



上海交通大学学位论文

# 大模型的风格化文本生成及应用研究

姓 名：王凯俊

学 号：520021910282

导 师：孙世轩

学 院：电子信息与电气工程学院

专业名称：计算机科学与技术

申请学位层次：学士

2024 年 05 月

**A Dissertation Submitted to  
Shanghai Jiao Tong University for Bachelor Degree**

**RESEARCH ON STYLIZED TEXT GENERATION  
AND APPLICATIONS OF LARGE LANGUAGE  
MODELS**

**Author : Wang Kaijun**

**Supervisor: Sun Shixuan**

School of Electronic Information and Electrical Engineering

Shanghai Jiao Tong University

Shanghai, P.R. China

May, 2024

## 上海交通大学

### 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：王凯俊  
日期：2024 年 5 月 15 日

## 上海交通大学

### 学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

公开论文

内部论文，保密  1 年 /  2 年 /  3 年，过保密期后适用本授权书。

秘密论文，保密 \_\_\_\_ 年（不超过 10 年），过保密期后适用本授权书。

机密论文，保密 \_\_\_\_ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打“√”）

学位论文作者签名：王凯俊

日期：2024 年 5 月 15 日

指导教师签名：孙世轩

日期：2024 年 5 月 15 日

## 摘要

大语言模型的出现颠覆了众多领域的范式，在自然语言生成领域，大语言模型强大的文本“涌现”能力使得文本生成任务拥有更大的发展潜力。传统的自然语言生成任务，如机器翻译、对话系统、文本摘要等，其应用场景相对固定，而大语言模型的出现使得文本生成任务的应用场景更加广泛，在本文中，将考虑基于大语言模型完成对文本生成任务应用的深入探讨。

本文归纳和分析了大语言模型（如 GPT 系列）的发展历程、基本结构及其在自然语言处理领域的广泛应用，特别是在自然语言生成领域。在此基础上，研究首先回顾了传统的自然语言生成任务，然后探讨了大语言模型推动下的新任务和评估标准。文章针对文本生成及其质量评估进行了详细研究，基于大语言模型对生成和评估两方面进行详尽的探讨和实验。在文本生成部分，本文对该任务根据实际应用需求划分为两大阶段任务，相对基本且普适的简单文本生成和考虑到实际应用需求的特定文本生成；后者又分为具体的实时文本生成，长文本生成和风格化文本生成。在文本评估部分，本文给出了基于大语言模型的文本质量评估方法，并对该方法进行了验证。对于以上的两方面的方法，本文还设计一系列实验检测其性能和可行性，使用多个数据集，包括 ROC 故事、Dailymail 数据集、百大人工智能人物等，验证了各种文本生成方法的有效性和大语言模型在生成实时文本、长文本和风格化文本方面的评估性能。

总之，本文提出了多种文本生成及评估方法，通过多数据集的广泛实验，本文不仅确认了各种文本生成方法的特性，还提高了对基于大语言模型的文本评估技术的理解，为未来研究提供了实证基础。利用上述方法，目前成果已经成功应用到公众号的文章自动生成中，已经在线运行。

**关键词：**大语言模型，自然语言生成，文本生成，文本质量评估，文本风格化

## ABSTRACT

The emergence of large language models has revolutionized paradigms across numerous fields. In the domain of natural language generation, the robust text "Emergent" capability of large models significantly enhances the potential for text generation tasks. Traditional natural language generation tasks, such as machine translation, dialogue systems, and text summarization, have relatively fixed application scenarios. However, the advent of large language models has broadened these scenarios considerably. This paper delves into a detailed discussion on the application of large language models to text generation tasks.

This paper summarizes and analyzes the development history, basic structures, and broad applications of large language models (such as the GPT series), especially in the field of natural language generation. Based on this, the research first revisits traditional natural language generation tasks and then explores new tasks and evaluation standards driven by large models. A detailed study on text generation and its quality assessment is conducted, discussing and experimenting extensively with both aspects based on large language models. For text generation, this paper divides the task into two main stages based on practical application needs: basic and universally applicable simple text generation, and specific text generation tailored to practical requirements. The latter is further divided into real-time text generation, long text generation, and stylized text generation. In the text evaluation section, the paper introduces a text quality assessment method based on large models and validates this method. For both aspects, the paper designs a series of experiments to test their performance and feasibility, using multiple datasets including the ROC stories, Dailymail dataset, and a list of top 100 AI personalities, to confirm the effectiveness of various text generation methods and the assessment capabilities of large models in generating real-time text, long text, and stylized text.

In summary, the paper proposes various methods for text generation and evaluation, and through extensive experiments across multiple datasets, it not only identifies the characteristics of various text generation methods but also enhances understanding of text evaluation technologies based on large language models, providing an empirical foundation for future research. Utilizing these methods, the current results have already been successfully applied to the automatic generation of articles for official accounts and are operational online.

**Key words:** LLMs, NLG, Text Generation, Text Evaluation, Text Stylization

# 目 录

摘要 .....	D
<b>ABSTRACT .....</b>	<b>E</b>
<b>第一章 绪论 .....</b>	<b>1</b>
1.1 研究背景 .....	1
1.2 本文研究主要内容 .....	1
1.3 文章结构安排 .....	2
<b>第二章 大语言模型调研 .....</b>	<b>4</b>
2.1 本章引言 .....	4
2.2 大模型发展历程 .....	4
2.2.1 神经语言模型 (NLM) .....	4
2.2.2 预训练语言模型 (PLM) .....	4
2.2.3 大语言模型 (LLM) .....	5
2.3 大模型结构及特点 .....	5
2.3.1 预训练语言模型 (PLM) 的延伸 .....	5
2.3.2 上下文学习 (in-context learning) .....	6
2.3.3 思维链条 (Chain-of-Thought) .....	6
2.3.4 基于人类反馈的强化学习 (RLHF) .....	7
2.4 大模型评估领域调研 .....	7
2.4.1 自然语言处理 (NLP) 领域 .....	7
2.4.2 社会科学领域 .....	8
2.4.3 工程领域 .....	9
2.4.4 医学领域 .....	9
2.4.5 代理应用领域 .....	10
2.5 本章小结 .....	10
<b>第三章 大模型背景下的自然语言生成 (NLG) .....</b>	<b>11</b>
3.1 本章引言 .....	11

3.2 NLG 传统任务 .....	11
3.2.1 机器翻译 .....	11
3.2.2 文本摘要 .....	11
3.2.3 对话生成 .....	12
3.2.4 故事生成 .....	12
3.3 NLG 评估方法 .....	12
3.4 NLG 评估的方向设定 .....	13
3.4.1 基于大语言模型本身的评估指标 .....	13
3.4.2 基于大语言模型提示词的评估指标 (Prompting LLMs) .....	13
3.4.3 基于大语言模型微调的评估指标 (Fine-tuning LLMs) .....	14
3.4.4 基于人机协同的评估指标 (Human-LLM Collaborative Evaluation) .....	14
3.5 本章小结 .....	15
<b>第四章 简单文本生成方法探究 .....</b>	<b>16</b>
4.1 本章引言 .....	16
4.2 文本基本生成方法 .....	16
4.2.1 文本直接生成 .....	17
4.2.2 基于角色的文本直接生成 .....	17
4.2.3 基于模版的文本直接生成 .....	17
4.2.4 基于 Meta 的文本阶段生成 .....	18
4.3 文章生成实验 .....	18
4.4 本章小结 .....	19
<b>第五章 特定文本生成方法 .....</b>	<b>20</b>
5.1 本章引言 .....	20
5.2 实时文章生成 .....	20
5.3 长文章生成 .....	21
5.3.1 基于模版的生成方法 .....	21
5.3.2 基于 Function calling 的 Meta 生成方法 .....	23
5.3.3 实验设置 .....	23
5.4 风格化文本生成 .....	23

5.4.1	实验设置 .....	24
5.4.2	生成方法 .....	26
5.4.3	评估方法 .....	27
5.5	本章小结 .....	28
<b>第六章</b>	<b>生成文本质量评估方法及验证 .....</b>	<b>29</b>
6.1	本章引言 .....	29
6.2	生成文本评估方法 .....	29
6.3	实验一：故事续写 .....	29
6.3.1	实验设定 .....	29
6.3.2	实验结果和分析 .....	31
6.4	实验二：文本总结 .....	33
6.4.1	实验设定 .....	33
6.4.2	实验结果和分析 .....	35
6.5	本章小结 .....	38
<b>第七章</b>	<b>实验结果分析 .....</b>	<b>39</b>
7.1	基本文本生成结果和分析 .....	39
7.2	一般文本生成结果和分析 .....	40
7.2.1	实时长文本结果分析 .....	41
7.2.2	风格化文本生成结果分析 .....	41
7.3	本章小结 .....	44
<b>第八章</b>	<b>全文总结 .....</b>	<b>46</b>
8.1	内容总结 .....	46
8.2	可能的限制 .....	46
8.3	研究展望 .....	47
<b>参 考 文 献 .....</b>	<b>48</b>	
<b>致 谢 .....</b>	<b>51</b>	

# 第一章 绪论

## 1.1 研究背景

伴随着 ChatGPT 的发布，大语言模型的热潮席卷了整个计算机领域。从 GPT1, GPT2, GPT3 这些学术使用的模型，到可以商用、好用的应用。大模型的热潮席卷了工业界和学界<sup>[1-3]</sup>。各个公司或者院校都在根据自身需要推出自己的大模型，这其中包括 OpenAI 的 GPT (Generative Pre-trained Transformer) 系列，Meta 开源的 LLaMA 系列，国内的百度的文心一言系列等<sup>[4]</sup>。

大语言模型的出现颠覆了很多领域的交互范式，也解决了相当数量的问题。在自然语言处理领域，大语言模型的出现使得很多任务的效果得到了显著提升，比如机器翻译、文本生成、文本分类等<sup>[5-6]</sup>。不仅限于自然语言处理领域，包括社会科学领域、工程领域、医学领域等都有大语言模型的相关应用。

对于本文中重点探讨的自然语言生成领域，传统的任务包括机器翻译、摘要总结、对话生成、故事生成等。在大语言模型出现以后，这些任务的效果得到了显著提升，同时由于大语言模型强大的文本生成能力，这一领域的实际应用场景也得到了极大的拓展。

在这种背景下，了解调研大语言模型的具体发展及结构，研究文本生成的方法和评估方法对于更好的使用大语言模型，更好的解决问题，更好的应用大语言模型是非常有必要的。

## 1.2 本文研究主要内容

本文首先对大语言模型的发展历程，基本结构和应用及相应的评估方法进行调研介绍。对于大语言模型的发展历程，从最开始的神经语言模型到预训练语言模型，再到大语言模型进行详尽的调研。对于其结构及相应特点，以 ChatGPT 为例，介绍了其上下文学习，思维链条，基于人类反馈的强化学习等特点。对于大语言模型评估及应用领域，介绍了在自然语言处理，社会科学领域，医学领域，以及代理应用领域的相关评估及应用。在调研清楚大语言模型的具体应用范围以及其结构特点后，选择自然语言生成领域作为重点研究领域。

对于自然语言生成 (NLG) 领域，首先调研其传统的任务，对机器翻译，文本摘

要，对话生成，故事生成等自然语言生成的传统任务进行调研，了解并确定其应用范围及评估方法。在此基础上发现其文本生成在大语言模型产生后的全新任务和评估指标的在这些任务的不足，根据这些发现，选择文本生成及文本质量评估作为本文的具体研究任务。

对于文本生成及文本质量评估任务，本文将分阶段完成对文本生成的探究，具体分为三个阶段：探究简单文本生成方法，探究特定文本生成方法，以及探究文本质量评估方法，大致流程如图1-1所示。

首先将对比较简单的文本生成进行探究，提出四种文本生成方法，分别是文本直接生成，基于角色的文本直接生成，基于模版的文本直接生成，基于 Meta 的文本阶段生成。这些方法可以作为文本生成的基本方法，可以以模块的形式组合成更复杂的文本生成方法。

在探究简单文本生成方法的基础上，考虑到实际应用的场景，本文将进一步探究特定种类文本的生成方法，包括实时文本，长文本，以及风格化文本的生成。这些文本种类都是符合实际应用场景的文本，对于这些文本的生成方法的探究，可以更好地应用到实际应用场景中。

在探究文本生成方法后，本文对文本生成质量的评估方法进行探究。在文本生成质量评估方法中，本文将提出基于大语言模型的评估方法，并设计实验验证其可行性，并将其实际运用到对文本生成的评估中。

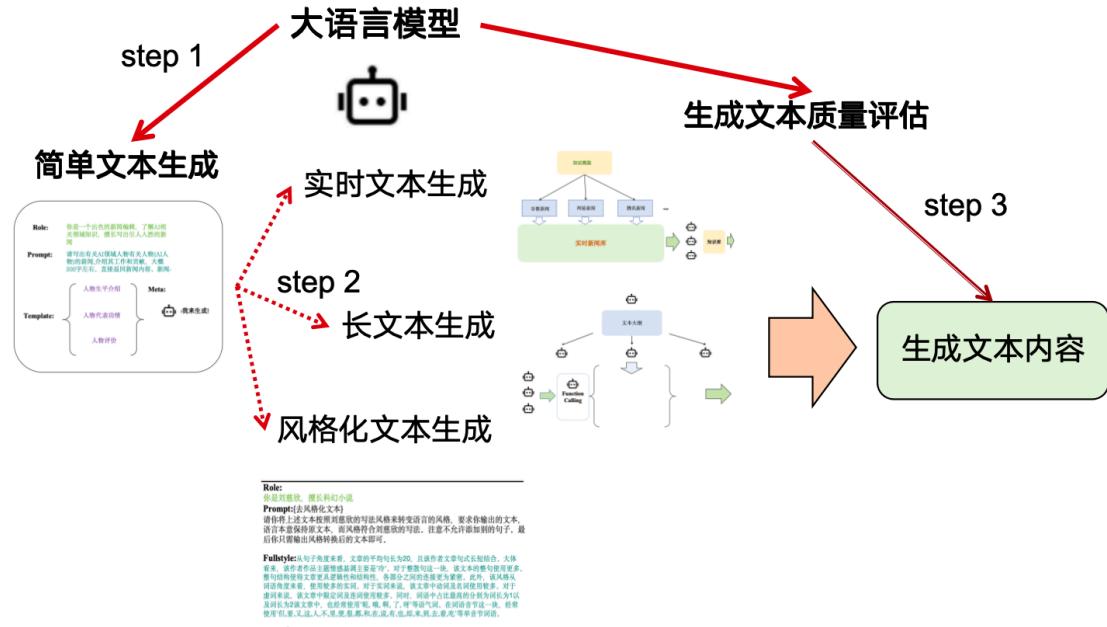
最后，本文将对以上文本生成方法进行实验。对于简单文本生成的方法，使用《时代》周刊评选的 100 大人工智能人物作为生成内容列表，对四种生成方法进行实验，并使用基于大语言模型的评估方法和基于比较的评估方法对生成文本进行评估。对于实时文本和长文本的生成方法，将通过为期两个月的实时文章投放公众号来测试其文本质量。对于风格化文本的生成方法，使用《三体》和《阿 Q 正传》作为生成内容，通过与原文和去风格化文本的对比，来评估生成文本的质量。

在以上的实验中，本文采用多种数据集进行实验测试，通过多种文本的多次实验，明确各种文本生成方法的特性，对基于大语言模型的文本评估性能有更明确的测定，为进一步探索更高效、更优质的文本提供了有效的尝试和实践。

### 1.3 文章结构安排

本章主要介绍了本文的研究背景，以及本文的研究内容。本文的内容安排如下：

- 第一章介绍本文的研究背景，以及本文的研究内容；



图中为本文研究文本生成方法的示意图，分为三个步骤，分别为探究简单文本生成方法，探究特定文本生成方法，以及探究文本质量评估方法；最终完成对文本的生成评估全流程。

图 1-1 文本生成探究示意图  
Figure 1-1 Graph of Text Generation Inquiry

- 第二章对大语言模型的发展历程，结构，应用及评估进行综合调研；
- 第三章将对自然语言生成这一特定领域介绍相关任务及评估方法；
- 第四章将对生成文本方法，并进行具体生成简单文本（人物介绍文本）实验；
- 第五章将基于第四章的方法，对特定种类文本：实时文本，长文本，以及风格化文本采用延伸的生成及评估方法，并进行相应的实验；
- 第六章将介绍基于大语言模型的文本评估方法，并通过实验验证其可行性；
- 第七章会对第四、五章的实验结果进行展示，并利用第六章的评估方法进行详细的分析；
- 第八章将对全文进行总结，阐明文章中可能的限制，并写明研究展望。

