生成式文本质量的自动评估方法综述

兰天[1]，马梓奥[1]，周杨浩[1]，徐晨[2]，毛先领[1]
[1]北京理工大学，计算机学院，北京市，100081
[2]北京理工大学，医工技术学院，北京市，100081
lantiangmfty@gmail.com

摘要

人工评估，作为生成式文本质量评价的金标准，成本太高；自动评估，核心思想在于要使其评估结果与人工评估高度相关，从而实现对生成式文本质量的自动化分析和评价。随着自然语言处理领域相关技术的迭代进步，使得生成式文本质量的自动评估技术，已然经历了多次技术范式的迭代。然而，学界至今依然缺乏对生成式文本质量自动评估技术的系统化总结。因此，本文将首先系统地对已有的生成式文本自动评估方法进行归纳总结，然后分析了生成式文本自动评估方法的主要发展趋势，最后为了使读者更加宏观地了解自动评估整体，对自动评估领域整体的未来研究方向进行了探讨和展望。

关键词： 文本生成 ； 自动评估 ； 综述

A Survey of Automatic Evaluation on the Quality of Generated Text

Tian Lan[1], Ziao Ma[1], Yanghao Zhou[1], Chen Xu[2], Xian-Ling Mao[1]
[1]School of Computer Science and Technology, Beijing Institute of Technology, 100081
[2]School of Medical Technology, Beijing Institute of Technology, 100081
lantiangmfty@gmail.com

Abstract

Human evaluation, as the gold standard for assessing the quality of generated text, is prohibitively expensive. Automatic evaluation, on the other hand, aims to achieve high correlation with manual evaluation, thereby enabling automated analysis and assessment of generated text quality. With the iterative advancement of technologies in the field of natural language processing, the automatic evaluation of generated text quality has undergone several paradigm shifts. However, there is still a lack of systematic summarization of these automatic evaluation techniques in the academic community. Therefore, this paper first systematically summarizes the existing methods for automatic evaluation of generated text. It then analyzes the main development trends of these automatic evaluation methods. Finally, to provide a more comprehensive understanding of automatic evaluation, the paper discusses and anticipates future research directions in the field of automatic evaluation.

Keywords: Text Generation , Automatic Evaluation , Survey

第二十三届中国计算语言学大会论文集，第169页-第196页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

随着以ChatGPT和GPT-4为代表的大规模语言模型（Large Language Models, LLMs）的迅猛发展 (OpenAI, 2023)，自然语言处理（NLP）领域正经历着前所未有的变革 (Lee et al., 2023; Zhao et al., 2023a)。大规模语言模型在文本生成 (Su et al., 2022b)、机器翻译 (Vaswani et al., 2017)、对话系统(Lan et al., 2020; Lan et al., 2023) 等多个NLP应用场景中展现出卓越的性能，极大地推动了生成式人工智能技术的发展。然而，大规模语言模型文本生成能力的提升不可避免地对其生成内容的自动化评估带来了巨大的挑战 (Zheng et al., 2023; Liu et al., 2023a)。因此，如何有效地自动化评估文本生成模型的能力，以确保其可靠性和实用性，已成为NLP领域面临的一项重要挑战 (Fu et al., 2023; Li et al., 2024b; Pan et al., 2023)。

