索增强生成方法(RAG)等关键增强技术的应用为文本生成模型效果带来的提升(第三节)。第四节讨论了基于 LLM 的文本生成评估的几个评估数据集以及在文本生成在 LLM 时代评估面临的挑战，并介绍了新兴的评估方法。最后，本文总结了基于 LLM 的文本生成技术在文本生成、摘要、机器翻译等任务中的应用，并展望了未来的研究方向。

2. 文本生成与 LLM 基础

在 LLM 时代，文本生成技术的发展与 LLM 息息相关。基于 LLM，文本生成系统能够处理复杂的语义结构、长距离依赖以及多样化的上下文输入。基于 Transformer 架构[25]，文本生成系统通过自注意力机制能够精确捕捉文本中的细微语境差异，实现更加流畅和连贯的生成效果。随着 GPT、BERT 等模型的预训练规模不断扩大，基于 LLM 的文本生成技术在多个领域的任务中展现了卓越的性能，包括机器翻译、文本摘要和对话生成。它不仅可以生成高质量的自然语言输出，还能有效应对噪声数据和不完整输入，展现出极高的鲁棒性。同时，随着模型规模和数据集的扩展，文本生成系统通过上下文学习(ICL) [26]和思维链推理[27] (Chain-of-Thought, CoT)等涌现能力的出现，进一步提升了处理复杂任务的能力。

2.1. 文本生成

文本通常被建模为一个由多个标记组成的序列 $\mathbf{y}=\langle \boldsymbol{y}_1,\cdots,\boldsymbol{y}_i,\cdots,\boldsymbol{y}_n\rangle$，其中每个标记 $\boldsymbol{y}_i$ 来自词汇表 V。文本生成任务的目标是生成逻辑合理且可读的自然语言文本。在大多数情况下，文本生成依赖于某些输入数据(如文本、图像、表格或知识库)，这些输入数据记为 $\mathbf{x}$。此外，生成的文本还需要满足预期的语言属性，例如流畅性、自然性和连贯性。我们将这些属性记为 P。基于上述符号，文本生成任务可以形式化描述为：

$$\mathbf{y}=f_M\left(\mathbf{x},P\right)$$

其中，生成模型 f_M 根据输入 $\mathbf{x}$ 和属性集合 P 生成输出文本 $\mathbf{y}$。在本文中，我们主要关注 LLM 的文本生成模型 f_M。

随着 LLM 的发展，文本生成任务得到了突破性进展。LLM 通过大规模预训练，在各种文本生成任务中表现出卓越的能力，特别是在捕捉复杂的语言模式和长距离依赖方面。

具体而言，基于输入数据 $\mathbf{x}$ 的类型和属性集合 P 的不同，文本生成可以细化为以下几类代表性任务：

1) 无条件文本生成：当未提供明确的输入数据 $\mathbf{x}$ 或输入为随机向量时，文本生成退化为语言建模或无条件文本生成[14] [28]。在这种情况下，LLM 无需明确的上下文即可生成连贯自然的文本，满足流畅

性和自然性等基本语言属性。GPT 系列模型在这类任务中表现出色，能够生成从故事到技术文档的多种类型文本。

2) 基于属性的文本生成：当输入数据 **x** 是离散属性(如主题词或情感标签)时，任务转化为主题到文本生成[29]或属性驱动的文本生成[30]。此时，LLM 可以根据输入属性控制生成内容。例如，GPT-3 能够根据给定的情感标签生成特定情感风格的文本或根据关键词生成相关主题的段落。

3) 文本到文本生成：最常见的输入类型是文本序列，涉及的任务包括机器翻译、文本摘要和对话生成。LLM 在这些任务中展现了强大的能力，能够生成高度语义一致的翻译、精准概括的摘要以及与对话历史高度相关的回复。例如，基于 Transformer 架构的 GPT 和 BERT 模型，通过对大量文本进行预训练，具备了处理复杂上下文并生成连贯文本的能力。

下文将对 LLM 基础知识，包括 LLM 基础构建模块 Transformer、扩展法则以及涌现能力进行介绍。

2.2. LLM 基础知识

LLM 由 Transformer 作为基本构建块堆叠而成，Transformer 的基础架构如图 1 所示。Transformer 是一种专为序列数据设计的深度神经网络，核心是自注意力机制，使模型理解每个词在句中的作用，而不受词序影响。得益于此，生成的文本更加贴近人类表达。此外，Transformer 的并行处理特性大大减少了训练时间，增强了可扩展性。LLM 使用多种架构，如编码器–解码器、因果编码器和前缀编码器[25] [31]这些架构可以通过专家混合(Mixture of Experts (MoE)) [32]进一步扩展和优化，其中模型的性能显著提升，尤其是当增加专家数量或参数规模时。

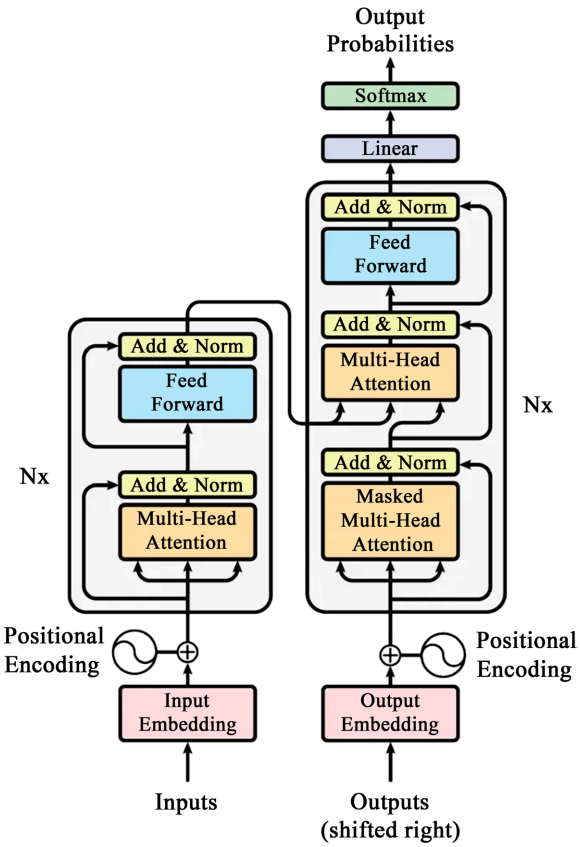

Figure 1. Transformer model architecture [12]
图 1. Transfomer 模型架构图[12]