## 第二章 大语言模型调研

### 2.1 本章引言

大模型在最近的两年中席卷了工业界、学术界，以及大众的日常生活。在本章的内容，将从大语言模型的发展历程开始调研，通过对大模型的结构、应用和相关评估的调研，梳理出较为清晰的脉络。

本章将分为三大部分进行调研介绍，分别为大语言模型的发展历程，大语言模型的结构特点，以及大语言模型评估领域的调研。

### 2.2 大模型发展历程

对于大语言模型的发展，本文采用对语言模型的理解过程为顺序介绍，从最开始的神经语言模型，到预训练语言模型，再到大语言模型。整个内容主要介绍了从神经网络在本世纪兴起后的语言模型的发展历程，供读者理清大模型的发展脉络。

#### 2.2.1 神经语言模型（NLM）

对于神经语言模型（Neural language models），最有代表性的是 word2vec<sup>[7-8]</sup>。神经语言模型通过像循环神经网络等结构，描述语言序列的概率<sup>[9]</sup>。在这种模型的建模下，神经网络被用来学习单词的分布表示，这种方法证明是有效的，但这种模型的网络架构是相对简单的，无法从深度层面上学习到更多的语言特征<sup>[9-11]</sup>。

#### 2.2.2 预训练语言模型（PLM）

预训练语言模型（Pre-trained language models），相对神经语言模型的显著特点是整体架构更加大。这一类语言模型通过“预训练”的方法，完成对通用任务的模型训练。对于特定的各种任务，通过“微调”的方式实现对特定下游任务的适配<sup>[1-2,12]</sup>。对于这类模型，最有代表性的是 Transformer 模型架构<sup>[12]</sup>。这种模型建立在编码器-解码器架构上，基于自注意力机制实现模型对单词的感知，这种结构也深深地影响到后续大语言模型的架构设计。受到 Transformer 模型架构的启发，GPT1、GPT2 和 BERT 引入类似的架构，并加以改动<sup>[1-2,13]</sup>。在这个阶段，模型的体量都在向“大”靠拢，在以上模型的研究中均有发现越大似乎越好的趋势，并且并没有明显的阈值。

### 2.2.3 大语言模型 (LLM)

大语言模型可以认为是由预训练语言模型“量变到质变”的模型。研究人员发现通过提升模型的参数大小和训练数据大小，对于处理下游任务的性能在不断提升，收到这种现象的激励，GPT3 和 PaLM 等模型提出了<sup>[3,14-15]</sup>。这种类型相比之前的模型更“大”，训练难度更大，如果需要微调来处理下游任务，代价是巨大的。但是，这种大语言模型展示了相比之前模型不具有的“涌现”能力——不需要复杂的微调，也可以直接用于下游的各种复杂任务，这些能力是“小”规模模型所不拥有的。

受到大语言模型能力的鼓舞，很多研究团队致力于大语言模型的研发，其中以 OpenAI 的 GPT 系列和 Meta 的 LLaMA 系列作为闭源和开源模型的代表。对于个人研究者或者实验室来说，从零训练大语言模型是不现实的，但是，使用这些训练好的模型，或者微调已有的大模型是相对方便的，可以通过访问接口的方式进行（如 GPT-4 的 API，文心一言的 API 等）。通过这种方式，可以实现对大语言模型能力的应用和探究。

现在的大语言模型在逐渐向更广的领域发展，除了自然语言处理领域，大语言模型正在向多模态领域、多语言领域、多任务领域等方向发展，这些领域的发展将会进一步推动大语言模型的架构和能力的提升<sup>[16-19]</sup>。

## 2.3 大模型结构及特点

在本部分，将以 ChatGPT 这一商用的大语言模型为介绍对象，介绍其中的几个核心结构和衍生的特点。ChatGPT 是一个基于 GPT-3 的模型，下面将从四个方面介绍 ChatGPT 的特殊性。

### 2.3.1 预训练语言模型 (PLM) 的延伸

ChatGPT 的第一个核心技术是预训练语言模型 (PLM)。预训练语言模型在之前的介绍中被描述为大语言模型之前的流行类型，但实际上大语言模型在相当程度上是预训练语言模型的延伸。总体的训练思维是通过利用大量的文本数据来学习语言的结构和用法。预训练语言模型使用一种称为自监督学习的技术进行训练，在训练过程中，模型最初被暴露于文本并学习预测文本的部分内容。这个初始训练是无监督的，意味着它不需要标记的输入输出对，只需要大量的文本。预训练语言模型为 ChatGPT 提供了基础技术，使其能够根据接收到的输入生成连贯且语境相关的文本。

这种能力源于模型在预训练阶段获得的对语言细微差别、语法和语义的理解。一旦预训练完成，模型便可以在更具体的任务上进行微调或其他技术进行适应，仅需极少的额外示例即可执行任务。

### 2.3.2 上下文学习 (in-context learning)

ChatGPT 的一个强大能力是上下文学习 (in-context learning)<sup>[3,20]</sup>。这种能力使模型能够在没有显式微调的情况下通过输入的上下文来理解和生成响应<sup>[3]</sup>。这是通过利用模型在预训练阶段学到的大量信息和模式来实现的。

在上下文学习中，模型使用其预训练期间获得的知识来推断输入中的隐藏意图和信息，从而在不需要额外训练数据的情况下，对新任务做出反应。例如，当给 ChatGPT 提供一个关于天气的问题时，它能够根据问题的上下文生成有关天气的合适回答，即使在训练中未特别针对天气问题进行优化。上下文学习的一个关键特点是其对“示例”（或提示）的使用。通过向模型展示一系列相关的示例，模型可以学习如何在类似情况下生成相应的输出。这种方式使 ChatGPT 能够进行所谓的“零次学习”（zero-shot learning）或“少次学习”（few-shot learning），在这些学习模式中，模型利用少量或没有示例就能完成任务<sup>[3]</sup>。这种技术的强大之处在于它使 ChatGPT 能够灵活应对各种语言处理任务，即使在训练数据中未直接涵盖这些任务时也能表现出色。此外，上下文学习也支持模型在对话中保持连贯性和上下文相关性，从而能更自然地与用户进行交互。

### 2.3.3 思维链条 (Chain-of-Thought)

ChatGPT 的一个核心技术是思维链条 (Chain-of-Thought)。这是一种在处理复杂问题时模拟人类思考过程的技术，特别适用于需要推理和逻辑判断的任务。通过这种技术，模型不仅生成最终答案，还能生成解释其答案的中间步骤，从而使得生成的内容更加透明和可解释。更重要的是，通过思维链条的方式，可以更好地激发大语言模型的知识提取能力，使其能够更好地处理复杂问题。这种技术的应用使 ChatGPT 能够更好地理解问题的背景和上下文，从而更准确地回答问题<sup>[21]</sup>。思维链条的工作方式是在模型的回答中包含一系列逻辑步骤，这些步骤详细描述了从问题到答案的思维过程。例如，在解答数学问题时，模型不仅给出最终答案，还会展示计算的每一步，如何逐步解决问题，这与人类解决问题的方式类似。这种方法的优点是增强了模型的解释性，使得用户不仅看到答案，还能理解模型是如何得出这个答案的。这对于

提高用户对模型输出的信任度非常重要，特别是在需要高度准确性和可靠性的应用场景中。思维链条还有助于模型在复杂情况下做出更加准确的决策。通过模拟人类的推理过程，模型能够更好地处理那些需要多步逻辑推理的问题，如多层次问题解答、复杂查询或者需要广泛背景知识的情况。总的来说，思维链条通过增加解释步骤来提高模型的输出质量，使其在解决复杂问题时更加精确和可靠，同时也提高了人机交互的自然性和用户满意度。

### 2.3.4 基于人类反馈的强化学习（RLHF）

ChatGPT 使用了基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）。这是一种高级的学习技术，通过这种技术，模型能够根据人类提供的反馈调整其行为，以产生更符合人类偏好和期望的输出。在强化学习中，模型通过试错的方式在一个环境中学习，以最大化获得的奖励<sup>[22]</sup>。在 ChatGPT 的开发中，这种环境通常是与人类互动的对话场景，奖励则是基于人类如何评价模型生成的回复。换句话说，如果一个回复得到了正面的人类评价，那么模型在未来生成类似回复的概率会增加<sup>[23-24]</sup>。

RLHF 的关键步骤包括：1. 数据收集：首先需要收集人类与模型的对话数据。这些数据不仅包括问题和回答，还包括人类对这些回答的评价。2. 奖励建模：根据人类给予的反馈创建一个奖励模型。这个模型学习哪些类型的回答是被视为好的或者不好的，从而用于训练 ChatGPT 生成更优质的回答。3. 策略优化：使用强化学习算法调整模型的行为，以最大化从奖励模型获得的预期奖励，通过这种奖励机制，模型会学习生成那些能够获得更高奖励的回答。

## 2.4 大模型评估领域调研

在本部分，将对大语言模型的具体应用及评估的各个领域任务进行调研介绍，既是作为调研大语言模型的重要部分，也是对于下一步进行文本生成部分的重要参考。具体领域包括：自然语言处理领域，社会科学领域，工程领域，医学领域，代理应用领域等。

### 2.4.1 自然语言处理（NLP）领域

在自然语言处理领域（Natural language processing），大语言模型的应用场景相当广泛，主要可以分为自然语言理解（NLU），自然语言生成（NLG），以及多语言等方面。

面<sup>[6]</sup>。下面将对这几个分领域进行简要介绍。

#### 2.4.1.1 自然语言理解 (NLU)

自然语言理解 (Natural language understanding) 是评估 LLMs 在理解文本方面能力的基础，主要包括以下任务：

- 情感分析：情感分析通常涉及通过文本的语言特征来识别情绪倾向，比如积极、消极或中性<sup>[5]</sup>。
- 文本分类：文本分类则是将文本数据按预设的类别进行组织，例如将新闻文章依照其主题进行分类<sup>[25]</sup>。
- 自然语言推理 (NLI)：自然语言推理 (Natural language inference) 目的是评估一个假设是否能逻辑上从一个前提推导出来，这一过程需要模型对语言逻辑的深刻理解<sup>[26]</sup>。

#### 2.4.1.2 自然语言生成 (NLG)

自然语言生成 (Natural language generation) 是测试语言模型在生成人类语言方面的一种方法。具体任务包括：

- 机器翻译：将一种语言的文本翻译成另一种语言。
- 摘要总结：要求从较长的文本中提取关键信息，生成简洁的总结<sup>[5]</sup>。
- 对话生成：关注如何在对话系统中自然流畅地回应用户的提问。
- 故事生成：生成连贯的故事情节，这一任务对模型的逻辑推理能力和故事情节的连贯性要求较高。

#### 2.4.1.3 多语言任务

评估模型在处理多种语言时的能力，尤其是那些资源较少的语言，例如阿拉伯语等<sup>[27]</sup>。

### 2.4.2 社会科学领域

在现代社会科学研究中，大语言模型 (LLMs) 的应用逐渐展现出其独特的价值。这些模型不仅能够处理和分析大量数据，还能在多个社会科学领域内提供深刻的见解。例如，在政治学领域，大语言模型能够生成关于政治意识形态的复杂讨论，展示其对政治理论和实践的深入理解<sup>[28]</sup>。在计算社会科学 (CSS) 的任务中，大语言模型的应用尤为广泛<sup>[29]</sup>。它们在分类任务如误导信息识别、立场分类和情绪分类中的表

现，揭示了它们处理社会科学数据的强大能力。此外，这些模型还能在生成任务中发挥作用，如模拟社会现象或预测社会动态，这些能力使得大语言模型成为理解和解析社会科学问题的强有力工具。法律应用是大语言模型评估应用的另一个重要领域。这些模型能够自动化生成法律文档和案例总结，展示了它们在理解复杂法律概念和用语方面的能力<sup>[30]</sup>。这不仅能够帮助法律专业人士提高工作效率，还能在确保信息准确传递的同时，辅助法律教育和公共法律服务。心理学研究也开始利用大语言模型进行跨学科探索。通过结合发展心理学和比较心理学的方法，研究者们评估了大语言模型在理解人类心理和行为方面的能力<sup>[31]</sup>。这种新颖的评估方法不仅提高了模型的认知能力，还为心理学研究提供了新的视角和工具。

### 2.4.3 工程领域

在工程领域，大语言模型的应用表现令人瞩目。在代码生成任务中，它们不仅能够根据给定指令生成高质量的编程代码，还能理解并执行复杂的编程任务<sup>[32]</sup>。此外，大语言模型在软件工程任务中也表现出色，能够进行代码漏洞检测和信息检索等高级功能。尽管大语言模型在工程领域上取得了一定进展，但它们在处理高级科学问题和专业工程任务时仍显示出一定的局限性<sup>[33]</sup>。这些评估不仅揭示了大语言模型在执行特定任务时的能力，也提供了宝贵的见解，帮助研究者更好地理解这些模型的实用性和潜在改进方向。通过不断的评估和优化，大语言模型有望在未来在工程领域发挥更大的作用。

### 2.4.4 医学领域

在医疗领域中，大语言模型展示出了其处理复杂和专业化任务的强大能力。这些模型不仅在医疗咨询方面提供支持，帮助患者和医疗专业人员获取可靠的健康信息，还在医学辅助和医学检查领域中扮演了关键角色。在医疗咨询中，大语言模型能够回答关于健康和疾病的具体问题，处理复杂的医疗术语，并提供基于证据的建议<sup>[34]</sup>。作为医学辅助工具，这些模型支持医疗流程，例如病历管理、药物交互作用查询以及为临床决策提供数据支持<sup>[35]</sup>。此外，大语言模型在解读医学图像和分析实验室报告结果方面也展现出其潜力，这需要模型不仅处理文字信息，还能有效整合视觉和数值数据。

尽管这些技术的潜力巨大，但在在医学领域的大语言模型需要考虑很多除了本身性能外的很多因素，准确性和可靠性是评估的重中之重，模型的每一个建议和诊断

都可能直接影响患者的健康和治疗效果；考虑到处理的数据通常涉及个人健康信息，遵守隐私和伦理标准因此也成为了重要的考虑指标；评估大语言模型时，还需要重点关注其对专业术语的处理能力，以及模型在处理真实世界医疗数据时的表现<sup>[6]</sup>。

#### 2.4.5 代理应用领域

大语言模型（LLMs）在代理应用方面正越来越多地被用作智能代理，以自动执行各种任务并与用户进行交互<sup>[36]</sup>。这些模型被设计成能够理解和响应用户的需求，从而执行诸如日程管理、邮件响应和其他工作流自动化的任务<sup>[37]</sup>。在智能家居系统中的应用尤其引人注目，大语言模型需要根据用户的习惯和偏好调整其行为，显示出高度的情境适应性。

评估这类应用时，研究人员通常关注模型响应的速度和准确性，这直接影响用户的满意度和整体体验。此外，模型的适应能力也是一个重要评估指标，它衡量了模型面对变化的环境或用户需求时的学习和适应能力。通过这些评估，可以深入理解大语言模型在实际操作中的表现，从而为进一步的技术改进提供指导。这些智能代理的成功实施不仅可以提高用户体验，还能使日常任务自动化程度更高，实现个性化服务。

### 2.5 本章小结

本章大语言模型的发展历程，核心的结构和特点，以及应用评估领域进行了调研介绍，为下一步的文本生成任务提供了重要的参考。

## 第三章 大模型背景下的自然语言生成（NLG）

### 3.1 本章引言

基于以上对于大语言模型的调研，在本章内容将重点探究自然语言生成的方面。本文章将对自然语言生成的具体任务和对应的评估进行深入的探索。

本章节将对在大语言模型产生后的背景下，对自然语言生成这一细分领域进行调研，并确定探究的具体实验和方法。

### 3.2 NLG 传统任务

在调研介绍大语言模型部分，已经简要介绍了自然语言生成的任务，这里将对自然语言生成的任务进行更详细的介绍。本部分将介绍四个有代表性的自然语言生成任务，分别是机器翻译（Machine Translation, MT）、文本摘要（Text Summarization, TS）、对话生成（Dialogue Generation, DG）和故事生成（Story Generation, SG）。