目前，自动评估技术是解决这一挑战的具有前景的重要技术方案 (Tao et al., 2017; Pan et al., 2023; Lan et al., 2024)。自动评估技术旨在设计和构建和人工评估高度相关的自动评估方法，以实现对生成文本的质量的可靠评估 (Lan et al., 2020; Tao et al., 2017; Fu et al., 2023)。自动评估技术不仅能够帮助研究人员快速分析优化模型，还能够进一步推动大规模语言模型的自我提升（LLM self-improvement），即通过自动评估方法生成的反馈优化并增强大规模语言模型生成结果 (Yuan et al., 2024b; Xu et al., 2024)。因此自动评估技术具有重要的研究意义。目前，生成式文本的自动评估技术已经取得了显著的进展。如图1所示，从2014年至今，有关自动评估指标论文的发表数量呈现出逐年快速增长的趋势。然而，当前关于生成式文本的自动评估技术仍缺乏系统的总结。这种状况不仅导致了针对具体NLP任务中自动评估方法的选择和应用上的混乱，同时也限制了对生成式文本自动评估技术发展趋势的总结，以及对未来研究方向的深入洞察。为了解决该问题，本文系统地整理和综述了生成式文本的自动评估方法，分析了生成式文本自动评估发展过程中四轮技术范式的转变 (Papineni et al., 2002; Zhang* et al., 2020a; Lan et al., 2020; Fu et al., 2023)。基于上述讨论，本文概括了自动评估技术发展的主要趋势，即自动评估方法在通用性、可解释性和多维度性方面的持续改进。最终，本文对自动评估领域的未来重要的发展趋势和研究方向进行探讨和展望，包括以下四个方面内容：（1）小模型自动评估能力的提升；（2）评估可解释评估的质量；（3）评估技术的应用；（4）多模态生成内容的自动评估。

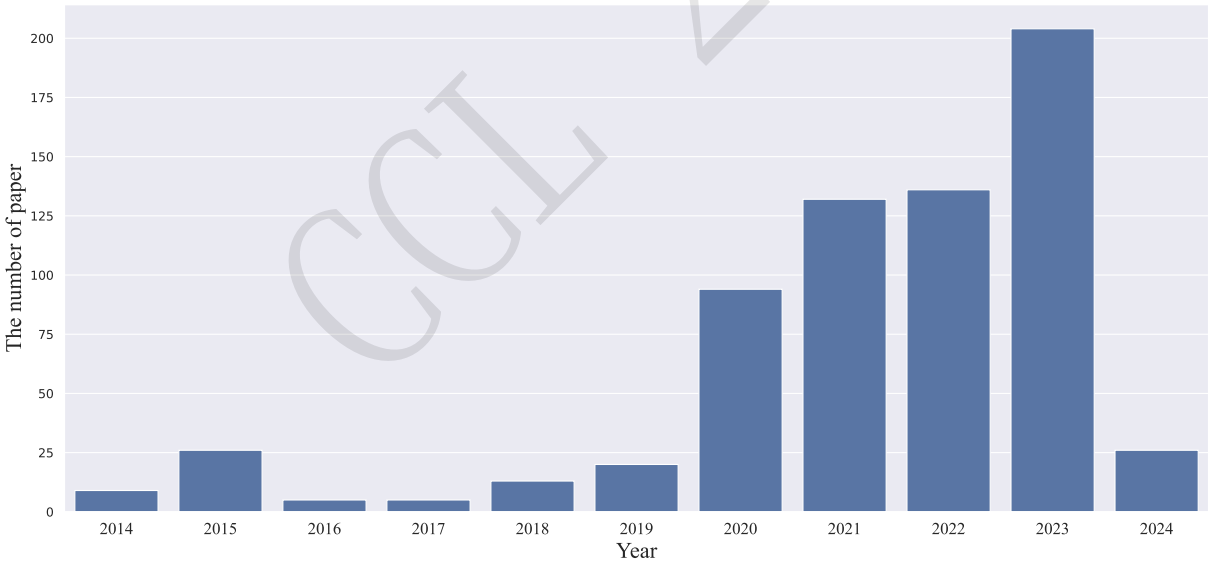

Figure 1: 生成式文本自动评估指标相关论文发表的数量呈现逐年递增的趋势（2014-2024）。

本文的后续内容安排如下：第2节将介绍生成式文本自动评估技术的相关背景。第3节归纳总结了生成式文本的自动评估技术的四轮技术范式的特点。基于归纳总结的内容，第4节对生成式文本自动评估技术的发展趋势进行了总结。最终，第5节将自动评估技术未来的研究趋势进行深入探讨和展望。

