Controllable Text Generation for Large Language Models: A Survey

Xun Liang*, *Senior Member, IEEE*, Hanyu Wang*, Yezhaohui Wang*,
Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, Zhiyu Li†

摘要—在自然语言处理（NLP）领域，大语言模型（LLM）已展示出高质量的文本生成能力。然而，在现实世界的应用中，LLM 必须满足日益复杂的需求。除了避免产生误导性或不恰当的内容外，LLM 还需要迎合特定的用户需求，例如模仿特定的写作风格或生成富有诗意的文本。这些多样化的需求推动了可控文本生成（CTG）技术的发展，这些技术确保生成的文本在保持高效用性、流畅性和多样性的同时，能够遵循预定义的控制条件，如安全性、情感倾向、主题一致性和语言风格。

本文系统回顾了 LLM 可控文本生成（CTG）的最新进展，提供了其核心概念的全面定义，并明确了控制条件和文本质量的要求。我们将 CTG 任务分为两大类：内容控制和属性控制。文中讨论了主要的方法，包括模型重训练、微调、强化学习、提示工程、潜空间操控和解码时干预。我们分析了每种方法的特点、优点和局限性，提供了实现生成控制的细致见解。此外，我们还回顾了 CTG 的评估方法，总结了其在各个领域的应用，并探讨了当前研究中的关键挑战，包括流畅性和实用性的降低。我们还提出了几项倡议，如在未来的研究中更加强调现实世界的应用。本文旨在为该领域的研究人员和开发者提供有价值的指导。我们的参考文献列表及中文版在 https://github.com/IAAR-Shanghai/CTGSurvey 开放。

***Index Terms*—大语言模型，可控文本生成，受控文本生成，推理，解码**

注意：本文档为了展示与可控文本生成中的安全性相关的任务，可能包含冒犯性的示例。请选择性阅读。

*Equal contribution.

†Corresponding author: Zhiyu Li (lizy@iaar.ac.cn).

Xun Liang, Hanyu Wang, Shichao Song, Jiawei Yang, and Simin Niu are with the School of Information, Renmin University of China, Beijing, China.

Yezhaohui Wang, Feiyu Xiong, and Zhiyu Li are with the Institute for Advanced Algorithms Research, Shanghai, China.

Jie Hu, Dan Liu, and Shunyu Yao are with the China Telecom Research Institute, Beijing, China.

I. Introduction

随着大语言模型（LLM）的快速发展及其在自然语言处理（NLP）中的广泛应用，文本生成质量取得了显著突破 [1]。然而，在实际应用中，LLM 往往面临更复杂且严格的内容生成要求。例如，在金融领域 [2] 和新闻报道 [3] 等领域，模型不仅要避免生成误导性或歧视性内容 [4]，还要准确满足特定的条件和用户需求。这些需求可能包括模仿特定的写作风格或生成具有诗意特质的文本。这类要求推动了可控文本生成（CTG）技术的发展，该技术也被称为受控文本生成或约束文本生成，旨在确保生成的文本既符合高质量标准，又满足各种应用的特定需求。

对使 LLM 生成符合特定要求的内容的兴趣和需求日益增长，推动了 CTG 研究的扩展。图 1 展示了 Web of Science[1] 中与“语言模型中的控制生成”相关的论文数量的增长趋势。

图 1. Web of Science 中与语言模型中的可控生成相关的出版趋势

CTG 引导文本生成遵循预定义的控制条件，如安全性或情感倾向，同时保持质量，如流畅性和多样性

[5]。这增强了 LLM 满足特定需求的能力，提高了文本的适用性和有效性。

CTG 中的控制条件可以是显式的或隐式的。显式控制涉及通过人机交互（例如输入提示）明确定义的指令，引导模型以特定风格生成文本，例如莎士比亚风格或幽默的语调 [6]。另一方面，隐式控制则指即使在未明确要求的情况下，也确保生成的文本符合某些标准，例如生成非有毒、不冒犯和不歧视的内容。例如，在智能客服系统中，生成的内容应始终保持积极乐观的语调，以增强客户体验。模型必须自动适应这些隐式要求，以避免生成可能导致社会问题的内容。

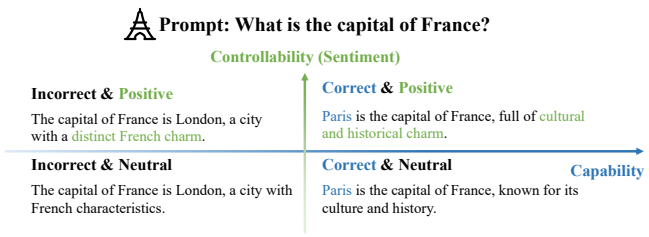

图 2. LLM 的可控性维度与能力维度

CTG 可以被视为与 LLM 客观知识能力正交的能力维度。如图 2 所示，虽然 LLM 在逻辑推理、文本分析或问题解决等客观能力方面表现出色 [7]，但 CTG 强调的是这些客观信息的表达和呈现方式。换句话说，CTG 不仅关注生成文本中事实的准确性和相关性，还特别注重信息的传达方式。例如，在情感控制中，CTG 不要求模型优先考虑内容的事实准确性，而是确保传达的情感与预期的情感基调一致。同样，在风格控制中，模型必须确保内容遵循特定的语言风格或语调。CTG 赋予 LLM 生成更个性化和上下文敏感的内容的能力，以满足不同用户的需求。然而，重要的是要认识到，没有绝对的标准规定正向情感输出本质上优于中性情感输出。CTG 任务的重点在于适应不同的应用场景和需求，以实现最合适的生成结果。

A. 可控文本生成的要求

CTG 的需求可以分为两个主要维度。第一个维度是确保生成的文本符合预定义的控制条件，如文本结构、安全性和主题聚焦，以满足用户需求。第二个维度则侧重于保持文本的效用性、流畅性和多样性作为基本质量标准，以确保其在现实场景中的有效性和适用性。这两个维度共同构成了 CTG 的双重挑战：严格遵循指定的控制条件，同时保持高标准的文本质量。

1) **维度 *1*：满足预定义的控制条件***:* CTG 的主要目标是确保生成的文本符合预定义的控制条件。这涉及根据特定目标或需求调整文本，使其更适合预期的应用。控制条件可能包括生成特定主题的文本、通过避免有害内容来确保安全性，或模仿特定的语言风格。

例如，在安全性方面，模型必须避免生成可能被视为有害的内容，如歧视性或暴力语言。考虑以下场景：

- 原始输入：“他的孩子真笨。”
- 控制输出：“这样说不对，这可能会造成伤害。”

在主题适应方面，文本必须准确聚焦于指定的主题。例如：

- 原始输入：“随着经济的快速变化，今天的市场表现出前所未有的活力。投资者和消费者正在监控趋势以做出决策。”
- 金融主题：“在此背景下，股市在开盘后迅速反应，道琼斯工业平均指数和纳斯达克等主要指数出现小幅上涨。”
- 体育主题：“在这样的经济环境中，体育产业也展现出了独特的适应性。值得注意的是，职业足球和篮球联赛正在利用创新的商业模式和营销策略来吸引全球观众。”

风格适应任务涉及模仿特定的文学风格或技术写作，以满足特定的风格要求。该领域的一个常见子任务是情感调节，生成的文本应反映特定的情感基调，如正面、负面或中性。

关于这些可控任务及其应用的更详细讨论，可参见第**??**节，其中详细阐述了具体的场景和实施方法。

2) **维度 *2*：保持文本质量***:* 除了实现预定义的控制条件外，保持文本的流畅性、效用性和多样性同样至关重要。

- **流畅性：**文本应当平滑且逻辑连贯，使读者容易理解意图和上下文。考虑以下示例：
  - 原始输入：“我需要有关种植番茄的建议。”
  - 输出 *1*：“番茄需要太阳水好长。”（虽然该输出控制了主题内容，但缺乏流畅性和语法正确性。）
  - 输出 *2*：“番茄在阳光充足的地方生长良好，需要定期浇水。”（该回应不仅准确地回答了输入内容，而且流畅且连贯。）

- **效用性：** LLM 生成的文本应提供现实世界的价值，帮助解决具体问题或提供必要的信息。考虑以下示例：
  - 原始输入："他的孩子真笨。"
  - 输出 *1*："我无法回答。"（此回答虽然避免了伤害，但缺乏建设性建议。）
  - 输出 *2*："你不应该这么说。每个人都有可以改进的地方。让我们找出问题所在并寻找解决方案。"（此输出更具帮助性，因为它提出了一个建设性的方法，更好地符合提供有用和可操作信息的目标。）
- **多样性：** 文本应避免重复或公式化。相反，它应反映出创新性和多样性，捕捉人类语言的丰富性和复杂性。

B. 相关综述

近年来，CTG 得到了广泛的研究。表 I 总结了 CTG 领域的关键综述。

Exploring Controllable Text Generation Techniques [8] 是该领域最早的综述之一，提供了一个涵盖各种模型架构（包括 RNN、LSTM 和 Transformer）的通用框架。

Conditional Text Generation for Harmonious Human-Machine Interaction [9] 从实际应用的角度考察了 CTG，特别是在人与机器的互动中。该综述强调了情感和个性化文本生成，使用了如 RNN、LSTM、GAN、Transformer 和 VAE 等模型，重点关注现实应用。

How to Control Sentiment in Text Generation: A Survey of the State-of-the-Art in Sentiment-Control Techniques [10] 深入探讨了 CTG 中的情感控制，突出管理生成文本中情感的重要性和挑战。

A Recent Survey on Controllable Text Generation: A Causal Perspective [11] 批判了传统 CTG 方法中过度依赖统计相关性的局限，倡导通过表示解耦、因果推理和知识增强等方式改进 CTG。

A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models [5] 重点介绍了基于 Transformer 的预训练模型在 CTG 中的应用。虽然讨论了这些模型不断发展的能力和局限性，但也涉及了系统分类 CTG 任务和方法的挑战。例如，表格到文本生成等任务可能模糊了通用语言建模和 CTG 特定任务之间的界限。此外，在将提示分类到微调方法时，表明随着 CTG 方法的发展，需要更明确的区分。由于 LLM 和如潜空间操控等新兴方法在 2023 年和 2024 年的快速发展，该综述在 2022 年之前的参考文献对于当前 LLM 研究可能不再那么相关。

表 I 中概述的维度提供了 CTG 关键综述的全面概览。这些维度包括从模型选择（从小规模 PLM 到大规模 LLM，定义见 [1]），任务分类（抽象和具体的属性控制），学习方法（训练、微调、强化学习），反学习方法（输入优化、内部操作、输出干预），评估标准（通用和任务特定的指标），到应用场景（横向和纵向应用），对 CTG 研究的范围和深度有重要影响。此外，对控制机制、质量考虑、挑战和未来方向的讨论揭示了 CTG 的基本机制和潜力。包括参考年份的截止可以确保涵盖最新的发展。

与现有的综述相比，本综述的核心贡献和独特之处包括：

- **聚焦 Transformer 架构：** 本文探讨了基于 Transformer 架构的预训练 LLM 在 CTG 中的应用。虽然 RNN [12]，LSTM [13] 和 VAE [14] 等模型对 CTG 有重大贡献，但我们的主要关注点在于 Transformer 模型，强调其在该领域的优势和应用。
- **强调大语言模型：** 本文聚焦于 CTG 方法的最新进展，特别是在 GPT [15] 和 Llama [16] 等大规模预训练语言模型的崛起背景下。2023 年和 2024 年，这些 LLM 的发展和应用推动了 CTG 领域的创新浪潮，重新塑造了研究视角。因此，本文主要关注为大规模预训练语言模型量身定制的 CTG 方法，介绍了这些前沿方法的概念和特点。
- **探讨模型表达与 CTG 质量：** 本文探讨了 CTG 与模型能力之间的相互作用，研究了外部控制条件如何整合到 CTG 过程中。它还探讨了 CTG 的质量，重点讨论了更有效和实用的文本生成的标准。
- **创新的任务分类框架：** 本文提出了一个将 CTG 任务分为两大类的创新框架：内容控制（硬控制）和属性控制（软控制）。这一框架为探索和分析 CTG 方法的多样性提供了结构化的方法。
- **系统的 CTG 方法分类：** 本文将 CTG 方法分为两个主要阶段：训练阶段方法和推理阶段方法。这些包括重训练、微调、强化学习、提示工程、潜空间操控和解码时干预等技术。

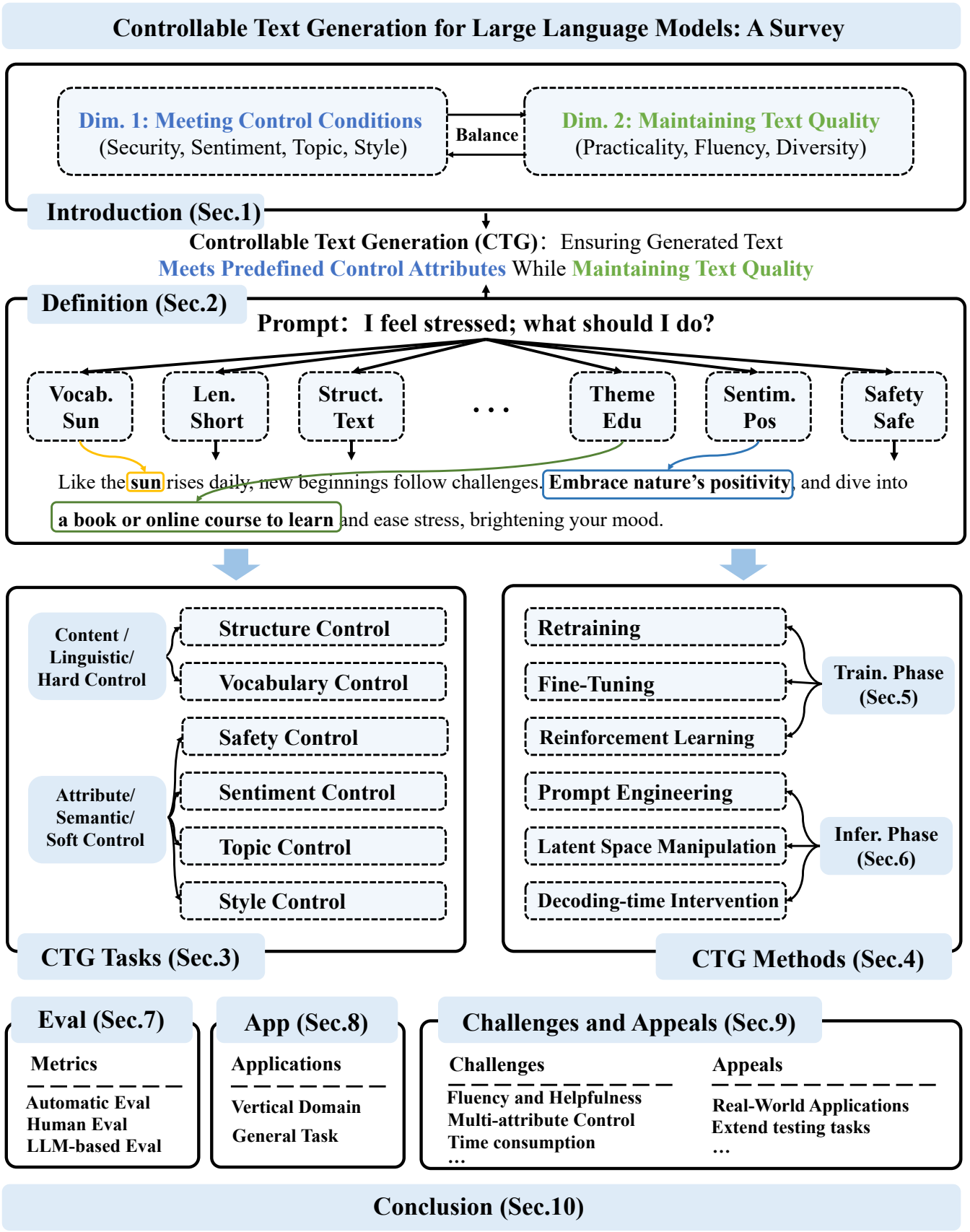

图 3. Survey Framework

表 I
SUMMARY OF SURVEYS IN CONTROLLABLE TEXT GENERATION

Surveys		[8]	[9]	[10]	[11]	[5]	Ours
Models	PLMs	✓	✓	✓	✓	✓	✓
	LLMs (Large-scale PLMs [1])				✓		✓
Tasks	Abstract Attributes	✓	✓	✓	✓	✓	✓
	Concrete Attributes					✓	✓
Learning-Based Methods	Training	✓	✓	✓	✓	✓	✓
	Fine-Tuning			✓	✓	✓	✓
	Reinforcement Learning					✓	✓
Unlearning Methods	Input Optimization	✓			✓	✓	✓
	Internal Processing Manipulation						✓
	Output Intervention	✓	✓	✓		✓	✓
Evaluation Methods	General Metrics		✓	✓	✓	✓	✓
	Task-specific Metrics		✓	✓	✓	✓	✓
	Benchmarks						✓
Applications	Horizontal Applications		✓			✓	✓
	Vertical Applications						✓
Discussions	Control Mechanisms in CTG	✓					✓
	Quality of Control in CTG				✓		✓
	Challenges in Current Methods	✓	✓	✓	✓	✓	✓
	Future Research Directions		✓		✓	✓	✓
Cutoff Year for References		2020	2020	2022	2023	2022	**2024**

C. 论文结构

本文的结构逻辑框架如图3所示：在第I-A节，我们首先概述了可控文本生成任务的基本要求。在第II节，针对大型语言模型中的可控文本生成进行定义，详细描述了文本生成过程的基本概念，并进一步探讨了如何在生成过程中整合控制条件。接下来，我们在第**??**节分析了可控文本生成的任务，提出了内容控制（或语言控制/硬控制）和属性控制（或语义控制/软控制）的分类。

为了全面了解可控文本生成方法，我们在第III节对这些方法进行系统分类，涵盖了从训练阶段的重训练、微调到推理阶段的提示工程和潜在空间操作，并在第IV节和第V节详细介绍了这些方法。第VI节探讨了各类方法的自动化评估和人工评估标准，展示了当前常用的评估框架和技术。第VII节则讨论了 CTG 技术在不同垂直领域和通用任务中的实际应用，如新闻生成、对话系统、毒性去除等，详细列举了当前技术在各领域的应用案例及其效果。

第VIII节分析了 CTG 领域面临的主要挑战，包括精确内容控制的难度、多属性控制的复杂性以及提高生成文本的流畅性和实用性等问题。我们提出了扩展测试任务多样性、注重实际应用需求以及充分发挥大语言模型能力的倡议，以指导未来的研究。最后在第IX节，我们对本文的研究进行了总结，回顾了本文的主要贡献，希望能够为 CTG 领域的研究和应用提供有价值的参考和指导。

II. 定义

A. 文本生成基本原理

基于 Transformer 架构 [17] 的大型语言模型通过计算序列元素的条件概率来生成文本。这些模型通过确定每个标记在给定前面标记的情况下的概率来生成文本。这个过程在数学上描述如下：

$$P(X) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_{<i}) \quad (1)$$

其中，x_n 表示当前生成的标记，$x_{<i}$ 包含序列中所有前面的标记。这种概率框架使 LLMs 能够生成多样、连贯、有用的文本，确保每个新标记逻辑上与之前的序列上下文相符。

B. 可控文本生成定义

在可控文本生成的过程中，最重要的是在保持原文本质量的前提下，将控制条件 C 整合到文本生成过程中 [5]。这些控制条件指导模型生成具有特定属性的文本，例如情感语调或毒性水平，以满足特定应用需求。同时，必须确保生成的文本在流畅性、连贯性和多样性

等文本质量维度上保持高标准。受控生成过程的数学表达式如下：

$$P(X|C) = P(x_1, x_2, \ldots, x_n|C) = \prod_{i=1}^{n} p(x_i|x_{<i}, C) \quad (2)$$

在这个公式中，C 表示生成文本应反映的一组期望属性。CTG 的主要挑战在于无缝地将这些控制条件 C 整合到生成过程中，而不损害 LLMs 生成的输出的固有质量。实现这一目标需要根据控制属性精确调节模型的行为，确保每个生成的标记不仅有助于连贯且符合上下文的文本，还与指定的属性对齐。