通常，LLM 是指包含数百亿(或更多)参数的 Transformer 语言模型[12]，这些模型是在大规模文本数据上进行训练的[33]，例如 GPT-3，PaLM，Galactica [34]和 LLaMA [35]。LLM 展现了理解自然语言和解决复杂文本生成任务的强大能力，图 2 概述了一些最具代表性的 LLM，以及在推动 LLM 成功和突破其极限方面作出贡献的相关工作。为了对 LLM 的工作原理有一个快速的了解，本部分将介绍 LLM 的基本背景，包括扩展法则、涌现能力等。

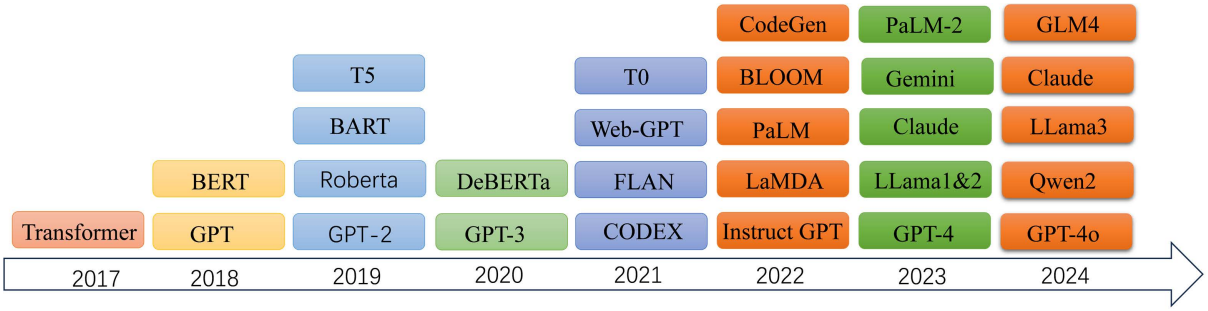

Figure 2. Timeline of representative large language models
图 2. 目前具有代表性的大语言模型时间轴

语言模型的扩展法则：目前，LLM 主要基于 Transformer 架构，其中多头注意力层在深层的神经网络中堆叠。现有的 LLM 采用类似的 Transformer 架构和与小型语言模型相同的预训练目标(如语言建模)。然而，LLM 大幅度扩展了模型参数规模、数据规模和总计算量(数量级)。大量研究表明，扩展参数规模和训练数据集规模可以大幅提高 LLM 的模型能力[14] [26] [36]。因此，建立一个定量的方法描述扩展法则具有重大意义。在此，我们介绍两个 Transformer 语言模型的代表性扩展法则：KM 扩展法则[37]和 Chinchilla 扩展法则[38]。

KM 扩展法则 2020 年，Kaplan 等人(OpenAI 团队)首次提出了神经语言模型的性能与模型规模(N)、数据集规模(D)和训练计算量(C)之间的幂律关系。在给定计算预算 c 的条件下，研究人员依据实验提出了三个基本公式来描述扩展法则[39]：

$$\begin{aligned} L(N) &= \left(\frac{N_C}{N}\right)^{\alpha_N}, \alpha_N \sim 0.076, N_C \sim 8.8\times10^{13} \\ L(D) &= \left(\frac{D_C}{D}\right)^{\alpha_D}, \alpha_D \sim 0.095, D_C \sim 5.4\times10^{13} \\ L(C) &= \left(\frac{C_C}{C}\right)^{\alpha_C}, \alpha_C \sim 0.050, C_C \sim 3.1\times10^{8} \end{aligned} \tag{1}$$

Chinchilla 扩展法则 Hoffmann 等人(Google DeepMind 团队)提出了一种扩展法则的替代形式来指导 LLM 最优计算量的训练。该工作通过变化更大范围的模型大小(7000 万到 160 亿个参数)和数据大小(50 亿到 5000 亿个 token)进行了严格的实验，并拟合出了类似的扩展法则，但具有不同的系数，如下所示：

$$L(C) = \left(\frac{C_C}{C}\right)^{\alpha_C}, \alpha_C \sim 0.050, C_C \sim 3.1\times10^{8} \tag{2}$$

其中 $E = 1.69$，$A = 406.4$，$B = 410.7$，$\alpha = 0.34$ 和 $\beta = 0.28$。通过在约束条件 $C \approx 6ND$ 下优化损失 $L(N, D)$，得到了将计算预算最优地分配给模型大小和数据大小的方法：

$$N_{opt}(C) = G\left(\frac{C}{6}\right)^{a}, \quad D_{opt}(C) = G^{-1}\left(\frac{C}{6}\right)^{b} \tag{3}$$

其中，$a=\frac{\alpha}{\alpha+\beta}$，$b=\frac{\beta}{\alpha+\beta}$，$G$ 是由 A、B、α 和 β 计算得出的扩展系数。如[40]所给结论：随着给定计算预算的增加，*KM* 扩展法则更偏向于将更大的预算分配给模型大小，而 Chinchilla 扩展法则认为模型大小和数据大小应该以相同的比例增加，即在公式(3)中的 a 和 b 取相近的值。

虽然存在一些限制性的假设，但是这些扩展法则为理解扩展效应提供了直观视角，使得在训练过程中能够预测 LLM 的性能[41]。然而，一些能力(如 ICL)无法根据扩展法则进行预测，只有当模型超过一定规模时才能被观察到，即 LLM 的涌现能力。

大语言模型的涌现能力：在文献[42]中，LLM 的涌现能力被正式定义为“在小模型中不存在但在大型模型中产生的能力”，这是区别 LLM 与先前 PLM 的最显著特征之一[42]。进一步介绍了当涌现能力出现时的一个显著特点：当规模达到一定水平时，性能显著提高，超出随机水平。类比而言，这种涌现模式与物理学中的相变现象有密切联系[43]。原则上，涌现能力可以与一些复杂任务相关联，但我们更关注可以用来解决各种任务的普遍能力[44]。以下是 LLM 涌现能力几个典型的具体体现。

1) 上下文学习：ICL 能力是由 GPT-3 正式引入的：假设已经为语言模型提供了一个自然语言指令和/或几个任务演示，它可以通过完成输入文本的单词序列的方式来为测试实例生成预期的输出，而无需额外的训练或梯度更新[26]。在 GPT 系列模型中，1750 亿的 GPT-3 模型在一般情况下表现出强大的 ICL 能力，但 GPT-1 和 GPT-2 模型则没有。此外，这种能力还取决于具体的下游任务。例如，130 亿参数的 GPT-3 可以在算术任务(例如 3 位数的加减法)上展现出 ICL 能力，但 1750 亿参数的 GPT-3 在波斯语问答任务上无法很好地工作[38]。