2 背景

2.1 生成式文本评估的分类体系

生成式文本的质量评估是衡量现有文本生成模型性能的关键手段 (Fu et al., 2023)。根据评估方法的不同，现有的生成式文本评估技术可以分为两大类：一类是基于标准答案的评估，另一类是无标准答案的评估 (Cobbe et al., 2021; Lan et al., 2020)。

2.1.1 有标准答案的评估

数据集	任务	数据规模
GSM8K (Cobbe et al., 2021)	小学数学题	8.79K
MATH (Hendrycks et al., 2021)	竞赛数学题	12.5K
HumanEval (Chen et al., 2021)	代码题	164
MBPP (Austin et al., 2021)	代码题	500
MMLU (Hendrycks et al., 2020)	人文社科等57个主题	14K
HellaSwag (Zellers et al., 2019)	常识推理	10K
AI2 Reasoning Challenge (Clark et al., 2018)	多项选择问答	2.59K
DROP (Dua et al., 2019)	段落理解问答	96K
C-Eval (Huang et al., 2023b)	中文52学科多项选择题	14K

Table 1: 有标准答案的评估基准数据集。

有标准答案的评估方法依赖于包含了人工标注的标准答案的测试数据集，以此来评价模型生成内容的质量。这种方法在评估近期流行的大规模语言模型，如GPT-4、InternLM2 (Team, 2023)、DeepSeek (DeepSeek-AI, 2024)和Qwen (Bai et al., 2023a)系列模型在解决特定下游任务的能力方面得到了广泛应用。[0] 例如，在模型解决数学问题和编程问题的任务中，通过将生成的答案与标准答案进行比较，以判断解题的正确性。部分常用的有标准答案的评估数据集如上表1所示。

2.1.2 无标准答案的评估

无标准答案的评估主要适用于那些没有固定标准回答的开放式文本生成任务，如开放式对话 (Tao et al., 2017)和故事生成 (Guan and Huang, 2020)等场景。

在这些任务中，人工评估是最常见且最可靠的评估方法 (Tao et al., 2017)。人工评估的优点在于人工结果的可靠性，因为人工标注能够直接反映人类的判断和偏好。然而，人工评估也存在明显的局限性，包括耗时、成本高昂以及实验结果的不可重复性 (Lan et al., 2020; Lan et al., 2023)。为解决无标准答案评估中人工评估的上述问题，自动评估技术逐渐开始受到研究人员的重视。自动评估的目标是开发一种自动化的评估模型或方法，使其评分结果与人工评估具有较高的相关性。然而，目前自动评估与人工评估的相关性仍有显著差距，难以完全替代人工评估。因此，如何有效提高自动评估方法与人工评估的一致性，成为当前研究的重点 (Fu et al., 2023; Pan et al., 2023)。随着自动评估技术的不断发展，它已经经历了从启发式评估、基于语义向量的评估、基于学习的评估到基于大规模语言模型的评估的四轮技术范式的转变。总体来看，随着自动评估技术范式的演变，自动评估方法的通用性、可解释性和评估维度的多样性都得到了显著的提升。

2.2 评价自动评估方法的性能

根据自动评估输出格式的类型，已有的自动评估工作可以分为两类：（1）**判别式自动评估**：这种方法直接为生成式文本一个预测的质量分数；（2）**生成式自动评估** (Pan et al., 2023)：这种方法生成针对待评估文本的可解释性分析内容，指出文本中的错误并提供改进意见 (Lan et al., 2024)。随着大规模语言模型的快速发展，生成式自动评估开始成为主要趋势，即大规模语言模型通过使用思维链（chain-of-thought） (Wei et al., 2023)的方式逐步生成针对待评估文本内容质量的分析以及对应的质量分数标签，实现更可解释、更细粒度、信息更丰富的自动评价。