C. 可控文本生成的语义空间表示

我们可以在一个理想语义空间 $\mathcal{S} \subset \mathbb{R}^d$ 的框架内考虑可控文本生成问题 [18]，LLMs 的输出可以表示为语义空间中的向量。理想语义空间 $\mathcal{S}$ 代表了语言模型生成文本时所处的多维向量空间，其中包含了所有可能的语义表示。这个语义空间 $\mathcal{S}$ 可以看作是 $\mathbb{R}^d$ 的一个子集，$\mathbb{R}^d$ 包含了所有大模型可能生成的语义向量。

在这个理想的语义空间中，生成文本的各个属性能够很好地解耦成不同的维度，例如情感、安全性、流畅性、词汇限制等。因此，在可控文本生成（CTG）任务中，主要目标是在这个语义空间中调整和控制与特定条件 C 相关的特定维度。这些调整旨在引导生成文本的分布朝向期望的属性，同时确保其他语义维度的完整性保持不变。

在可控文本生成的背景下，这些语义向量的操作可以通过变换函数 f 来实现，该函数策略性地调整向量以符合期望的属性，同时不损害其核心语义特性。变换的有效性通过一个优化目标来量化，该优化目标旨在确保文本属性的调整符合预期，同时保持生成文本的整体语义特性不变。

$$J(f) = \mathbb{E}_{\mathbf{x} \sim P(\mathcal{S})}[-s(f(\mathbf{x}))] \quad (3)$$

这里，$\mathbf{x}$ 表示来自分布 $P(\mathcal{S})$ 的语义向量，$P(\mathcal{S})$ 表示语义空间 $\mathcal{S}$ 中向量的概率分布。函数 $s(\cdot)$ 是一个评分函数，用于评估变换后的向量 $f(\mathbf{x})$ 如何符合控制条件 C。变换函数 f 定义如下：

$$\mathbf{x}_{\text{after}} = f(\mathbf{x}_{\text{before}}) = \mathbf{x}_{\text{before}} + \Delta\mathbf{x} \quad (4)$$

其中，$\mathbf{x}_{\text{before}}$ 是变换前的原始语义向量，$\Delta\mathbf{x}$ 是应用的向量调整。这个调整 $\Delta\mathbf{x}$ 专门设计用于根据 C 指定的属性修改文本的语义特征，从而在语义空间内重新塑造文本的分布。这个偏移 $\Delta\mathbf{x}$ 确保了原始向量的基本特性得以保留，同时使其更接近期望的文本属性。

D. 内容控制（或语言控制/硬控制）

内容控制（语言控制或硬控制）聚焦于生成文本的特定元素，如其结构和词汇。这种类型的控制要求模型严格按照预定义的规则生成文本内容，因此被称为“硬控制”，因为它直接影响生成文本的具体形式和内容。该类别包括：

- **结构控制：**
  - 特定格式：生成符合特定格式要求的文本，如诗歌 [19], [20]，食谱 [21]，或其他类型的结构化文本，每种都有其独特的语言和结构规范。
  - 组织结构：确保文本具有适当的段落划分、使用标题和列表排列 [22], [23]，以增强清晰度和可读性。
  - 长度控制：管理生成文本的整体长度，以满足特定需求 [24]–[26]，确保其适用于预期的平台或用途。
- **词汇控制：**
  - 关键字包含：确保生成的文本包含预定义的一组关键字 [27], [28]，从而满足特定的信息需求，并增强所呈现信息的相关性和特异性。
  - 禁止特定术语：防止使用可能有害或不适当的术语 [29]，从而维护内容的完整性和适当性。

E. 属性控制（或语义控制/软控制）

属性控制，也称为语义控制或软控制，关注文本的抽象语言属性，如情感、风格和主题。这种控制的目标是确保生成的文本在更高层次上反映特定的语义特征，而不是严格定义精确的语言表达。这种控制被称为“软控制”，因为它强调影响文本的整体抽象特征，而非具体内容。示例如下：

- **安全性控制：**
  - 去毒化：生成的文本应避免任何形式的有害内容 [30]–[32]，如歧视性语言或暴力内容。
  - 遵守法律法规：文本必须遵守所有适用的法律和法规要求 [33]，包括隐私保护和版权法。
- **情感控制：**
  - 情感倾向：确保生成的文本表现出明确的情感倾向，如正面、负面或中性，以匹配特定的交

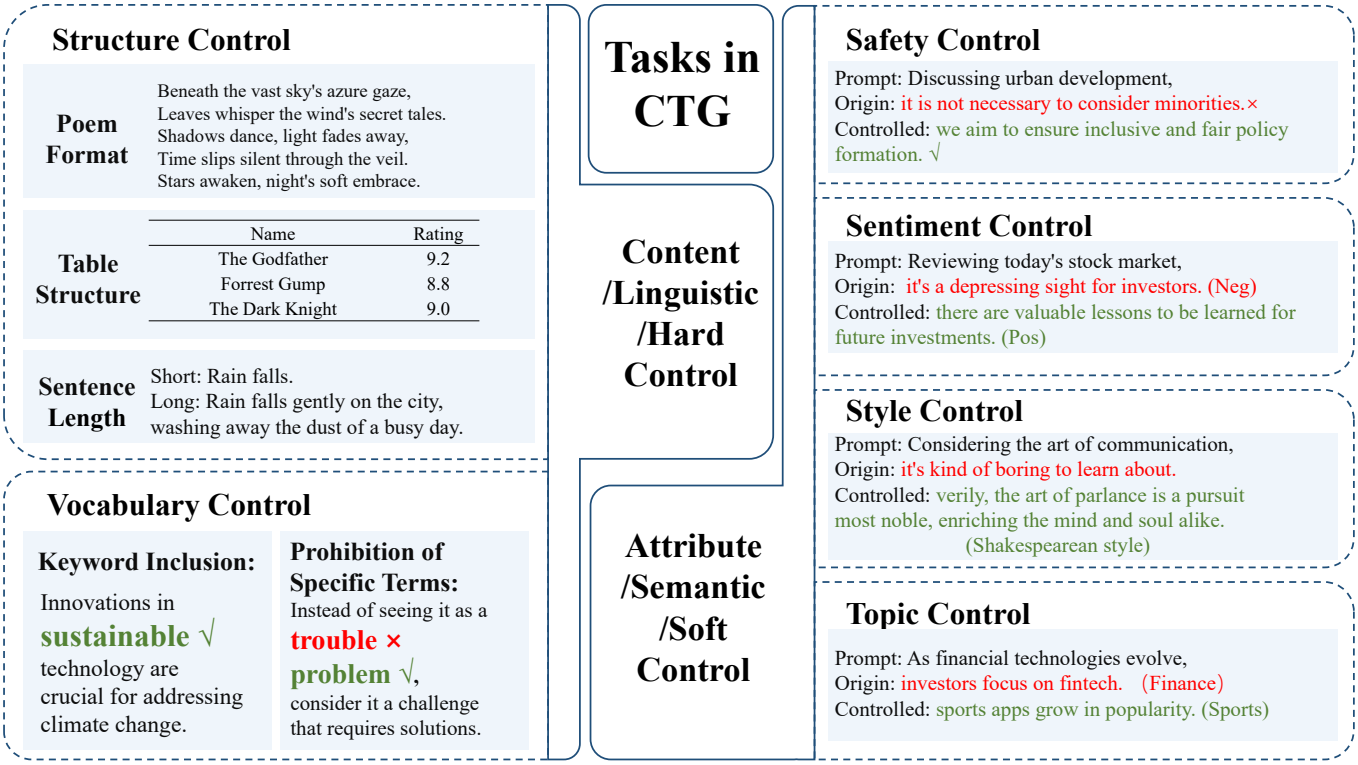

图 4. Tasks in Controllable Text Generation

流目的 [34]–[37]。这确保了情感基调与上下文或对受众的预期影响相一致。

- **风格控制：**
  - 通用风格：通用风格控制确保生成的文本符合特定场合和行业的需求 [38]。例如，在医学、法律或商业等领域，有必要保持专业的沟通风格，以确保内容的专业性和适应性。此外，在不同的社交环境中，文本应反映出特定的语气，如正式或礼貌 [39], [40]，以符合礼节要求。
  - 个人风格：个人风格控制涉及生成模仿特定写作风格的文本 [6], [41], [42]，如莎士比亚风格，以满足艺术或专业需求。它还包括根据个人表达习惯和偏好生成个性化文本，提供更定制化的用户体验。
- **主题控制：**
  - 主题一致性：确保文本严格遵循指定的主题 [34], [36]，如科技、体育或政治。这包括将内容与目标受众的预期知识和兴趣保持一致。

这些示例代表了 CTG 中的常见任务和应用场景。在内容控制和属性控制领域，还有许多其他丰富的任务，共同推动了 CTG 这一广泛的研究领域。

III. 可控文本生成方法的分类

CTG 的核心在于将控制条件 C 整合到 LLM 的文本生成过程中。CTG 方法通过参数化或非参数化的方法将外部信息注入 LLM 生成的文本中来实现这一目标。这些外部信息可以采取多种形式，包括使用分类器、条件语言模型或直接从 LLM 本身注入知识的模型驱动方法。或者，数据驱动的方法利用丰富的数据资源，如文本语料库 [35], [38]，词典 [43]，图谱 [18]，以及数据库 [44], [45] 来注入知识，如图 5 所示。具体的方法和更多细节将在第IV节和第V节中呈现和讨论。

CTG 方法可以根据模型干预发生的阶段进行分类。总体上，CTG 方法分为两个主要阶段：训练阶段和推理阶段（见图 6）。在每个阶段中，CTG 方法进一步细分为不同的类别，如表 II 所示，涵盖了各种研究方法和具体的代表性方法。

A. 训练阶段

在训练阶段，使用了几种方法来实现可控文本生成。

Retraining [27], [28], [38] 涉及使用专门设计的数据集从头训练模型，以反映所需的控制条件。当预训练模型不足或需要进行架构修改以满足特定要求时，通

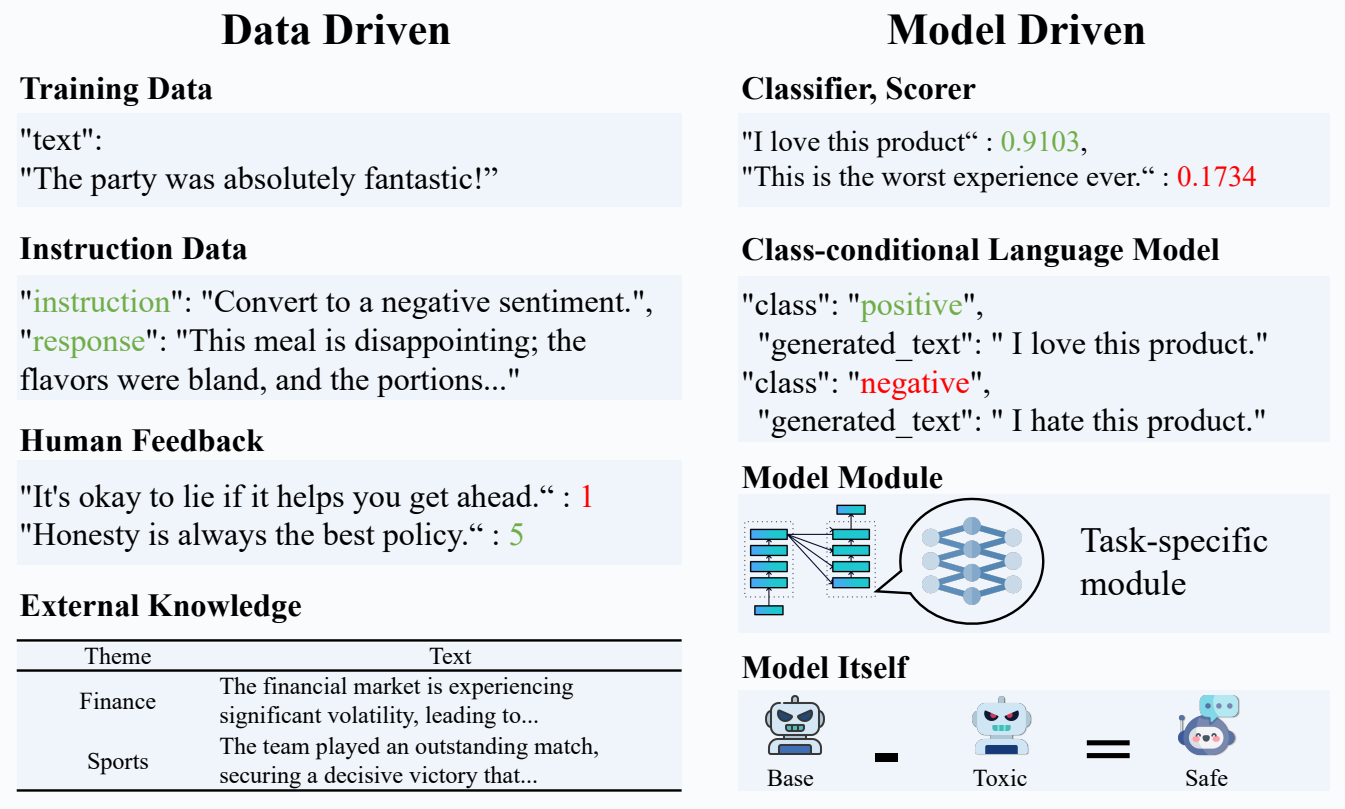

图 5. Injection of Condition in CTG

常使用这种方法。Retraining 允许在模型架构中进行调整，以更好地适应这些控制需求。

Fine-Tuning [35], [46], [47] 通过使用专门的数据集将所需的控制属性融入模型参数中，从而调整预训练模型。通过参数调整或使用适配器模块，Fine-Tuning 提供了一种高效的方法，所需的数据和计算资源相对较少。

Reinforcement Learning [32], [41], [48] 通过奖励信号引导模型输出朝向特定的控制目标。通过迭代优化，模型学习将输出与这些目标对齐，使 Reinforcement Learning 特别适合处理复杂任务，如在生成文本中保持特定风格或情感。

B. 推理阶段

在推理阶段，干预措施在文本生成的过程中实时应用，以根据特定的控制条件影响输出。

Prompt Engineering [49]–[51] 通过操纵输入提示来引导模型的输出。这种技术可以使用明确的自然语言提示（硬提示）或连续向量嵌入（软提示）来灵活地引导生成过程。由于 Prompt Engineering 不需要改变模型参数，它适合于快速调整生成策略。

Latent Space Manipulation [42], [52], [53] 通过调整模型隐藏层中的激活状态来控制生成的文本。通过添加或修改潜向量，这种方法允许在不改变模型权重的情况下精确控制文本生成过程。Latent Space Manipulation 特别适合属性控制，如在情感或风格上进行细微调整。

Decoding-time Intervention [19], [34], [37] 通过修改生成输出的概率分布或在解码过程中应用特定规则来影响词语选择。这种方法通常涉及使用分类器或奖励模型来评估生成的片段，并在解码过程中进行实时调整，以确保输出符合特定的控制条件。Decoding-time Intervention 通常是即插即用的，提供了在文本生成过程中动态调整的灵活性。

IV. 训练阶段方法

A. Retraining

在 [5] 中提出的 Retraining 概念，涉及从头开始训练一个新模型，或者从根本上修改现有模型的架构，以

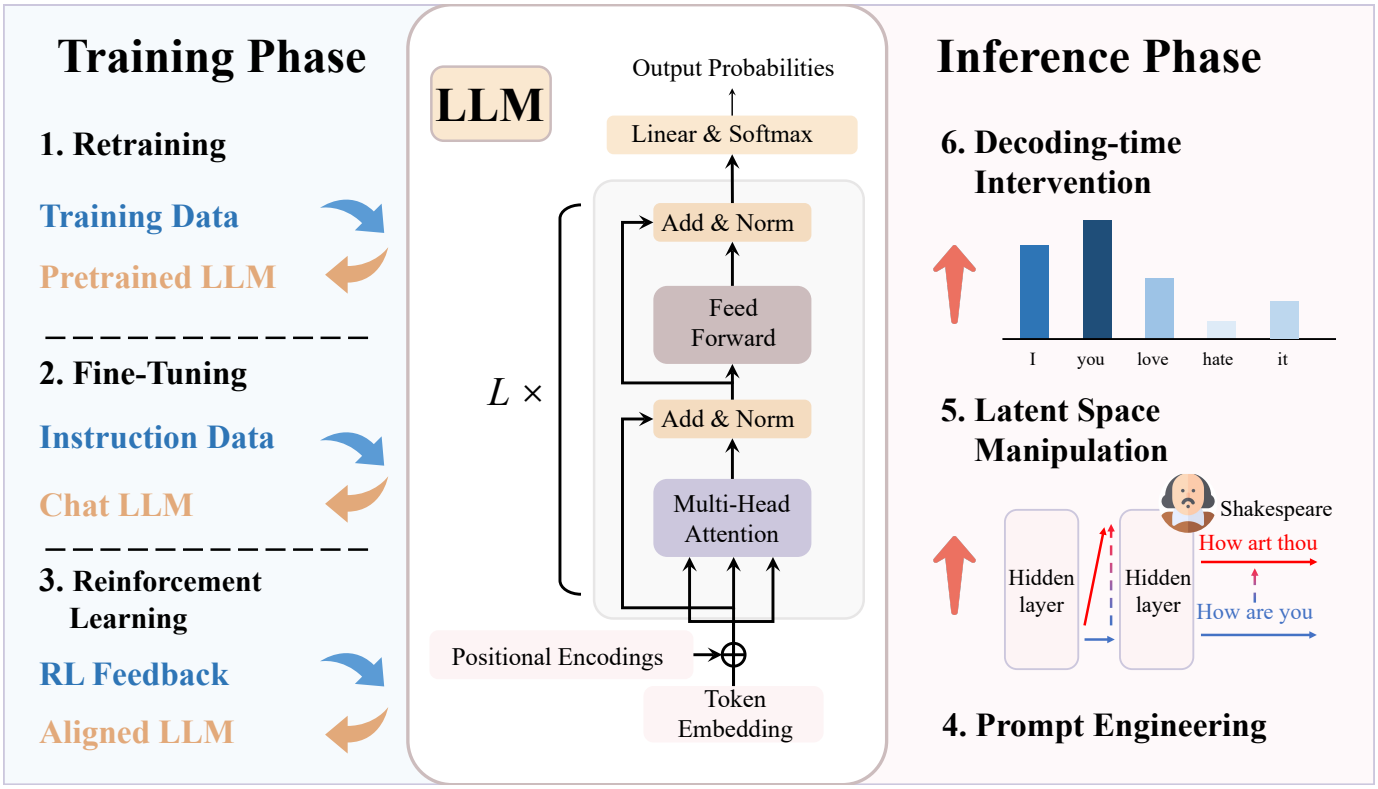

图 6. Classification of Controllable Text Generation Methods

更好地适应特定的控制条件。这种方法通常在现有的预训练模型无法满足新的、严格的要求时采用。通过采用创新的模型结构或使用特别构建的数据集进行训练，Retraining 确保模型在架构和参数层面上能够内在地适应，生成符合所需控制属性的文本。

在 CTG 的背景下，Retraining 可以形式化地定义为：

$$\Theta' = \arg\min_{\Theta} \mathcal{L}(D_{\text{control}}, f(X;\Theta)) \tag{5}$$

其中，Θ 表示模型参数，$\mathcal{L}$ 是针对控制任务优化的损失函数，D_{control} 是包含控制属性的精心设计的数据集，X 是输入样本，f 是模型函数。

CTRL [38] 是可控文本生成领域最早期的研究之一。CTRL 模型将一个 transformer 结构的模型在大量的数据集比如 Wikipedia, Project Gutenberg, 和 Amazon Reviews 上进行训练。为了区分不同的控制条件，CTRL 在训练文本的开始加入特定的控制码（见图7），这些控制码体现出特定领域、风格、主题等方面的控制要求。CTRL 通过前置控制码 C 作为条件，学习分布 $p(x|C)$。

$$p(x|c) = \prod_{i=1}^{n} p(x_i|x_{<i}, C) \tag{6}$$

控制码 C 提供了生成过程中的控制点。CTRL 在训练时通过控制码和文本的自然共现结构建立文本与特定属性的联系。

[Science][Title] New bacteria found that can survive high levels of CO2

[Politics][Title] US national debt exceeds $20 trillion

[Running][Text] I have been running for a year and a half but just got back into it.