#### 3.2.1 机器翻译

机器翻译任务的核心是将句子或文档从源语言转换为目标语言，同时保留相同的语义。这一任务涉及将文本从一种语言翻译成另一种语言，同时保持原始内容的语义完整性。评估通常关注翻译的流畅性、适当性和源文本的忠实度。评估方法可能包括比较生成的翻译与人工翻译或预先定义的标准翻译。

#### 3.2.2 文本摘要

对于文本摘要（总结）任务，这一任务要求生成较长文本的简洁且相关的摘要。在评估摘要时，重点是摘要的连贯性、相关性以及是否能够准确反映原文的主要信息。这种类型的评估通常会涉及比较生成的摘要与一组参考摘要。

SummEval 是广泛使用的评估基准之一<sup>[38]</sup>。该评估包括由 16 个模型从 CNN/DailyMail 测试集随机抽样的源新闻文章中生成的摘要，每个摘要都经过五名独立的众包工作人员和三名独立专家在李克特表（Likert）上的注释沿着四个维度从 1 到 5 进行评分：连贯性、一致性、流畅性和相关性。

### 3.2.3 对话生成

对话生成任务目的在于在对话中生成类似人类的响应<sup>[25]</sup>。评估标准通常包括响应的自然性、与前一对话回合的相关性、以及对话的吸引力，此外，还评估对话的连贯性和上下文适应性。相对有代表性的对话语料库是 Topical-Chat 和 PersonaChat<sup>[39-40]</sup>。

### 3.2.4 故事生成

故事生成（续写）任务要求根据给定的故事开头或写作要求创建连贯且与上下文相关的叙述或故事。评估侧重于故事的创造性、连贯性和情感参与度。对于大部分的故事生成评估基准都是包含故事和人工标注的评分，如 ROCStories<sup>[41]</sup>。

## 3.3 NLG 评估方法

在本部分将介绍关于自然语言生成的文本的各种评估方法，并分析这些方法的利弊，确定本文的有效评估方法。

最直观的判断评估指标是通过人类专家来进行评估，这种方式更符合文本生成的应用场景，也更加有效，但使用人类专家的时间等成本太高，无法作为大批量的文本评估指标。

传统的评估指标依赖于对模型输出和对应的参考内容之间的重叠程度来衡量文本的质量，比如 BLEU、ROUGE 等方法。这种方式虽然在一些简单的任务中有效，但当处理复杂任务时，评估的结果和人类实际判断的结果相关性较低，通过相对表面的匹配无法深入、可靠地评估文本。

伴随着深度学习的兴起，像 BERTScore、BARTScore 等基于模型的评估指标提出，用来评估文本生成的总体质量。这类方法可以在综合层面上评估生成文本的整体质量，可以综合评价文本生成的流畅度、连贯性等。虽然这类方法比传统的指标要好，但对于现在大语言模型生成的内容的评估性能不能令人满意，且这些方法的应用范围也十分有限。

最近一两年，由于大语言模型的兴起，在自然语言生成能力上有前所未有的增强，同时基于大语言模型的文本评估也逐渐兴起。由于大语言模型生成文本的质量已经远远超出之前传统的生成方法的质量，通过以上的方法指标无法有效描述，基于大语言模型的文本评估逐渐兴起。这类方法在评价尺度上更接近人类，有研究表明这类方法的评估结果与人类评估的相关性更高；在方法的应用范围上更广，几乎可

以评估自然语言生成的所有任务；但是，目前还没有系统的理论性的评估方法设置，对于评估的结构好坏收到很多设定的影响，在下面将调研介绍可能的大模型评估的研究方向：

## 3.4 NLG 评估的方向设定

### 3.4.1 基于大语言模型本身的评估指标

对于使用大语言模型完成评估的探索，有研究者直接考虑根据大语言模型本身的结构，基于语义嵌入或基于文本生成概率，考虑基于大语言模型本身的评估指标。

基于大型语言模型本身的评估指标主要分为两种类型：

1. 基于嵌入的指标：这类指标利用语言模型生成的语义嵌入来评估生成文本和参考文本之间的相似度。这种方法的优势在于能够捕捉到文本的深层语义信息，而不仅仅是表面的词汇匹配。例如，使用 BERTScore 这类基于预训练语言模型的嵌入来计算目标文本和参考文本之间的相似度。高度的相似性指示目标文本与相关要求更为一致，表明文本质量较高。

2. 基于概率的指标：这类指标通过考察语言模型生成某文本的条件概率来评估文本质量。基本思想是，如果一个模型能够以较高的概率生成某段文本，那么这段文本的质量被认为是较高的。这种方法的灵活性较高，可以适应不同的评估需求。

### 3.4.2 基于大语言模型提示词的评估指标（Prompting LLMs）

除了以上基于大语言、模型本身结构和输出方式的研究探索，还有的更多是对于大语言模型本身的特殊能力的利用。通过提示词已经被广泛认为可以提升大语言模型的性能，进而提高对应的评估性能。如何通过特定设计的提示词直接向现有的大语言模型命令以进行评估，成为该研究方向的重要问题。

对于这种评估方法的探究，充分利用了语言模型出色的指令理解和文本生成能力，使得评估工作可以通过精心构造的提示来完成。通过这种方式，评估者可以将评估任务、评估标准、输入内容等元素完整地表达在提示中，从而指导语言模型生成评估结果。在实际操作中，这包括了使用不同的评估方法如评分、比较、排序和布尔问答。评分通常采用定量的方法，如 1 到 5 或 1 到 100 的评分范围，用于评估生成文本的多个方面，例如对于对话生成的评估<sup>[42]</sup>。比较方法则是选择两个选项中较好的一个，而排序则涉及同时决定多个选项的顺序。布尔问答则要求模型对特定问题回答

“是”或“否”，用于评估文本的语法正确性或事实性等。此外，评估过程中的任务指令通常以任务描述的形式出现，清晰指示评估者应如何阅读或操作不同部分以完成注释任务。输入内容类型主要取决于评估标准，通常包括需要评估的目标文本和可能需要的其他内容，如源文档、参考文献和外部知识。评估标准则定义了在特定质量方面评估文本的好坏，例如流畅性、忠实性等。通过这种设置，基于大语言模型提示词的评估方法不仅展示了大型语言模型在自动文本评估方面的巨大潜力，还显示了如何将其与人类评估相结合，实现更为精细和可靠的评估结果。

### 3.4.3 基于大语言模型微调的评估指标（Fine-tuning LLMs）

大语言模型可以通过微调的方式提升其在专项任务上的表现。在这方面的研究通过使用特定的高质量数据对 LLM 进行微调，以增强其在自然语言生成（NLG）评估中的表现。这些数据通常涉及具体的任务场景、输入、目标文本和评估结果，反映了人类交互的需求。

微调过程广泛采用了如 LLaMA 这样的开源大型语言模型，并在一些研究中如 Prometheus 通过采取平衡数据分布、随机交换样本对比等策略，提高模型训练的效率和评估的稳健性<sup>[43]</sup>。评估方法方面，微调的大语言模型使用了详尽的评估准则，例如 AUTO-J 精心设计了针对 332 种不同任务的评估标准，并支持无参考和有参考的评估，这增加了其在实际应用中的灵活性和价值。

尽管微调大语言模型在 NLG 评估中显示了优异的性能，但仍存在一些局限性，如数据构建过程依赖于特定模型（如 GPT）进行关键评估标注，可能仍会带来与 GPT 相关的偏差。此外，重复的微调过程和高昂的计算成本也是实际应用中的挑战。

### 3.4.4 基于人机协同的评估指标（Human-LLM Collaborative Evaluation）

人机协同评估（Human-LLM Collaborative Evaluation）被描述为一种结合人类直觉、判断能力与大语言模型的分析处理能力的方法，目的是提供更加精确和全面的文本评估。这个方向的研究者大部分来自于人机交互领域。

这种评估方式利用大语言模型的高效处理能力与人类评估员的深刻洞察力，共同作用于评估流程，以提高评估的可靠性和深度。在这种协同评估中，大语言模型首先生成初步的评估结果和解释，随后人类评估员进行审查和细化，修正可能存在的错误并增强评估的准确性。这不仅减轻了人类在复杂评估任务中的负担，还利用了大语言模型在处理大量数据和保持一致性方面的优势。此外，这种方法也扩展到了更

广泛的评估任务，如测试和调试模型，以及对模型进行审核，确保其符合公平性和安全性的标准。通过这种方式，能够实现评估过程的高效性与经济性，并通过生成的解释提高评估的透明度，使得评估过程更加可信。然而，这种方法也存在挑战，包括对询问格式的敏感性，可能需要额外的工作来优化大语言模型的提示编写。此外，评估结果的可信度还依赖于能够准确判断 LLM 的可靠性，这仍然需要一定程度的人类监督。总之，基于人机协同评估方式通过充分利用人类与大语言模型各自的优势，为面对复杂或主观评估任务提供了一个高效且深入的解决方案。

### 3.5 本章小结

本章介绍了自然语言生成的传统任务，包括机器翻译、文本摘要、对话生成和故事生成。然后介绍了自然语言生成的评估方法，包括基于人类专家的评估、传统的评估指标。最后，介绍了基于大语言模型的评估方法，包括基于大语言模型本身的评估指标、基于大语言模型提示词的评估指标、基于大语言模型微调的评估指标和基于人机协同的评估指标。

## 第四章 简单文本生成方法探究

### 4.1 本章引言

在本章及第五章内容，将正式设计并探究文本生成内容。本章内容将从简单的文本生成开始，介绍基本的生成方法和评估方式；第五章内容则将对文本生成主题进行进一步的探索，探究对于长文本、风格化文本、实时文本的内容生成方法。

在本部分，将首先介绍本文使用的文本生成的基本设定和相应的方法。本文将重点探究自然文本生成的文章质量。关于文章质量的分析，在大语言模型兴起前研究是不够深入的，原因可以从两方面分析：一是之前文本生成的水平较差，对这种文本的质量评估是没有必要的；二可以认为之前对于文本生成的评估指标也是停留在表面，无法使用自动化的方法对文本进行和人类高相关性的文章质量评估。

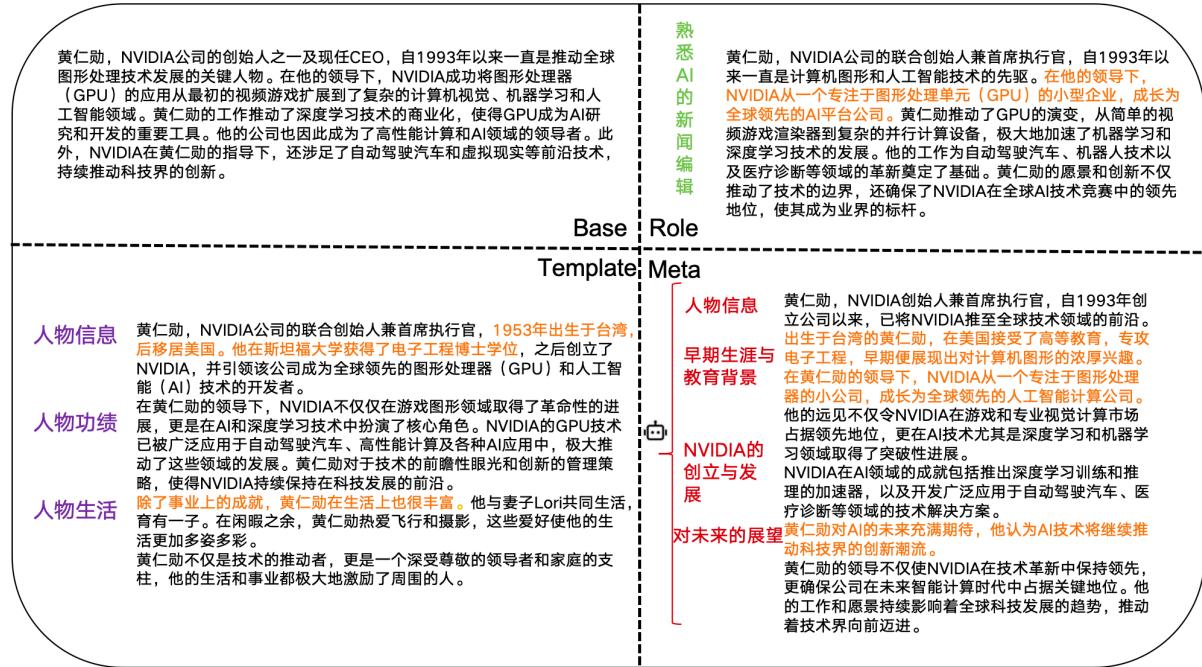
伴随着大语言模型的兴起，以上的两个文本质量分析研究不深入的原因都在一定程度上被解决<sup>[6]</sup>。因此，对于文本质量这一个研究领域，在无论文本生成还是对应的评估部分都有了十足的进步。在本次探究中，将深入研究该领域，完成完整的生成到评估的生成过程，促进文本生成的领域向自动化、高质量化发展。

在第四章内容中，将介绍基于大语言模型的文本生成方法，并对其进行基本实验，并使用相应的评估方法加以评价和简要分析。具体的实验结果在第七章内容展示并分析。

### 4.2 文本基本生成方法

在本部分，将介绍实验使用的生成方法。在本文中，主要使用的是 OpenAI 推出的 GPT 系列的商用大语言模型，通过调用 API 的方式调用相应的服务。

在本部分实验中，选择生成新闻作为文本的生成任务，本部分选择的生成的新闻具体内容源自网络中实时产生的热点新闻内容。对于生成文本的质量评估和分析，将主要采用基于大语言模型的评估方式（具体评估方式见第六章内容）。下面介绍新闻生成任务的实验设置，可以通过图4-2来了解实验的基本流程。



图中分别展示了四种生成方法生成的黃仁勛人物介绍文本，左上部分为最基本方法的生成结果，剩余部分分别为三种改进方法的生成结果，橙色部分为相应的效果提升部分。

图 4-1 简单文本生成方法生成文本示意图

Figure 4-1 Schematic diagram of the text generated by the simple text generation method

#### 4.2.1 文本直接生成

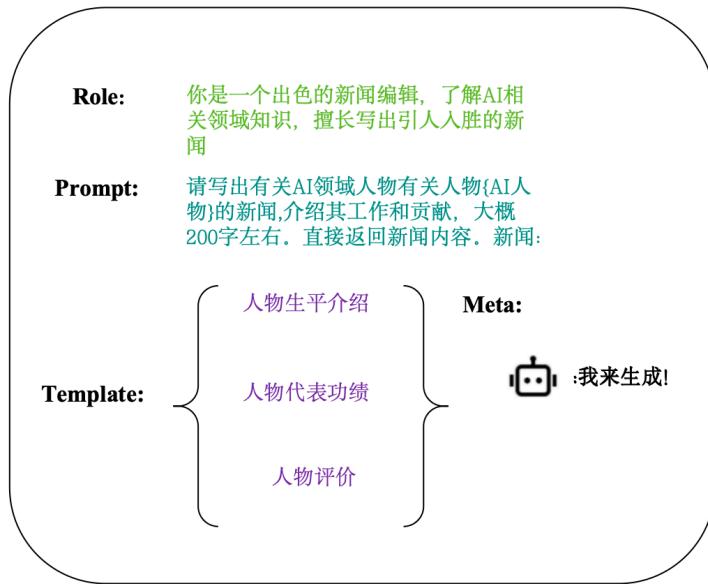
在目前的商用大语言模型中，绝大部分提供了方便的使用接口。本方法是最基本的生成方法。直接调用 OpenAI 中的模型列表的模型。本方式通过添加提示词 (prompt) 的方式，一次性生成相应要求的新闻文本。

#### 4.2.2 基于角色的文本直接生成

相比最基础的文本生成方法，本方法加入了大模型的角色 (role)，通过在给定具体提示词，给定生成任务前，加入对大模型的角色指令，可以激发大语言模型在特定领域任务的知识，从而可以得到更加优质的输出。

#### 4.2.3 基于模版的文本直接生成

本方法是在基于角色的文本生成方法的基础上，加入了模版的生成方式。通过给定模版，可以更好地控制生成文本的内容和格式，从而可以得到更加符合要求的文



图中蓝色字体代表文本直接生成方法，在此基础上，绿色字体代表基于角色的文本生成方法；紫色部分代表基于模版的文本生成方法；机器人部分代表基于 Meta 的文本生成方法。