2) 指令遵循：通过使用自然语言描述的混合多任务数据集进行微调(称为指令微调，即 Instruction Fine-Tuning, SFT)，LLM 在未见过的、以指令形式描述的任务上表现出色[45]-[47]。通过指令微调，LLM 能够在没有使用显式示例的情况下遵循新的任务指令，因此它具有更好的泛化能力[43]。中的实验证明，当模型大小达到 680 亿时，经过指令微调的 LaMDA-PT [48]开始在未见过的任务上显著优于未微调的模型，但对于 80 亿或更小的模型大小则不会如此。最近的一项研究[49]发现，PaLM 至少在 620 亿的模型大小上才能在四个评估数据集(即 MMLU、BBH、TyDiQA 和 MGSM)的各种任务上表现良好，尽管较小的模型可能足够完成某些特定任务(例如 MMLU)。

3) 逐步推理：对于小型语言模型而言，通常很难解决涉及多个推理步骤的复杂任务，例如数学问题。然而，通过使用 CoT 提示策略，LLM 可以通过利用包含中间推理步骤的提示机制来解决这类任务，从而得出最终答案。这种能力可能是通过在代码上进行训练而获得。一项实证研究表明，当应用于模型大小大于 600 亿的 PaLM 和 LaMDA 变体时，CoT 提示可以提高模型在算术推理基准任务上的性能，而当模型大小超过 1000 亿时，其相对于标准提示的优势更加明显。此外，CoT 提示的性能改进在不同的任务上也存在差异，例如对于 PaLM 来说，GSM8K > MAWPS > SWAMP [27]。

3. 基于 LLM 的文本生成关键技术

前文介绍了文本生成与 LLM 的基础知识，在 LLM 蓬勃发展的背景下，基于 LLM 的文本生成技术的性能和应用能力得到了突破性进展。本章将聚焦于基于 LLM 的文本生成技术中的核心技术，重点介绍支撑 LLM 的增强技术及其在文本生成中的作用与影响。这些技术不仅推动了生成模型在自然语言处理领域的进步，也为文本生成在创作、人机对话和翻译等场景中带来了质的飞跃。

3.1. 基于 LLM 文本生成基础技术

基于 LLM 的文本生成模型通过大规模预训练技术，在海量数据中捕捉复杂的语言模式和深层语义结构，显著提升了对语言的理解深度，克服了传统方法在语义表示和泛化能力上的局限。基于指令微调，

基于 LLM 的文本生成系统能够在特定任务上精准调整文本生成模型参数，显著增强其适应性和表达准确性，有效降低生成偏差，并引入伦理和合规性以保证生成内容质量。提示工程进一步优化基于 LLM 的文本生成系统，通过精细设计输入提示和生成策略，实现生成内容与任务目标的高度一致，克服一致性、领域特定知识和实时性等挑战，确保文本生成的精准性和可控性。

3.1.1. 预训练

预训练阶段是基于 LLM 文本生成技术的关键环节，该阶段通过在大规模语料库上计算数十亿参数，使模型编码语言的通用知识，从而为后续的文本生成任务奠定坚实基础。在预训练中，文本生成模型通过学习语言结构和语义模式，逐步形成逻辑连贯的生成能力。预训练任务主要包括语言建模和去噪自编码两种方法：

1) 语言建模：语言建模(LM)是文本生成的重要任务之一，尤其适用于仅包含解码器的 LLM(如 GPT3 和 PaLM)。给定一个 token 序列 $x=\{x_1,\cdots,x_n\}$，LM 任务旨在基于序列中前面的 $token\, x_{<i}$，自回归地预测目标 $token\, x_i$。通常的训练目标是最大化以下似然函数：

$$\mathcal{L}_{LM}(x)=\sum \log P(x_i \mid x_{<i}). \tag{4}$$

2) 去噪自编码：去噪自编码任务(DAE)也被广泛用于 PLM [50] [51]。DAE 任务的输入 $\mathbf{x}_{/\tilde{\mathbf{x}}}$ 是一些有随机替换区间的损坏文本。然后，训练语言模型以恢复被替换的 $token\, \tilde{\mathbf{x}}$。形式上，DAE 的训练目标如下：

$$\mathcal{L}_{DAE}(\mathbf{x})=\sum \log P(\mathbf{x} \mid \mathbf{x}_{/\tilde{\mathbf{x}}}). \tag{5}$$

由于 DAE 任务在实现上比 LM 任务更为复杂且效果提升不明显，它并没有被广泛用于 LLM 预训练。

3.1.2. 指令微调

指令微调通过明确的任务指令，显著优化了基于 LLM 的文本生成效果[52]。基于 LLM 的文本生成系统通过指令微调，仅需少量任务提示训练，便能灵活应对新颖的提示，在特定生成任务中展现更高的准确性和一致性。与预训练不同，指令微调实现了任务定向优化，不仅缩短了生成时间，还提升了内容质量。例如，在摘要生成中，经过微调的模型能生成更简洁、逻辑清晰的摘要，更适应不同文本的需求。指令微调还增强了系统的灵活调整能力，使模型在情感分析、对话生成等特定任务中生成更专业、精准的内容，比通用模型更符合任务要求。

3.1.3. 提示工程

提示工程是(Prompt Engineering (PE)) LLM 文本生成中核心技术之一，旨在为每个任务设计最合适的提示模板，以优化生成效果。PE 可以通过人工或算法设计实现，主要关注提示的位置和类型：例如将提示置于模板中间为完形填空型，置于末尾为前缀型。前缀提示适用于生成任务，特别是在自回归模型中，因其符合左至右的预测过程。PE 分为人工和自动模板设计。人工设计依赖设计者的经验，能精准引导模型输出，但可能错失最佳提示。自动提示包括离散提示和连续提示：离散提示在字典空间中搜索文本串，但解释性较弱；连续提示在编码空间优化，灵活性更高，生成的文本更自然且符合上下文需求。

相比 Fine-tuning，PE 在 LLM 中更具优势。Fine-tuning 对每个任务需重新训练，成本高且迁移性差，而提示工程通过模板适配，实现 LLM 在少样本任务中快速生成内容，尤其适用于大模型，二者对比如图 3 所示。提示工程还通过多步推理增强生成逻辑性，例如，CoT 方法提供示例推理，有助于模型应对复杂任务，提升生成内容的质量。