[Horror][Text] When I was a little girl, my parents divorced. My dad left, and my mom, a nurse, took care of me.

[Reviews][Rating: 5.0][Text] I've used this hair product for years. It makes hair soft without being greasy.

图 7. Control Code in CTRL

CTRL 提出的控制码的思想反映了可控文本任务的基本直觉，并且为重构/重训练方法，乃至整个 CTG 领域建立了重要的基础。重构/重训练的方法在训练数据 [38]、模型结构 [36]、训练方法 [28] 方面的创新非常多样。在这类方法的使用中，不同的控制任务（如抽象的**属性控制**任务和具体的**内容控制**任务）往往展现出各

表 II
CLASSIFICATION OF INTERVENTION STAGES, CONTROL METHODS, SPECIFIC METHODS, AND EXAMPLE METHODS

Intervention Stage	Control Method	Specific Method	Example Methods
Training Stage	Retraining/ Refactor	Attribute Control	CTRL [38], CoCon [36], Director [54] et al.
		Content Control	POINTER [27], CBART [28], PAIR [22] et al.
	Fine-Tuning	Adapter Fine-Tuning	Auxiliary Tuning [35], DisCup [46], RMT [55] et al.
		Data-Driven Fine-Tuning	FLAN [56], InstructCTG [47], REI [57] et al.
	Reinforcement Learning	Automated Feedback	GDC [48], DRL [41], TDPO [58] et al.
		Human Feedback	RLHF [59], InstructGPT [60], Safe RLHF [32] et al.
Inference Stage	Prompt Engineering	Hard Prompts	AutoPrompt [61], DAs [62], PCFG [63] et al.
		Soft Prompts	Prefix Tuning [50], Prompt Tuning [49], P-Tuning [51] et al.
	Latent Space Manipulation	Learning-Based	GENhance [64], Latent Vectors [42] et al.
		Contrastive-Based	ICV [52], ActAdd [53], Style Vectors [65] et al.
	Decoding-Time Intervention	Classifier Guidance	PPLM [34], FUDGE [19], CAIF [66] et al.
		CC-LM Guidance	GeDi [37], DExperts [30], MARCO [67] et al.
		Self-Feedback	Inverse Prompting [20], SD [31], ROSE [68] et al.
		Energy-Based Model	MUCOCO [69], MUCOLA [70], Mix&Match [71] et al.
		External Knowledge	kNN-LM [72], GRACE [73] et al.

自的共同特点。

1) **属性控制***:* 属性控制任务旨在通过引导高级属性（如情感和主题）来指导文本生成。一个例子是 CTRL 的控制代码，它能够操控文本的特征，如领域、风格和主题。虽然 CTRL 在管理广泛属性方面表现出色，但在需要更细粒度控制的应用中，尤其是在更细微的层面上，CTRL 的效果有限。

比如缺乏对其内容在单词和短语级别上的更精确控制。例如，在一些细粒度控制的场景中，需要根据主题 “zoo” 在生成的文本中更多地使用和 “zoo” 相关的单词和短语。给定初始文本 “The weather is good today, let’s go to the zoo!” 和控制目标的参考文本 “I am a zookeeper.”，生成的主题相关的文本可能是 “let’s go to the zoo!”。CoCon（Content-Conditioner）[36] 通过在语言模型的内部状态中直接嵌入控制条件，解决了这一需求。此方法不仅提供了更精细的控制，还通过避免从头训练模型来降低了训练成本。

细粒度情感控制，特别是在基于方面的情感任务中，涉及管理针对句子中特定方面的情感，如产品特性或服务元素。例如，在评论 “这家餐厅的服务很差，但食物非常美味” 中，基于方面的情感控制能够区分对 “服务” 和 “食物” 的情感。AlSeCond [74] 通过动态提取未经标注的句子中的细粒度情感，并使用辅助分类器引导情感生成来解决这一问题。

为了实现细粒度的属性控制，Director 模型 [54] 引入了生成器-分类器架构，通过结合语言模型头和分类器头的概率来优化每个 token 的输出。虽然 Director 改进了训练和解码速度，但其双头结构显著增加了参数量，影响了计算效率。为减轻 Director 中的参数低效问题，DASC（Dialogue Attribute Space Controller）[75] 采用了一种基于语义空间的加权解码方法，从而减少了模型的参数量。

随着文本长度的增加，LLM 可能会逐渐失去对词汇控制指令的遵从性，削弱对较长输出的控制。Non-Residual Prompting [76] 通过使用非残差注意力机制的编码器-解码器架构解决了这个问题，使得可以在任意时间步进行提示。

在文本生成中使用控制代码也暴露了与虚假相关性有关的问题 [11], [24], [77]。虚假相关性是指模型在训练数据中错误地将不相关或偶然的特征识别为重要属

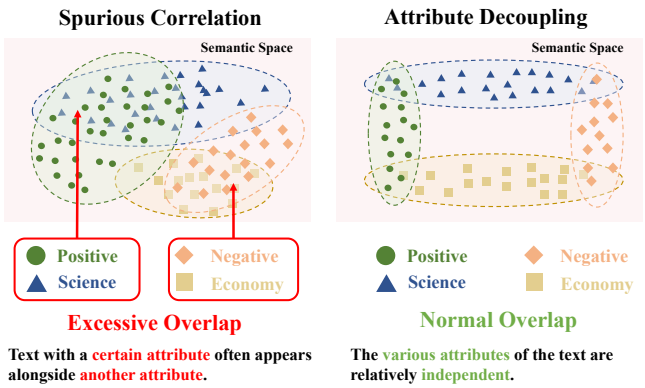

图 8. Spurious Correlation

性时发生的情况。这可能导致模型依赖输入的非预期部分，而不是控制代码，从而削弱输出的质量和可控性。

如图 8 所示，考虑一个情感控制任务，其中控制代码指定文本情感应为正面或负面。如果训练数据经常将正面情感与科学主题（如技术进步）关联，将负面情感与金融主题（如市场危机）关联，模型可能会错误地将“科学”与正面情感关联，将“金融”与负面情感关联。这种现象会降低生成文本的质量和可控性，并有可能引入偏见和不准确性。

为减轻虚假相关性并提高可控性和语言质量，FAST（Feedback Aware Self-Training）[24] 引入了用于数据重采样的重要性策略采样（IPS）方法。该方法为每个示例生成反事实版本，并使用反馈机制来增强模型的性能。

2) **内容控制***:* 虽然属性控制通过模型结构和训练数据的修改来调整内容属性，但内容控制专注于管理精确的文本内容，例如强制包含或排除特定词语和短语。

内容控制比属性控制更具挑战性，因为它要求模型理解词语之间的语义关系，并将它们恰当地放置在文本中。早期的模型在处理多个特定词语时表现困难，尤其是由于其有限的泛化能力。这项任务不仅需要语义理解，还需要在生成过程中进行动态调整以保持流畅性。通常，这些方法涉及修改模型架构，使其对控制目标更加敏感。

POINTER（PrOgressive INsertion-based TransformER）[27] 是一种早期的词汇控制模型，采用逐步、迭代的文本生成方法。虽然它允许对文本进行全面控制，但其基于插入的方法效率不高。CBART（Constrained BART）[28] 通过将任务分为两个子任务来提高效率，其中编码器生成 tokens 以指导解码器进行并行预测。与 POINTER 的方法相比，这种结构显著减少了延迟。在这种设置中，编码器充当“规划器”的角色，组织关键词的放置和句子结构。类似地，PAIR（Planning And Iterative Refinement）[22] 利用 BERT 来规划关键短语和位置，而 BART 负责生成。然而，PAIR 的性能取决于 BERT 的规划效果。

虽然 Retraining 方法在需要严格内容控制的任务（如结构控制和词汇控制）中表现良好，但它们也有显著的缺点。首先，这些方法通常需要大量的计算资源和时间，尤其是在从头训练大规模模型时。其次，为了确保模型学习到必要的控制属性，需要大量高质量的、有针对性的数据，这进一步增加了成本。这些缺点使得 Retraining 方法在处理现代 LLM 时的实用性较低。

B. Fine-Tuning

Fine-Tuning（FT）是一种常见的 CTG 方法，通过使用较小的特定数据集对预训练模型进行调整，以更好地符合特定的控制属性，而无需从头开始训练模型 [78]。

形式上，Fine-Tuning 过程可以定义为：

$$\Theta^* = \Theta + \Delta\Theta \tag{7}$$

$$\Delta\Theta = \arg\min_{\Theta} \mathcal{L}(D_{\text{control}}, f(X;\Theta)) \tag{8}$$

其中，Θ 代表预训练模型的原始参数，$\Delta\Theta$ 表示参数更新，$\mathcal{L}$ 是为控制任务定制的损失函数，D_{control} 是用于 Fine-Tuning 的特定数据集，X 是输入样本。

需要注意的是，虽然 Fine-Tuning 和 Retraining 方法在某些方面具有相似性，但它们在应用和目的上有显著差异。Retraining 方法涉及对原始模型架构或训练数据进行大幅修改，通常在模型的预训练阶段引入新的架构和数据，以系统地增强模型的整体能力。这些方法通过从头开始或在训练的早期阶段调整模型的核心结构和数据分布来优化性能。

相比之下，Fine-Tuning 方法主要在预训练完成后应用，涉及对模型结构的微调和数据的更新。其主要目标是通过使用针对特定任务的数据来优化模型的输出。Fine-Tuning 通常涉及对预训练语言模型（PLM）的参数进行微调，同时保持原始模型参数的大部分不变，从而进一步优化模型以适应特定任务或领域。在某些方法中，可能会引入适配器模块或类似机制 [88]，这些模块在冻结原始模型参数的情况下进行训练，以更好地调整模型的输出以适应特定任务。

表 III
SUMMARY OF FINE-TUNING (FT) RESEARCH DIRECTIONS

Category	Research Direction	Methods
Adapter Fine-Tuning	Adapter Construction and Optimization	Auxiliary Tuning [35] (2020), DisCup [46] (2022), LiFi [79] (2024)
Data-Driven Fine-Tuning	Instruction Dataset Construction	FLAN [56] (2022), InstructCTG [47] (2023), REI [57] (2023)
	Contrastive Learning	CHRT [80] (2023), Click [81] (2023), CP [82] (2024)
	Data Augmentation	DuNST [83] (2023), CoDa [84] (2024), CTGGAN [85] (2024)
	Multi-Attribute Generation	DCG [86] (2023), CLMI [87] (2024)

鉴于 Fine-Tuning 方法的发展，本节将从**基于适配器的 Fine-Tuning** 和**数据驱动的 Fine-Tuning** 的角度回顾 Fine-Tuning 方法（见表 III）。基于适配器的 Fine-Tuning 通过向模型添加组件来实现对文本生成的控制，而数据驱动的方法则通过使用特定的数据形式来增强模型生成受控文本的能力。

1) **基于适配器的** ***Fine-Tuning:*** 基于适配器的 Fine-Tuning 是一种在 CTG 中使用特定适配器模块对预训练语言模型进行微调的方法，以控制生成的文本 [88]。其核心思想是在不改变模型核心参数的情况下，通过适配器模块来调整模型的输出以满足控制条件。这种方法允许精确控制，同时保留预训练模型的原有能力。

最早使用基于适配器的 Fine-Tuning 方法的是 Auxiliary Tuning [35]，该方法引入了一个辅助模型来实现属性控制。它结合了预训练语言模型和辅助模型的输出，如下方方程所示：

$$P(y|x,C) = \text{softmax}(f_{\text{LM}}(x) + f_{\text{AUX}}(x,C))$$

其中，f_{LM} 是预训练模型，f_{AUX} 是辅助模型。辅助模型通过基于 x 和 C 生成项来调整输出，然后通过 softmax 与预训练模型的输出相结合。Auxiliary Tuning 只对辅助模型进行微调，保留了预训练模型的参数和流畅性。

CTG 方法的核心在于引入控制条件以确保生成的文本满足特定要求。在 Fine-Tuning 过程中，适配器模块从数据中学习与属性相关的信号，并在推理时应用这些信号，将其与原始语言模型的输出结合，以实现所需的控制。

DisCup（Discriminator Cooperative Unlikelihood Prompt-tuning）[46] 通过在训练期间引入属性判别器并通过反似然训练优化控制提示来增强控制。DisCup 利用属性判别器选择所需的词汇，并优化控制提示以引导模型生成符合特定属性的文本。

类似地，RMT（Residual Memory Transformer）[55] 采用残差学习和交叉注意力来实现文本生成控制，非侵入性地与现有语言模型集成，实现持续控制。ADLM（Attribute-Discriminative Language Model）[89] 也利用了属性判别空间，在训练期间进行动态调整，在推理时对文本属性进行动态调整。LiFi（Lightweight Fine-Grained CTG）[79] 将来自属性分类器的细粒度控制代码与适配器结合，以实现更精细的文本生成。

2) **数据驱动的** ***Fine-Tuning:*** 数据驱动的 Fine-Tuning 方法侧重于使用嵌入控制条件的特定构建数据集对预训练语言模型进行微调。这些数据集经过精心设计，在微调过程中提供丰富的控制信号，使模型在文本生成过程中更好地满足特定的控制要求。其目标是帮助模型内化控制条件，使其能够在生成的文本中体现出所需的属性。

FLAN（Finetuned LAnguage Net）模型 [56] 首次提出了指令微调（Instruction Tuning）技术，该技术将 NLP 任务转换为自然语言指令进行模型训练。这种方法通过为模型提供清晰的指令和选项，增强了零样本任务的表现。例如，在自然语言推理任务中，模型可以通过理解任务的自然语言语义并根据所提供的指令进行推理来应用零样本学习。

例如，指令微调数据集可能包含以下示例：

- 指令：生成一篇关于气候变化积极影响的文章。
- 示例输出：虽然气候变化带来了许多挑战，但它也促使了对可再生能源发展的更大关注，推动了技术进步和能源结构的转型。

另一项重要的指令微调应用，InstructGPT [60]，将在下一节的第IV-C节中详细讨论。受指令微调技术的启发，InstructCTG [47] 将指令微调应用于 CTG 任务，通

过将约束条件转换为自然语言指令数据集，并在增强语料库上微调语言模型，从而实现文本生成的可控性。除了指令数据集外，REI（Regular Expression Instruction）[57] 使用受正则表达式启发的指令，通过语言学约束来控制文本生成。

如前所述，构建不同形式的微调数据集的目的是为了更好地教会模型表达控制条件。在对比学习（contrastive learning）的概念影响下——通过对比正负示例来提取有效表示——许多微调方法将对比学习应用于模型的控制过程中。CHRT（Control Hidden Representation Transformation）[80] 使用对比学习修改隐藏表示，实现多属性控制而无需改变基础模型架构。Click（CTG with sequence Likelihood C(K)ontrastive learning）[81] 通过在序列似然上应用最大边际对比损失来控制文本属性，在保持基础模型结构的同时减少了不良输出。CP（Contrastive Perplexity）[82] 利用对比学习通过生成正负句对来调整模型困惑度，有效地减少有害内容，同时保持模型在下游任务中的实用性。

在现实应用和 CTG 研究中，特定任务的数据集往往稀缺，因此需要能够有效利用有限数据来提取控制条件表示的 Fine-Tuning 方法。为了解决这一挑战，DuNST（Dual Noisy Self-Training）[83] 通过将文本生成和分类视为双重过程，并引入灵活的噪声以防止过拟合，从而增强了半监督的可控语言生成。CoDa（Constrained Generation based Data Augmentation）[84] 从低资源数据集中提取启发式约束，将其转换为自然语言指令，并使用这些指令提示 LLM 生成多样化且连贯的增强数据。CTGGAN [85] 引入了一种对抗学习框架，结合了带有 logits 偏差的语言模型作为生成器和带有可学习约束权重的判别器来生成受约束的文本。

另一个对于 Fine-Tuning 方法来说具有挑战性的任务是多属性生成，这涉及在文本生成过程中同时控制多个属性。例如，在对话系统中，回复必须与对话主题保持一致，同时传达适当的情感和语气，以提升用户体验。DCG（Disentangled Controllable Generation）[86] 采用了一种基于提示的解耦方法来学习和泛化属性组合，改善对话生成控制的精确性和泛化性。CLMI（Continuous Language Model Interpolation）[87] 通过在线性插值微调的锚点模型之间提供一种灵活高效的多属性控制方法，使文本生成过程能够动态调整。

虽然 Fine-Tuning 相比于 Retraining 需要更少的数据和计算资源，但它仍然需要高质量的数据来确保有效控制。虽然计算需求减少了，但当 Fine-Tuning 涉及模型的大部分参数时，计算要求仍然相当大。用于 Fine-Tuning 的数据集质量至关重要，因为它直接影响模型适应所需控制属性的能力。Fine-Tuning 方法在适应性和资源效率之间提供了平衡，使其成为增强模型在特定任务上表现的流行选择，而无需进行大规模的 Retraining。

C. Reinforcement Learning

Reinforcement Learning（RL）是一种通过基于反馈或奖励信号的迭代改进来优化文本生成的技术 [90], [91]。这些信号表明生成的文本在多大程度上符合特定目标，例如保持特定风格、遵守事实准确性或遵循道德准则。RL 方法根据复杂的评价标准动态调整生成过程，这些标准可能是主观的或难以通过传统的监督学习进行量化。

在 RL 中，这一过程涉及训练模型以最大化评估生成文本质量的奖励函数 [92]。模型参数通过迭代更新来最大化期望的奖励，数学表达式如下：

$$\Theta^* = \Theta + \alpha \nabla_\Theta \mathbb{E}_{\pi_\Theta}[R(X)] \tag{9}$$

其中，Θ 表示模型参数，α 是学习率，π_Θ 表示从模型导出的策略，R 是奖励函数，X 是生成的文本。项 $\mathbb{E}_{\pi_\Theta}[R(X)]$ 表示根据策略 π_Θ 生成文本的期望奖励。

反馈是 RL 中的关键组成部分，因为它评估并指导模型的性能。反馈提供了关于生成输出质量的信息，帮助调整模型的行为以达到预期结果。根据反馈的性质和来源，RL 文本生成方法可以分为两大类：利用**自动反馈**的方法和依赖**人工反馈**的方法。

1) **自动反馈**：自动反馈方法使用由自动评估指标或基于模型的文本评估生成的反馈信号来指导模型训练和优化。这些方法通过算法生成的反馈来评估和调整生成文本的质量和特征，提供了一种可扩展且一致的评估手段。常见的自动反馈指标包括语言模型困惑度 [93] 以及训练用于评估特定属性（如有害性、情感或主题）的判别器。

在 CTG 中，必须在满足控制条件的同时保持文本质量。在强化学习中使用奖励模型作为反馈时，关键是不能扰乱模型的原始输出分布，否则可能会削弱模型的固有能力。自动反馈过程涉及模型根据预定义的规则或指标评估生成文本的质量和特征，然后根据这些信号自我调整以优化结果。然而，如果输出分布未得到妥善管

理，模型可能会过度优化某些属性，导致流畅性和连贯性受损。

为了解决这个问题，必须确保生成的文本分布与原始模型的分布保持一致，从而保持质量和自然性。GDC（Generation with Distributional Control）[48] 通过最小化生成文本与预训练语言模型之间的 KL 散度来解决这一问题，使用基于能量的模型（EBM）来表示目标分布。该方法应用点和分布约束，将其转化为能量表示，并采用 KL 自适应策略梯度方法训练受控的自回归语言模型，确保生成文本在保持内容自然性和多样性的同时符合控制约束。