图 4-2 文本生成方法示意图

Figure 4-2 Schematic diagram of the text generation method

本。

#### 4.2.4 基于 Meta 的文本阶段生成

以上的三种方法都是调用大语言模型直接生成文本对象，这种方式与实际人在创作的过程不一致。在人类创作文章过程中，首先会产生创作文章的大纲，然后再依据大纲创作整篇文章；本方法借鉴了这种人类创作的过程，首先生成新闻的整体大纲，然后根据已有的大纲，创作实际的文章内容。

### 4.3 文章生成实验

在本章节，为了验证大语言模型生成文章内容各方法的效果，选择生成人物的介绍文本，作为生成文章的内容，具体生成效果可以参考图4-1。选择该任务的原因是，人物介绍文本生成是一种应用范围广泛的任务，且由于介绍方式多样，难以找到合适的传统评估指标来评估生成文本的质量。关于使用数据集，本实验选择 2023 年《时代》周刊评选的 100 大人工智能人物作为生成内容列表，共有 87 组人物样本。在实

验中，将使用以上的四种生成方法，分别生成对应的人物介绍文本，通过对比生成文本的质量，可以验证各种生成方法的效果。

对于生成文本的质量评估，除了上基于大语言模型的评分方法，在本实验中还加入了基于比较的评估方法。具体的评估方法是，将基于样本中的人物，在网络上搜索并找到相关的人物介绍新闻，在检测质量通过之后，将其作为参考，交给大语言模型和大模型生成的文本进行比较。最终根据胜利率（WinRate）来对文本生成方法进行评估。

具体的实验结果和分析将在第七章内容中详细介绍。

#### 4.4 本章小结

本章介绍了基于大语言模型的基本生成方法，在文中共介绍四种，在此基础上，进行了人物介绍文本生成实验来测试这四种方法的生成效果。

## 第五章 特定文本生成方法

### 5.1 本章引言

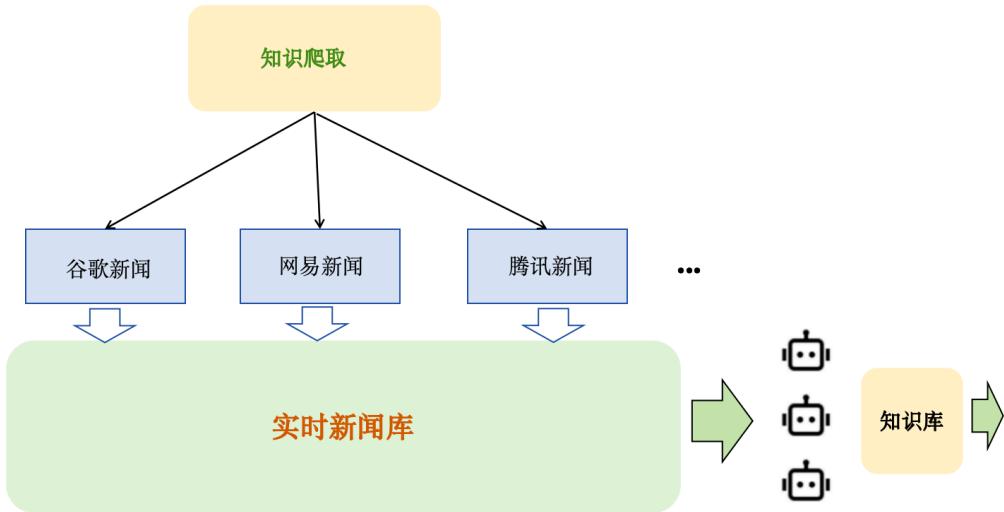
在本章中，将对第四章的文章生成及评估方法进行进一步延伸，对更加一般的文章生成进行探究。在本章将探讨几个文本生成的情况，包括实时的文章生成，长文章的生成，以及风格化的文本生成，具体将在 5.2, 5.3 以及 5.4 部分中具体展示。

### 5.2 实时文章生成

在本部分，将对具有实时性的新闻生成进行进一步探讨。目前大语言模型可以实现对文章的有创造力的输出，然而，受限于训练时的语料库，大语言模型无法获取实时的新闻等文本信息，不仅极大地限制了其应用范围，同时也会产生幻觉等一系列问题。

在实际的大语言模型应用中，例如微软的必应等产品已经支持了实时内容的检索和知识提取。虽然获取并提取实时内容的技术已经被使用在商业产品中，但是如何将这些内容整合到大语言模型中，仍然是一个值得探讨的问题，本部分将尝试复现这一过程，作为已有文本生成方法的补充。下面将对具体的实时文章生成方法进行介绍。

要实现文章的实时生成，最关键的部分就是实时内容的获取和对应的知识提取。现有的支持实时信息检索的大模型应用，都有对应的一套搜索体系。在简单的实验环境下中，可以通过爬虫技术获取实时的新闻内容，然后通过大语言模型进行知识提取，最终生成实时的文章内容。在本实验中，将使用爬虫爬去网易新闻，腾讯新闻等新闻网站的实时新闻内容，同时，借用 *G\_news* 库，获取谷歌新闻的实时新闻内容，最终这些新闻会汇总成一个实时新闻的数据库。对于这些实时信息，还需要从这些内容中提取出有效的新闻信息，这一部分将使用分层知识提取的方法，通过多层知识的提取蒸馏，最终将有效的内容提取出来，提取过程如图5-1所示。



图中左侧部分代表使用爬虫及 Gnews 库获取实时新闻内容，右侧部分代表实时新闻知识提取过程。

图 5-1 实时信息知识提取图

**Figure 5-1 Real-time information knowledge extraction diagram**

### 5.3 长文章生成

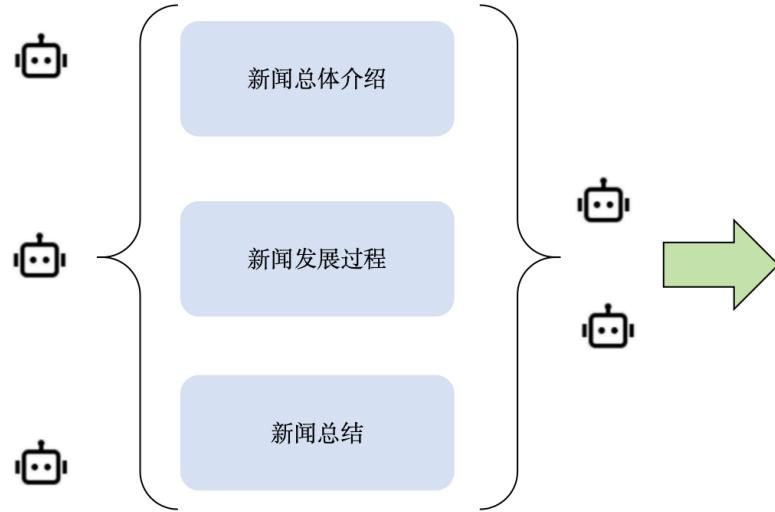
在本部分，将对长文章文本的生成进行进一步探讨。目前大语言模型对于字符较大的输入或输出，效果会比较短字符的结果要差；对于现有的商用 API，对于输入的字符有最大字符限制，其输出字符数也有一定限制。在这种背景下，使用第四章中的方法生成文本是不可行的，在本部分将对长文章文本的生成进行探讨。

首先介绍本实验的场景设置：本实验的任务场景是根据现有热门新闻生成一篇长文章内容，关于现有热门新闻的知识提取，已经在实时文本生成部分中实现，在本实验中，将使用这些知识作为输入，生成一篇长文章内容。

关于长文章的生成，在本实验中采取两种基于大语言模型的生成方法，分别是基于模版的生成方法和基于 Function calling 的 meta 生成方法。具体的生成方法可以通过图5-2和图5-3展示。

#### 5.3.1 基于模版的生成方法

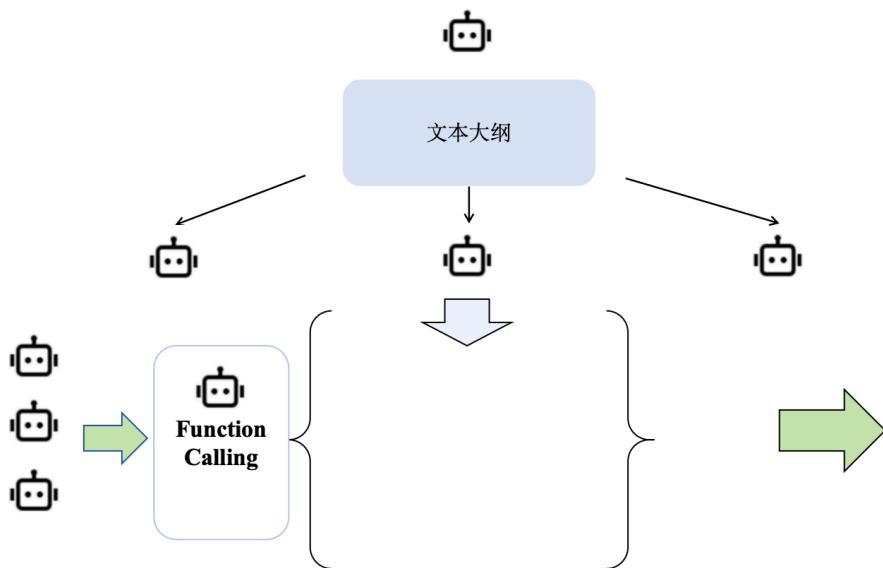
本方法是在第四章的基础上进行延伸，对于长文章的生成，需要更多的内容和结构，且在本实验场景故在实际实验中，单次生成所有的文本是不可行的。故对于长文



图中中心部分为已有的生成模版，两边的机器人代表多个大模型代理，通过填补对应内容完成长文章。

图 5-2 基于模版的生成方法图

Figure 5-2 Template-based generation method diagram



图中上侧部分大模型生成文本大纲，通过分层对细节部分进行补充；下侧部分，通过 Function calling 部分，调用对应的代理函数，完成长文章的生成。

图 5-3 基于 Function calling 的 Meta 生成方法

Figure 5-3 Meta generation method based on Function calling

本，模版将以 json 格式存储，每次生成一部分文本，然后将生成的文本与模版进行合并，最终生成长文章。在这种情况下，不再是单次生成全部文本，而是分批生成文本，最终合并成长文章。

### 5.3.2 基于 Function calling 的 Meta 生成方法

本方法是第四章基于大模型的 Meta 方法的延伸，对于长文章的内容生成，基于 Meta 的方法符合长文章人类的写作习惯和思维需求，然而，在实际实验中发现，基于 Meta 的方法直接作用到长文章的效果并不理想，原因在于 Meta 方法生成的文本内容大纲还是比较简单且宽泛的，如果基于这种模版直接生成长文章，会导致生成的文本内容过于简单，缺乏细节和深度。在这种情况下，考虑到 Meta 方法的优势和 GPT 的调用函数能力，在实验中采用了基于 Function calling 的 Meta 生成方法。

具体的方法是，将 Meta 部分分层处理，首先先生成大纲，然后根据大纲的内容，分部分生成更加具体的内容要求；基于细分的内容要求，调用 Function calling 的方法，由大模型自动判别需要生成的内容和方法，最终生成长文章。

### 5.3.3 实验设置

关于实时文章和长文章的生成，本次实验将综合进行测试，具体的实验设置如下：

本个实验的实验设置比较新奇，考虑到实时文章和长文章生成的特殊性，本次实验将根据实时的新闻内容，使用不同生成方式，生成实时长文章内容。对于已经生成的实时长文章，会将其内容推送到微信公众号，以供用户阅读。在实际实验中，将使用基于模版的生成方法和基于 Function calling 的 meta 生成方法，通过不同长文章内容的阅读量，来对比两种生成方法的效果。同时，也通过这种方式，实现实时长文章生成的实际应用。

## 5.4 风格化文本生成

对于风格化文本的生成，这一特殊性的内容，其生成难度和评估难度相比生成一般的文本要更大。

首先明确本文讨论的风格化文本中的风格是相对抽象的概念，故在实际实验的过程中，认为风格是某个作者的写作风格，通过其文学作品体现出来。接下来，将对具体的风格化文本实验进行介绍：

对于本实验，需要选择文本的生成和评估方法。对于文本内容，利用基于大语言模型的生成方法是可以实现的，但是对于文本风格生成后的评估方法，需要进一步的研究。传统的评估指标在这一问题上不再适用，而使用基于大模型的评估方法，在实际使用中发现，在没有参考文本的情况下，对于风格化程度的评判不敏感。

基于以上的分析，本文采取以下的实验设置：

首先，选取作者的文学作品片段作为参考文本，用来作为最后的评估比较。对于这些文学作品片段，需要经过去风格单元，获得去风格化后的文本部分，这些去风格化的文本部分作为风格化文本生成需要的原始文本。这些原始文本将作为大语言模型的输入，经过多种风格化生成方法，最终获得一系列生成的风格化文本。在生成风格化文本后，通过比较生成文本和参考文本的相似度，来评估生成文本的风格化程度；当然，除了考虑文本的风格化程度外，还需要将原始文本和生成文本进行比较，来评估生成文本的内容保留程度。

以上是基本的实验设置，下面将介绍具体实验采用的数据集，生成方法，以及评估方法。

#### 5.4.1 实验设置

---

##### 《阿Q正传》原文：

那破布衫是大半做了少奶奶八月间生下来的孩子的衬尿布，那小半破烂的便都做了吴妈的鞋底。在未庄再看见阿Q出现的时候，是刚过了这年的中秋。人们都惊异，说是阿Q回来了，于是又回上去想道，他先前那里去了呢？阿Q前几回的上城，大抵早就兴高采烈的对人说，但这一次却并不，所以也没有一个人留心到。

---

##### 《阿Q正传》去风格化文本：

那件破旧的衬衫，大部分被剪成了给少奶奶八月份生的孩子用的尿布，剩下的破布则变成了吴妈的鞋垫。等到阿Q再次出现在未庄的时候，已经是那年中秋过后了。大家都很惊讶，说阿Q回来了。这让大家又开始好奇，他之前去了哪里？以往阿Q每次去城里，总是会兴高采烈地告诉大家，但这次他却没有提起，因此也没人特别注意到他的去向。

---

图 5-4 去风格化文本和与原文效果对比图

Figure 5-4 Comparison of de-stylized text and effect with original text

本实验的数据集选取的是两位风格迥异的文学作家的风格化内容，分别是鲁迅和刘慈欣。在使用文本方面，使用的是鲁迅的《阿Q正传》和刘慈欣的《三体》作为参考文本。选择这两个文学作家及作品的原因是，两位作家的文学风格都比较显著，文学作品类型相差明显，具体的风格也有很大的差异，选择这两个文学作家的作品，可以更好地体现风格化文本生成的效果。

选取以上文学作品后，需要对文本做预处理，将原始的作品内容切分成若干段落，每个段落作为一个样本。考虑到成本等因素，在实际实验中，选择文章内容的片段作为数据集，最终划分鲁迅的《阿Q正传》的节选部分为120个样本，刘慈欣的《三体》的节选部分为25个样本。

在数据集准备好后，需要对数据集进行去风格化处理，在本实验中，采用的是基于大语言模型的去风格化方法，通过给定标准无风格内容，将原始文本内容去风格化。最终去风格化的结果可以通过图5-4展示。具体的实验流程如图5-5所示。