[0] https://cdn.openai.com/papers/gpt-4.pdf

数据集	任务	数据规模
SummEval (Fabbri et al., 2021)	文本摘要	1.6K
WMT-22 (zh-en) (Freitag et al., 2022)	中英翻译	33.75K
WebNLG-2020 (Zhou and Lampouras, 2020)	RDF-to-Text	2.8K
LFQA (Jiang et al., 2023)	长文档问答	1.6K
GSM8K† (Jiang et al., 2023)	数学题评估	2.6K
OpenMEVA (Guan et al., 2021)	故事生成	1K
BAGEL (Mairesse et al., 2010)	Data-to-Text	202
CommonGen† (Lin et al., 2019)	生成式问答	2.8K
CriticBench (Lin et al., 2024)	推理类任务	3.83K

Table 2: 评价判别式自动评估方法的常用Meta-Evaluation数据集。†标识的数据集表示在原数据集（如GSM8K数据集）上构建的meta-evaluation数据集。

评估**判别式自动评估**的性能，首先需要收集人工标注的meta-evaluation数据集，这些数据集包含特定NLP下游任务的输入、待评估生成文本以及针对这些文本的多组人工标注质量分数。通过计算自动评估方法预测的分数与人工标注分数之间的相关性系数来衡量自动评估指标与人类评价之间的相关性。常用的相关性评估方法有Spearman、Pearson和Kendall相关性系数 (Zar, 2005; Fu et al., 2023)。目前，已有大量针对不同下游NLP任务的meta-evaluation数据集，常用的meta-evaluation数据集罗列如表2所示。

评估**生成式自动评估**方法的性能目前通常采用先进的大规模语言模型来完成，如GPT-4 (Li et al., 2024a; Cui et al., 2023; Sun et al., 2024)。然而，近期工作指出，GPT-4在评估生成式自动评估的质量和真实人类评估具有较大差别 (Wang et al., 2023c; Zhang et al., 2024)，因此如何准确评价生成式评估内容依然是一个较为困难的问题。

3 现有自动评估研究技术的发展

如图2所示，本文通过综合调研现有的自动评估技术，根据其发展先后顺序，将已有生成式文本自动评估方法归类为如下四个技术范式：（1）启发式的自动评估方法（Heuristic Evaluation）；（2）基于语义向量的自动评估方法（Embedding-based Evaluation）；（3）基于学习的自动评估方法（Learning-based Evaluation）；（4）基于大规模语言模型的自动评估方法（LLM-based Evaluation）。本节将对这四类自动评估范式的特点及其代表性工作进行系统分析和归纳。

3.1 启发式自动评估方法（Heuristic Evaluation）

启发式自动评估范式在NLP任务的早期阶段非常普遍。这种方法通常依赖于特定下游NLP任务中的规则或特征来构建评估方法，通过设计可解释的公式、算法或评估流程来计算质量分数。以下是一些常见的自然语言下游任务中使用的启发式评估方法介绍。

3.1.1 机器翻译

在机器翻译任务中，n-gram级别的字面信息是最为关键的特征。因此，启发式自动评估方法主要围绕n-gram特征构建。BLEU (Papineni et al., 2002)是一种广泛采用的方法，它通过计算生成翻译与高质量参考翻译之间n-gram的准确率来衡量二者的相似度。为了解决BLEU指标无法考虑翻译中重要词汇的问题，NIST (Doddington, 2002)引入了n-gram的信息量影响，考虑了翻译中罕见但重要的词汇对翻译质量的作用。由于BLEU仅计算了生成翻译n-gram相对于参考翻译的准确率，未考虑召回率，这影响了二者相似度的准确衡量。METEOR (Banerjee and Lavie, 2005)对此进行了改进，引入了召回率，并通过WordNet扩展了参考译文中每个n-gram的同义词，从而更好地考虑了翻译准确但不完全匹配的情况。类似地，chrF (Popović, 2015)和chrF++ (Popović, 2017)通过统计生成译文和参考译文之间字符级别n-gram的匹配程度来计算二者的相似度。除了n-gram级别的字面信息，文本的编辑距离也常用于评估生成翻译的质量。TER（Translation Edit Rate） (Snover et al., 2006; Moramarco et al., 2022)是一种基于编辑距离的评估指标，它通过计算将生成的机器翻译转换为参考翻译所需的最小编辑操作次