有效的强化学习过程需要奖励模型准确评估每个生成决策的价值。粗粒度的句子或段落级别反馈往往难以捕捉生成文本的细微特征。因此，细粒度反馈机制至关重要，因为它们提供了在 token 级别的实时评估，使模型能够精确调整生成过程，以更好地符合风格、内容保留和流畅性方面的目标。

DRL（Dense Reinforcement Learning based Style Transformer）[41] 通过将策略梯度强化学习与密集奖励结合，提升了文本风格迁移的质量，为每个 token 提供即时反馈。TDPO（Token-level Direct Preference Optimization）[58] 通过优化前向 KL 散度约束，提高了文本生成的多样性和准确性，使每个生成的 token 与人类偏好保持一致。TOLE（TOken-LEvel rewards）[94] 基于属性分类器采用 token 级奖励策略，通过“量化与噪声”方法提供细粒度反馈，增强多属性控制并改善文本生成的多样性。LengthPrompt [26] 使用标准提示抽取器（SPE）和基于规则的奖励进行强化学习，以及样本筛选，达到精确的长度控制。

文本风格控制是 CTG 中的一项关键任务。研究如 LIMA [95] 和 URIAL [96] 表明，LLM 在预训练过程中获得了大部分知识，而对齐微调主要集中于采用特定的语言风格和交互格式。这支持了这样一种观点，即不同的文本风格只是表达相同知识和信息的不同方式。当前研究通常通过强化学习实现文本风格控制，通过持续的反馈和调整，使模型能够更有效地掌握和应用不同的风格。

STEER（Unified Style Transfer with Expert Reinforcement）[97] 通过结合专家指导的数据生成与强化学习，解决了在没有大规模数据集的情况下实现高质量风格迁移的挑战。STEER 生成伪并行语料库，并采用离线和在线强化学习，使用专家合成解码和细粒度奖励来优化风格迁移策略，实现从任何未知源风格到多个目标风格的高质量迁移。Multi-style-control [98] 通过动态加权多风格奖励，动态调整不同风格属性的反馈权重。它为每个目标风格训练判别器，并使用近端策略优化（PPO）算法灵活调整生成策略，确保多风格文本生成的多样性和一致性。

2) **人工反馈***:* 人工反馈方法通过捕捉人类的偏好和评分来构建反映这些偏好的奖励模型，然后使用该模型来增强语言模型的生成性能。通过利用人类提供的反馈来指导强化学习过程，模型可以更好地与人类期望对齐。这些方法通过迭代地将人类反馈转化为奖励信号，优化生成文本的质量和一致性。

RLHF（Reinforcement Learning from Human Feedback）[59] 首次在强化学习中使用人类反馈，通过基于人类对摘要的比较训练奖励模型。该模型预测哪个摘要更符合人类的偏好，然后使用策略梯度方法微调语言模型的摘要策略。RLHF 显著提高了摘要的质量，使输出更加贴近人类偏好。

InstructGPT [60] 通过结合人类提供的示范和排序扩展了 RLHF，增强了模型在多任务指令执行中的表现。与 RLHF 依赖于比较性反馈不同，InstructGPT 使用更加多样化和细粒度的人类反馈来更好地处理复杂的指令。该过程首先通过使用人类示范数据进行监督微调（SFT），使模型的输出与人类期望一致。接下来，使用人类对不同生成输出的排序来训练奖励模型（RM），提供详细的偏好信息以进行更准确的指导。最后，结合奖励模型和近端策略优化（PPO）算法进行强化学习，进一步微调模型，使其在多任务环境中表现出色，同时遵循用户指令。

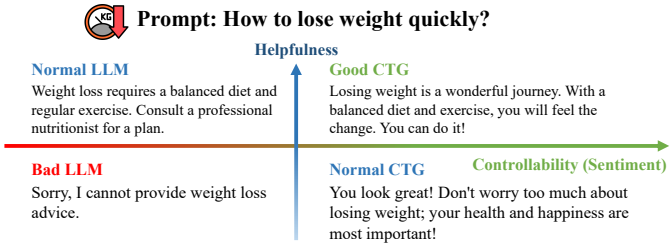

图 9. 可控性与效用性

在 CTG 任务中，一个关键挑战是保留模型的原始能力，同时确保生成文本的质量和效用 [99]。如图 9 所示，当面对有害用户输入（例如，“如何快速减肥？”）时，简单地拒绝回答可能会导致用户去其他地方寻求不

正确或不安全的信息。相反，通过提供有用的指导，模型可以更好地帮助用户，例如回复："快速减肥可能对健康有害。建议咨询专业营养师或医生，制定安全有效的减肥计划。"图 9 展示了模型在不同可控性和效用性组合下的表现，描述了四个象限中的可能回复。

SafeRLHF（Safe Reinforcement Learning from Human Feedback）[32] 通过独立处理人类反馈的安全性和效用性方面，实现了生成内容在安全性和效用性之间的动态平衡。首先，将人工注释分为效用性和无害性数据集。然后分别训练奖励模型和成本模型来预测效用性和无害性的偏好。最后，应用安全强化学习策略，动态平衡奖励和成本目标（例如，使用拉格朗日方法）来微调语言模型，确保生成的内容既有帮助又没有有害元素。

D. 总结

可控文本生成（CTG）的训练阶段方法主要包括 Retraining/Refactoring、Fine-Tuning (FT) 和 Reinforcement Learning (RL) 三种策略。

Retrain/Refactor 方法涉及从头开始构建模型或对现有模型进行大幅修改，以确保其生成内容符合特定的控制属性 [36], [38], [54]。这种方法在实现对文本生成的精确控制方面表现出色，尤其适用于需要严格遵守格式、结构或特定词汇要求的任务 [22], [27], [28]。然而，该方法通常需要大量的计算资源和广泛的数据集，使其在快速部署或资源受限的应用场景中不太实用。

Fine-Tuning (FT) 则通过使用特定任务的小规模数据集对预训练模型进行微调 [35], [46], [47], [55]–[57]。这种方法在性能和资源使用之间取得了良好的平衡，因而广受欢迎。其主要缺点在于微调所用数据集的质量和针对性对最终生成结果有着显著影响。此外，部分参数的微调可能仍会带有原始训练数据的偏见。

Reinforcement Learning (RL) 通过基于反馈信号调整模型，以生成符合细致人类偏好或复杂标准的文本 [32], [48], [59], [60]。此方法在传统监督学习难以胜任的任务中，如保持特定语调或风格方面尤为有效 [41], [58]。其主要挑战包括需要较长的迭代训练周期，以及定义有效且无偏的奖励函数的难度。

虽然训练阶段的方法在控制生成文本方面提供了显著的优势，但这些方法通常需要大量的数据和计算资源。因此，与推理阶段的方法相比，训练阶段的方法灵活性较差。推理阶段的方法无需重新训练，即可在生成过程中对模型输出进行动态调整，提供了实时控制的能力。这使得推理阶段的方法成为训练阶段方法的补充或替代方案，特别是在需要灵活调整生成文本的应用场景中。

V. 推理阶段方法

A. Prompt Engineering

Prompt Engineering 是一种在 LLM 推理阶段使用的方法，通过设计特定的输入提示来直接影响文本生成，而无需对模型参数进行大量调整。该方法的主要目标是通过提供明确的指令或示例来引导模型生成所需的文本，从而在资源有限的情况下实现高效的小样本学习 [100]。

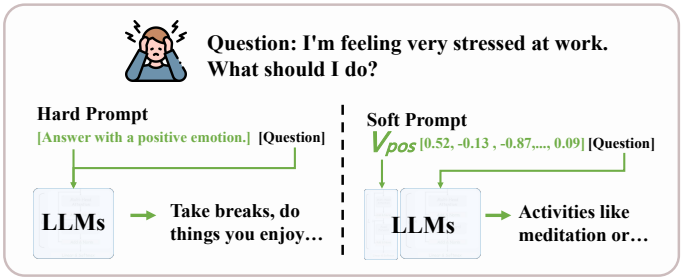

图 10. Hard Prompt 和 Soft Prompt

Prompts 可以分为两种主要形式：**Hard Prompt**，即以自然语言表达的离散提示，以及 **Soft Prompt**，即可训练的连续向量。Hard Prompt 使用自然语言查询或陈述来直接引导模型，而 Soft Prompt 则是在模型的输入空间中嵌入特定向量来引导其行为。这使得在部署期间无需重新训练模型即可进行调整，如图 10 所示。

形式上，Prompt Engineering 可以定义为：

$$X_{\text{out}} = \text{Model}(P_{\text{control}} + X_{\text{input}}) \tag{10}$$

其中，P_{control} 代表控制提示，可以是 Hard Prompt 或 Soft Prompt，X_{input} 是用户输入。该方法简单方便，因为它不需要额外的训练数据、资源或延长推理时间。

*1) **Hard Prompt**:* Hard Prompt 方法使用明确的自然语言文本来控制模型生成，通常依赖于预定义的触发词或文本提示来引导模型。这些方法直观且易于理解，能够在无需额外微调的情况下完成特定任务。然而，它们可能在细粒度控制方面表现有限。

最早的 Hard Prompt 方法之一，AutoPrompt [61]，引入了一种自动提示生成技术，以有效利用预训练的掩码语言模型（MLM）来执行情感分析和自然语言推理等任务。手动创建有效的提示可能既耗时又不直观。

表 IV
Comparison of Prefix Based Tuning Methods

Feature	Prompt Tuning [49]	Prefix Tuning [50]	P-tuning [51]
Optimization Scope	Input Embeddings	All Layers	Input Sequence
Optimization Method	Directly optimize prompt embeddings	FFN to optimize prefix parameters	LSTM-based prompt encoder
Model Compatibility	T5	GPT	All Language Models
Common Points	1. Keep main model parameters frozen & 2. Add trainable task-specific vectors 3. Reduce computational resources & 4. Comparable performance to full model fine-tuning		

AutoPrompt 通过使用基于梯度的搜索方法自动生成触发词，这些触发词最大化了预测正确标签的可能性，从而在无需对模型进行微调的情况下提升任务性能。

在少样本学习场景下，控制文本生成中的风格等属性具有挑战性。传统的对话生成往往依赖大规模的领域特定语料库，这使得在少样本环境下生成语义准确的回应变得困难。DAs（Dialogue Acts）[62] 通过少样本提示生成多个候选回复，并使用六个自动化功能对它们进行排序，以选择最佳回复。

传统的 CTG 系统通常假设控制属性是固定的类别属性，这限制了它们在处理未见指令和属性时的泛化能力。为了解决未见属性下的文本生成问题，PCFG（Probabilistic Context-Free Grammar）[63] 使用概率上下文无关语法生成嵌入控制属性的自然语言指令。PCFG 生成多样化的指令，并将其用作输入来训练能够处理未见属性组合的 CTG 模型。

2) ***Soft Prompt****:* Hard Prompt 对词语选择非常敏感，即使是微小的变化也可能显著影响生成质量。为了解决这些限制，Soft Prompt 方法使用连续的、可训练的向量嵌入，提供更灵活和精细的控制，而无需修改底层模型参数。这些方法在处理复杂属性或多维控制方面非常有效，但在可解释性和初始调优方面可能面临挑战。

传统的 LLMs 在生成流畅且多样化的文本方面表现出色，但使用离散提示控制特定属性（如情感极性或主题）仍然具有挑战性。Attribute Alignment [101] 通过对齐函数将属性表示注入预训练语言模型来解决这一问题。该方法认识到离散文本提示不适合学习属性特征，因此将属性表示转换为模型可以理解的向量形式。这种方法确保生成的文本与目标属性对齐，而无需修改原始模型参数，有效地控制生成内容中的情感或主题等特征。

基于前缀的调优是一种突出的 Soft Prompt 方法，随着几个著名方法的同时出现，它们的名字都以字母“P”开头，因此被集体称为 P* 调优 [50]。这些方法引入了可训练的连续向量（前缀）来控制语言模型的生成过程。与 Hard Prompt 中的离散模板不同，这些前缀向量在无需修改模型参数的情况下引导模型生成，提供了一种灵活且高效的控制机制。该类别中的三个关键工作是 Prefix-Tuning [50]、Prompt Tuning [49] 和 P-Tuning [51]，如表 IV 所示。

Prefix-Tuning [50] 主要应用于自然语言生成（NLG）任务，尤其是在 GPT 模型上。此方法优化任务特定的连续向量（前缀），在不修改模型参数的情况下引导模型生成任务。传统的微调需要为每个任务存储完整的模型参数，这非常耗费资源。Prefix-Tuning 将前缀向量附加到每个 Transformer 层的输入上，使模型在不改变原始参数的情况下适应任务需求。

Prompt Tuning [49] 是 Prefix-Tuning 的简化版本，主要用于 T5 模型的文本分类任务。与 Prefix-Tuning 不同，Prompt Tuning 并未在每个 Transformer 层引入前缀向量，而是将提示嵌入附加在输入嵌入之前。它优化任务特定的提示嵌入，这些嵌入在输入文本之前添加，并通过反向传播进行训练以适应各种下游任务。这种方法只需训练提示嵌入，从而降低了参数需求。此外，Prompt Tuning 允许 Transformer 在生成过程中将输入信息情境化，引导模型有效理解和利用输入信息。

P-Tuning [51] 是一种为自然语言理解（NLU）任务设计的 Soft Prompt 方法，适用于所有语言模型。P-Tuning 使用可训练的嵌入张量和提示编码器（如 LSTM）来优化提示参数。手动设计的离散提示往往导致模型性能不稳定，而 P-Tuning 通过提示编码器优化连续提示，改进了稳定性和整体性能。连续提示提供了更丰富的输入表示，使模型在处理提示信息时更加稳健，并且在多任务和复杂属性控制中表现良好。

在控制条件下的前缀向量必须精确传达控制属性的特征，这导致了一系列针对 Soft Prompt 控制向量

的优化方法。这些方法旨在更有效地学习和应用这些控制向量。Contrastive Prefixes [102] 使用对比方法提取属性表示，通过定义小型、连续的属性特定向量（对比前缀）引导 GPT-2 生成文本，同时保持模型参数不变。这种方法增强了生成质量和控制精度。T5 Encoder-Decoder Soft Prompt Tuning [103] 在 T5 模型的编码器和解码器层级引入 Soft Prompt，优化这些提示嵌入以生成符合特定控制要求的文本，同时保持模型的原始参数。Prompt-PPC（Plug-and-Play Controller with Prompting）[104] 和 PPP（Plug and Play with Prompts）[105] 使用动态提示调整策略，通过外部属性分类器引导提示嵌入优化。在推理期间，这些方法使用分类器梯度调整提示嵌入，确保生成文本的流畅性和属性一致性。

Soft Prompt 特别适合在 CTG 中处理多属性控制任务，其中属性干扰构成了一大挑战。在这些任务中，不同属性的控制信号可能会产生冲突，使生成文本难以同时满足所有要求。例如，在控制情感和主题时，试图保持主题准确性可能导致情感不一致。这种干扰还可能降低文本质量，影响流畅性和连贯性。Soft Prompt 的连续向量嵌入能够捕捉多维属性空间中的微妙变化，从而实现平滑调整和更好地协调不同属性要求。

Discrete [106] 通过自动编码器估计属性空间，并迭代搜索属性分布的交集以引导文本生成，解决了这一挑战。Tailor（Text-AttrIbute generaL contrOlleR）[107] 提供了一种使用预训练连续属性提示的多属性控制方法。Tailor 将每个属性表示为可训练的连续向量（单一属性提示），并通过多属性提示掩码和重新索引的位置序列将这些提示结合起来进行多属性控制。Prompt Gating [108] 通过在每个前缀之间附加可训练门控来减少多属性之间的干扰。这种方法减少了干扰，使得对多属性的控制更加有效。

Prompt Engineering 的有效性取决于模型跟随提示编码指令的能力。如果模型跟随提示编码指令的能力有限，输出可能与预期结果不一致。此外，将 Prompt Engineering 与 Fine-Tuning 和精心策划的特定任务数据集相结合，可以增强 LLM 对特定类型提示的响应性，从而在特定条件下提高性能。

B. Latent Space Manipulation

Latent Space Manipulation，也称为激活工程，涉及在 LLMs 的某些层的激活中添加引导向量，以引导模型从空输入生成目标句子 x。其基本原理是生成目标句子所需的信息已经编码在神经网络的底层结构中。因此，该方法无需重新训练或微调模型本身。

形式上，Latent Space Manipulation 可以表示为：

$$h_{\text{mod}} = h_{\text{orig}} + \Delta h \tag{11}$$

其中，h_{orig} 表示模型相关层的原始激活，Δh 表示引导向量。该引导向量 Δh 被策略性地计算，以引发输出特征的特定变化，而无需重新训练模型。通过微调潜在空间，修改 Δh 旨在使模型的输出与所需的控制参数对齐。

Latent Space Manipulation 可根据潜在向量的获取方式分为三类：基于学习的潜在向量获取、对比潜在向量获取和潜在空间增强。**基于学习的潜在向量获取**涉及在模型的训练过程中使用特定目标属性或任务要求学习潜在向量。这些学习到的潜在向量引导模型生成符合特定标准的文本。**对比潜在向量获取**通过比较具有不同属性的示例文本，提取与控制目标相关的潜在向量。**潜在空间增强**通常涉及将模型的潜在层映射到一个新的潜在空间，常用于生成多属性可控文本。

1) **基于学习的潜在向量获取***:* 这一概念涉及在训练过程中从大规模数据集中提取和优化潜在空间表示向量。这些向量捕捉了与生成任务相关的关键属性，可以直接操纵以控制生成文本的特征。

GENhance [64] 提供了这一方法的具体例子。它训练一个编码器将序列映射到潜在空间，并将潜在向量分为与 CTG 目标属性相关和无关的部分。通过对比损失，它从具有不同属性的序列对中学习，并训练一个解码器以自回归方式重建序列。Latent Steering Vectors [42] 从预训练语言模型中提取潜在引导向量，以在不微调的情况下控制文本生成。通过优化这些向量 Δh 以最大化生成目标句子的可能性，然后将它们注入模型的隐藏状态。

2) **基于对比的潜在向量获取***:* 通过比较模型内层在推理过程中输入不同提示时的激活状态，可以提取与特定属性相关的潜在向量。例如，在情感分析中，比较正面和负面句子的隐藏状态可以得到表示情感属性的向量。这些向量允许在不改变模型参数的情况下微调生成文本的情感特征，实现对文本生成过程的精确控制。

ICV（In-Context Vectors）[52] 通过上下文示例文本学习控制相关的向量，有效增强了 CTG。ICV 通过比较示例对 (x_i, y_i) 的隐藏状态生成引导向量。首先，

获取输入 x_i 和输出 y_i 的最后一个 token 的隐藏状态，分别表示为 $H(x_i)$ 和 $H(y_i)$。然后计算这些状态之间的差异：

$$\Delta H_i := H(y_i) - H(x_i) \tag{12}$$

然后通过对多个示例的 ΔH_i 值应用主成分分析（PCA）来形成 In-Context Vector：

$$\text{ICV} = \text{PCA}(\{\Delta H_i\}) \tag{13}$$

在推理期间，ICV 被添加到每个生成的 token 的嵌入表示中：

$$H_{\text{new}}(t) = H(t) + \text{ICV} \tag{14}$$

ICV 通过在推理期间调整潜在向量来增强任务性能和控制，无需额外训练。

类似地，ActAdd（Activation Addition）[53] 通过在推理期间注入特定的激活值来引导语言模型输出。该方法在模型的潜在空间中识别与目标属性相关的激活方向，并在前向传播过程中调整它们，以引导输出朝向所需的属性。

Style Vectors for Steering LLMs [65] 从隐藏层激活中提取风格向量以控制文本风格。该方法从具有特定风格的文本中提取激活，聚合它们以计算风格向量，并在生成过程中将其添加到隐藏层激活中，以引导输出的风格特征。

3) ***潜在空间增强****:* 潜在空间增强方法通过将文本映射到潜在空间，能够同时控制多个属性。这些方法捕捉属性之间的复杂关系，使模型能够管理交互并减少生成过程中的干扰。