尽管 PE 提升了文本生成效果，但模型在内容准确性和一致性上仍面临挑战，如“幻觉”现象，即生成内容语义通顺但包含错误或虚构信息。对此，RLHF 和 RAG 等增强技术被开发，以改善 LLM 的文本生成质量。

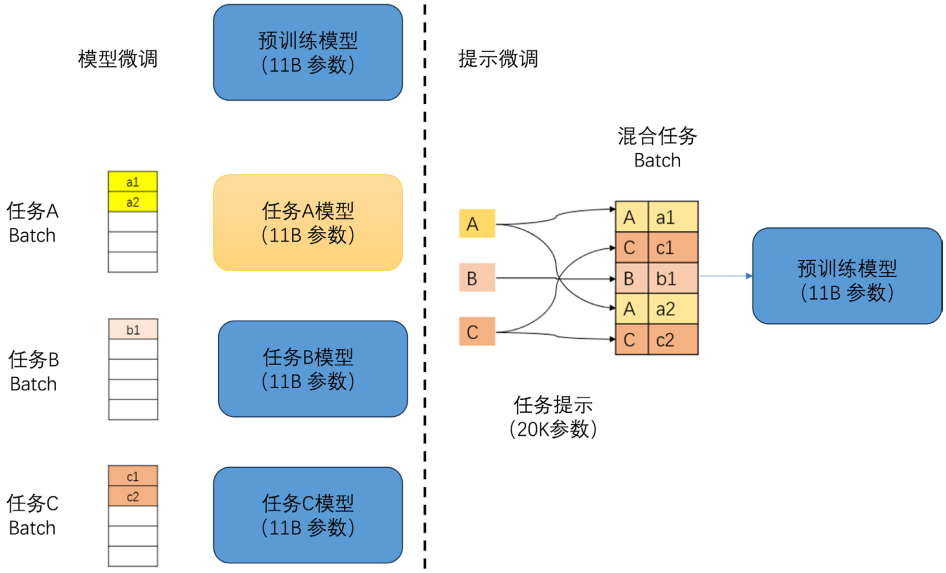

Figure 3. Comparison of pre-training + fine-tuning model and prompt-tuning model
图 3. 预训练 + 微调模型与提示微调模型的对比图

3.2. 基于 LLM 的文本生成增强技术

3.2.1. 基于人类反馈的强化学习

基于 LLM 的文本生成系统通过 RLHF 显著增强了内容质量与可靠性。RLHF 利用人类反馈数据，以奖励模型为生成内容打分，为文本生成系统提供明确的优化方向。具体而言，如图 4 所示，RLHF 包括以下两步[42]：

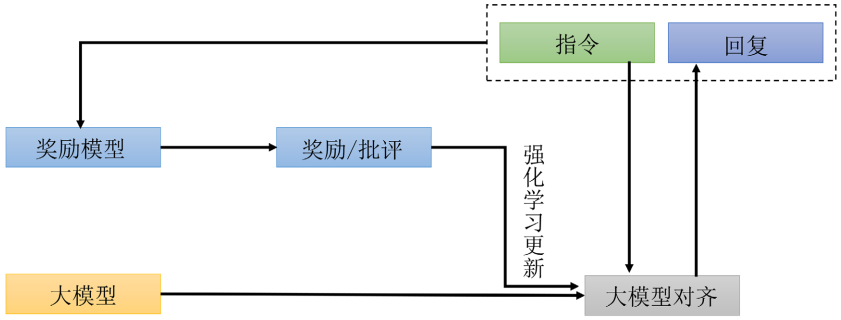

Figure 4. Basic workflow of RLHF
图 4. RLHF 基本流程图

奖励模型的训练：奖励模型是 RLHF 的核心部分。通过人类反馈数据，如对多个生成输出的排序，奖励模型学习如何根据人类的偏好为生成内容打分。这一评分系统帮助模型区分高质量和低质量输出，从而为强化学习提供监督信号，指导模型生成更符合人类期望的文本。

基于 PPO 的强化学习：在奖励模型训练完成后，使用近端策略优化算法(Proximal Policy Optimization, PPO)对模型进行训练。通过反复生成文本、评价奖励并更新模型参数，PPO 算法逐步优化语言模型的输出。为避免模型过度拟合奖励模型，常引入 KL 散度作为正则化项，保持生成结果的多样性和创造性。

在基于 LLM 的文本生成模型中应用 RLHF 带来的提升效果显著。结合人类反馈的生成策略有效减少了“幻觉”现象，并提升了文本的伦理安全性，尤其在医疗、法律等高准确性需求的领域表现突出。

此外，RLHF 提升了文本生成的泛化能力，使系统能够在新场景和未见数据中生成符合人类需求的合适内容，从而显著增强文本生成的灵活性与广泛适用性。

3.2.2. 知识检索增强

RAG 通过引入检索机制，扩展了基于 LLM 的文本生成模型知识获取范围，尤其在需要实时信息和知识密集的任务中效果显著。RAG 的技术框架由检索模块和生成模块组成，分别为文本生成提供了丰富的外部知识来源和高质量的输出文本支持，如图 5 所示，主要包括检索和生成模块。