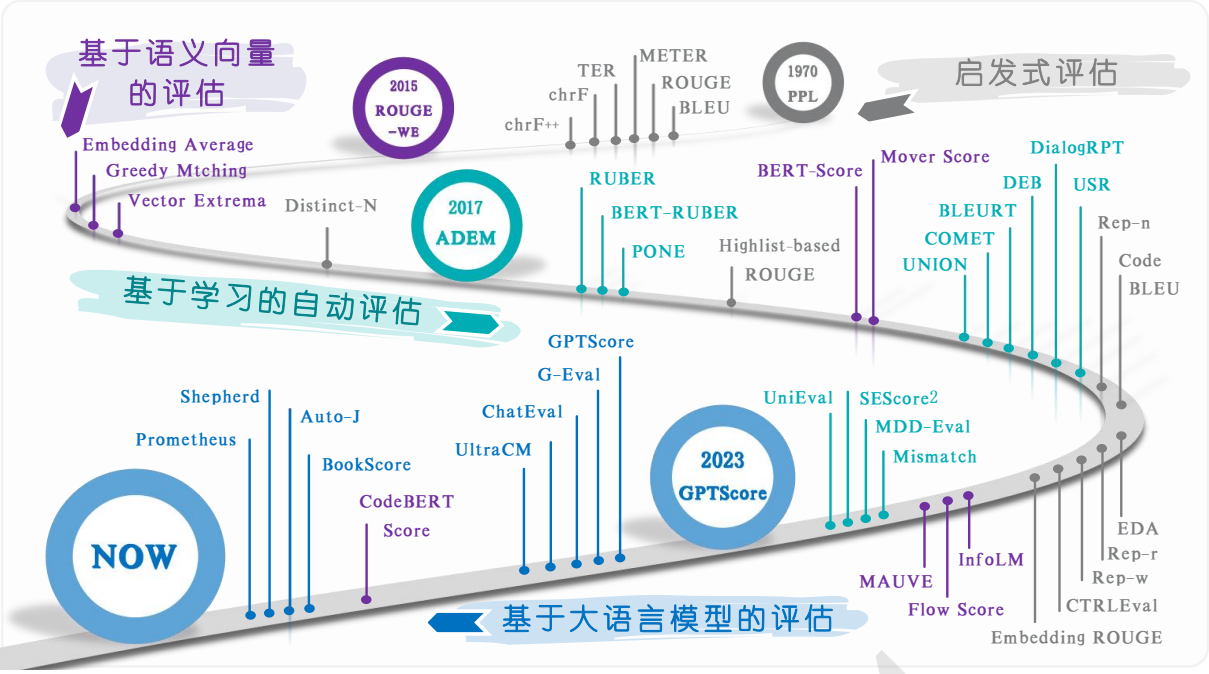

Figure 2: 生成式文本的自动评估指标经历了从启发式评估（Heruistic Evaluation）、基于语义向量的评估方法（Embedding-based Evaluation）、基于学习的自动评估方法（Learning-based Evaluation）到基于大规模语言模型（LLM-based evaluation）的四轮技术范式发展过程。

数来衡量翻译的质量。这些编辑操作包括插入、删除、替换和移动单词或字符。TER值越低，表示机器翻译的质量越高，因为它与参考翻译的差异越小。因此，TER提供了一种量化翻译误差的方法，帮助研究者评估和改进机器翻译系统的性能。与BLEU等基于准确率计算的指标相比，基于编辑距离的指标能够考虑到生成翻译中的错译和漏译问题。