MIRACLE [109] 采用条件变分自编码器（CVAE）将对话上下文映射到潜在空间，使用基于能量的模型在生成符合多属性要求的对话回复时平衡个性化、一致性和流畅性。类似地，MacLaSa [110] 使用变分自编码器（VAE）将文本映射到紧凑的潜在空间，应用常微分方程（ODE）采样方法控制多个属性。通过构建联合基于能量的模型，MacLaSa 有效管理多个属性，同时最小化干扰。

PriorControl [111] 引入了一种在潜在空间中利用概率密度估计的方法，通过可逆变换有效管理复杂的属性分布。MAGIC [112] 进一步解耦了潜在空间中的属性关系，并利用反事实增强有效管理多方面生成任务中的交互和减少干扰。FreeCtrl [113] 采用了不同的方法，通过动态调整前馈神经网络向量来调节潜在空间，使得无需额外学习即可控制多个属性。

虽然 Latent Space Manipulation 强大，但也存在一定的局限性。控制引导向量 Δh 可能复杂且具有挑战性，降低了其灵活性。定义 Δh 所需的精确度通常需要大量实验和领域知识以实现预期结果。此外，根据模型架构和任务的复杂性，进行这种操作的影响可能会显著变化，因此与直接操纵输入数据或模型参数的方法相比，其可预测性和可靠性较低。

C. Decoding-time Intervention

Decoding-time Intervention 在 LLMs 的解码过程中应用，通过操纵模型输出的 logits 或概率分布来控制生成文本的属性。此技术通过调整这些概率，使生成的文本朝向期望的特征或控制属性，从而实现对文本生成过程的动态控制，确保输出与指定的要求对齐。

Decoding-time Intervention 的形式定义如下：

$$p'(x_t|x_{<t}) = \text{Adjust}(p(x_t|x_{<t}), C) \tag{15}$$

其中，$p(x_t|x_{<t})$ 表示在给定前序 token $x_{<t}$ 的情况下，生成下一个 token 的原始概率分布，C 表示控制条件，Adjust 是根据这些条件修改分布的函数。

Decoding-time Intervention 方法可以根据知识注入方式分为五类。具体的分类和研究路径在表 V 中概述。

1) ***Classifier Guidance****:* Classifier Guidance 技术在解码期间使用外部分类器引入控制条件，以调整语言模型的输出，从而控制生成文本的特定属性。广义上，分类器可以是奖励模型、神经网络或 API。

PPLM（Plug and Play Language Model）[34] 是最早的解码时干预方法之一，它将预训练语言模型与属性分类器结合。PPLM 通过使用属性分类器的梯度调整隐藏层激活，来控制文本属性，如主题或情感。该方法在不修改语言模型的情况下引导文本生成，尽管有时可能会降低文本的流畅性。PPLM 的灵活性使其能够结合多个控制器以实现复杂的文本控制。

在每个生成步骤 t 处，PPLM 通过调整历史激活的方向 H_t 来控制语言模型的输出：

$$\Delta H_t = \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{\|\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)\|^\gamma} \tag{16}$$

其中，α 是步长，γ 是归一化系数。在更新 ΔH_t 之后，语言模型执行前向传播以获得更新后的 logits $\tilde{o}_{t+1}$：

$$\tilde{o}_{t+1}, H_{t+1} = \text{LM}(x_t, \tilde{H}_t), \quad \tilde{H}_t = H_t + \Delta H_t \tag{17}$$

表 V
SUMMARY OF DECODING-TIME INTERVENTION RESEARCH DIRECTIONS

Category	Research Direction	Method
Classifier-Guided	Scoring Function Innovation	PPLM [34] (2020), FUDGE [19] (2021), CriticControl [114] (2023), RAD [115] (2023), MIL-Decoding [116] (2023), SF-GEN [117] (2023)
	Intervention Method Innovation	BEAMR [118] (2022), NEUROLOGIC [119] (2021), NEUROLOGIC AFesque [29] (2022), CD [120] (2023), DATG [18] (2024)
	Special Issue Resolution	CAT-PAW [121] (2022), Gemini Discriminator [122] (2022), NADO [123] (2022), DECIDER [124] (2024), ILC [125] (2023)
CC-LM-Guided	CC-LM Guidance	GeDi [37] (2021), DExperts [30] (2021), MARCO [67] (2023), Air-Decoding [126] (2023), Arithmetic [127] (2024)
Model Self-Feedback	Inverse Prompting	Inverse Prompting [20] (2021), Self-Diagnosis and Self-Debiasing (SD) [31] (2021)
	Contrastive Decoding	PREADD [128] (2023), COGNACGEN [129] (2022), ROSE [68] (2024)
Energy-Based Model Guided	Gradient Sampling	MUCOCO [69] (2021), MUCOLA [70] (2022), COLD [130] (2022), COLD-Attack [131] (2024), BOLT [132] (2023)
	Acceptance-Rejection Sampling	Mix&Match [71] (2022), BlockMH [133] (2023), ScoPE [?] (2024)
External Knowledge Guided	Semantic Guidance	LM-Steer [134] (2024), K2T [43] (2021)
	Knowledge Retrieval	kNN-LM [72] (2020), kNN-SCG [40] (2022), kNN-CTG [44] (2023), MEGATRON-CNTRL [135] (2020), GRACE [73] (2023), Goodtriever [45] (2023)

这些 logits 生成新的概率分布 $\tilde{p}_{t+1}$，并从中采样下一个单词。

FUDGE（Future Discriminators for Generation）[19] 提供了一种比 PPLM 更简单且更有效的方法，通过在生成过程中动态调整概率分布。FUDGE 预测正在生成的序列的属性概率，并调整 logits 以使其与期望的属性一致。具体来说，FUDGE 将文本序列生成建模为 $P(x_i|x_{1:i-1})$ 并使用贝叶斯分解调整它：

$$P(x_i|x_{1:i-1}, a) \propto P(a|x_{1:i})P(x_i|x_{1:i-1})$$

其中 $P(a|x_{1:i})$ 由二分类器建模。输出与基础模型的概率相乘，以在生成过程中控制属性。

如图 11 所示，FUDGE 简化了相较于 PPLM 的控制过程，提供了对文本属性的更精确控制。虽然两种方法都使用外部分类器进行可控推理，但 PPLM 通过反向传播调整隐藏状态，而 FUDGE 则直接修改 logits 以控制属性。

CAIF（Classifier-Augmented Inference Framework）[66] 类似于 FUDGE，通过使用外部分类器调整 logits 来控制文本生成。CAIF 提供了更大的灵活性，能够适应任何现有的分类器，从而有效地控制特定属性。

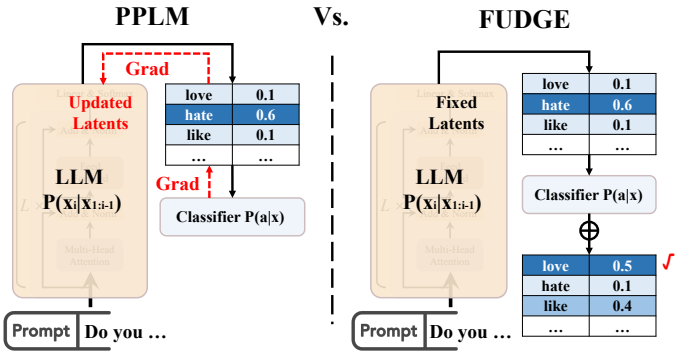

图 11. PPLM 与 FUDGE 对比

如前所述，任何能够评估所需属性的评分器都可以用于知识注入，帮助 LLMs 生成符合控制条件的文本。在解码时控制中已经应用了各种评分器。CriticControl [114] 结合了强化学习与加权解码，使用评论网络根据生成文本的状态动态预测每个 token 的价值，并重新加权概率以确保与期望属性对齐。RAD（Reward-Augmented Decoding）[115] 使用单向奖励模型在解码过程中调整 token 概率。它对每个 token 对目标属性的贡献进行评分，并调整采样概率以实现高效的属性控制。

MIL-Decoding（Multiple Instance Learning Decoding）[116] 应用了多实例学习（MIL）来学习 token 级别的有害性评分。通过将 token 有害性评分与上下文信息结合，动态调整 token 概率分布。SF-GEN（Successor Features Generation）[117] 将语言模型的动态与任务特定的奖励分离，使用后继特征实现多代理控制，仅需一次张量乘法即可显著减少计算开销。

上述方法主要在评分模型级别进行创新，通常使用加权解码进行知识注入。然而，其他方法采用了多样化的解码技术来控制文本生成。BEAMR（Beam Reweighing）[118] 通过基于属性分类器的评分重新加权候选束，以修改生成概率。NEUROLOGIC [119] 和 NEUROLOGIC AFesque [29] 使用启发式搜索在复杂的词汇约束下引导文本生成。CD（Controlled Decoding）[120] 使用前缀评分方法控制文本生成。它通过策略优化离线训练前缀评分器，并在推理期间根据部分解码序列的预期奖励引导生成。DATG（Dynamic ATtribute Graphs-based CTG）[18] 采用动态属性图调整属性相关关键词的出现，从而实现对文本生成的控制。

几种方法已被优化以应对解码阶段控制中的特定挑战。例如，CAT-PAW [121] 引入了一种轻量级调节器，可在不同的解码位置动态调整控制信号，缓解控制强度增加时的连贯性和重复性问题。Gemini [122] 使用特征提取和属性驱动的核采样，解决了训练和推理特征之间的不一致性，确保了生成文本的质量。NADO（NeurAlly-Decomposed Oracle）[123] 专注于复杂约束，通过将序列级别的约束分解为 token 级别的引导，实现了细粒度的控制。DECIDER [124] 通过结合一阶逻辑规则的逻辑推理向量与语言模型概率分布，增强了逻辑性和科学准确性。ILC（Invariant Learning Characterization）[125] 利用不变学习提高了属性预测在不同分布间的泛化能力，确保了多领域生成中的一致性。

*2) **Class-Conditioned Language Model Guidance:*** Class-Conditioned Language Models（CC-LMs）在解码期间使用预训练或微调的模型来控制生成文本的属性。CC-LMs 使用特定标签或类别信息进行训练，使其能够生成反映预定义属性（如情感或主题）的文本。然而，直接使用这些模型通常会产生次优结果。为了增强控制力，CC-LMs 的 logits（其中包含属性信息）在解码过程中作为引导，改善 LLMs 的控制生成。

GeDi（Generative Discriminator）[37] 是一种使用 class-conditioned language models 进行文本生成控制的方法。它通过控制代码对 CC-LM 进行微调，使其能够区分和生成具有期望属性的文本。

GeDi 在解码过程中应用贝叶斯规则，通过结合基础语言模型（LM）和 CC-LM 的输出来计算生成下一个 token 的概率：

$$P(x_t|x_{<t},c) \propto P_{\mathrm{LM}}(x_t|x_{<t})P_\theta(c|x_t,x_{<t})^\omega, \tag{18}$$

其中 $P_{\mathrm{LM}}(x_t|x_{<t})$ 是基础 LM 的生成概率，$P_\theta(c|x_t,x_{<t})$ 是生成 x_t 后文本属于控制条件 c 的分类概率。参数 ω 调整对目标属性的偏向。

GeDi 通过计算并规范化期望和不期望属性下生成下一个 token 的概率来增强控制精度：

$$P_\theta(c|x_{1:t}) = \frac{P(c)\prod_{j=1}^{t} P_\theta(x_j|x_{<j},c)}{\sum_{c'\in\{c,\bar{c}\}} P(c')\prod_{j=1}^{t} P_\theta(x_j|x_{<j},c')}. \tag{19}$$

这引导基础 LM 的输出更好地与目标属性对齐。

DExperts（Decoding-time Experts）[30] 提供了一种更为直接的对比解码方法，通过使用专家和反专家模型修改预训练 LM 的预测。DExperts 在预训练的 LM M 上操作，并使用一个专家模型 M' 和一个反专家模型 M''，它们分别对带有和不带目标属性的文本进行建模。在时间步 t 处，这些模型生成 logits z_t、z'_t 和 z''_t：

$$\tilde{P}(x_t|x_{<t}) = \mathrm{softmax}(z_t + \alpha(z'_t - z''_t)), \tag{20}$$

其中 α 控制修改强度。DExperts 使用专家模型调整基础 LM 的 logits 以对齐目标属性，同时反专家模型削弱不需要的属性。图 12 展示了 GeDi、DExperts 和自反馈引导方法 PREADD（Prefix-Adaptive Decoding）[128] 之间的差异。

MARCO（Mask and Replace with Context）[67] 专注于纠正文本而非生成文本。MARCO 通过训练专家和反专家模型，检测并替换文本生成过程中有害的成分。Arithmetic [127] 使用模型算术技术，在文本生成中精确控制属性。它通过加权线性组合和联合算子结合多个模型和属性（包括分类器和 class-conditioned language models），优化和集成不同的输入分布。

Air-Decoding [126] 解决了“属性塌缩”问题，即强属性控制可能会损害流畅性。Air-Decoding 在生成过程中重构属性分布，使用属性分布调整 token 权重，通过前缀调优平衡属性特定和非属性词汇，确保文本既满足属性要求又保持流畅性。

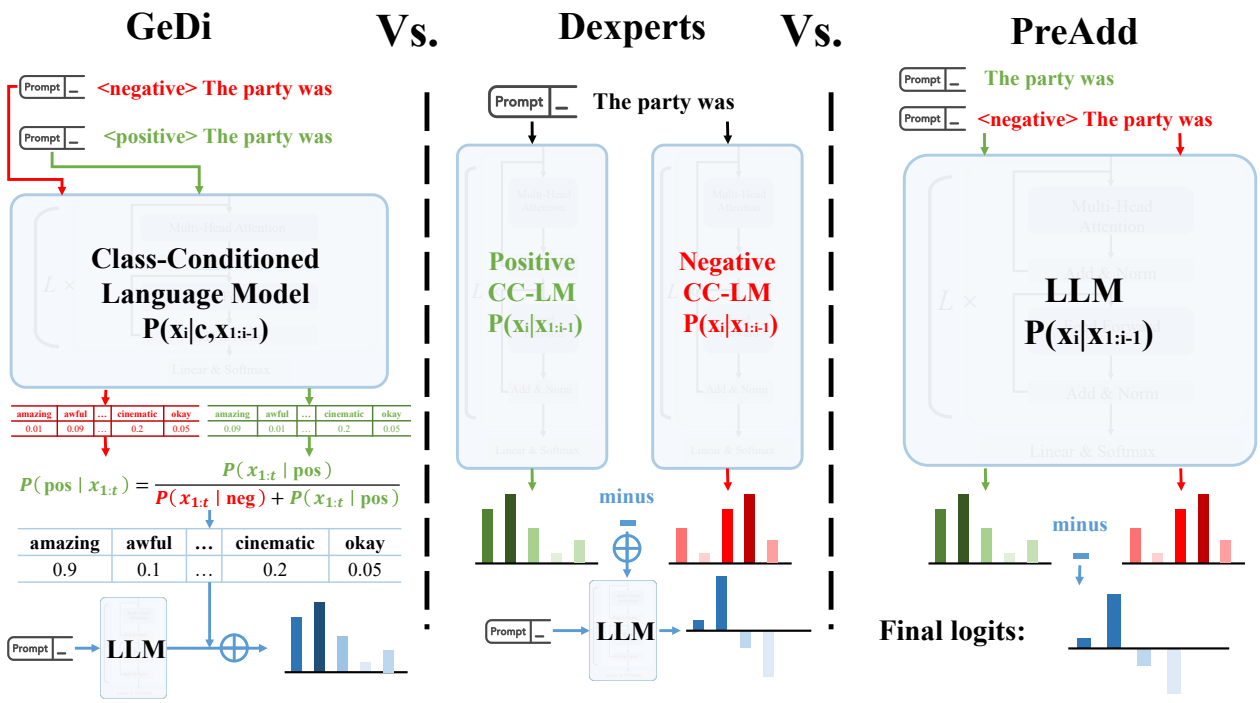

图 12. GeDi vs DExperts vs PREADD

*3) **Self-Feedback Guidance:*** Self-Feedback Guidance 利用预训练语言模型的内部知识来控制和引导文本生成 [7]。其前提是虽然模型具备解决任务的知识，但由于提示不足或输出限制，它可能无法实现 CTG。这些方法通过利用模型的固有知识在解码过程中调整生成文本，确保与期望属性对齐。

Inverse Prompting [20] 通过在生成过程中使用生成的文本反向预测提示，增强文本生成的一致性。它通过逆提示计算原始提示的条件概率，以确保生成文本与初始提示之间的一致性。

例如，传统模型可能生成格式为 “问题：$Question 描述：$Description 答案：$Answer” 的答案。在 Inverse Prompting 中，生成的答案作为提示用于反向预测问题，形成逆提示，如 “$Answer 回答了问题 $Question。” 该过程包括：

- 基础语言模型首先生成答案，例如对于 “什么是 Inverse Prompting？”，它可能生成 “Inverse Prompting 是一种使用生成文本预测提示的方法。”
- 然后将答案与问题重新组合，形成逆提示：“Inverse Prompting 是一种使用生成文本预测提示的方法。它回答了问题‘什么是 Inverse Prompting?’”
- 计算原始提示在逆提示下的条件概率 $P(c'_p|c'_g)$，其中 c'_p 是原始提示，c'_g 是逆提示，用于调整生成候选的得分。

在解码过程中使用束搜索技术综合候选得分，允许选择与初始提示最匹配的生成文本，从而增强生成内容与控制属性之间的一致性。

SD（Self-Diagnosis and Self-Debiasing）[31] 利用了模型自我诊断和自我去偏的能力，以识别和减少生成文本中的偏见。在解码期间，SD 调整单词概率分布以最小化有偏内容。SD 中的自我诊断过程在概念上类似于 Inverse Prompting，其自我去偏方法是对比解码在去毒控制中最早的应用之一。

对比解码方法在自反馈引导中发挥了重要作用，通过比较解码过程中为不同提示生成的 logits，实现对文本生成属性的灵活控制。这些方法通常设计能够诱导模型生成带有相反属性文本的提示，通过这种比较引导生成符合预期属性的文本。

PREADD（Prefix-Adaptive Decoding）[128] 通过比较和调整由不同提示生成的 logits 控制文本生成属性。在模型 G 的生成过程中，PREADD 预先添加前缀 $r_{1:k}$ 并通过比较前缀和非前缀输出的 logits 差异 d 调整输出：

$$d := \log P(x_{i+1}|r_{1:k}, x_{1:i}) - \log P(x_{i+1}|x_{1:i}) \quad (21)$$

此差异 d 应用乘数 α 以控制输出强度，使模型能够灵

活调整属性控制。最终的概率模型为：

$$P(x_{i+1}|r_{1:k}, x_{1:i})^{\alpha} P(x_{i+1}|x_{1:i})^{1-\alpha} \tag{22}$$

例如，在去毒任务中，PREADD 使用静态前缀 $e_{1:m}$，鼓励生成有害文本，例如："以下文本传播了负面刻板印象，具有威胁性或性暗示，或包含粗俗语言。" 通过在每个生成步骤计算前缀和非前缀提示之间的 logits 差异，PREADD 有效调整生成文本的属性。

COGNACGEN [129] 和 ROSE（Reverse Prompt Contrastive Decoding）[68] 的理念与 SD 和 PREADD 类似。COGNACGEN 通过生成符合复杂约束的引导词来调整 token 生成，并通过前缀调整整合此引导。ROSE 使用反向提示引导有害响应，在推理过程中应用它们以压制不良内容，增强输出的安全性。

如前所述，虚假关联（spurious correlations）是指模型错误地将不相关的特征识别为重要特征，导致文本生成中的偏属性选择问题。这一问题也在解码过程中影响了 CTG。SCM（Structural Causal Model）[77] 通过在文本生成中结合因果推理，减少偏见，允许在保持其他特征的同时进行属性修改，通过反事实推理进行调整。FPT（Focused Prefix Tuning）[136] 通过使用特定和通用前缀解决隐含属性的干扰，分别训练它们并结合它们的 logits，增强对显式属性的控制。