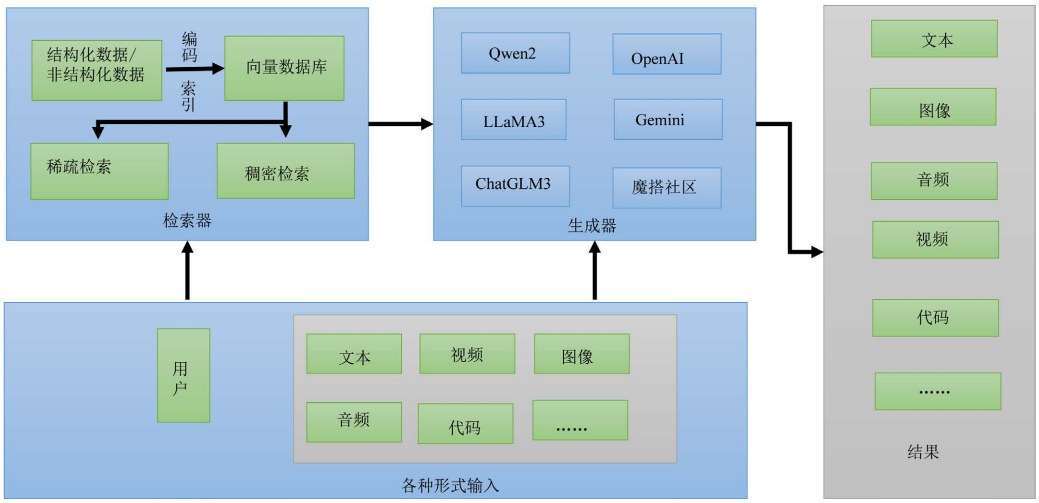

Figure 5. General workflow of RAG
图 5. RAG 通用流程图

检索模块：RAG 的检索模块首先从外部知识库中获取与输入查询相关的文档或文本片段。一般采用稠密检索(Dense Retrieval)或稀疏检索(Sparse Retrieval)技术来完成此任务。稠密检索技术(如 DPR，Dense Passage Retrieval)将查询和文档转化为向量，并通过向量相似度匹配检索相关信息。稀疏检索则更多依赖于关键词匹配，如 BM25 等经典算法。

生成模块：在检索到相关信息后，生成模块将这些信息与输入的查询结合，通过生成模型(如 T5、BART)生成最终的文本输出。生成模型会根据检索结果对其进行补充和完善，从而提高生成内容的准确性和上下文一致性[53]。

通过引入 RAG，文本生成过程能够动态检索和利用最新的知识，显著减少了幻觉现象，确保内容的时效性和准确性。这种动态更新机制特别适用于法律、医学等领域，满足了实时信息更新的需求，使文本生成任务在复杂环境中的表现更加可靠和流畅。

综上所述，RLHF 和 RAG 在优化文本生成时效性、准确性以及安全伦理性等方面发挥了关键作用。RLHF 通过人类反馈数据动态调整生成内容，确保文本输出的合理性和适应性；RAG 则通过外部知识检索机制扩展了内容覆盖，使生成的文本更具实时性和上下文关联性。这两项技术的协同作用显著提升了文本生成系统在多样性、连贯性和知识覆盖方面的表现，推动了文本生成技术的进一步发展。

4. 基于 LLM 的文本生成技术评估

随着 LLM 文本生成技术的普及，科学系统的评估体系变得尤为重要，以全面衡量生成内容的质量。高质量的评估体系不仅能揭示模型的优势和不足，还能为优化提供有力支持。不同于传统文本生成评估，LLM 生成内容的评估需涵盖多维标准，包括准确性、流畅性和伦理合规性等。由于 LLM 具有高度灵活

性和适应性，评估过程也更加复杂化，需结合自动化指标与人工反馈，以确保全面和公正。本章将介绍几种常用的基于 LLM 的文本生成评估数据集，并分析评估中的挑战，为优化评估体系提供参考。

4.1. 基于 LLM 的文本生成评测数据集

在本节中，我们将介绍 LLM 背景下文本生成技术主流的常用评测数据集，包括 LBD [54]、WMT’22 [55]、XSum [56]和 HumanEval [57]等。

LAMBADA (LBD)是一个专注于长文本理解和摘要生成的自然语言处理数据集。该数据集包含大量的长篇文章及其对应的摘要，旨在评估模型在理解和压缩长篇内容方面的能力。LBD 数据集的文本来源于多个领域，如科学、历史、新闻等，对模型的领域适应性和长文本处理能力提出了挑战。

WMT’22 是一个用于衡量机器翻译模型性能的数据集，特别关注于跨语言的词汇映射问题。WMT 数据集包含了多种语言对的翻译实例，每个实例都包含了源语言句子、目标语言句子以及单词级别的对齐信息。该数据集不仅用于评估翻译质量，还可以用于研究多语言模型的词汇对齐能力。

XSum 是一个极端文本摘要数据集，专门用于评估模型的极端摘要生成能力。数据集中的每个文档都配有一个单句摘要，这个摘要是对文档内容的极度精简。XSum 数据集的文档主要来自英国的广播公司(BBC)，覆盖了广泛的主题。该数据集对模型提取关键信息并生成简洁摘要的能力提出了高要求。

HumanEval 是一个用于评估 PLM 代码生成能力的评测基准。它包含了多个编程问题，这些问题需要模型不仅理解自然语言描述，还能生成相应的代码。HumanEval 数据集覆盖了多种编程语言，旨在推动模型在编程和算法任务上的应用。表 1 是部分大语言模型在各数据集上面的表现：

Table 1. Performance of selected LLMs on various datasets
表 1. 部分大语言模型在各数据集上面的表现

模型名称	文本生成			
	LBD	WMT	XSum	HumanEval
ChatGPT	55.81	36.44	21.71	79.88
Claude	64.47	31.23	22.86	51.22
Davinci003	69.98	37.46	18.19	67.07
Davinci002	58.85	35.11	19.15	56.70
Vicuna (7B)	60.12	18.06	13.59	17.07
Alpaca (7B)	60.45	21.52	8.74	13.41
ChatGLM (6B)	33.34	16.58	13.48	13.42
LLaMA (7B)	66.78	13.84	8.77	15.24
Falcon (7B)	66.89	4.05	10.00	10.37
Pythia (12B)	60.49	5.43	8.87	14.63
Pythia (7B)	50.96	3.68	8.23	9.15

4.2. 基于 LLM 的文本生成评估存在的问题

基于 LLM 的文本生成评估面临多重挑战。传统的评估指标如 BLEU 和 ROUGE，虽然在过去广泛用于文本生成质量的衡量，但这些方法主要依赖于词汇级别的 n-gram 匹配，聚焦于生成文本与参考文本在

词汇和短语层面的相似性。然而，基于 LLM 的文本生成系统生成的文本通常包含复杂的语义、逻辑关联和上下文依赖，这使得传统指标难以准确评估其语义深度、逻辑一致性和语言流畅性，尤其是在长文本生成、跨语境理解和复杂推理等任务中表现出明显不足[58]。

同时，随着基于 LLM 的文本生成任务逐渐扩展到问题回答和常识推理等领域，评估基准也更偏向于答案的准确性，忽视了对生成文本连贯性、自然度和上下文相关性的全面评价。由于 LLM 具备生成高质量、流畅自然文本的能力，单一的正确性指标难以充分反映其整体表现。此外，现有的自动化评估工具与人类主观判断之间的相关性偏低，进一步突显了传统方法的局限性。研究表明，基于 LLM 的文本生成系统生成的文本在自动评估中得分偏低，而在人类评估中表现优异，揭示了现有评估工具难以全面反映 LLM 生成文本的质量[30]。