3.1.2 开放式文本生成

开放式文本生成，如故事生成和对话生成等，是NLP领域中应用非常广泛的一类生成任务 (Zhang et al., 2018; Xu et al., 2022a; Yang et al., 2023)。这类任务的质量评估通常从多个角度进行，包括多样性、流畅性等 (Pascual et al., 2021; Xu et al., 2021)。

困惑度（perplexity）是评估生成文本流畅度中最常用的指标。它通过使用预训练的语言模型，如BERT/GPT等，来计算生成文本的预测概率 (Su et al., 2022c; Lan et al., 2022)。为了评估生成文本的多样性，Distinct-N (Li et al., 2016)指标通过统计不重复的1-gram和2-gram的比例来计算生成对话回复的多样性。类似的，Rep-n (Welleck et al., 2020)，Rep-w，Rep-r (Li et al., 2023a; Fu et al., 2021)指标通过评估生成文本中不重复n-gram的比例来评价文本的多样性。为了解决Distinct-N指标在长文本评估中的偏差，Expectation-Adjusted Distinct （EAD） (Liu et al., 2022)通过引入不重复token的期望来调整Distinct-N指标的计算过程。CTRLEval (Ke et al., 2022)通过设计多种不同的文本填充任务，计算基于上下文（包括前缀和属性标签）的生成文本的生成概率，以评估生成文本的连贯性、一致性和属性相关性。

3.1.3 文本摘要

ROUGE系列指标 (Lin, 2004)是摘要生成质量评估中应用最广泛的评估指标。ROUGE指标通过将生成摘要和参考摘要分割为n-gram，统计生成摘要n-gram相对于参考摘要的召回率，以此来衡量生成摘要与参考摘要的相似度。然而，ROUGE指标在计算过程中要求严格匹配n-gram，这可能导致对语义相近但字形不匹配的生成结果产生误判。为了解决这个问题，EmbeddingROUGE (Tsann et al., 2022)通过将n-gram的匹配对象扩展到参考摘要中语义相似的n-gram列表，从而放宽了精确匹配的约束。除了与参考文档匹配外，Highlight-based ROUGE (Hardy et al., 2019)通过计算生成摘要与源文档中标注的显著句子之间的n-gram重叠来评估摘要的质量。这种方法侧重于检测生成摘要是否包含了源文档中最重要、最相关的信

息，这对于确保摘要能够准确捕捉和传达原文核心内容至关重要。

3.1.4 代码生成

为了有效评估生成代码的质量，CodeBLEU (Ren et al., 2020)指标对BLEU指标进行了改造，在BLEU的计算过程中加入了三个计算分数，以分别考虑代码生成场景下的三种特征：（1）BLEU_{weight}对代码中的关键词施加了更高的权重，从而更准确地评估生成代码中的关键词相似性；（2）Match_{df}通过比较代码运行过程中的数据流图之间的相似度来评估生成代码和参考代码的相似度；（3）Match_{dst}将生成代码和参考代码分别表示为抽象语法树（AST），评估语法树之间的相似度。

3.1.5 其他NLP任务

除了上述NLP任务以外，针对更多NLP任务的启发式自动评估方法罗列在表3中。

启发式评估方法	针对任务
PARENT (Dhingra et al., 2019)	Table-to-text
SARI,FKBLEU (Xu et al., 2016)	文本简化
L'AMBRE (Pratapa et al., 2021)	文本形态变化
KDA (Moon et al., 2022)	多项选择问答