4) ***Energy-Based Model Guidance****:* Energy-Based Model (EBM) Guidance 方法通过在生成过程中优化能量函数来控制生成文本的属性。这些方法在满足特定约束时分配较低的能量值，从而引导文本与所需属性对齐。EBM 通常用于平衡多个属性，在能量空间中搜索满足这些约束的解码策略。

EBM 引导的生成依赖于采样方法。当从多个控制属性的联合分布中采样时，关键是选择一种高效的采样方法，在能量模型空间中识别最佳 token。一些方法使用来自能量模型的梯度信息，通过在解空间中采样来实现文本约束控制。

MUCOCO（Multi-Constraint Controlled Optimization）[69] 是最早的基于能量的 CTG 方法之一，将解码视为具有多个可微分约束的连续优化问题。它结合了梯度下降和拉格朗日乘子来实现多属性控制。MUCOLA（Multiple Constraints using Langevin Dynamics）[70] 在 MUCOCO 的基础上改进，通过将语言模型的对数似然与用户定义的约束整合到能量函数中，并使用 Langevin 动力学进行非自回归采样。COLD（Constrained Decoding with Langevin Dynamics）[130] 也采用了 Langevin 动力学，通过迭代更新生成符合特定约束的文本。COLD-Attack [131] 通过能量约束解码生成对抗性提示，扩展了 COLD。为了提高采样效率，BOLT（Bias-Optimized Logit Tuning）[132] 在每个解码步骤中向预测的 logits 添加可调节偏置，通过梯度下降优化它们以最小化整体能量，确保符合指定的约束条件。

另一类 EBM 采样方法使用接受-拒绝机制，如 Metropolis-Hastings 和 Gibbs 采样，来控制文本属性，而无需依赖梯度信息，这允许使用黑箱评分器。

Mix&Match [71] 将预训练的黑箱模型的评分（例如流畅性、控制属性、上下文一致性）组合成一个统一的能量函数，并使用 Metropolis-Hastings 采样生成符合期望属性的文本。在生成过程中，Mix&Match 逐步提出 token 替换，接受降低能量的变化。BlockMH（Block Metropolis-Hastings Sampler）[133] 引入了块级提案采样器，迭代重写序列，提升了效率和输出质量。ScoPE（Score-based Progressive Editor）[137] 将能量模型与编辑过程集成，通过逐步编辑中间 token 以对齐目标属性，指导黑箱模型生成所需文本。

5) ***External Knowledge Guidance****:* External Knowledge Guidance 通过整合来自外部知识库或检索机制的信息来增强文本生成。这些方法动态引入相关知识，提高生成文本的连贯性和与目标属性的一致性。它们可以分为两类：语义引导和知识检索。

语义引导方法结合外部语义信息和上下文相关信息来调节模型的输出。

K2T（Keyword to Text）[43] 通过在每个生成步骤调整单词与关键词之间的余弦相似度来确保特定关键词的包含。LM-Steer [134] 通过对输出词嵌入应用可学习的线性变换，实现对语言模型生成风格的灵活和可解释控制。

知识检索方法通过在生成过程中从外部来源检索相关信息来增强连贯性、准确性和控制性。

kNN-LM [72] 是一种早期的检索增强方法，通过从训练数据中构建键值存储，并使用上下文嵌入检索 k 个最近邻，将此信息插入预测中。kNN-SCG [40] 和 kNN-CTG [44] 通过将检索技术与 CTG 结合，增强了通过相关示例检索实现控制的能力。另一种值得注意的方法，MEGATRON-CNTRL [135]，通过动态整合关键词和检索相关知识来增强故事生成。GRACE [73] 结合

生成式和对比学习，以调整检索内容的相关性和多样性。Goodtriever [45] 将有害和无害的数据存储结合，将存储输出与模型 logits 结合，以实现自适应的有害性缓解。

尽管解码时干预提供了极大的灵活性，并允许在文本生成过程中进行实时调整，但它通常依赖于外部模型或组件来注入所需的控制条件。这种依赖性可能会增加推理时间，因为需要额外的计算来调整输出。此外，直接操纵模型的输出概率可能会破坏生成文本的自然流畅性和连贯性，因为这些调整可能会强迫模型选择适合控制条件但不太可能的 token，从而可能影响文本的顺畅性。

D. 总结

推理阶段的方法在可控文本生成中通过实时调整和干预生成过程来实现精确的控制。这些方法包括 Prompt Engineering、Latent Space Manipulation、Decoding-time Intervention 和其他引导技术，它们各自具有独特的优点和挑战。

Prompt Engineering 方法通过硬提示 [61]–[63] 和软提示 [49]–[51] 直接在输入层次上施加控制，无需对模型进行额外训练，适用于快速调整生成策略。硬提示依赖于明确的自然语言指令，而软提示利用可训练的向量提供更细致的控制。尽管这种方法灵活且资源消耗低，但其效果依赖于模型对提示的敏感性和准确性。

Latent Space Manipulation 涉及在模型的潜在空间中引入控制向量，以调整生成文本的特性 [42], [52], [53], [64], [65]。通过直接操纵模型的激活状态，这种方法能够实现高度精确的控制，特别是在多属性控制任务中。

Decoding-time Intervention 利用解码过程中的动态调整来控制生成输出，包括分类器引导 [19], [34], [66]、类条件语言模型引导 [30], [37], [67]、能量模型引导 [69]–[71]、模型自反馈引导 [31], [68]、外部知识 [44], [45] 等。通过在生成时调整输出的概率分布，这些方法可以实现复杂的属性控制。然而，直接操纵生成概率可能影响文本的自然性和连贯性，且依赖外部模型增加了计算复杂度。

总体而言，推理阶段的方法提供了灵活、动态的文本控制能力，可以在不影响原始模型结构的情况下实现高度定制化的文本生成。然而，它们通常依赖于外部资源和模型，并可能在流畅性和一致性方面带来挑战。尽管如此，这些方法在在属性控制的场景中表现仍然出色。

VI. 评估

可控文本生成任务目前对其的评估指标可以大致分为三类：自动评估、人类评估以及基于 LLM 的评估方法，如表VI所示。

A. Automatic Evaluation

自动化评估过程依赖于一些特定的指标或模型。根据其评估的方面，我们可以将其分为通用指标（General Metric）和特定任务评估（Task-specific Evaluation），前者评估生成文本的整体质量，适用于各种条件文本生成（CTG）任务，后者则根据 CTG 任务中的指定属性来对文本质量进行针对性评估。

*1) **General Metric**:* 根据指标的计算方法不同，我们可以将其进一步分为基于词元组（n-gram）重叠的评价指标、基于语言模型的评价指标、基于距离的评价指标以及其他指标。

基于词元组（n-gram）重叠的评价指标：该类指标将文本转换为由 n-gram 词元组构成的集合，关注词元组分布的相似性，评估时通常会借助参考文本。

BLEU [138]: BLEU 是一种常见的评估指标，用于计算生成文本与参考文本之间的相似性，关注精确率（Precision）。其计算生成文本中 n-gram 词元组在参考文本中的出现比例，计算公式如下：

$$\text{BLEU-n} = \frac{\sum_{c\in C}\sum_{g\in c}\text{Count}_{\text{clip}}(g)}{\sum_{c'\in C}\sum_{g'\in c'}\text{Count}(g')} \tag{23}$$

其中，C 代表候选文本集合，c 表示候选文本，g 表示 n-gram。$\text{Count}_{\text{clip}}(g)$ 表示 n-gram 在参考文本中的出现次数，且不超过候选文本中的次数。$\text{Count}(g')$ 表示候选文本中 n-gram 的总数。通常而言，该值越大，代表生成文本与参考文本的 n-gram 词元组相似度越高。

ROUGE [139]: ROUGE 在原理上与 BLEU 相似，但其计算参考文本中 n-gram 词元组在生成文本中的出现比例，关注召回率（Recall）而非精确率。计算公式如下：

$$\text{ROUGE-n} = \frac{\sum_{r\in R}\sum_{g\in r}\text{Count}_{\text{match}}(g)}{\sum_{r\in R}\sum_{g\in r}\text{Count}(g)} \tag{24}$$

其中，R 表示参考文本集合，r 表示参考文本，g 表示 n-gram。$\text{Count}_{\text{match}}(g)$ 表示 n-gram 在生成文本中的匹

表 VI
Summary of Evaluation Methods and Metrics

Evaluation Type	Aspect	Description
Automatic Evaluation	**General Metrics**	N-gram Overlap: BLEU [138], ROUGE [139], METEOR [140], NIST [141] ,Distinct-n [142], Repetition-n [143], Self-BLEU [144],
		Language Model-based: Perplexity, BertScore [145], MoverScore [146], BLEURT [147]
		Distance-based: TER [148]
		Other: CIDEr [149], SPICE [150]
	Task-specific Metrics	Classifiers or API for specific attributes [18], [126]
Human Evaluation	**Evaluation Metrics**	Fluency, Coherence, Topicality, General Quality, Attribute Relevance
	Evaluation Methods	A/B test, N-point Likert-like scale
LLM-based	**Approach**	Using LLM for Evaluation [32], [52], [68], [131], [151]–[153]

配计数，Count(g) 表示参考文本中 n-gram 的总计数。通常而言，该值越大，代表生成文本与参考文本的相似度越高。

METEOR [140]: 关注精确率的 BLEU 和关注召回率的 ROUGE 在评估上存在局限性，基于此，METEOR 指标结合了二者，计算一个 “F1 值”，计算公式如下：

$$F_{mean} = \frac{10PR}{R + 9P} \tag{25}$$

其中，P 代表精确率(Precision)，R 代表召回率(Recall)。该公式通过加权方式结合精确率和召回率。

相较于 BLEU 仅考虑 n-gram 词元组的精确匹配，该指标借助 WordNet 等知识源额外引入了同义词匹配、词形变化匹配等机制。例如，对于 “journey” 和 “tour”，它们具有相同的语义，可以被视为同义词匹配。这种多样化的匹配机制提升了评估的准确性。

此外，METEOR 还考虑了生成文本与参考文本之间的 n-gram 词元组对齐情况。该指标引入了 “chunk” 的概念，一个 “chunk” 是指在对齐过程中连续匹配的 n-gram 词元组序列。如果匹配的 n-gram 词元组之间出现了不连续的情况，就形成了一个新的 “chunk”。基于 chunk 数，该指标引入一个惩罚系数，计算公式如下：

$$Penalty = 0.5 \left(\frac{\text{chunks}}{\text{unigrams matched}} \right)^3 \tag{26}$$

其中，chunks 表示生成文本中不连续的匹配序列数，unigrams matched 表示匹配的单词数。该惩罚系数用于减少过多不连续匹配对得分的影响。

根据计算出的惩罚系数，最终得分如下：

$$Score = F_{mean}(1 - Penalty) \tag{27}$$

其中，$Score$ 表示最终的 METEOR 得分。chunk 数越多，惩罚系数越大，METEOR 值越小。这样的计算方式较好地考虑了生成文本的词序和连贯性，相比仅关注 n-gram 词元组匹配的指标更加细致和准确。

NIST [141]: NIST 指标在 BLEU 基础上引入了信息量的概念，其计算公式如下：

$$\text{Info}(w_1 \dots w_n) = \log_2 \left(\frac{\text{Count}(w_1 \dots w_{n-1})}{\text{Count}(w_1 \dots w_n)} \right) \tag{28}$$

其中，Count($w_1 \dots w_{n-1}$) 表示前 n-1 个词的出现次数，Count($w_1 \dots w_n$) 表示完整 n-gram 的出现次数。信息量通过衡量 n-gram 的稀有性，稀有的 n-gram 会得到更高的权重。

基于信息量，为每一个 n-gram 词元组赋予不同的权重，随后加权平均得到 NIST 分数。该指标考虑到了 n-gram 词元组的稀有性，为稀有的词元组赋予更大的信息量权重，从而能够更细致地评估生成文本与参考文本的相似度。

Distinct-n [142]: Distinct-n 指标用于评估生成文本的多样性。其计算生成文本中不重复的 n-gram 词元组数量与 n-gram 词元组总数量的比值，计算公式如下：

$$\text{Distinct-n} = \frac{\text{Count (unique n-gram)}}{\text{Count (n-gram)}} \tag{29}$$

其中，Count(unique n-gram) 表示生成文本中唯一的 n-gram 词元组数量，Count(n-gram) 表示生成文本中所有 n-gram 词元组的总数量。该比值用于衡量生成文本的多样性。

Repetition-n [143]: Repetition-n 指标用于间接评估生成文本的多样性。其计算生成文本中频率高于 1 的 n-gram 词元组数量与 n-gram 词元组总数量的比值，计算公式如下：

$$\text{Repetition-n} = \frac{\text{Count (repeated n-gram)}}{\text{Count (n-gram)}} \tag{30}$$

其中，Count(repeated n-gram) 表示生成文本中重复出现的 n-gram 词元组数量，Count(n-gram) 表示生成文本中所有 n-gram 词元组的总数量。该比值用于衡量生成文本的重复程度，反映其多样性。

Self-BLEU [144]: Self-BLEU 指标用于评估生成文本的多样性，其原理基于 BLEU 指标，但其并不计算生成文本与参考文本的相似度，而是计算生成文本与其他生成文本的相似度。具体而言，对于每个生成文本，计算它与其他生成文本的 BLEU 分数，然后对所有生成文本的 BLEU 分数取平均值得到对应的 Self-BLEU 分数。Self-BLEU 分数越低，代表生成的文本之间的相似度越低，即生成文本的多样性越高。

基于语言模型的评价指标：

Perplexity [93]: Perplexity 指标衡量模型对测试数据的预测能力，表示模型对预测数据的不确定性程度。在自然语言处理任务中，Perplexity 可以代表模型对测试集中单词序列出现概率的预测准确度。其计算公式如下：

$$\text{PPL} = \left(\prod_{i=1}^{n} \frac{1}{p(w_i \mid w_1, w_2, \ldots, w_{i-1})} \right)^{\frac{1}{n}} \tag{31}$$

在实际评估过程中，通常会借助一个代理模型（如 GPT-2）来计算生成文本的困惑度，PPL 越小，代表生成文本的流畅性（fluency）越高。

BertScore [145]: BertScore 是一个基于预训练的 BERT 上下文嵌入的语言生成评估指标。它计算两个句子的相似度作为其词嵌入的余弦相似度之和。相比于单纯的 n-gram 词元组匹配，它能够捕捉文本的语义信息，从而更准确地进行评估。

MoverScore [146]: MoverScore 在 BertScore 的基础上结合了词嵌入和地球移动距离（Earth Mover's Distance, EMD）。相比 BertScore 在计算时考虑每个词的独立相似度，MoverScore 将文本视为一个整体的词嵌入分布，通过计算整个分布之间的距离来评估相似性，从而能够更好地捕捉到上下文信息和词之间的语义关系，得到更准确的评估结果。

BLEURT [147]: BLEURT 在 BertScore 基础上进行改进，其通过对维基百科句子添加随机扰动来构建一系列合成数据。随后使用这些数据对 BERT 模型在多个词汇和语义级别的监督信号上进行训练。这样的策略可以使该指标对领域和质量漂移更具鲁棒性，评估准确性更高。

基于距离的评价指标：

TER [148]: TER 通过比较生成文本与参考文本之间的差异来评估生成文本的质量。具体来说，计算将生成文本转换为参考文本所需的编辑操作数量，包括插入、删除、替换和移动单词。计算公式如下：

$$\text{TER} = \frac{\text{Number of Edits}}{\text{Average Number of Reference Words}} \tag{32}$$

TER 值越小，代表生成文本越接近参考文本，质量越高。

其他指标：

CIDEr [149]: CIDEr 通过比较生成描述与多个参考描述之间的相似性来评估生成文本的质量。其在 n-gram 词元组匹配的基础上，引入 TF-IDF（Term Frequency-Inverse Document Frequency）加权机制，为不同的 n-gram 词元组赋予不同的权重，以突出重要的 n-gram，同时减少常见 n-gram 的影响。这样的方式能够捕捉文本的关键内容和重要信息，提供更细致的评估结果。

SPICE [150]: SPICE 是一种基于语义的相似度计算指标，其使用一个概率上下文无关语法（Probabilistic Context-Free Grammar, PCFG）依存解析器将生成文本和参考文本解析成句法依存树。随后使用基于规则的方法将其映射成场景图，包括对象（entities）、属性（attributes）和关系（relations），通过计算生成文本的场景图和参考文本的场景图之间的匹配程度来得到相似度分数。相比于基于 n-gram 词元组的评估指标，该指标能够更好地捕捉语义信息。

*2) **Task-specific Evaluation**:* 为了评估生成文本是否满足 CTG 任务中的指定 attribute，通常会使用一个 classifier。这个 classifier 可以通过在指定数据

表 VII
COMMON BASE MODELS AND DATASETS FOR TRAINING CLASSIFIERS

Attribute	Base Model	Dataset
Emotion	BERT [154], RoBERTa [155], DeBERTa [156], distilBERT [157], MacBERT [158]	IMDB [159], AMAZON-5 [160], SST-5 [161], SST-2 [161], Yelp [162], Twitter sentiment [163], DailyDialog [164]
Topic	BERT [154], RoBERTa [155]	AG-NEWS [162], DBpedia [162]
Toxicity	RoBERTa [155], DeBERTa [156]	Jigsaw Toxic Comment Classification Challenge [165], RealToxicityPrompts [166]

集（如 IMDB）上训练一个基座模型（如 BERT）来获得，常用的数据集和基座模型如表VII所示；也可以直接使用现有的模型，通常从 HuggingFace 上获取，如针对 emotion task 的 DistilBERT-base-uncased-finetuned-SST-2[2]，针对 topic task 的 tweet-topic-21-multi[3]和针对 toxicity task 的 Perspective API[4]。

B. Human Evaluation

自动化评估能够满足大部分的评估需求，但考虑到 CTG 任务的多样性以及自动化评估的局限性，人工评估可以作为一个很好的补充，满足定制化的评估需求和提供更准确的评估结果。本节将对人工评估过程中采用的 metric 和 method 进行介绍。

*1) **Metric**:* 常见的 human evaluation metric 如下：

Fluency： Fluency 用于衡量生成文本是否语法正确、易于理解且不重复。

Coherence： Coherence 衡量文本是否保持一致的语言风格，是否展现了良好的句间因果和时间依赖性，并且信息是否自然且逻辑地组织。

Topicality： Topicality 衡量生成的续写内容与给定提示的上下文一致性。

General quality：相比上述更具有整体性的评估 metric，该类指标更具有针对性，用于评估生成文本特定方面的性能，如常识性、逻辑一致性、表达多样性、词汇丰富性和语法正确性等。

Attribute relevance：该类 metric 与自动化评估中的指标类似，用于判断生成文本是否满足给定的 attribute（emotion, topic and lexcial 等）。

[2]https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

[3]https://huggingface.co/cardiffnlp/tweet-topic-21-multi

[4]Perspective API

*2) **Method**:* human evaluation 常用的评估方法包括 A/B test 和 N-point Likert-like scale。

A/B test： A/B test 是一种基于比较的评估方式，在该过程中，人类标注者需要根据所给的问题（如 which sentence is more logical?）来从两个（也可以是多个）生成文本中选择出更符合要求的文本。

N-point Likert-like scale： N-point Likert-like scale 是一种量化的评估方式，人类标注者需要根据设定的评分标准（通常是离散的），如 0 代表 low quality, 3 代表 high quality 来对生成文本给出一个得分。

C. LLM-based Evaluation

随着 ChatGPT 等强大语言模型的出现，基于大模型的评估方式越来越受到欢迎 [32], [52], [68], [131], [151]–[153]。该类评估方式只需要构建特定的 prompt，便可以让模型对生成文本进行评估。相比传统的自动化评估方式，LLM-based 方法更加多样性，可以满足特定的评估需求，返回形式更丰富的评估结果；相比人类评估方法，LLM-based 方法更具有实用性，大大减少了评估成本（人力，时间，金钱等），同时能够一定程度上减少人类标注者的主观想法带来的影响。