具体的结果分析将在第七章进行介绍。

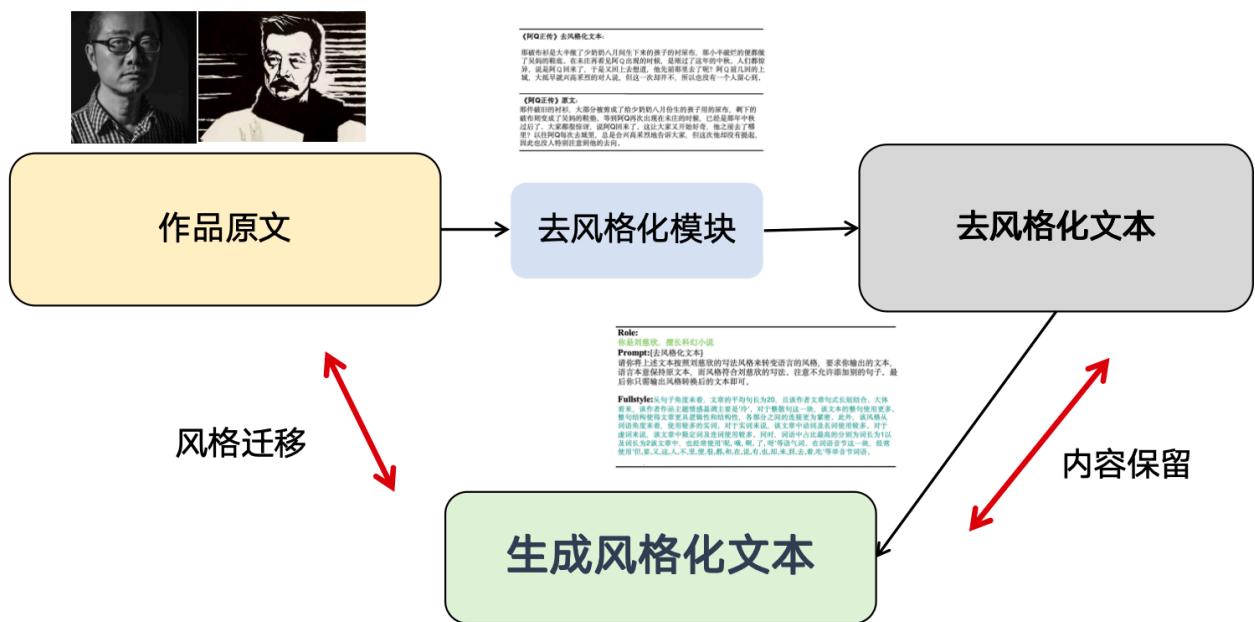


图 5-5 风格化文本生成实验流程图

Figure 5-5 Stylized Text Generation Experiment Flowchart

### 5.4.2 生成方法

在生成方法上，本实验采用了基于大语言模型的生成方法。在实际实验中，使用了以下几种生成方法，可以认为是第四章文本生成方法的延伸，可以通过图5-6展示。这里使用《阿Q正传》的部分片段作为示例，展示其生成效果，如图5-7所示。

---

**Role:**

你是刘慈欣，擅长科幻小说

**Prompt:{去风格化文本}**

请你将上述文本按照刘慈欣的写法风格来转变语言的风格，要求你输出的文本，语言本意保持原文本，而风格符合刘慈欣的写法。注意不允许添加别的句子。最后你只需输出风格转换后的文本即可。

**Fullstyle:**从句子角度来看，文章的平均句长为20，且该作者文章句式长短结合。大体看来，该作者作品主题情感基调主要是'冷'。对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。此外，该风格从词语角度来看，使用较多的实词。对于实词来说，该文章中动词及名词使用较多。对于虚词来说，该文章中限定词及连词使用较多。同时，词语中占比最高的分别为词长为1以及词长为2该文章中，也经常使用'呢，哦，啊，了，呀'等语气词。在语音节这一块，经常使用'但，要，又，这，人，不，里，便，很，都，和，在，说，有，也，却，来，到，去，着，吃'等单音节词语。

---

图中绿色部分为基于角色的生成方法，蓝色部分为基于文法的生成方法的指令的一部分，对于基于阅读的生成方法，由于篇幅原因，未在图中展示。

图 5-6 风格化文本生成方法图

Figure 5-6 Stylized Text Generation Methodology Diagram

#### 5.4.2.1 基于阅读的生成方法

在本实验中，基于阅读的生成方法是通过添加“示例”的方式，激活大语言模型的风格化生成能力。在具体实践上，将添加鲁迅和刘慈欣的原文（即《阿Q正传》和《三体》的原文非测试内容）作为示例，作为大语言模型风格化文本的范例。

#### 5.4.2.2 基于角色的生成方法

在本实验中，基于角色的生成方法是第四章基于角色的文本直接生成的延伸，在调用过程中添加“role”，激活大语言模型在对于特定角色的特定风格的相关知识。在实际实验中，将使用这种方法生成鲁迅和刘慈欣的风格化文本。



图中上半部分分别为文本原文和去风格化文本；下半部分为三种方法风格化生成后的文本，橙色部分为风格化显著的内容。

图 5-7 风格化文本生成结果示意图  
Figure 5-7 Graph of Stylized text generation result

#### 5.4.2.3 基于文法的生成方法

在本实验中，将根据 Tao 的方法，生成基于文法的风格描述，这种方式可以更好地从实际用词上还原、生成对应作者的写作风格<sup>[44]</sup>。下图为对于该方法的一些实际描述示例：

#### 5.4.3 评估方法

在评估方法上，本实验采用了 BLEU 和 BertScore 两种评估方法。考虑到在实际实验中，有对应参考文本，故使用传统评估方法进行评估。以下是评估方法的简单介绍：

- **BLEU:** BLEU (Bilingual Evaluation Understudy) 是一种评估机器翻译质量的方法，通过计算机器翻译输出与人类翻译之间的相似度来工作。它主要关注 n-gram 的匹配程度，其中 n-gram 是连续 n 个词的序列<sup>[45]</sup>。
- **BertScore:** BERTScore 利用预训练的语言模型（如 BERT）来评估文本生成任务（如翻译或文本摘要）的质量。它计算的是生成文本和参考文本中的词语之间的相似度，使用的是 BERT 嵌入向量的余弦相似度<sup>[46]</sup>。

## 5.5 本章小结

本章主要对特定文本生成方法进行了探讨，包括实时文章生成，长文章生成，以及风格化文本生成，提出了特定任务的生成方法。通过不同的文本生成实验，了解大语言模型的生成能力和对应的生成方法性能。

## 第六章 生成文本质量评估方法及验证

### 6.1 本章引言

在介绍文本生成的方法后，将对生成的文本进行评估，本部分将介绍一种基于大语言模型的文本评估方法，通过大语言模型对生成的文本进行评估，最终通过两个实验验证评估方法的有效性。

对于无论是人创造的还是模型生成的文本，要评估文本的质量，通过专家评估是比较推崇的做法。然而，对于大量的模型生成的文章，使用专家进行评估会耗费大量时间和金钱成本。因此，有一些评估指标被提出，用来作为评估文本在特定文本生成领域的质量。

### 6.2 生成文本评估方法

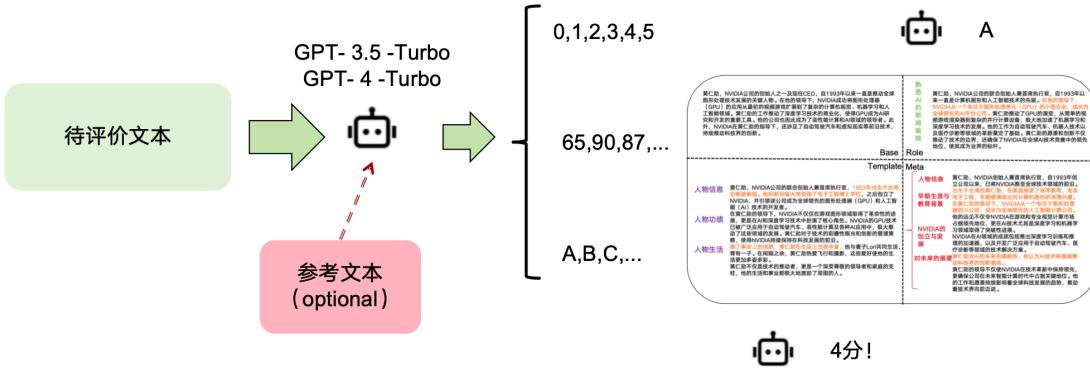
对于本文重点探讨的文本质量这一任务，就作者的调研，目前没有完备的评价指标或体系对生成的文本进行评估。在这个背景下，考虑到大语言模型在自然语言领域的强大能力，大语言模型的文本评估作为一种可能的方式被提了出来。在本文中，将利用大语言模型的文本分析能力，通过大语言模型评估，对生成的文本进行打分来评估对应文本的质量。在本文中评估方法是基于角色的文本生成方法的变种，大致流程如图6-1所示。

在给出了评估方法后，需要验证其可行性，在接下来将使用两个自然语言生成的实验来验证以上的评估方法。

### 6.3 实验一：故事续写

#### 6.3.1 实验设定

本实验是对于基于大语言模型的评估方法的验证，实验的具体任务是故事续写，即给定一个故事的开头，模型需要续写出一个完整的故事，基本。在这个实验中，将使用 **OpenMEVA-ROC** 作为实验的数据集，该数据集中包含了 200 个故事的开头，以及对应的专家续写的故事作为标准参考，每个故事都有 5 个机器模型生成对应的续写故事（分别为 **Seq2Seq** 模型，**Plan&Write** 模型，**Fusion** 模型，**GPT-2** 模型，以及



图中左边部分为基于角色的文本生成方法的变种评估模块，右边部分为示例。

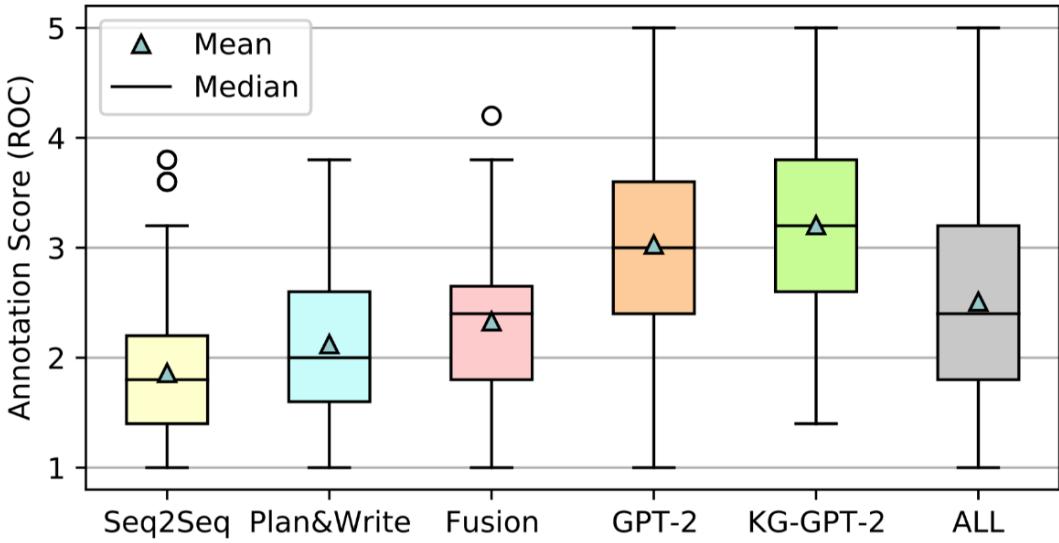
**图 6-1 基于大语言模型的文本评估示意图**  
**Figure 6-1 Graph of text evaluation based on large language model**

KnowledGe-enhanced GPT-2 模型），共计有 1000 个故事续写样例。对于每个故事续写样例，都有多个评估指标的评分和专家评分，评分情况可通过图6-2体现，通过计算各个评估指标的评分和专家评分的相关系数，可以验证评估方法的有效性。

下面是评估指标的简单介绍：

- **BLEU:** BLEU 是一种广泛使用的机器翻译质量评估指标<sup>[45]</sup>。它通过比较机器翻译输出与一个或多个人工翻译的参考文本来工作。BLEU 评分通过计算短语匹配的精确度（通常是 n-gram 匹配）来评估翻译的质量。
- **PPL:** PPL，或困惑度，是用来评估语言模型性能的指标<sup>[47]</sup>。它衡量模型预测样本的不确定性，困惑度越低表示模型对语言的掌握越好，即模型预测下一个词的能力越强。
- **R<sub>u</sub>-BERT:** RUBER-BERT 是一个用于评估对话系统生成文本质量的指标<sup>[48]</sup>。RUBER-BERT 是 RUBER 指标的一种变种，使用 BERT (Bidirectional Encoder Representations from Transformers) 模型来增强其评估能力<sup>[13,49]</sup>。
- **UNION:** UNION 指标主要应用于故事生成和其他创意写作领域，不仅评估文本的语法正确性和流畅性，还特别注重文本的常识性和逻辑连贯性<sup>[50]</sup>。

以上的几个评估指标都是在自然语言生成领域中比较常见的评估指标，通过计算这些指标的评分和专家评分的相关系数，在本次实验中，将加入大语言模型评估的评分（包括 GPT-3.5-Turbo 和 GPT-4-Turbo），通过对比相关系数的大小，可以验证大



图中分别是 5 种模型生成的文本的评分分布，横坐标为评分，纵坐标为人工标注的评分分布。

图 6-2 ROC 故事人工标注得分图

Figure 6-2 Graph of manual labeling scores for ROC stories

语言模型评估的有效性。

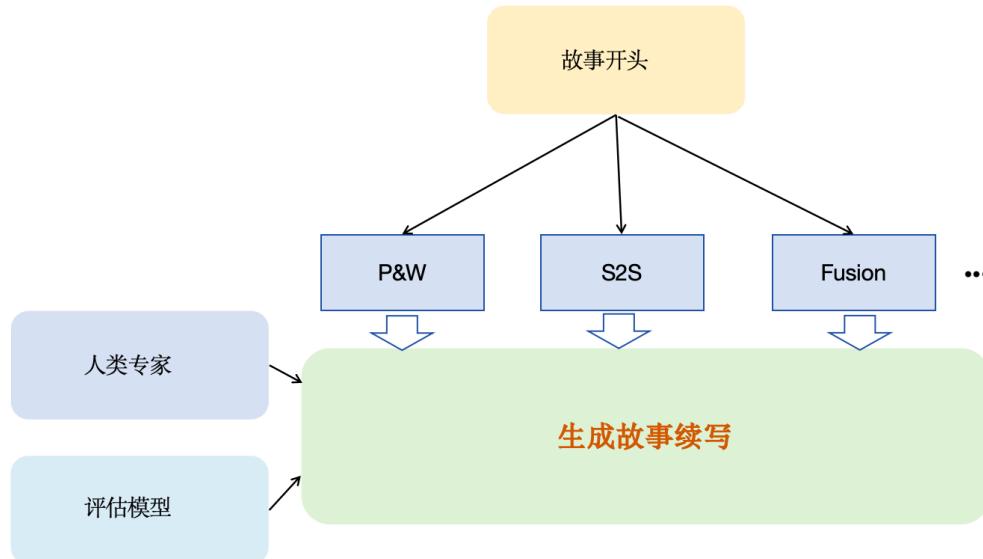
对于人类评分，评分范围为 0-5 的连续整数，每个样例都有 5 个人类评分，最终的评分为 5 个人类评分的平均值；对于传统评估指标评分，将由原始值映射到 0-5 的整数；对于基于大语言模型的评分，评分范围为 0-5 的连续整数。关于相关系数的计算，将使用皮尔逊相关系数<sup>[51]</sup>。

### 6.3.2 实验结果和分析

表6-1是最终各评估指标的相关系数结果，图6-4是相关系数的可视化结果。

从表格整体上看，包括 GPT-3.5-Turbo 和 GPT-4-Turbo 的基于大语言模型的评分在 5 种模型生成的文本评估中，有 4 种与人类专家评分相关系数最高。这说明基于大语言模型的评估方法在故事续写任务上有着较好的效果。

对于理论上相对效果应该更好的 GPT-4-Turbo 模型，在基于 Fusion 生成的故事续写文本上的相关系数处于五种评估指标的倒数第二，相关系数甚至低于 GPT-3.5-Turbo 的评分，对于这一奇怪的现象，原因分析为 Fusion 模型生成的文本在质量上差异较大，而 GPT-4-Turbo 评分相对稳定，导致相关系数较低。