Table 3: 针对其他NLP任务的启发式自动评价方法。

3.2 基于语义向量的评估方法（Embedding-based Evaluation）

启发式自动评估方法严重依赖于特定下游NLP任务中的关键特征来计算生成文本和参考文本之间的相似度。然而，这些特性通常都是字符级别特征，如n-gram特征等。这些特征很难捕捉到复杂的语义信息。因此，在开放式文本生成任务等场景中，评价那些在字面上无法精确匹配但语义相似度高的评估文本时，其评估结果往往与真实人工评估的相关性存在较大差距。为了解决启发式自动评估方法无法考虑文本语义信息的问题，基于语义向量的自动评估方法开始逐渐被引入。基于语义向量的评估方法的计算过程可以形式化定义如下：

$$\text{Score} = \mathcal{M}\left(E\left(g\right), E\left(c\right), E\left(r\right)\right) \tag{1}$$

其中E代表具体的语义嵌入模型，早期的语义嵌入模型为词向量模型，例如Word2Vec (Mikolov et al., 2013)和GloVe (Pennington et al., 2014)等。随着预训练模型如BERT (Devlin et al., 2019)和RoBERTa (Liu et al., 2019)的快速发展，基于预训练语言模型计算得到的语义嵌入包含有更丰富的上下文信息，进一步提升了基于语义向量的评估方法的可靠性。其中g表示待评估文本。c, r表示上下文文本和参考文本，它们在基于语义向量的自动评估方法中通常用于衡量和生成文本的相关性，从而衡量生成文本的质量。需要注意的是，部分基于语义向量的评估方法可能不包含上下文文本或者参考文本，只依赖二者中某一部分具体信息计算和生成文本的相关性。$\mathcal{M}$代表度量语义相似度或者语义距离的计算方法，目前常用的计算方法有点积相似度、余弦相似度、WMS距离以及KL散度等。需要注意的是，大部分基于语义向量的自动评估方法直接计算生成文本和参考文本或上下文信息的语义相似度，无需针对嵌入模型进行额外的微调过程，因此往往都是无监督的。

根据调研，现有的基于语义向量的自动评估方法可以分为如下三类：（1）无需上下文但是基于参考文本的方法（Context-free and Reference-based）；（2）基于上下文但是无需参考文本的方法（Context-based and Reference-free）；（3）基于上下文以及参考文本的方法（Context-based and Reference-based）。

无需上下文但是基于参考文本的方法（Context-free and Reference-based） 无需上下文但是基于参考文本的方法方法旨在直接计算生成文本与参考文本之间的语义相似性，而不考虑上下文信息。例如，ROUGE-WE (Ng and Abrecht, 2015)通过计算参考和生成文本中单词的Word2Vec (Mikolov et al., 2013)词向量的点积相似度来扩展ROUGE指标。此外，Vector Extrema、Embedding Average、Greedy Matching (Liu et al., 2016)通过使用Word2Vec和ELMo模

型获取生成对话回复中单词的语义向量，然后通过不同的聚合策略得到对话回复整体的语义向量，然后通过计算和参考文本的语义相似度来衡量生成文本质量。例如Embedding Average计算句子中所有单词的语义嵌入向量的平均值（即平均池化）以获取句子语义向量。最终的生成对话回复和参考回复的语义向量通过余弦相似度计算得到。BERTScore (Zhang* et al., 2020a)是一种具有代表性的基于语义向量的自动评估方法。其首先使用BERT模型 (Devlin et al., 2019)来获取文本的语义向量，随后通过计算参考文本和生成文本中token对之间的余弦相似性的最大值的加权求和来计算得到最终的自动评估分数，其中逆文档频率（IDF）作为重要性权重。除了BERTScore指标以外，还有大量方法通过计算参考文本和生成文本之间的余弦相似性作为生成文本的质量分数 (Zhao et al., 2023b; Su et al., 2023a; Akula and Garibay, 2022; Lo, 2019)。除了文本语义向量之间的相似度以外，参考文本和生成文本之间的距离度量也广泛用于评估文本之间的相似度。例如，MoverScore (Zhao et al., 2019)和Sentence Mover's Similarity （SMS） (Clark et al., 2019)使用BERT作为嵌入模型，计算生成文本和参考文本之间的Word Move Distance。此外，散度度量也广泛用于基于语义向量的自动评估指标。例如，InfoLM (Colombo et al., 2021)使用不同类型的散度函数来比较参考文本和生成文本的离散概率分布之间的差异。MAUVE (Pillutla et al., 2021)通过计算参考文本和生成文本的混合分布之间的KL散度来衡量两者之间的差异。