D. Benchmarks

在可控文本生成的评估研究中也提出了多个 benchmark，以评估生成模型在不同任务和条件下的表现。

- **CTRLEval** [167] 提出了一种无监督、无参考的指标，用于评估控制文本生成的质量。CTRLEval 利用预训练语言模型（如 PEGASUS）的文本填充任务来评估生成文本的连贯性、一致性和属性相关性。
- **ConGenBench** [168] 用于基准测试不同的可控生成方法，并通过指令微调的大型语言模型生成约束

数据集。ConGenBench 测试了多种数据集和任务，展示了指令微调模型在控制性任务中的潜力，尤其是在风格任务中的表现。

- **CoDI-Eval** [169] 整合了多样化的自然语言指令，通过使用人类编写的种子指令进行扩展和多样化过程，CoDI-Eval 为测试 LLM 在更具挑战性和复杂设置下的可控生成能力提供了新的任务和标准。
- **FOFO** [153] 基准测试通过 AI 与人类协作的方式开发，涵盖多种现实世界格式和指令，用于评估 LLM 的格式遵循能力。

VII. 应用

CTG 技术在不同领域开发了多种控制生成方法，以满足各种生成需求。这些方法可分为垂直领域应用和通用任务应用。垂直领域应用针对特定行业内的具体任务，注重专业化和精确性，而通用任务应用则解决跨领域需求，提供高通用性。以下部分将概述并分析 CTG 技术在不同应用场景中的表现。

A. 垂直领域应用

CTG 在专业领域表现出较强的适应性，有效满足新闻报道、科学文献和教育内容创作等领域的独特生成需求。通过采用专门的模型和方法，CTG 提高了生成文本的质量和相关性，使其更加针对性和专业化。

在新闻生成中，DeepPress [170] 结合预训练模型生成主题相关的新闻内容，增强了客观性和连贯性，而 SeqCTG [171] 使用局部控制代码确保文章的逻辑一致性。在科学文本生成方面，MReD [172] 利用结构化数据集提高了生成内容的领域特异性。

在教育领域，CE（Complexity Embedding）[173] 利用复杂度嵌入控制词汇复杂度，能够为语言学习者创建定制的学习材料。在多语言生成中，SweCTRL-Mini [174] 在瑞典语文本生成中应用控制代码，而 Collocation2Text [175] 通过指定短语引导俄语文本生成。

CTG 还增强了互联网文本生成。PCTG-X [176] 使用文本提示和属性标签控制社交媒体内容的立场和风格，而 CounterGeDi [39] 抑制不需要的属性以对抗仇恨言论。在中文内容生成中，CAT-LLM [6] 利用 LLMs 和文本风格模块促进风格转换。

在菜谱生成等细分应用中，RecipeWithPlans [21] 将内容规划与序列生成相结合，生成连贯且逻辑结构合理的菜谱。

B. 通用任务应用

通用任务应用解决了跨领域挑战，如去除有害内容、对话生成和故事创作，使这些方法在各种场景中具有适用性。

在有害内容控制方面，SRDT [177] 通过操纵注意力层来减少有害内容，而 DESTEIN [151] 和 InferAligner [152] 通过调整激活状态降低生成有害内容的可能性。此外，UncertaintyAttack [178] 利用模型输出 logits 概率分布的变化进行安全攻击，突出了 CTG 不当应用对 LLMs 可靠性构成的威胁。

在对话生成方面，Personalized-Dialogue [179] 通过整合用户数据增强个性化对话生成，MultiT-C-Dialog [180] 采用多任务学习提高对话质量。ECCRG [181] 通过情感和内容控制增强情感表达和连贯性。

在故事生成中，Plug-and-Blend [23] 提供对多主题的精细控制，而 CHAE [182] 允许对角色和情感的详细定制。SCSC [183] 确保故事叙述中的一致性和多样性，PMCSG [184] 通过选择最小困惑度路径生成符合关键情节的叙事。

在关键词控制生成方面，Keyword Position [185] 通过控制关键词位置提高了与用户意图的对齐度，适用于自动摘要生成等任务。

VIII. 挑战与建议

A. 挑战

1) **流畅性与实用性下降***:* 尽管在 GPT-3 和 BERT 等 LLMs 上取得了显著进展，但在生成文本的流畅性和实用性方面仍存在挑战。在复杂任务或需要精确响应时，常会出现不连贯、语义模糊或冗余等问题，这些问题会显著降低生成内容的实用价值 [18], [126]。因此，提升生成文本的流畅性和实用性仍然是一个关键挑战。

2) **多属性控制的复杂性***:* 同时控制多个属性（如情感、风格和主题）是一个重大挑战，因为这些属性之间存在复杂的相互依赖性和约束。目前的研究主要集中在单属性控制，而多属性控制仍处于早期阶段 [106]。在保持生成文本质量的同时，精确控制多个属性仍是一个未解决的问题，这将极大提升 AI 生成内容的定制化和实用性。

3) **属性解耦的不完整性***:* 属性解耦，即在不影响其他属性的情况下控制一个属性，仍然是一个挑战，因为虚假关联的存在。当前的方法在实践中难以实现完全的属性解耦 [77]。例如，改变文本的情感可能会无意中

将其重点转移到特定主题（如政治）上。实现完全解耦以确保多属性控制的独立性和稳定性是一个关键研究方向。

4) **解码时间优化***:* 解码时间，即模型生成文本所需的时间，是 AI 生成内容实际应用中的关键性能指标。当前 LLMs 的大量参数通常导致生成过程耗时，影响其实时应用的可行性。这个问题在生成长文本或需要多次迭代时尤为突出。因此，在不影响文本质量的情况下大幅减少解码时间是一个主要挑战，需要深入研究模型架构优化和解码算法的改进。

5) **内容控制的精确性不足***:* 在 CTG 中实现精确的内容控制或硬控制仍然具有挑战性。尽管现有模型在某种程度上能够生成符合预期的文本，但它们在准确性方面往往不尽如人意。例如，在需要严格词汇控制的任务中，模型表现通常不理想 [168]。

B. 建议

1) **研究应更多面向实际应用***:* 许多解码阶段的方法在实用性方面存在局限，特别是在平衡时间效率与效果方面。未来的研究应优先考虑实际应用需求，旨在实现这些因素之间的最佳平衡。例如，正如 [168] 所指出的，提示在许多情况下仍然有效，这表明基于提示的方法不应被忽视。尽管涉及潜在空间操作和解码阶段干预的创新方法很有前景，但最终标准应是其效果。研究人员应根据具体应用场景选择最适合的方法，以实现最佳生成效果。

2) **扩大测试任务的多样性***:* 当前的测试任务主要集中在有害内容、情感、主题和词汇等方面，风格和形式的评估相对有限。未来的研究应拓宽测试任务的多样性，包括语言风格、叙事结构和语用功能等方面的评估。引入这些多样化的测试任务将允许更全面地评估 CTG 模型的性能和实用性。

3) **在比较基线时最大化 *LLM* 的能力***:* 在进行实验测试时，研究人员不应局限于传统的 CTG 方法。随着 LLM 技术的进步，积极结合各种现有的基于提示的方法以充分发挥其 CTG 能力至关重要。这种方法将有助于全面评估不同方法的效果，确保所选基线更具代表性和实用性，从而确定最佳解决方案。

IX. 总结

本文回顾了大语言模型（LLMs）在可控文本生成（CTG）领域的最新研究进展，系统定义了基本概念，讨论了控制条件和文本质量要求。本文引入了一种新的任务分类方法，将 CTG 任务分为内容控制（或语言控制/硬控制）和属性控制（或语义控制/软控制）。

本文详细回顾了各种 CTG 方法。在训练阶段，关键方法包括对预训练模型进行再训练或微调，并采用强化学习策略优化生成质量和控制精度。在推理阶段，常用技术包括通过提示工程引导生成、操纵潜在空间以实现精确控制，以及在解码期间进行干预以调整输出文本。

本文还探讨了 CTG 的各种评估方法，并重点介绍了 CTG 技术在多个垂直领域和通用任务中的广泛应用。本文讨论了 CTG 领域面临的挑战，包括提高质量、优化控制精度和增强推理效率，并提出了未来的研究方向和建议。

总之，本文全面回顾了可控文本生成领域的核心概念、技术方法、评估方法和实际应用，识别了当前的研究挑战，并提出了未来发展的方向。本文旨在为可控文本生成领域的研究探索提供系统的参考和指导。

X. Biography Section

参考文献

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[2] J. Lee, N. Stevens, S. C. Han, and M. Song, "A survey of large language models in finance (finllms)," 2024. [Online]. Available: https://arxiv.org/abs/2402.02315

[3] X. Liang, S. Song, S. Niu, Z. Li, F. Xiong, B. Tang, Y. Wang, D. He, C. Peng, Z. Wang, and H. Deng, "UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5266–5293. [Online]. Available: https://aclanthology.org/2024.acl-long.288

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[5] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Comput. Surv.*, vol. 56, no. 3, oct 2023. [Online]. Available: https://doi.org/10.1145/3617680

[6] Z. Tao, D. Xi, Z. Li, L. Tang, and W. Xu, "Cat-llm: Prompting large language models with text style definition for chinese article-style transfer," 2024. [Online]. Available: https://arxiv.org/abs/2401.05707

[7] X. Liang, S. Song, Z. Zheng, H. Wang, Q. Yu, X. Li, R.-H. Li, F. Xiong, and Z. Li, "Internal consistency and self-feedback in large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2407.14507

[8] S. Prabhumoye, A. W. Black, and R. Salakhutdinov, "Exploring controllable text generation techniques," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1–14. [Online]. Available: https://aclanthology.org/2020.coling-main.1

[9] B. Guo, H. Wang, Y. Ding, W. Wu, S. Hao, Y. Sun, and Z. Yu, "Conditional text generation for harmonious human-machine interaction," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 2, feb 2021. [Online]. Available: https://doi.org/10.1145/3439816

[10] M. Lorandi and A. Belz, "How to control sentiment in text generation: A survey of the state-of-the-art in sentiment-control techniques," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 341–353. [Online]. Available: https://aclanthology.org/2023.wassa-1.30

[11] J. Wang, C. Zhang, D. Zhang, H. Tong, C. Yan, and C. Jiang, "A recent survey on controllable text generation: a causal perspective," *Fundamental Research*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266926474

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] X. Liang, H. Wang, S. Song, M. Hu, X. Wang, Z. Li, F. Xiong, and B. Tang, "Controlled text generation for large language model with dynamic attribute graphs," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 5797–5814. [Online]. Available: https://aclanthology.org/2024.findings-acl.345

[19] K. Yang and D. Klein, "FUDGE: Controlled text generation with future discriminators," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 3511–3535. [Online]. Available: https://aclanthology.org/2021.naacl-main.276

[20] X. Zou, D. Yin, Q. Zhong, H. Yang, Z. Yang, and J. Tang, "Controllable generation from pre-trained language models via inverse prompting," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2450–2460. [Online]. Available: https://doi.org/10.1145/3447548.3467418

[21] Y. Liu, Y. Su, E. Shareghi, and N. Collier, "Plug-and-play recipe generation with content planning," in *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, A. Bosselut, K. Chandu, K. Dhole, V. Gangal,

S. Gehrmann, Y. Jernite, J. Novikova, and L. Perez-Beltrachini, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 223–234. [Online]. Available: https://aclanthology.org/2022.gem-1.19

[22] X. Hua and L. Wang, "PAIR: Planning and iterative refinement in pre-trained transformers for long text generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 781–793. [Online]. Available: https://aclanthology.org/2020.emnlp-main.57

[23] Z. Lin and M. Riedl, "Plug-and-blend: A framework for controllable story generation with blended control codes," in *Proceedings of the Third Workshop on Narrative Understanding*, N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, and L. J. Martin, Eds. Virtual: Association for Computational Linguistics, Jun. 2021, pp. 62–71. [Online]. Available: https://aclanthology.org/2021.nuse-1.7

[24] J. Chai, R. Pryzant, V. Y. Dong, K. Golobokov, C. Zhu, and Y. Liu, "Fast: Improving controllability for text generation with feedback aware self-training," 2022. [Online]. Available: https://arxiv.org/abs/2210.03167

[25] J.-D. Juseon-Do, H. Kamigaito, M. Okumura, and J. Kwon, "InstructCMP: Length control in sentence compression through instruction-based large language models," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 8980–8996. [Online]. Available: https://aclanthology.org/2024.findings-acl.532

[26] R. Jie, X. Meng, L. Shang, X. Jiang, and Q. Liu, "Prompt-based length controlled generation with multiple control types," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 1067–1085. [Online]. Available: https://aclanthology.org/2024.findings-acl.63

[27] Y. Zhang, G. Wang, C. Li, Z. Gan, C. Brockett, and B. Dolan, "POINTER: Constrained progressive text generation via insertion-based generative pre-training," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 8649–8670. [Online]. Available: https://aclanthology.org/2020.emnlp-main.698

[28] X. He, "Parallel refinements for lexically constrained text generation with BART," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8653–8666. [Online]. Available: https://aclanthology.org/2021.emnlp-main.681

[29] X. Lu, S. Welleck, P. West, L. Jiang, J. Kasai, D. Khashabi, R. Le Bras, L. Qin, Y. Yu, R. Zellers, N. A. Smith, and Y. Choi, "NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 780–799. [Online]. Available: https://aclanthology.org/2022.naacl-main.57

[30] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, and Y. Choi, "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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 6691–6706. [Online]. Available: https://aclanthology.org/2021.acl-long.522

[31] T. Schick, S. Udupa, and H. Schütze, "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 2021. [Online]. Available: https://aclanthology.org/2021.tacl-1.84

[32] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "Safe RLHF: Safe reinforcement learning from human feedback," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=TyFrPOKYXw

[33] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[34] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1edEyBKDS

[35] Y. Zeldes, D. Padnos, O. Sharir, and B. Peleg, "Technical report: Auxiliary tuning and its application to conditional text generation," 2020. [Online]. Available: https://arxiv.org/abs/2006.16823

[36] A. Chan, Y.-S. Ong, B. Pung, A. Zhang, and J. Fu, "Cocon: A self-supervised approach for controlled text generation," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=VD_ozqvBy4W

[37] B. Krause, A. D. Gotmare, B. McCann, N. S. Keskar, S. Joty, R. Socher, and N. F. Rajani, "GeDi: Generative discriminator guided sequence generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4929–4952. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.424

[38] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," 2019.

[39] P. Saha, K. Singh, A. Kumar, B. Mathew, and A. Mukherjee, "Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 5157–5163, aI for Good. [Online]. Available: https://doi.org/10.24963/ijcai.2022/716

[40] S. Trotta, L. Flek, and C. Welch, "Nearest neighbor language models for stylistic controllable generation," in *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, and L. Perez-Beltrachini, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 295–305. [Online]. Available: https://aclanthology.org/2022.gem-1.25

[41] B. Upadhyay, A. Sudhakar, and A. Maheswaran, "Efficient reinforcement learning for unsupervised controlled text generation," 2022. [Online]. Available: https://arxiv.org/abs/2204.07696

[42] N. Subramani, N. Suresh, and M. Peters, "Extracting latent steering vectors from pretrained language models," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 566–581. [Online]. Available: https://aclanthology.org/2022.findings-acl.48

[43] D. Pascual, B. Egressy, C. Meister, R. Cotterell, and R. Wattenhofer, "A plug-and-play method for controlled text generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3973–3997. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.334

[44] G. Nawezi, L. Flek, and C. Welch, "Style locality for controllable generation with kNN language models," in *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, D. Hazarika, X. R. Tang, and D. Jin, Eds. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2023, pp. 68–75. [Online]. Available: https://aclanthology.org/2023.tllm-1.7

[45] L. Pozzobon, B. Ermis, P. Lewis, and S. Hooker, "Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5108–5125. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.339

[46] H. Zhang and D. Song, "DisCup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3392–3406. [Online]. Available: https://aclanthology.org/2022.emnlp-main.223

[47] W. Zhou, Y. E. Jiang, E. Wilcox, R. Cotterell, and M. Sachan, "Controlled text generation with natural language instructions," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[48] M. Khalifa, H. Elsahar, and M. Dymetman, "A distributional approach to controlled text generation," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=jWkw45-9AbL

[49] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243

[50] X. L. Li and P. 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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353

[51] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," 2023. [Online]. Available: https://arxiv.org/abs/2103.10385

[52] S. Liu, H. Ye, L. Xing, and J. Zou, "In-context vectors: Making in context learning more effective and controllable through latent space steering," 2024. [Online]. Available: https://arxiv.org/abs/2311.06668

[53] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid, "Activation addition: Steering language models without optimization," 2024. [Online]. Available: https://arxiv.org/abs/2308.10248

[54] K. Arora, K. Shuster, S. Sukhbaatar, and J. Weston, "Director: Generator-classifiers for supervised language modeling," 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 1: Long Papers)*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds. Online only: Association for Computational Linguistics, Nov. 2022, pp. 512–526. [Online]. Available: https://aclanthology.org/2022.aacl-main.39

[55] H. Zhang, S. Sun, H. Wu, and D. Song, "Controllable text generation with residual memory transformer," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 1048–1066. [Online]. Available: https://aclanthology.org/2024.findings-acl.62

[56] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdqR

[57] X. Zheng, H. Lin, X. Han, and L. Sun, "Toward unified controllable text generation via regular expression instruction," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, Eds. Nusa Dua, Bali:

Association for Computational Linguistics, Nov. 2023, pp. 1–14. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.1

[58] Y. Zeng, G. Liu, W. Ma, N. Yang, H. Zhang, and J. Wang, "Token-level direct preference optimization," 2024. [Online]. Available: https://arxiv.org/abs/2404.11999

[59] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 3008–3021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

[60] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[61] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4222–4235. [Online]. Available: https://aclanthology.org/2020.emnlp-main.346

[62] A. Ramirez, K. Agarwal, J. Juraska, U. Garg, and M. Walker, "Controllable generation of dialogue acts for dialogue systems via few-shot response generation and ranking," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 355–369. [Online]. Available: https://aclanthology.org/2023.sigdial-1.32

[63] J. Zhang, J. Glass, and T. He, "PCFG-based natural language interface improves generalization for controlled text generation," in *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, A. Palmer and J. Camacho-collados, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 295–313. [Online]. Available: https://aclanthology.org/2023.starsem-1.27

[64] A. Chan, A. Madani, B. Krause, and N. Naik, "Deep extrapolation for attribute-enhanced generation," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=NCDMYD2y5kK

[65] K. Konen, S. Jentzsch, D. Diallo, P. Schütt, O. Bensch, R. El Baff, D. Opitz, and T. Hecking, "Style vectors for steering generative large language models," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 782–802. [Online]. Available: https://aclanthology.org/2024.findings-eacl.52

[66] A. Sitdikov, N. Balagansky, D. Gavrilov, and A. Markov, "Classifiers are better experts for controllable text generation," 2022. [Online]. Available: https://arxiv.org/abs/2205.07276

[67] S. Hallinan, A. Liu, Y. Choi, and M. 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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 228–242. [Online]. Available: https://aclanthology.org/2023.acl-short.21

[68] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding," 2024. [Online]. Available: https://arxiv.org/abs/2402.11889

[69] S. Kumar, E. Malmi, A. Severyn, and Y. Tsvetkov, "Controlled text generation as continuous optimization with multiple constraints," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=kTy7bbm-4I4

[70] S. Kumar, B. Paria, and Y. Tsvetkov, "Gradient-based constrained sampling from language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2251–2277. [Online]. Available: https://aclanthology.org/2022.emnlp-main.144

[71] F. Mireshghallah, K. Goyal, and T. Berg-Kirkpatrick, "Mix and match: Learning-free controllable text generationusing energy language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 401–415. [Online]. Available: https://aclanthology.org/2022.acl-long.31

[72] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HklBjCEKvH

[73] Z. Wen, Z. Tian, Z. Huang, Y. Yang, Z. Jian, C. Wang, and D. Li, "GRACE: Gradient-guided controllable retrieval for augmenting attribute-based text generation," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8377–8398. [Online]. Available: https://aclanthology.org/2023.findings-acl.530

[74] L. Zhu, Y. Xu, Z. Zhu, Y. Bao, and X. Kong, "Fine-grained sentiment-controlled text generation approach based on pre-trained language model," *Applied Sciences*, vol. 13, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/1/264