图中右侧部分为数据集中生成续写故事的部分，人类专家和评估模型通过续写故事进行评分，得到结果后获得相关系数。

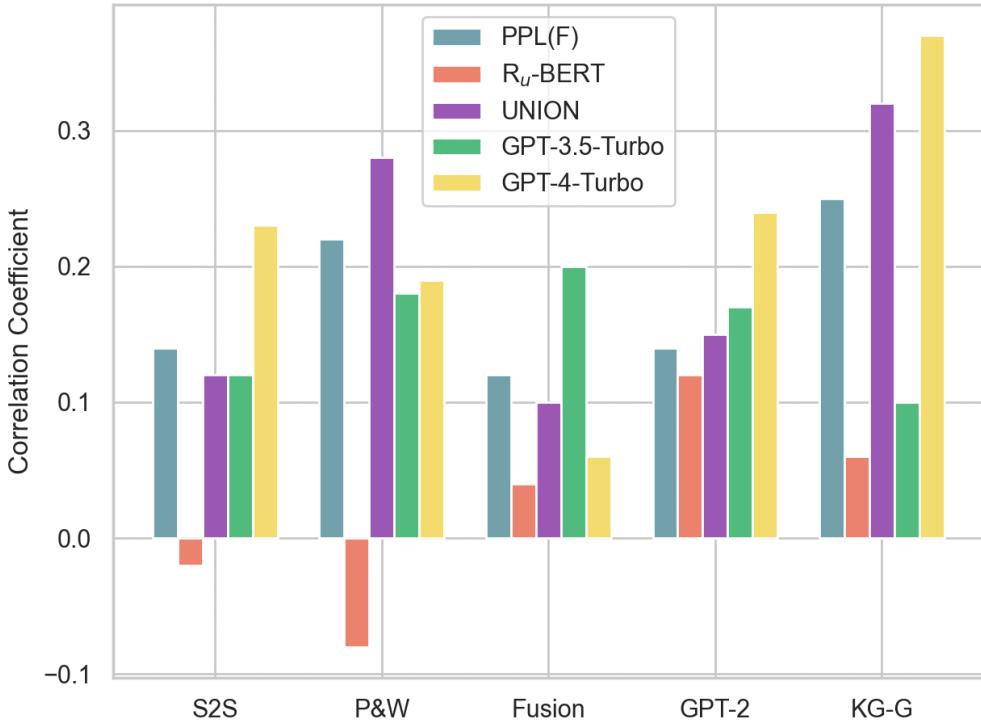
图 6-3 故事续写实验示意图

Figure 6-3 Schematic diagram of the story continuation experiment

表 6-1 ROC 故事相关系数实验数据

Table 6-1 ROC story correlation coefficient experimental data

Metrics	S2S	P&W	Fusion	GPT-2	KG-G
PPL(F)	0.14	<b>0.22</b>	0.12	0.14	0.25
R <sub>u</sub> -BERT	-0.02	-0.08	0.04	0.12	0.06
UNION	0.12	0.28	0.10	0.15	0.32
GPT-3.5-Turbo	0.12	0.18	<b>0.20</b>	0.17	0.10
GPT-4-Turbo	<b>0.23</b>	0.19	0.06	<b>0.24</b>	<b>0.37</b>



图中不同颜色代表不同种评估指标的评分和人工评分的皮尔逊相关系数。

图 6-4 ROC 故事续写相关系数图

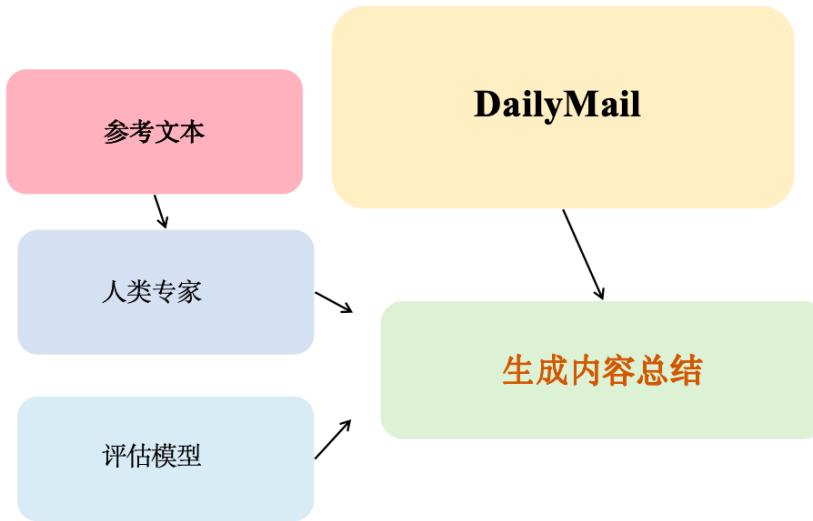
Figure 6-4 ROC story continuation correlation coefficient graph

## 6.4 实验二：文本总结

### 6.4.1 实验设定

本实验是对于基于大语言模型的评估方法的验证，实验的具体任务是文本总结，即给定一个文本，模型需要生成一个总结。在这个实验中，将使用 **CNN/DailyMail** 语料库<sup>[52]</sup>。这个数据集原本用于问答任务，但有研究者在文本总结任务中也使用了该数据集<sup>[53]</sup>。该数据集有一系列新闻文章和相对应的人工创建的要点摘要组成。在具体的评估方面，R. Fabbri 将该数据集的样例文章用 16 种语言模型借助参考文本生成了对应的总结，并使用人工标记了每个生成总结的分数<sup>[38]</sup>，这个实验作为本次验证实验的基础。在本次实验中，将使用 Fabbri 的模型总结和对应分数，使用 DailyMail 部分的新闻部分，共 87 个样例，经过 16 个模型生成共 1392 个文本总结，通过计算各个评估指标的评分和专家评分的相关系数。

在这个任务中，将总结内容的分析分为了四个维度——连贯性 (Coherence)，一致



图中右侧部分为数据集中通过 16 中模型生成的内容总结，左侧人类专家和除 GPT-3.5-Turbo 和 GPT-4-Turbo 外的模型结合参考文本进行评分，得到结果后获得相关系数。

**图 6-5 文本总结实验示意图**  
**Figure 6-5 Schematic diagram of the text summary experiment**

性 (Consistency)，流畅性 (Fluency)，以及相关性 (Relevance)。对于每个评估指标，都将对文本总结的以上四个维度进行评分。下面是评价维度和评估指标的简单介绍<sup>[38,54]</sup>：

- **连贯性 (Coherence)**: 连贯性是指文本总结的内容是否连贯，偏向所有句子的整体质量。
- **一致性 (Consistency)**: 一致性是指文本总结的内容是否与原文一致，用于评价总结和总结来源之间的事实一致性。
- **流畅性 (Fluency)**: 流畅性是指文本总结的内容是否流畅，偏向单个句子的质量。
- **相关性 (Relevance)**: 相关性是指文本总结的内容是否与原文相关，用于评价原文重要部分的使用程度。
- **ROUGE-1**: ROUGE-1 是一种广泛使用的文本总结质量评估指标<sup>[55]</sup>。它通过比较生成的总结和参考总结的单个词的重叠来评估总结的质量<sup>[55]</sup>。
- **ROUGE-2**: ROUGE-2 是 ROUGE-1 的扩展，通过比较生成的总结和参考总结的两个词的重叠来评估总结的质量<sup>[55]</sup>。
- **BertScore**: BertScore 是一种基于 BERT 模型的文本质量评估指标<sup>[46]</sup>。它通过比较生成的总结和参考总结的 BERT 模型的输出来评估总结的质量<sup>[46]</sup>。

- **BLEU**: BLEU 是一种广泛使用的机器翻译质量评估指标<sup>[45]</sup>。它通过比较机器翻译输出与一个或多个手工翻译的参考文本来工作。BLEU 评分通过计算短语匹配的精确度（通常是 n-gram 匹配）来评估翻译的质量<sup>[45]</sup>。

最终评估指标在文本总结任务上的可靠性可以通过计算这些指标在以上四个维度的评分和专家评分的相关系数体现，在本次实验中，将加入大语言模型评估的评分（包括 GPT-3.5-Turbo 和 GPT-4-Turbo），通过对比相关系数的大小，可以验证大语言模型评估的有效性。

对于人类评分，评分范围为 0-5 的连续整数，每个样例都有 5 个人类评分，最终的评分为 5 个人类评分的平均值；对于传统评估指标评分，将由原始值映射到 0-5 的整数；对于基于大语言模型的评分，评分范围为 0-5 的连续整数。关于相关系数的计算，将使用 kendall tau 相关系数<sup>[51]</sup>。

**表 6-2 DailyMail 相关系数实验数据**  
**Table 6-2 DailyMail correlation coefficient experimental data**

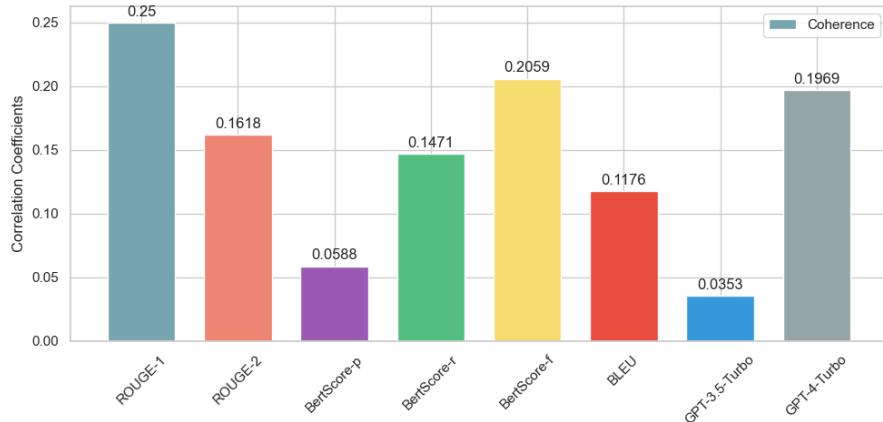
Metrics	Coherence	Consistency	Fluency	Relevance
ROUGE-1	<b>0.2500</b>	0.5294	<b>0.5240</b>	<b>0.4118</b>
ROUGE-2	0.1618	<b>0.5882</b>	0.4797	0.2941
BertScore-p	0.0588	-0.1912	0.0074	0.1618
BertScore-r	0.1471	<b>0.6618</b>	0.4945	<b>0.3088</b>
BertScore-f	<b>0.2059</b>	0.0441	0.2435	<b>0.4265</b>
BLEU	0.1176	0.0735	0.3321	0.2206
GPT-3.5-Turbo	0.0353	<b>0.5912</b>	<b>0.5743</b>	-0.0717
GPT-4-Turbo	<b>0.1969</b>	0.5727	<b>0.5178</b>	0.2450

#### 6.4.2 实验结果和分析

表6-2是最终各评估指标的相关系数结果，图6-6，图6-7，图6-8，图6-9是相关系数的可视化结果。

从表格整体上看，关于文本总结的四个维度的评估指标中，基于大语言模型的评估方法与人类专家的评分相关系数并没有明显优势。

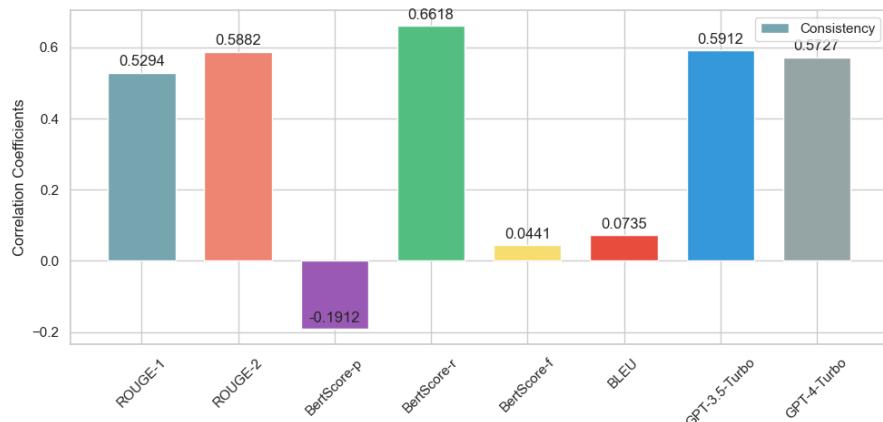
但事实上，除了基于大模型的评估方式外，其他的评估指标对于每个评估样本，都有 11 个额外的参考摘要，这些参考摘要在一定程度上加强了以上指标的可靠程度，且 GPT 系列的最终相关程度也是较高的。因此，可以认为基于大语言模型的评估方



图中不同颜色代表不同评估指标关于连贯性在 DailyMail 数据集上的评分和人工评分的 kendall tau 相关系数，纵坐标为相关系数。

**图 6-6 DailyMail Coherence 相关系数图**

**Figure 6-6 DailyMail Coherence Correlation Coefficient Chart**

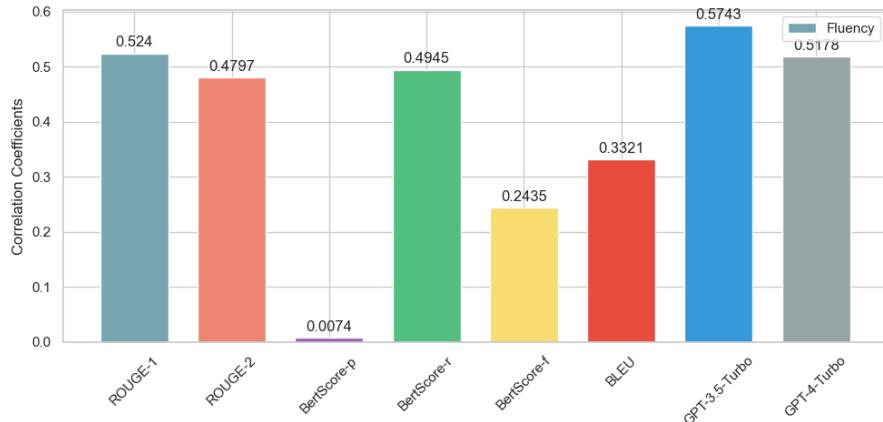


图中不同颜色代表不同评估指标关于一致性在 DailyMail 数据集上的评分和人工评分的 kendall tau 相关系数，纵坐标为相关系数。

**图 6-7 DailyMail Consistency 相关系数图**

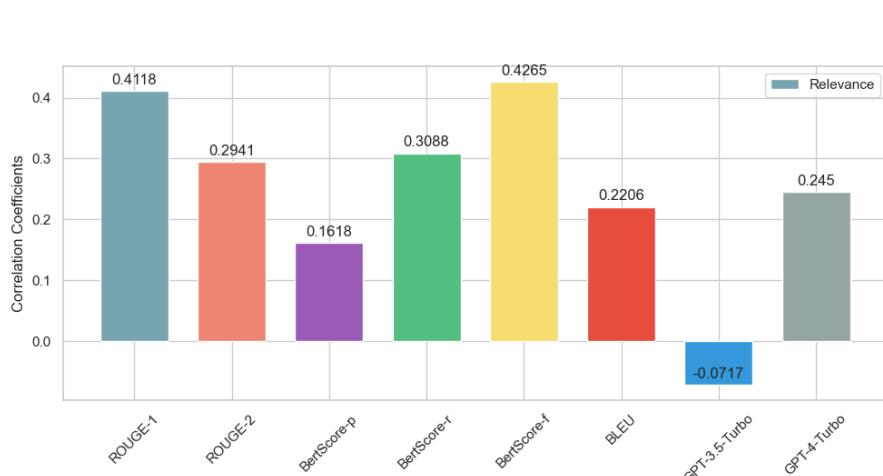
**Figure 6-7 DailyMail Consistency Correlation Coefficient Chart**

法在文本总结任务上也是有效的。



图中不同颜色代表不同评估指标关于流畅性在 DailyMail 数据集上的评分和人工评分的 kendall tau 相关系数，纵坐标为相关系数。

**图 6-8 DailyMail Fluency Correlation Coefficient Chart**



图中不同颜色代表不同评估指标关于相关性在 DailyMail 数据集上的评分和人工评分的 kendall tau 相关系数，纵坐标为相关系数。

**图 6-9 DailyMail Relevance Correlation Coefficient Chart**

从四个维度分开看，可以看到在连贯性和流畅性两个维度上，基于大语言模型的评估方法与人类专家的评分相关系数较高，而在一致性和相关性两个维度上，基于大语言模型的评估方法与人类专家的评分相关系数较低。分析原因同实验一，大语言模

型的评分相对稳定，导致相关系数较低。

通过以上两个实验，可以验证基于大语言模型的评估方法在文本生成任务上的有效性。

## 6.5 本章小结

本章介绍了基于大语言模型的文本评估方法，通过大语言模型对生成的文本进行评估，最终通过两个实验，可以验证基于大语言模型的评估方法在文本生成任务上的有效性。

## 第七章 实验结果分析

### 7.1 基本文本生成结果和分析

下面是对 2023 年《时代》周刊评选的 100 大人工智能人物生成人物介绍新闻的实验结果：

表 7-1 普通文本打分数据实验数据

Table 7-1 Experimental data

Method	Score	WinRate(%)
Base	67.8	40.23
Role	83.9	49.43
Template	<b>87.3</b>	<b>58.62</b>
Meta	85.9	57.47

从表格中可以看出，从平均得分来看，基于模板的生成方法在评分上有着比较高的得分，这说明基于模板的生成方法在生成普通文本的任务上有较好的效果。但在平均得分上，除了 Base 方法外，其他方法的平均得分都在 80 分以上，且三种方法相差分数不超过 4 分，这说明三种方法在生成普通文本的任务上有相对好的效果。