针对这些挑战，研究者提出了一些更适用于基于 LLM 的文本生成评估方法。例如，“原子事实”(Atomic Facts)作为新评估标准，将生成文本分解为独立的语义单元，通过减少主观性和消除歧义，使评估更具客观性和细致性，尤其适合信息密集型生成任务，如事实验证和知识生成[59]。此外，研究者探索了让 LLM 自身参与评估的可能性，即利用 LLM 评估生成文本的语义一致性和流畅性，实验表明此方法与人类评估的相关性较高，能够更好捕捉生成文本的质量。

进一步的研究还提出了多维度的评估方法，试图从上下文理解、逻辑推理和语言流畅度等角度进行综合评估，不再局限于简单的词汇匹配和句法规则，而是通过捕捉 LLM 生成文本的语义和逻辑连贯性，为文本生成质量提供了更全面、精准的评估框架[60]。这些新兴的评估方法为开发更符合基于 LLM 文本生成系统特点的评估工具提供了新思路，有望进一步提升 LLM 生成质量的全面性和准确性。

5. 基于 LLM 的文本生成技术应用

在快速发展的文本生成领域，基于 LLM 的文本生成技术正逐步改变人们与信息交互的方式。这一技术不仅提高了文本生成的效率和质量，还展现出在多个行业的广泛应用潜力。通过学习和分析大量文本数据，基于 LLM 的文本生成系统具有高度适应性，能够满足从对话生成到文档撰写的多样化需求，从而推动各行业的创新和发展。本节将深入探讨基于 LLM 的文本生成技术在各个具体应用中的变革性作用，展示其在不同任务中的独特优势。

文本生成：基于 LLM 的文本生成技术在文本生成领域实现了重大突破，并广泛应用于多个行业的自动化流程。在内容创作中，基于 LLM 的文本生成技术擅长生成高质量的营销文案，例如产品描述、广告和促销材料，还能够根据用户偏好生成个性化推荐[23]。此外，它能够基于 RAG 技术通过实时数据生成新闻报道或摘要，提升信息传递的速度和覆盖范围[61]。通过结合 RLHF，基于 LLM 的文本生成系统可以动态调整生成策略，以更好地满足用户需求，提升文本的相关性和吸引力。在金融领域，基于 LLM 的文本生成技术可以自动生成财务报告[62]；引入 RAG 后，LLM 可以动态获取最新的外部信息，确保生成内容的准确性和时效性。在教育领域，基于 LLM 的文本生成技术还可以生成个性化的学习材料和测验[30]，并通过 RLHF 不断优化内容适应性，以提高生成效果。此外，LLM 在创意写作方面应用广泛，例如叙事、短篇小说和诗歌创作，结合 RAG，它能够从外部知识库中检索相关信息，增强创作内容的深度和多样性。在客户服务领域，基于 LLM 的聊天机器人通过自然语言生成精准的响应，以解决客户问题、提供产品信息并支持技术故障排除。

文本摘要：作为文本生成技术的典型应用，文本摘要旨在从长篇文本中提取关键信息并生成简洁的概述。与传统基于规则或统计的方法相比，基于 LLM 的文本生成系统能够更准确地理解和总结长文档的内容[63]。在新闻领域，基于 LLM 的文本生成系统可以快速生成新闻文章的摘要，使读者无需通篇阅读即可掌握核心信息。在学术研究中，它被广泛用于将复杂的论文和文章浓缩成简短的摘要，帮助研究人

员高效筛选相关研究成果[64]。此外，内容聚合平台和网站利用它生成博客、帖子和文章的摘要，使用户能够根据兴趣快速决定阅读内容。在金融领域，通过基于 LLM 的文本生成系统生成财务报告和市场分析的摘要，帮助投资者迅速评估市场动态[25]；在医疗领域，它被用于总结患者病史、医疗对话和研究论文，简化医疗决策过程。

机器翻译：在全球化的推动下，语言翻译成为跨文化交流的核心，而基于 LLM 的文本生成技术显著提升了机器翻译的质量和效率。借助 LLM 的强大语言理解和生成能力，在线翻译服务(如谷歌翻译和 DeepL)实现了多语言之间的实时高效翻译，不仅提升了翻译的准确性，还减少了语义错误和文化偏差[62]。基于 LLM 的文本生成系统通过分析大规模多语言数据，能够自动捕捉不同语言间的细微差异，使翻译内容更自然流畅。通过引入 RLHF 技术进一步提高了翻译的准确性和用户满意度，通过不断吸收用户反馈，使模型更好地理解复杂语境和语义变化，从而生成符合目标语言表达习惯的文本。这不仅减少了翻译中的语法和表达错误，也显著提升了机器翻译的整体质量。同时，RAG 技术的引入使基于 LLM 的文本生成系统能够在翻译过程中动态获取外部信息，特别是在处理专有术语和富含文化背景的文本时表现出色。RAG 允许文本生成模型在翻译时实时检索并整合最新的外部知识库信息，确保翻译内容不仅依赖于模型已有的内部知识，还能借助外部数据源提高翻译的准确性和时效性。这在法律、医学和技术文档等专业领域的翻译中尤为关键，确保了内容的专业性和精准性。通过 RLHF 和 RAG 的协同作用，基于 LLM 的文本生成技术生成的翻译不仅更加符合人类语言习惯，还具备处理复杂语境和动态整合外部信息的能力，从而从根本上提升了机器翻译的效率和准确性[65]。

6. 未来展望

基于 LLM 的文本生成技术在 NLP 领域展现出强大潜力和广泛应用前景。然而，这一技术的发展仍面临诸多挑战，包括幻觉现象、灾难性遗忘、生成控制及伦理与偏见问题[66]-[69]。幻觉现象指基于 LLM 的文本生成系统在缺乏足够上下文或知识支撑时生成的表面合理但实际不准确的内容，影响了生成文本的真实性[58] [59]；灾难性遗忘限制了基于 LLM 的文本生成系统在多任务和长文本生成中的应用，导致其难以保持内容一致性[30]；而在生成风格及语义一致性方面的控制困难，影响了对特定文本输出的精确度[65]。此外，基于 LLM 的文本生成系统在生成过程中可能无意放大或重现训练数据中的偏见，带来伦理与公平性风险[70]。

为应对这些挑战，未来研究将聚焦于提升生成文本的事实性与精确性，尤其是在长文本生成中引入 RAG 和多轮验证机制，以确保生成内容符合最新知识，减少幻觉现象[56] [58]。在应对灾难性遗忘方面，弹性权重整合和生成回放技术为多任务或长时记忆环境中的基于 LLM 文本生成系统提供了更好的知识保留能力[59] [71]。同时，未来将进一步优化生成控制机制，借助条件生成模型实现对文本结构、风格和语义的细粒度控制，使用户能够动态调整生成内容[57]。此外，针对伦理与偏见问题，应开发自动审查机制和多样化的数据集，实时检测偏见并确保文本生成符合公正性和伦理标准[69] [70]。

参考文献

[1] Wikipedia (2013) Quantum EntanglementBrown. https://en.wikipedia.org/wiki/Quantum_entanglementBrown

[2] Brown, R.D. and Frederking, R. (1995) Applying Statistical English Language Modelling to Symbolic Machine Translation. *Proceedings of the Sixth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Leuven, 5-7 July 1995, 221-239.