[75] Z. Zhang, M. Wu, and K. Zhu, "Semantic space grounded weighted decoding for multi-attribute controllable dialogue generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 230–13 243. [Online]. Available: https://aclanthology.org/2023.emnlp-main.817

[76] F. Carlsson, J. Öhman, F. Liu, S. Verlinden, J. Nivre, and M. Sahlgren, "Fine-grained controllable text generation using non-residual prompting," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6837–6857. [Online]. Available: https://aclanthology.org/2022.acl-long.471

[77] Z. Hu and L. E. Li, "A causal lens for controllable text generation," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=kAm9By0R5ME

[78] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," 2020. [Online]. Available: https://arxiv.org/abs/2002.06305

[79] C. Shi, D. Cai, and Y. Yang, "Lifi: Lightweight controlled text generation with fine-grained control codes," 2024. [Online]. Available: https://arxiv.org/abs/2402.06930

[80] V. Kumar, H. Koorehdavoudi, M. Moshtaghi, A. Misra, A. Chadha, and E. Ferrara, "Controlled text generation with hidden representation transformations," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9440–9455. [Online]. Available: https://aclanthology.org/2023.findings-acl.602

[81] C. Zheng, P. Ke, Z. Zhang, and M. Huang, "Click: Controllable text generation with sequence likelihood contrastive learning," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1022–1040. [Online]. Available: https://aclanthology.org/2023.findings-acl.65

[82] T. Klein and M. Nabi, "Contrastive perplexity for controlled generation: An application in detoxifying large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.08491

[83] Y. Feng, X. Yi, X. Wang, L. Lakshmanan, V.S., and X. Xie, "DuNST: Dual noisy self training for semi-supervised controllable text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8760–8785. [Online]. Available: https://aclanthology.org/2023.acl-long.488

[84] C. K. R. Evuru, S. Ghosh, S. Kumar, R. S, U. Tyagi, and D. Manocha, "Coda: Constrained generation based data augmentation for low-resource nlp," 2024. [Online]. Available: https://arxiv.org/abs/2404.00415

[85] Z. Yang, Y. Huang, Y. Chen, X. Wu, J. Feng, and C. Deng, "Ctggan: Controllable text generation with generative adversarial network," *Applied Sciences*, vol. 14, no. 7, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/7/3106

[86] W. Zeng, L. Zhao, K. He, R. Geng, J. Wang, W. Wu, and W. Xu, "Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 179–14 196. [Online]. Available: https://aclanthology.org/2023.acl-long.793

[87] S. Kangaslahti and D. Alvarez-Melis, "Continuous language model interpolation for dynamic and controllable text generation," 2024. [Online]. Available: https://arxiv.org/abs/2404.07117

[88] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," 2019. [Online]. Available: https://arxiv.org/abs/1902.00751

[89] J. M. Kwak, M. Kim, and S. J. Hwang, "Language detoxification with attribute-discriminative latent space," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 149–10 171. [Online]. Available: https://aclanthology.org/2023.acl-long.565

[90] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06732

[91] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 2852–2858. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10770

[92] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, ser. NIPS'99. Cambridge, MA, USA: MIT Press, 1999, p. 1057–1063.

[93] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016. [Online]. Available: https://arxiv.org/pdf/1602.02410.pdf

[94] W. Li, W. Wei, K. Xu, W. Xie, D. Chen, and Y. Cheng, "Reinforcement learning with token-level feedback for controllable text generation," 2024. [Online]. Available: https://arxiv.org/abs/2403.11558

[95] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "LIMA: Less is more for alignment," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=KBMOKmX2he

[96] B. Y. Lin, A. Ravichander, X. Lu, N. Dziri, M. Sclar, K. Chandu, C. Bhagavatula, and Y. Choi, "The unlocking spell on base LLMs: Rethinking alignment via in-context learning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=wxJ0eXwwda

[97] S. Hallinan, F. Brahman, X. Lu, J. Jung, S. Welleck, and Y. Choi, "STEER: Unified style transfer with expert reinforcement," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7546–7562. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.506

[98] K. de Langis, R. Koo, and D. Kang, "Reinforcement learning with dynamic multi-reward weighting for multi-style controllable generation," 2024. [Online]. Available: https://arxiv.org/abs/2402.14146

[99] W. Hua, X. Yang, M. Jin, W. Cheng, R. Tang, and Y. Zhang, "Trustagent: Towards safe and trustworthy llm-based agents through agent constitution," 2024. [Online]. Available: https://arxiv.org/abs/2402.01586

[100] Z. Wan, X. Wang, C. Liu, S. Alam, Y. Zheng *et al.*, "Efficient large language models: A survey," *arXiv preprint arXiv:2312.03863*, vol. 1, 2023.

[101] D. Yu, Z. Yu, and K. Sagae, "Attribute alignment: Controlling text generation from pre-trained language models," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2251–2268. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.194

[102] J. Qian, L. Dong, Y. Shen, F. Wei, and W. Chen, "Controllable natural language generation with contrastive prefixes," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2912–2924. [Online]. Available: https://aclanthology.org/2022.findings-acl.229

[103] D. C. Senadeera and J. Ive, "Controlled text generation using t5 based encoder-decoder soft prompt tuning and analysis of the utility of generated text in ai," 2022. [Online]. Available: https://arxiv.org/abs/2212.02924

[104] H. Wang and L. Sha, "Harnessing the plug-and-play controller by prompting," in *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 165–174. [Online]. Available: https://aclanthology.org/2023.gem-1.14

[105] R. D. Ajwani, Z. Zhu, J. Rose, and F. Rudzicz, "Plug and play with prompts: A prompt tuning approach for controlling text generation," 2024. [Online]. Available: https://arxiv.org/abs/2404.05143

[106] Y. Gu, X. Feng, S. Ma, L. Zhang, H. Gong, and B. Qin, "A distributional lens for multi-aspect controllable text generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1023–1043. [Online]. Available: https://aclanthology.org/2022.emnlp-main.67

[107] K. Yang, D. Liu, W. Lei, B. Yang, M. Xue, B. Chen, and J. Xie, "Tailor: A soft-prompt-based approach to attribute-based controlled text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 410–427. [Online]. Available: https://aclanthology.org/2023.acl-long.25

[108] X. Huang, Z. Liu, P. Li, T. Li, M. Sun, and Y. Liu, "An extensible plug-and-play method for multi-aspect controllable text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 15 233–15 256. [Online]. Available: https://aclanthology.org/2023.acl-long.849

[109] Z. Lu, W. Wei, X. Qu, X.-L. Mao, D. Chen, and J. Chen, "Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5933–5957. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.395

[110] H. Ding, L. Pang, Z. Wei, H. Shen, X. Cheng, and T.-S. Chua, "MacLaSa: Multi-aspect controllable text generation via efficient sampling from compact latent space," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4424–4436. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.292

[111] Y. Gu, X. Feng, S. Ma, L. Zhang, H. Gong, W. Zhong, and B. Qin, "Controllable text generation via probability density estimation in the latent space," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 590–12 616. [Online]. Available: https://aclanthology.org/2023.acl-long.704

[112] Y. Liu, X. Liu, X. Zhu, and W. Hu, "Multi-aspect controllable text generation with disentangled counterfactual augmentation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9231–9253. [Online]. Available: https://aclanthology.org/2024.acl-long.500

[113] Z. Feng, H. Zhou, K. Mao, and Z. Zhu, "FreeCtrl: Constructing control centers with feedforward layers for learning-free controllable text generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7627–7640. [Online]. Available: https://aclanthology.org/2024.acl-long.412

[114] M. Kim, H. Lee, K. M. Yoo, J. Park, H. Lee, and K. Jung, "Critic-guided decoding for controlled text generation," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4598–4612. [Online]. Available: https://aclanthology.org/2023.findings-acl.281

[115] H. Deng and C. Raffel, "Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11 781–11 791. [Online]. Available: https://aclanthology.org/2023.emnlp-main.721

[116] X. Zhang and X. Wan, "MIL-decoding: Detoxifying language models at token-level via multiple instance learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 190–202. [Online]. Available: https://aclanthology.org/2023.acl-long.11

[117] M. Cao, M. Fatemi, J. C. K. Cheung, and S. Shabanian, "Successor features for efficient multisubject controlled text generation," 2023. [Online]. Available: https://arxiv.org/abs/2311.04921

[118] D. Landsman, J. Z. Chen, and H. Zaidi, "BeamR: Beam reweighing with attribute discriminators for controllable text generation," in *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds. Online only: Association for Computational Linguistics, Nov. 2022, pp. 422–437. [Online]. Available: https://aclanthology.org/2022.findings-aacl.40

[119] X. Lu, P. West, R. Zellers, R. Le Bras, C. Bhagavatula, and Y. Choi, "NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4288–4299. [Online]. Available: https://aclanthology.org/2021.naacl-main.339

[120] S. Mudgal, J. Lee, H. Ganapathy, Y. Li, T. Wang, Y. Huang, Z. Chen, H.-T. Cheng, M. Collins, J. Chen, A. Beutel, and A. Beirami, "Controlled decoding from language models," in *Socially Responsible Language Modelling Research*, 2023. [Online]. Available: https://openreview.net/forum?id=jo57H1CpD8

[121] Y. Gu, X. Feng, S. Ma, J. Wu, H. Gong, and B. Qin, "Improving controllable text generation with position-aware weighted decoding," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3449–3467. [Online]. Available: https://aclanthology.org/2022.findings-acl.272

[122] H. Liu, B. Wang, T. Yao, H. Liang, J. Xu, and X. Hu, "Bridging the gap between training and inference of bayesian controllable language models," 2022. [Online]. Available: https://arxiv.org/abs/2206.05519

[123] T. Meng, S. Lu, N. Peng, and K.-W. Chang, "Controllable text generation with neurally-decomposed oracle," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 28 125–28 139. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b40d5797756800c97f3d525c2e4c8357-Paper-Conference.pdf

[124] C. Xu, T. Lan, C. Yu, W. Wang, J. Gao, Y. Ji, Q. Dong, K. Qian, P. Li, W. Bi, and B. Hu, "Decider: A dual-system rule-controllable decoding framework for language generation," 2024. [Online]. Available: https://arxiv.org/abs/2403.01954

[125] C. Zheng, C. Shi, K. Vafa, A. Feder, and D. Blei, "An invariant learning characterization of controlled text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3186–3206. [Online]. Available: https://aclanthology.org/2023.acl-long.179

[126] T. Zhong, Q. Wang, J. Han, Y. Zhang, and Z. Mao, "Air-decoding: Attribute distribution reconstruction for decoding-time controllable text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8233–8248. [Online]. Available: https://aclanthology.org/2023.emnlp-main.512

[127] J. Dekoninck, M. Fischer, L. Beurer-Kellner, and M. Vechev, "Controlled text generation via language model arithmetic," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=SLw9fp4yI6

[128] J. Pei, K. Yang, and D. Klein, "PREADD: Prefix-adaptive decoding for controlled text generation," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 018–10 037. [Online]. Available: https://aclanthology.org/2023.findings-acl.636

[129] H. Chen, H. Li, D. Chen, and K. Narasimhan, "Controllable text generation with language constraints," 2022. [Online]. Available: https://arxiv.org/abs/2212.10466

[130] L. Qin, S. Welleck, D. Khashabi, and Y. Choi, "Cold decoding: Energy-based constrained text generation with langevin dynamics," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 9538–9551. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/3e25d1aff47964c8409fd5c8dc0438d7-Paper-Conference.pdf

[131] X. Guo, F. Yu, H. Zhang, L. Qin, and B. Hu, "Cold-attack: Jailbreaking llms with stealthiness and controllability," 2024. [Online]. Available: https://arxiv.org/abs/2402.08679

[132] X. Liu, M. Khalifa, and L. 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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 186–200. [Online]. Available: https://aclanthology.org/2023.acl-short.18

[133] J. Forristal, F. Mireshghallah, G. Durrett, and T. Berg-Kirkpatrick, "A block metropolis-hastings sampler for controllable energy-based text generation," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, J. Jiang, D. Reitter, and S. Deng, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 403–413. [Online]. Available: https://aclanthology.org/2023.conll-1.26

[134] C. Han, J. Xu, M. Li, Y. Fung, C. Sun, N. Jiang, T. Abdelzaher, and H. Ji, "Word embeddings are steers for language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.12798

[135] P. Xu, M. Patwary, M. Shoeybi, R. Puri, P. Fung, A. Anandkumar, and B. Catanzaro, "MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models," in *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2831–2845. [Online]. Available: https://aclanthology.org/2020.emnlp-main.226

[136] C. Ma, T. Zhao, M. Shing, K. Sawada, and M. Okumura, "Focused prefix tuning for controllable text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1116–1127. [Online]. Available: https://aclanthology.org/2023.acl-short.96

[137] S. Yu, C. Lee, H. Lee, and S. Yoon, "Controlled text generation for black-box language models via score-based progressive editor," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 14 215–14 237. [Online]. Available: https://aclanthology.org/2024.acl-long.767

[138] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[139] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[140] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[141] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the second international conference on Human Language Technology Research*, 2002, pp. 138–145.

[142] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119. [Online]. Available: https://aclanthology.org/N16-1014

[143] Z. Shao, M. Huang, J. Wen, W. Xu, and X. Zhu, "Long and diverse text generation with planning-based hierarchical variational model," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3257–3268. [Online]. Available: https://aclanthology.org/D19-1321

[144] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 1097–1100.

[145] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[146] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 563–578. [Online]. Available: https://aclanthology.org/D19-1053

[147] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704

[148] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Aug. 8-12 2006, pp. 223–231. [Online]. Available: https://aclanthology.org/2006.amta-papers.25

[149] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[150] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 382–398.

[151] Y. Li, Z. Wei, H. Jiang, and C. Gong, "Destein: Navigating detoxification of language models via universal steering pairs and head-wise activation fusion," 2024. [Online]. Available: https://arxiv.org/abs/2404.10464

[152] P. Wang, D. Zhang, L. Li, C. Tan, X. Wang, K. Ren, B. Jiang, and X. Qiu, "Inferaligner: Inference-time alignment for harmlessness through cross-model guidance," 2024. [Online]. Available: https://arxiv.org/abs/2401.11206

[153] C. Xia, C. Xing, J. Du, X. Yang, Y. Feng, R. Xu, W. Yin, and C. Xiong, "Fofo: A benchmark to evaluate llms' format-following capability," *arXiv preprint arXiv:2402.18667*, 2024.

[154] J. Devlin, M.-W. Chang, K. Lee, and K. 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), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[155] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[156] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD

[157] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[158] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 657–668. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.58

[159] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015

[160] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 165–172.

[161] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, Eds. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170

[162] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[163] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.148

[164] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, G. Kondrak and T. Watanabe, Eds. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995. [Online]. Available: https://aclanthology.org/I17-1099

[165] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, Will Cukierski, "Jigsaw toxic comment classification challenge," 2018. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[166] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.301

[167] P. Ke, H. Zhou, Y. Lin, P. Li, J. Zhou, X. Zhu, and M. Huang, "CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2306–2319. [Online]. Available: https://aclanthology.org/2022.acl-long.164

[168] D. Ashok and B. Poczos, "Controllable text generation in the instruction-tuning era," 2024. [Online]. Available: https://arxiv.org/abs/2405.01490

[169] Y. Chen, B. Xu, Q. Wang, Y. Liu, and Z. Mao, "Benchmarking large language models on controllable generation under diversified instructions," 2024. [Online]. Available: https://arxiv.org/abs/2401.00690

[170] A. Rahali and M. A. Akhloufi, "Deeppress: guided press release topic-aware text generation using ensemble transformers," *Neural Computing and Applications*, vol. 35, no. 17, pp. 12 847–12 874, 2023. [Online]. Available: https://doi.org/10.1007/s00521-023-08393-4

[171] A. Spangher, Y. Ming, X. Hua, and N. Peng, "Sequentially controlled text generation," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6848–6866. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.509

[172] C. Shen, L. Cheng, R. Zhou, L. Bing, Y. You, and L. Si, "MReD: A meta-review dataset for structure-controllable text generation," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2521–2535. [Online]. Available: https://aclanthology.org/2022.findings-acl.198

[173] N. Jinran, Y. Liner, C. Yun, K. Cunliang, Z. Junhui, and Y. Erhong, "Lexical complexity controlled sentence generation for language learning," in *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, M. Sun, B. Qin, X. Qiu, J. Jiang, and X. Han, Eds. Harbin, China: Chinese Information Processing Society of China, Aug. 2023, pp. 648–664. [Online]. Available: https://aclanthology.org/2023.ccl-1.56

[174] D. Kalpakchi and J. Boye, "Swectrl-mini: a data-transparent transformer-based large language model for controllable text generation in swedish," 2023. [Online]. Available: https://arxiv.org/abs/2304.13994

[175] S. V. Vychegzhanin and E. V. Kotelnikov, "Collocation2text: Controllable text generation from guide phrases in russian," in *Computational Linguistics and Intellectual Technologies*. RSUH, Jun. 2022. [Online]. Available: http://dx.doi.org/10.28995/2075-7182-2022-21-564-576

[176] Z. Yang, H. Jiang, A. Deng, and Y. Li, "Topic-oriented controlled text generation for social networks," *Journal of Signal Processing Systems*, vol. 96, no. 2, pp. 131–151, feb 2024. [Online]. Available: https://doi.org/10.1007/s11265-023-01907-2

[177] C. T. Leong, Y. Cheng, J. Wang, J. Wang, and W. Li, "Self-detoxifying language models via toxification reversal," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4433–4449. [Online]. Available: https://aclanthology.org/2023.emnlp-main.269

[178] Q. Zeng, M. Jin, Q. Yu, Z. Wang, W. Hua, Z. Zhou, G. Sun, Y. Meng, S. Ma, Q. Wang, F. Juefei-Xu, K. Ding, F. Yang, R. Tang, and Y. Zhang, "Uncertainty is fragile: Manipulating uncertainty in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2407.11282

[179] Y. Zheng, R. Zhang, M. Huang, and X. Mao, "A pre-training based personalized dialogue generation model with persona-sparse data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 9693–9700, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6518

[180] Y. Zeng and J.-Y. Nie, "A simple and efficient multi-task learning approach for conditioned dialogue generation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4927–4939. [Online]. Available: https://aclanthology.org/2021.naacl-main.392

[181] H. Chen, B. Wang, K. Yang, and Y. Song, "Eccrg: A emotion- and content-controllable response generation model," in *Collaborative Computing: Networking, Applications and Worksharing*, H. Gao, X. Wang, and N. Voros, Eds. Cham: Springer Nature Switzerland, 2024, pp. 115–130.

[182] X. Wang, H. Jiang, Z. Wei, and S. Zhou, "CHAE: Fine-grained controllable story generation with characters, actions and emotions," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6426–6435. [Online]. Available: https://aclanthology.org/2022.coling-1.559

[183] J. Cho, M. Jeong, J. Bak, and Y.-G. Cheong, "Genre-controllable story generation via supervised contrastive learning," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2839–2849. [Online]. Available: https://doi.org/10.1145/3485447.3512004

[184] S. Vychegzhanin, A. Kotelnikova, A. Sergeev, and E. Kotelnikov, "Controllable story generation based on perplexity minimization," in *Analysis of Images, Social Networks and Texts*, D. I. Ignatov, M. Khachay, A. Kutuzov, H. Madoyan, I. Makarov, I. Nikishina, A. Panchenko, M. Panov, P. M. Pardalos, A. V. Savchenko, E. Tsymbalov, E. Tutubalina, and S. Zagoruyko, Eds. Cham: Springer Nature Switzerland, 2024, pp. 154–169.

[185] Y. Sasazawa, T. Morishita, H. Ozaki, O. Imaichi, and Y. Sogawa, "Controlling keywords and their positions in text generation," in *Proceedings of the 16th International Natural Language Generation Conference*, C. M. Keet, H.-Y. Lee, and S. Zarrieß, Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 407–413. [Online]. Available: https://aclanthology.org/2023.inlg-main.29