从胜率来看，基于模版的生成方法和基于 Meta 的生成方法在胜率上有着较好的表现，两者的胜率相差不大。而对于 Base 方法和基于角色的生成方法，两种方法的胜率都没有超过 50%，说明这两种方法在和网络上的平均新闻的比较中，效果不如平均新闻。

从表格整体上来看，Base 方法似乎效果非常差，远远落后于其他三种生成方法，但通过对生成文本的实际检查，发现 Base 方法在生成部分的人物介绍文本的时候，会显示知识库中没有对应人物的知识，这导致了 Base 方法的得分较低，而其他三种方法在生成文本的时候，均没有这种情况出现。虽然 Base 方法在生成文本的时候，由于这一原因产生了部分低分文章，但考虑到这种知识无法提取的情况与低分的文本内容是有关联的，因此认为 Base 方法仍然是四种方法中效果最差的方法。

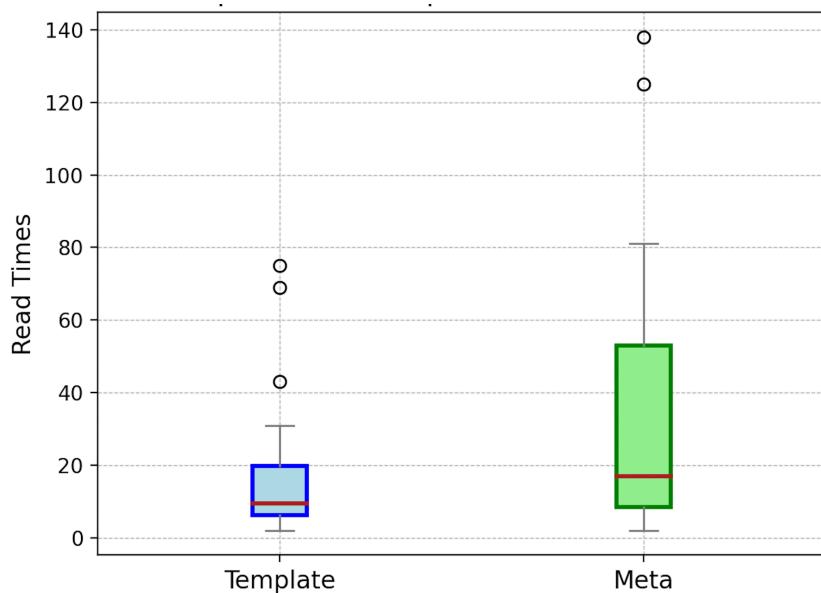
对于基于模版的生成方法和基于 Meta 的生成方法，在生成文本前，假定这两种方法的生成效果是最好的，最终的实验结果符合预期。然而，发现基于 Meta 的生成方法生成的文本在两种评价指标上，都没有超过基于模版的生成方法。考虑到在每次

评估的过程中，评估使用的大模型都是单次调用，而无法评价新闻风格、结构的单一性，在这种情况下，基于模版的生成方法在生成文本的时候，会有一定的优势，因此基于模版的生成方法在实验中的表现更好。

整体来看，基于大模型评分的实验结果表明，在评估文本质量的过程中，在没有参考的情况下，评分倾向于中庸，对不同的文章质量区分度不够明显。对于基于比较的评估，区分度要稍微好一些，但受到参考文本提取机制的影响，评分结果仍然有一定的偏差。

综上所述，基于大模型的文本生成方法中，基于模版的生成方法在生成普通文本的任务上有着较好的效果，而基于 Meta 的生成方法在这一任务上的效果略逊于基于模版的生成方法，但提供了文本结构的多样性。基于大模型的评分结果在一定程度上反映了文本生成的质量，但在实际应用中，需要对评分进行进一步的分析和调整，以适应不同的评估需求。

## 7.2 一般文本生成结果和分析



图中蓝色部分代表基于模版的生成方法，绿色部分代表基于 Function calling 的 Meta 生成方法，  
纵坐标为公众号文章阅读量。

图 7-1 实时长文章公众号阅读量

Figure 7-1 Real-time long article public readership

### 7.2.1 实时长文本结果分析

对于实时文本生成和长文本的生成，实验的结果通过公众号阅读量作为评价指标，如图7-1所示，实时长文本生成的效果图。

通过观察两种长文本生成的阅读量，发现基于 Meta 的生成方法在平均阅读量上有着较好的表现，并且有较高的阅读量峰值，这说明基于 Meta 的生成方法在生成长文本的任务上有着较好的效果。而基于模版的生成方法在平均阅读量上略逊于基于 Meta 的生成方法

我们承认，实验结果中的阅读量是一个相对主观的评价指标，可能会受到文章主题和系统推送机制的影响，故实验结果仅供参考。

### 7.2.2 风格化文本生成结果分析

表 7-2 三体实验风格迁移率

Table 7-2 Three-body experimental transfer style rate

Method	BLEU-1	BLEU—2	Precision	Recall	F1
readbased-3.5	22.3	11.2	63.39	66.7	64.96
roleplay-3.5	40.53	23.13	75.71	75.81	75.72
fullstyle-3.5	<b>46.21</b>	<b>27.18</b>	<b>77.61</b>	78.43	<b>77.98</b>
readbased-4	32.96	18.88	70.19	76.62	73.21
roleplay-4	41.94	23.1	74.5	76.08	75.26
fullstyle-4	29.33	19.43	66.99	<b>78.92</b>	72.22

表 7-3 三体实验内容保留率

Table 7-3 Three-body experimental content retention rate

Method	BLEU-1	BLEU—2	Precision	Recall	F1
readbased-3.5	34.5	27.43	68.02	71.27	69.56
roleplay-3.5	60.61	50.14	86.45	85.84	86.12
fullstyle-3.5	<b>84.22</b>	<b>79.43</b>	<b>92.4</b>	<b>92.43</b>	<b>92.41</b>
readbased-4	41.12	29.74	75.63	83.39	79.19
roleplay-4	53.24	36.47	81.03	82.46	81.73
fullstyle-4	33.63	26.51	70.96	79.41	77.86

对于风格化文本生成的结果，对于《三体》的风格化文本生成，实验结果如表7-

2和表7-3所示，对于《阿Q正传》的风格化文本生成，实验结果如表7-4和表7-5所示。

首先分析《三体》的风格化文本生成评估结果，从表7-2和表7-3中可以看出，基于文法的生成方法有着最突出的体现，其次是基于角色的生成方法，最后是基于阅读的生成方法。

首先分析基于阅读的生成方法在风格迁移率上和内容保留率上都评分较低的原因：基于阅读的生成方法在生成文本的时候，会受到参考文本的影响，而参考文本的内容和风格虽然都在同一部文学作品中，但与生成文本的内容和风格有一定的差异，对于特殊的作家，甚至差别会相当大，在这种情况下，基于阅读的生成方法表现较差。

表 7-4 阿 Q 实验风格迁移率

Table 7-4 Mr.Q experimental transfer style rate

Method	BLEU-1	BLEU—2	Precision	Recall	F1
readbased-3.5	20.86	10.05	61.52	64.81	63.06
roleplay-3.5	40.28	22.54	77.04	77.16	77.05
fullstyle-3.5	<b>45.8</b>	<b>27.17</b>	<b>78.42</b>	79.16	<b>78.75</b>
readbased-4	30.54	21.5	71.19	73.15	73.34
roleplay-4	42.42	25.57	76.82	75.88	79.12
fullstyle-4	39.17	23.45	74.13	<b>80.20</b>	78.45

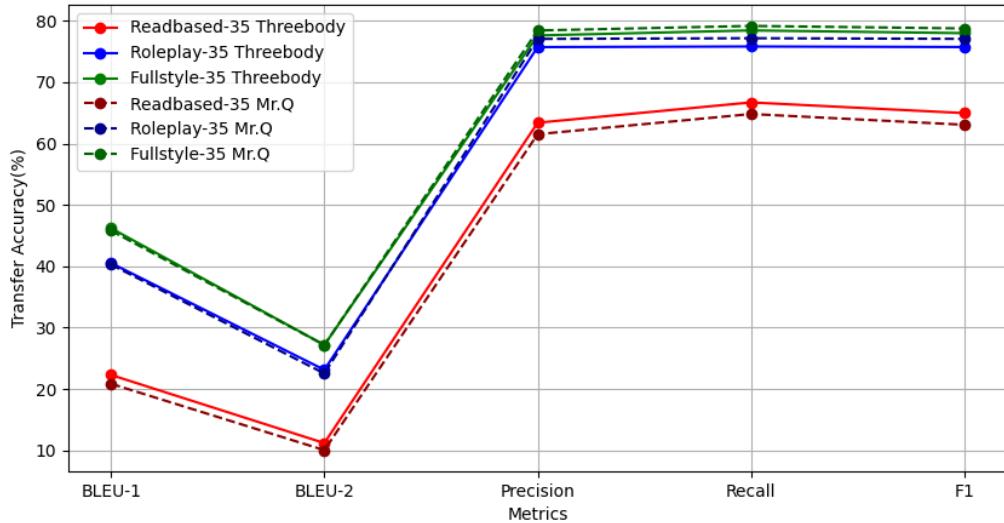
表 7-5 阿 Q 实验内容保留率

Table 7-5 Mr.Q experimental content retention rate

Method	BLEU-1	BLEU—2	Precision	Recall	F1
readbased-3.5	31.4	24.28	65.89	68.9	67.27
roleplay-3.5	70.58	62.55	91.23	90.5	90.8
fullstyle-3.5	<b>88.45</b>	<b>84.66</b>	<b>94.14</b>	<b>94.07</b>	<b>94.07</b>
readbased-4	40.22	38.27	73.44	75.89	74.66
roleplay-4	53.44	46.17	90.22	91.45	91.12
fullstyle-4	80.11	78.29	90.06	92.08	91.77

接下来分析基于角色的生成方法和基于文法的生成方法的评分差异：基于角色的生成方法在风格迁移率上和内容保留率上都有着较好的表现，但在内容保留率上，

基于文法的生成方法有着更好的表现，这说明基于角色的生成方法上，确实可以比较好地提取到作者的写作风格，并生成迁移率较好的文章，但同时也对原文本有一定程度上的语义丢失。



图中为基于 GPT-3.5-Turbo 下风格化生成方法的风格迁移效果，其中实线部分代表《三体》的效果，虚线部分代表《阿 Q 正传》的效果。

图 7-2 GPT3.5 风格迁移效果图

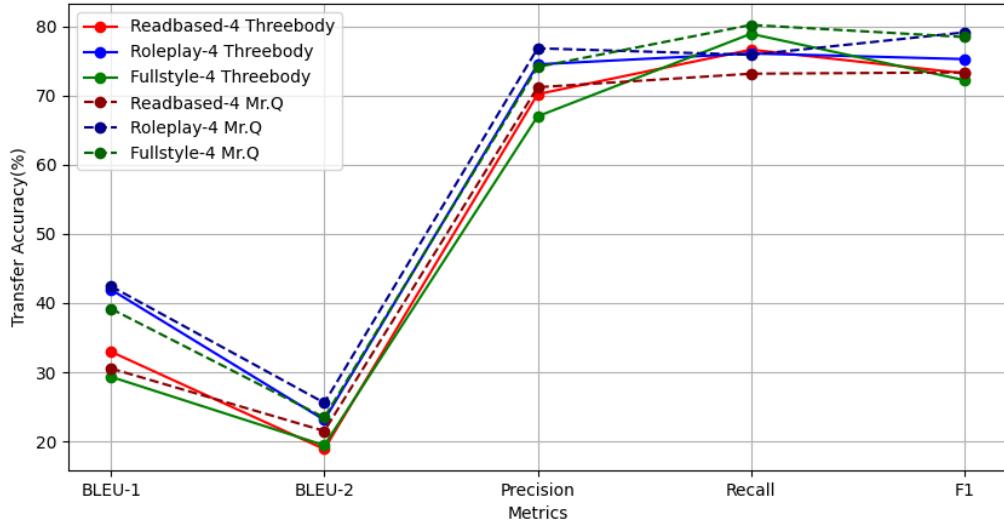
Figure 7-2 GPT3.5 style transfer effect diagram

对于基于文法的生成方法，在风格迁移率上和内容保留率上都有着较好的表现，然而，在实际的实验中发现，GPT-3.5-Turbo 的生成文本在大部分评价指标上都超过了 GPT-4-Turbo 的生成文本，这在一定程度上违反了直觉。

对于这一奇怪的现象，我们认为是由于 GPT-4-Turbo 更可以利用文法中提到的词汇使用，在这种情况下，很容易生成阅读上比较难理解的文本，在实际的评估测试中，也会使得 GPT-4-Turbo 的生成文本在评分上有所下降。

对于《阿 Q 正传》的风格化文本生成评估结果，从表7-4和表7-5中可以看出，总体的表现与《三体》的风格化文本生成评估结果相似，基于文法的生成方法有着最突出的体现，其次是基于角色的生成方法，最后是基于阅读的生成方法。

下面将对《三体》和《阿 Q 正传》的风格化文本生成评估结果进行分析，可以通过图7-2和图7-3来看出 GPT-3.5 和 GPT-4 的风格迁移效果图，通过图7-4和图7-5来



图中为基于 GPT-4-Turbo 下风格化生成方法的风格迁移效果，其中实线部分代表《三体》的效果，虚线部分代表《阿 Q 正传》的效果。

图 7-3 GPT4 风格迁移效果图

Figure 7-3 GPT4 style transfer effect diagram

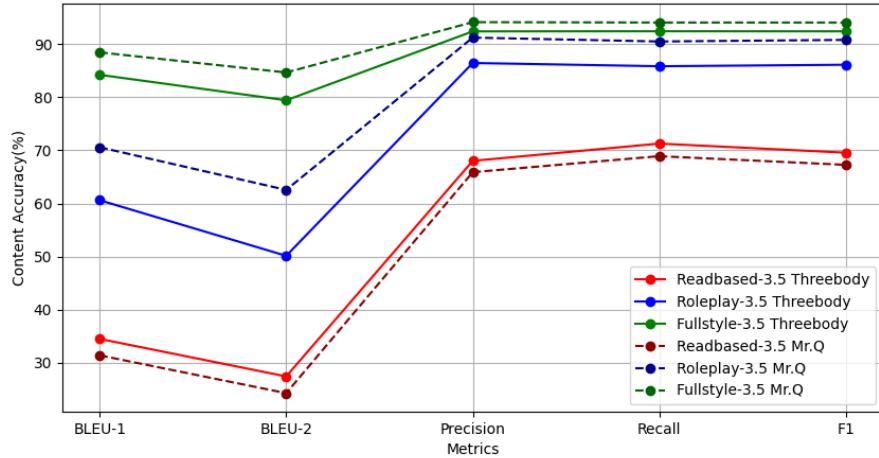
看出 GPT-3.5 和 GPT-4 的内容保留效果图。

通过几幅图可以看出，这两部文学作品的风格化文本生成效果总体趋势一致。但是，我们发现，对于《阿 Q 正传》的风格迁移率上，相比《三体》要高一些。分析原因，有一部分原因可能是模型出现的波动，同时也有相当一部分原因在于作者本身风格的差异：鲁迅在整个文学作品中，用词方式和行文习惯比较有个人特点，且使用个人特色用词较多，所以在基于文法的文本生成中，效果较好。

综上所述，发现考虑到实际文本作品的内容，基于文法的文本生成确实可以实现比较好的风格化，但需要确定恰当的指令和对应的大语言模型类型。

### 7.3 本章小结

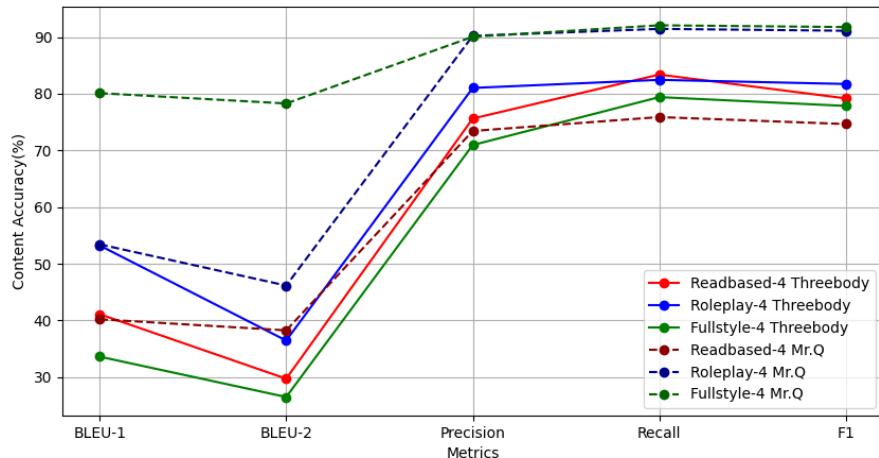
本章对第四、五章的实验结果进行展示并进行分析，对简单文本和特定复杂文本的生成方法及质量评估进行详细的讨论。