[3] Tao, T., Wang, X., Mei, Q. and Zhai, C. (2006) Language Model Information Retrieval with Document Expansion. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, 4-9 June 2006, 407-414. https://doi.org/10.3115/1220835.1220887

[4] Zhai, C. and Lafferty, J. (2001) Model-Based Feedback in the Language Modeling Approach to Information Retrieval. *Proceedings of the Tenth International Conference on Information and Knowledge Management*, Atlanta, 5-10 October 2001, 403-410. https://doi.org/10.1145/502585.502654

[5] Le Quoc, V. (2014) Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, **27**, 3104-3112.

[6] Li, J., Li, S., Zhao, W.X., He, G., Wei, Z., Yuan, N.J., *et al*. (2020) Knowledge-Enhanced Personalized Review Generation with Capsule Graph Neural Network. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, Virtual, 19-23 October 2020, 735-744. https://doi.org/10.1145/3340531.3411893

[7] Li, J., Zhao, W.X., Wen, J. and Song, Y. (2019) Generating Long and Informative Reviews with Aspect-Aware Coarse-to-Fine Decoding. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 1969-1979. https://doi.org/10.18653/v1/p19-1190

[8] Bahdanau, D. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. arXiv Preprint, arXiv: 1409.0473.

[9] See, A., Liu, P.J. and Manning, C.D. (2017) Get to the Point: Summarization with Pointer-Generator Networks. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, 30 July-4 August 2017, 1073-1083. https://doi.org/10.18653/v1/p17-1099

[10] Iqbal, T. and Qureshi, S. (2022) The Survey: Text Generation Models in Deep Learning. *Journal of King Saud University-Computer and Information Sciences*, **34**, 2515-2528. https://doi.org/10.1016/j.jksuci.2020.04.001

[11] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N. and Huang, X. (2020) Pre-Trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, **63**, 1872-1897. https://doi.org/10.1007/s11431-020-1647-3

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[13] Devlin, J. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv Preprint, arXiv:1810.04805.

[14] Radford, A., Wu, J., Child, R., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. Preprint, OpenAI Blog.

[15] Brown, T.B. (2020) Language Models Are Few-Shot Learners. arXiv Preprint, arXiv:2005.14165.

[16] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., *et al*. (2020) BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 7871-7880. https://doi.org/10.18653/v1/2020.acl-main.703

[17] Raffel, C., Shazeer, N., Roberts, A., *et al*. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, **21**, 1-67.

[18] Kaplan, J., McCandlish, S., Henighan, T., *et al*. (2020) Scaling Laws for Neural Language Models. arXiv Preprint, arXiv:2001.08361.

[19] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., *et al*. (2023) Large Language Models Encode Clinical Knowledge. *Nature*, **620**, 172-180. https://doi.org/10.1038/s41586-023-06291-2

[20] Chowdhery, A., Narang, S., Devlin, J., *et al*. (2023) Palm: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, **24**, 1-113.

[21] Wang, W., Bi, B., Yan, M., *et al*. (2019) Structbert: Incorporating Language Structures into Pre-Training for Deep Language Understanding. arXiv Preprint, arXiv:1908.04577.

[22] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., *et al*. (2023) A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from Gan to ChatGPT. arXiv Preprint, arXiv: 2303.04226.

[23] Sanner, S., Balog, K., Radlinski, F., Wedin, B. and Dixon, L. (2023) Large Language Models Are Competitive Near Cold-Start Recommenders for Language- and Item-Based Preferences. *Proceedings of the* 17*th ACM Conference on Recommender Systems*, Singapore, 18-22 September 2023, 890-896. https://doi.org/10.1145/3604915.3608845

[24] Liang, X., Wang, H., Wang, Y., *et al*. (2024) Controllable Text Generation for Large Language Models: A Survey. arXiv Preprint, arXiv:2408.12599.

[25] Bubeck, S., Chandrasekaran, V., Eldan, R., *et al*. (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv Preprint, arXiv:2303.12712.

[26] Anil, R., Dai, A.M., First, O., *et al*. (2023) Palm 2 Technical Report. arXiv Preprint, arXiv:2305.10403.

[27] Wei, J., Wang, X., Schuurmans, D., *et al*. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language

Models. *Advances in Neural Information Processing Systems*, **35**, 24824-24837.

[28] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) Improving Language Understanding by Generative Pre-Training. Preprint.

[29] Abdaljalil, S. and Bouamor, H. (2021) An Exploration of Automatic Text Summarization of Financial Reports. *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, Online, 19 August 2021, 1-7.

[30] Keskar, N.S., McCann, B., Varshney, L.R., *et al*. (2019) CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv Preprint, arXiv:1909.05858.

[31] Yao, S., Yu, D., Zhao, J., *et al*. (2024) Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Proceedings of the* 37*th International Conference on Neural Information Processing Systems*, New Orleans, 10-16 December 2023, 11809-11822.

[32] Fedus, W., Zoph, B. and Shazeer, N. (2022) Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, **23**, 1-39.

[33] Birhane, A., Kasirzadeh, A., Leslie, D. and Wachter, S. (2023) Science in the Age of Large Language Models. *Nature Reviews Physics*, **5**, 277-280. https://doi.org/10.1038/s42254-023-00581-4

[34] Taylor, R., Kardas, M., Cucurull, G., *et al*. (2022) Galactica: A Large Language Model for Science. arXiv Preprint, arXiv: 2211.09085.

[35] Touvron, H., Lavril, T., Izacard, G., *et al*. (2023) Llama: Open and Efficient Foundation Language Models. arXiv Preprint, arXiv: 2302.13971.