图中为基于 GPT-3.5-Turbo 下风格化生成方法的内容保留效果，其中实线部分代表《三体》的效果，虚线部分代表《阿 Q 正传》的效果。

图 7-4 GPT3.5 内容保留效果图

Figure 7-4 GPT3.5 content reserve effect diagram



图中为基于 GPT-4-Turbo 下风格化生成方法的内容保留效果，其中实线部分代表《三体》的效果，虚线部分代表《阿 Q 正传》的效果。

图 7-5 GPT4 内容保留效果图

Figure 7-5 GPT4 content reserve effect diagram

## 第八章 全文总结

### 8.1 内容总结

本文主要研究了基于大语言模型的文本生成方法及对应文本质量评估，在对如风格化文本生成等特定文本内容生成进行了深入研究探讨。基于对大语言模型的调研和对自然语言生成任务的调研，确定了本文的研究方向。在实际探究过程中，通过实验验证基于大语言模型的评估方法的可行性，并进行了详细分析；确定了评估方法后，对于文本生成方法进行了分层次的探究，从简单的人物介绍文本的生成到实时长文本生成，风格化文本的生成。在探究文本生成方法的过程中，提出了包括基于模版的生成方法，基于 Meta 的生成方法等效果较好的生成方法，并对这些方法在实际的应用中进行了相关拓展和实验验证。

### 8.2 可能的限制

在本次对文本生成及评估的探究中，有实现上有一些限制及改进的地方，在这里列出：

- 对于文本生成的实验，由于实验环境的限制，实验的数据集较小，对于结果可能有偶然性的影响。
- 在实际测试基于大语言模型的生成及评估方法时，本次实验使用的是 OpenAI 的 GPT 模型，并没有对其他大语言模型进行测试，可能需要添加更多种类的模型来进行对比和改进。
- 在进行实验的过程中，受限于实验条件，对于文本生成的实验，没有进行人工评分，而是使用基于大语言模型的评估方法，虽然通过实验验证了其评估方法的可行性，但是这种实验方法的评估结果仍然有一定的偏差。
- 在使用大语言模型进行相应实验时，发现大语言模型的输出相对难以控制，尽管采用调整温度的方法使模型输出更加稳定，但这种方法仍然无法完全解决模型输出的不确定性。
- 考虑到文本生成及质量评估这一任务目前并没有系统和理论的评价体系，在本次实验中使用的生成任务及评估指标都有一定主观性，受限于这种情况，并没有对实验结果进行理论性的分析。

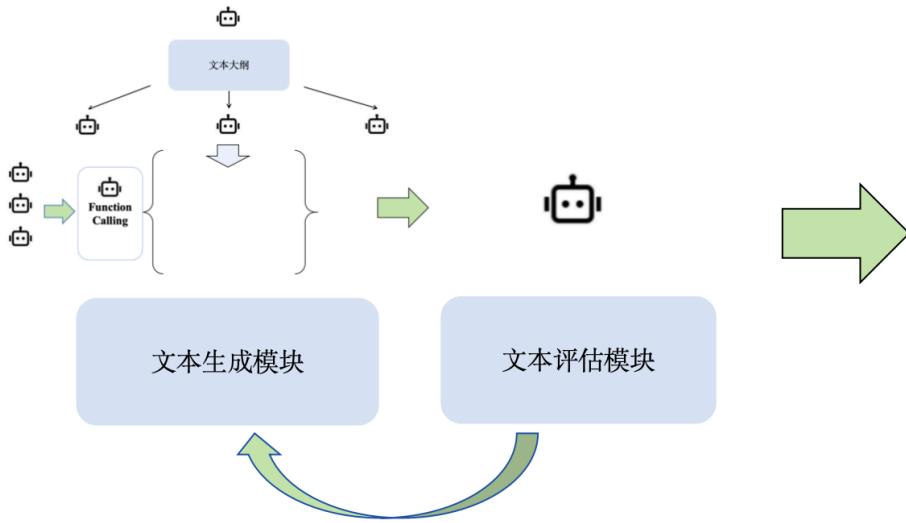


图 8-1 基于“审稿人”的文本生成图  
Figure 8-1 ”Reviewer” based text generation

### 8.3 研究展望

对于本文的相关工作，有一些可以进一步探究的方向，下面将列出一些可能的研究展望：

- 对于文本生成的实验，可以进一步扩大数据集，增加实验的数据量，以提高实验结果的稳定性。
- 在实验中，发现大语言模型在评估过程中评分倾向于中庸，对不同的文章质量区分度不够明显，可以进一步探究如何提高大语言模型的评分区分度。
- 基于本文的相关实验和研究，可以考虑基于大语言模型的文本评估方法，添加类似“审稿人”的制度，通过添加大语言模型评审方法，进一步提高文本生成的质量，大致思路如图8-1所示。

## 参 考 文 献

- [1] RADFORD A, NARASIMHAN K, SALIMANS I, Tim anImproving language understanding by generative pre-trainingd Sutskever, et al. Improving language understanding by generative pre-training[J]. OpenAI Technical Report, 2018, 1: 1-12.
- [2] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [3] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [4] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[EB/OL]. (2023-02-28). <https://arxiv.org/abs/2302.13971>.
- [5] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[EB/OL]. (2023-10-01). <https://arxiv.org/abs/2211.09110>.
- [6] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.
- [7] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26: 3111-3119.
- [8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16). <https://arxiv.org/abs/1301.3781>.
- [9] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model. [C] // Interspeech: vol. 2: 3. Makuhari, Japan, 2010: 1045-1048.
- [10] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Advances in neural information processing systems, 2000, 3: 932-938.
- [11] KOMBRINK S, MIKOLOV T, KARAFIÁT M, et al. Recurrent Neural Network Based Language Modeling in Meeting Recognition.[C] // Interspeech: vol. 11. Florence, Italy, 2011: 2877-2880.
- [12] VASWANI A, SHAZER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.
- [13] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11). <https://arxiv.org/abs/1810.04805>.
- [14] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23). <https://arxiv.org/abs/2001.08361>.
- [15] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways [J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.
- [16] HUANG S, DONG L, WANG W, et al. Language is not all you need: Aligning perception with language models[J]. Advances in Neural Information Processing Systems, 2024, 36: 3330-3356.
- [17] WU C, YIN S, QI W, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models[EB/OL]. (2023-03-08). <https://arxiv.org/abs/2303.04671>.
- [18] DRIESS D, XIA F, SAJJADI M S, et al. Palm-e: An embodied multimodal language model [EB/OL]. (2023-03-06). <https://arxiv.org/abs/2303.03378>.
- [19] CAO Y, LI S, LIU Y, et al. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt[EB/OL]. (2023-03-07). <https://arxiv.org/abs/2303.04226>.
- [20] DONG Q, LI L, DAI D, et al. A survey on in-context learning[EB/OL]. (2023-01-01). <https://arxiv.org/abs/2301.01001>

- iv.org/abs/2301.00234.
- [21] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.
- [22] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054.
- [23] XUE L, SUN C, WUNSCH D, et al. An adaptive strategy via reinforcement learning for the prisoner s dilemma game[J]. IEEE/CAA Journal of Automatica Sinica, 2017, PP(1): 1-10.
- [24] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
- [25] PEÑA A, MORALES A, FIERREZ J, et al. Leveraging large language models for topic classification in the domain of public affairs[C]//International Conference on Document Analysis and Recognition. Lausanne, Switzerland, 2023: 20-33.
- [26] QIN C, ZHANG A, ZHANG Z, et al. Is chatgpt a general-purpose natural language processing task solver?[EB]. (2023-11-19).
- [27] ABDELALI A, MUBARAK H, CHOWDHURY S A, et al. Benchmarking arabic ai with large language models[EB/OL]. (2024-02-05). <https://arxiv.org/abs/2305.14982>.
- [28] WU P Y, TUCKER J A, NAGLER J, et al. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting[EB/OL]. (2023-09-26). <https://arxiv.org/abs/2303.12057>.
- [29] ZHUANG Y, LIU Q, NING Y, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective[EB/OL]. (2023-10-28). <https://arxiv.org/abs/2306.10512>.
- [30] DEROY A, GHOSH K, GHOSH S. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?[EB/OL]. (2023-06-14). <https://arxiv.org/abs/2306.01248>.
- [31] FRANK M C. Baby steps in evaluating the capacities of large language models[J]. Nature Reviews Psychology, 2023, 2(8): 451-452.
- [32] ZHUANG Y, LIU Q, NING Y, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective[EB/OL]. (2023-10-28). <https://arxiv.org/abs/2306.10512>.
- [33] VALMEEKAM K, MARQUEZ M, SREEDHARAN S, et al. On the planning abilities of large language models-a critical investigation[J]. Advances in Neural Information Processing Systems, 2023, 36: 75993-76005.
- [34] GILSON A, SAFRANEK C W, HUANG T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment[J]. JMIR Medical Education, 2023, 9(1): e45312.
- [35] KHAN Y A, HOKIA C, XU J, et al. covllm: Large language models for covid-19 biomedical literature[EB/OL]. (2023-06-28). <https://arxiv.org/abs/2306.04926>.
- [36] QIN Y, HU S, LIN Y, et al. Tool learning with foundation models[EB/OL]. (2023-06-15). <https://arxiv.org/abs/2304.08354>.
- [37] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: Language Models Can Teach Themselves to Use Tools[EB/OL]. (2023-02-09). <https://arxiv.org/abs/2302.04761>.
- [38] FABBRI A R, KRYŚCIŃSKI W, MCCANN B, et al. Summeval: Re-evaluating summarization evaluation[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 391-409.
- [39] ZHANG S, DINAN E, URBANEK J, et al. Personalizing dialogue agents: I have a dog, do you have pets too?[EB/OL]. (2018-09-25). <https://arxiv.org/abs/1801.07243>.
- [40] GOPALAKRISHNAN K, HEDAYATNIA B, CHEN Q, et al. Topical-chat: Towards knowledge-grounded open-domain conversations[EB/OL]. (2023-08-23). <https://arxiv.org/abs/2308.11995>.

- [41] MOSTAFAZADEH N, CHAMBERS N, HE X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 839-849.
- [42] LIN Y T, CHEN Y N. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models[EB/OL]. (2023-05-23). <https://arxiv.org/abs/2305.13711>.
- [43] KIM S, SHIN J, CHO Y, et al. Prometheus: Inducing fine-grained evaluation capability in language models[EB/OL]. (2024-03-09). <https://arxiv.org/abs/2310.08491>.
- [44] TAO Z, XI D, LI Z, et al. CAT-LLM: Prompting Large Language Models with Text Style Definition for Chinese Article-style Transfer[EB/OL]. (2024-01-11). <https://arxiv.org/abs/2401.05707>.
- [45] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 311-318.
- [46] ZHANG T, KISHORE V, WU F, et al. Bertscore: Evaluating text generation with bert[EB/OL]. (2019-04-22). <https://arxiv.org/abs/1904.09675>.
- [47] GUAN J, HUANG F, ZHAO Z, et al. A knowledge-enhanced pretraining model for commonsense story generation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 93-108.
- [48] GHAZARIAN S, WEI J T Z, GALSTYAN A, et al. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings[EB/OL]. (2019-04-24). <https://arxiv.org/abs/1904.10635>.
- [49] TAO C, MOU L, ZHAO D, et al. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 32: 1. New Orleans, Louisiana, USA, 2018: 722-729.
- [50] GUAN J, HUANG F, ZHAO Z, et al. A knowledge-enhanced pretraining model for commonsense story generation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 93-108.
- [51] KENDALL M G. The treatment of ties in ranking problems[J]. Biometrika, 1945, 33(3): 239-251.
- [52] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[J]. Advances in neural information processing systems, 2015, 28: 1693-1701.
- [53] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[EB/OL]. (2016-02-19). <https://arxiv.org/abs/1602.06023>.
- [54] KRYŚCIŃSKI W, KESKAR N S, MCCANN B, et al. Neural text summarization: A critical evaluation[EB/OL]. (2019-08-23). <https://arxiv.org/abs/1908.08960>.
- [55] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. Barcelona, Spain: Association for Computational Linguistics, 2004: 74-81.

## 致 谢

本篇文章就此结束了，我的本科生涯也马上要画上句号。在这里，我想感谢所有在四年中帮助我的人。

我想感谢以下几位老师，在我的大学生活中，不论是科研，本次毕业设计，还是生活中，都教会了我很多（按拼音姓和名首字母排序，不分先后）：

感谢何丙胜老师，给我学习、研习的机会，在学习和工作中给了我友善的指导和支持。

感谢孙世轩老师，我的毕业设计导师，关于毕业设计帮助我很多，积极指导我完成毕业设计，在我遇到困难时给予我很多帮助。

感谢沈艳艳老师，在我迷茫的时候给我很多帮助，为我指明我的方向。

我还有感谢我的室友，我的家人，感谢他（她）们在困难时对我的支持和帮助。

最后，感谢所有在我大学生活中帮助过我的人，谢谢你们！

# RESEARCH ON STYLIZED TEXT GENERATION AND APPLICATIONS OF LARGE LANGUAGE MODELS

The thesis provides an exhaustive analysis of the evolution and impact of large language models (LLMs) in the realm of computer science, charting their journey from foundational models like GPT-1 to more advanced systems such as OpenAI's GPT series, Meta's LLaMA, and Baidu's Wenxin series. These models have not only revolutionized academic research but have also paved the way for robust commercial applications. They have fundamentally altered methodologies in natural language processing (NLP), achieving remarkable improvements in tasks such as machine translation, text generation, and text classification.

This research focuses primarily on the domain of natural language generation (NLG). It begins by examining traditional NLG tasks, identifying how LLMs have introduced innovative tasks and metrics that surpass previous capabilities. The study rigorously tests the feasibility of LLM-based evaluation methods through a diverse array of experiments, which are critical in establishing the efficacy of these models in real-world applications.

The experimental framework of the thesis employs four distinct approaches to text generation: direct text generation, role-based text generation, template-based generation, and a novel stage-based generation technique developed by Meta. The initial experiments focus on generating simple texts, such as character introductions, to ground the study in concrete examples before progressing to more complex applications.

Further investigation in the dissertation explores advanced text generation methodologies tailored for generating real-time texts, long-form texts, and stylized texts. Each category employs specific generation techniques and corresponding evaluation metrics designed to assess their effectiveness comprehensively. The study utilizes multiple datasets across several experiments, enabling a detailed examination of the characteristics of each text generation method. This rigorous testing framework provides a robust evaluation of the performance of LLM-based text generation, offering clear insights into the strengths and limitations of these methods.

In the concluding sections, the thesis synthesizes the research findings, affirming the chosen research directions based on the empirical evidence gathered. The text generation meth-

ods are analyzed in depth, extending from basic tasks to intricate challenges like generating real-time, long, and stylized texts. Innovative generation techniques, including template-based and Meta's stage-based methods, are not only proposed but also rigorously tested in practical settings, demonstrating their viability and effectiveness in enhancing text generation.

The limitations of the study are candidly addressed, including the constraints posed by the relatively small sizes of datasets used, the exclusive reliance on GPT models without comparative testing across other LLMs, the absence of human evaluative measures in favor of automated model-based assessments, and the challenge of managing the unpredictability of LLM outputs, despite efforts to mitigate these issues through temperature adjustments and other stabilization techniques.

The dissertation concludes with proposals for future research, suggesting that expanding dataset sizes could enhance the reliability of results. It also highlights the need for improving the discriminative capacity of LLM evaluations to better distinguish between texts of varying quality. Furthermore, the introduction of a "reviewer system" is proposed, which could incorporate advanced LLM-based evaluation strategies to further refine and enhance the quality of generated texts. These recommendations set a roadmap for future investigations, aiming to push the boundaries of what LLMs can achieve in the field of text generation.