[36] Dong, Q., Li, L., Dai, D., *et al*. (2022) A Survey on In-Context Learning. arXiv Preprint, arXiv: 2301.00234.

[37] Biderman, S., Schoelkopf, H., Anthony, Q.G., *et al*. (2023) Pythia: A Suite for Analyzing Large Language Models across Training and Scaling. *Proceedings of the* 40*th International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 2397-2430.

[38] Hoffmann, J., Borgeaud, S., Mensch, A., *et al*. (2022) Training Compute-Optimal Large Language Models. arXiv Preprint arXiv: 2203.15556.

[39] Rosenfeld, R. (2000) Two Decades of Statistical Language Modeling: Where Do We Go from Here? *Proceedings of the IEEE*, **88**, 1270-1278. https://doi.org/10.1109/5.880083

[40] Touvron, H., Martin, L., Stone, K., *et al*. (2023) LLAMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv Preprint, arXiv: 2307.09288.

[41] Achiam, J., Adler, S., Agarwal, S., *et al*. (2023) GPT-4 Technical Report. arXiv Preprint, arXiv: 2303.08774.

[42] Wei, J., Tay, Y., Bommasani, R., *et al*. (2022) Emergent Abilities of Large Language Models. arXiv Preprint, arXiv: 2206.07682.

[43] Huberman, B.A. and Hogg, T. (1987) Phase Transitions in Artificial Intelligence Systems. *Artificial Intelligence*, **33**, 155-171. https://doi.org/10.1016/0004-3702(87)90033-6

[44] Rae, J.W., Borgeaud, S., Cai, T., *et al*. (2021) Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv Preprint, arXiv: 2112.11446.

[45] Sanh, V., Webson, A., Raffel, C., *et al*. (2022) Multitask Prompted Training Enables Zero-Shot Task Generalization. arXiv Preprint, arXiv: 2110.08207.

[46] Ouyang, L., Wu, J., Jiang, X., *et al*. (2022) Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, **35**, 27730-27744.

[47] Wei, J., Bosma, M., Zhao, V.Y., *et al*. (2021) Finetuned Language Models Are Zero-Shot Learners. arXiv Preprint, arXiv: 2109.01652.

[48] Thoppilan, R., De Freitas, D., Hall, J., *et al*. (2022) LAMDA: Language Models for Dialog Applications. arXiv Preprint, arXiv: 2201.08239.

[49] Chung, H.W., Hou, L., Longpre, S., *et al*. (2024) Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, **25**, 1-53.

[50] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., *et al*. (2022) BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, **23**, bbac409. https://doi.org/10.1093/bib/bbac409

[51] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., *et al*. (2023) Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nature Biotechnology*, **41**, 1099-1106. https://doi.org/10.1038/s41587-022-01618-2

[52] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., *et al*. (2023) Self-Instruct: Aligning Language Models with Self-Generated Instructions. *Proceedings of the* 61*st Annual Meeting of the Association for Computational*

Linguistics (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 13484-13508. https://doi.org/10.18653/v1/2023.acl-long.754

[53] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., *et al*. (2021) Evaluating Large Language Models Trained on Code. arXiv E-Prints.

[54] Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N.Q., Bernardi, R., Pezzelle, S., *et al*. (2016) The LAMBADA Dataset: Word Prediction Requiring a Broad Discourse Context. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Berlin, 7-12 August 2016, 1525-1534. https://doi.org/10.18653/v1/p16-1144

[55] Kocmi, T., Bawden, R., Bojar, O., *et al*. (2022) Findings of the 2022 Conference on Machine Translation (WMT22). *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), Abu Dhabi, 7-8 December 2022, 1-45.

[56] Narayan, S., Cohen, S.B. and Lapata, M. (2018) Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 1797-1807.

[57] Chen, M., Tworek, J., Jun, H., *et al*. (2021) Evaluating Large Language Models Trained on Code. arXiv Preprint, arXiv: 2107.03374.

[58] Lewis, P., Perez, E., Piktus, A., *et al*. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, **33**, 9459-9474.

[59] Xi, Z., Chen, W., Guo, X., *et al*. (2023) The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv Preprint, arXiv: 2309.07864.

[60] Wu, N., Gong, M., Shou, L., Liang, S. and Jiang, D. (2023) Large Language Models Are Diverse Role-Players for Summarization Evaluation. *Natural Language Processing and Chinese Computing*, Foshan, 12-15 October 2023, 695-707. https://doi.org/10.1007/978-3-031-44693-1_54

[61] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K. and Hashimoto, T.B. (2024) Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, **12**, 39-57. https://doi.org/10.1162/tacl_a_00632

[62] Kocmi, T. and Federmann, C. (2023) Large Language Models Are State-of-the-Art Evaluators of Translation Quality. arXiv Preprint, arXiv: 2302.14520.

[63] Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., *et al*. (2023) Document-Level Machine Translation with Large Language Models. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 16646-16661. https://doi.org/10.18653/v1/2023.emnlp-main.1036

[64] Kobusingye, B.M., Dorothy, A., Nakatumba-Nabende, J. and Marvin, G. (2023) Explainable Machine Translation for Intelligent E-Learning of Social Studies. 2023 7*th International Conference on Trends in Electronics and Informatics* (*ICOEI*), Tirunelveli, 11-13 April 2023, 1066-1072. https://doi.org/10.1109/icoei56765.2023.10125599

[65] Dathathri, S., Madotto, A., Lan, J., *et al*. (2019) Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv Preprint, arXiv: 1912.02164.

[66] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., *et al*. (2023) Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, **55**, 1-38. https://doi.org/10.1145/3571730

[67] Maynez, J., Narayan, S., Bohnet, B. and McDonald, R. (2020) On Faithfulness and Factuality in Abstractive Summarization. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1906-1919. https://doi.org/10.18653/v1/2020.acl-main.173

[68] French, R. (1999) Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences*, **3**, 128-135. https://doi.org/10.1016/s1364-6613(99)01294-2

[69] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the* 2021 *ACM Conference on Fairness, Accountability, and Transparency*, Virtual, 3-10 March 2021, 610-623. https://doi.org/10.1145/3442188.3445922

[70] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., *et al*. (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 33-44. https://doi.org/10.1145/3351095.3372873

[71] Guo, Z., Jin, R., Liu, C., *et al*. (2023) Evaluating Large Language Models: A Comprehensive Survey. arXiv Preprint, arXiv: 2310.19736.