基于上下文但是无需参考生成文本的方法（Context-based and Reference-free） 此类方法旨在直接评估生成文本与给定上下文之间的相似性，无需高质量参考文本，因此更适合于生成空间较大的开放式生成任务中。例如，FlowScore (Li et al., 2021)借助DialoFlow对话生成模型，通过计算生成对话回复与基于对话历史预测得到的semantic influence来评估生成对话回复的质量。这种方法能够更直接地捕捉到生成文本与上下文之间的关系，从而提供了一种无需参考文本的评估途径。

基于上下文以及参考文本的方法（Context-based and Reference-based） 基于上下文和参考生成文本的方法旨在同时评估生成文本与上下文以及参考文本之间的相关性。对于开放式对话生成，Frechet Bert Distance（FBD）和Precision-Recall Distance（PRD） (Xiang et al., 2021)通过编码对话历史与生成对话来获取包含对话上下文的语义信息。针对代码生成任务，CodeBERTScore (Zhou et al., 2023)自动评估方法首先将编程问题分别与参考答案和生成答案进行拼接，然后按照BERTScore的计算方式得到最终的评分。这种方法能够有效地结合编程问题的上下文信息，以及生成答案与参考答案之间的语义差异，为代码生成任务提供了一种精准的评估手段。

3.3 基于学习的评估方法（Learning-based Evaluation）

由于仅考虑语义相似度，基于语义向量的模型在处理以下两种情况时存在局限：（1）语义信息可能缺乏一些基本特征，如语法正确性和多样性。因此，即使生成文本与参考文本具有较高的语义相似度，也可能包含严重的错误。（2）对于一些高熵任务，如对话生成和开放式文本生成，有限的参考文本可能无法覆盖所有高质量生成内容的空间，因此自动评估方法容易误判与参考文本不相似的高质量生成内容。针对基于语义向量的自动评估方法的这两类问题，研究人员通过训练神经网络来模仿人类标注者进行人工评估的过程，即基于学习的自动评估方法。具体的，基于学习的评估方法计算过程可以形式定义如下：

$$\text{Score} = \text{Model}(g|c \oplus r) \tag{2}$$

其中，c, r分别表示上下文信息和高质量的参考生成内容。需要注意的是，在某些任务中，上下文信息高质量的参考生成内容并不是必须的。$\oplus$表示文本的拼接操作。其中g是待评估生成文本。Model通常是一个编码文本信息并进行分类的深度神经网络模型。

根据是否利用NLP任务中的上下文信息以及是否依赖高质量参考文本，基于学习的评估方法可以分为以下四类：（1）基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）；（2）无需上下文但是基于参考生成文本的方法（Context-free and Reference-based）；（3）无需上下文以及参考生成文本的方法（Context-free and Reference-free）；（4）基于上下文以及参考生成文本的方法（Context-based and Reference-based）。大部分现有的基于学习的评估方法都在构造的正负样本数据上进行判别式训练。其中，正样本

是指针对任务输入生成的高质量文本，而负样本则是低质量文本。通过训练模型区分高质量与低质量文本，实现对生成内容质量的预测。因此，构建包含正负样本的训练数据是基于学习评估方法的核心。根据数据来源的不同，上述四类基于学习的评估方法可以进一步细分为以下三类：（1）**有监督方法**：通过人工标注的方式筛选和构建训练数据中的负样本。尽管构建数据集的成本较高，但由此产生的数据集噪声较少，质量较高；（2）**自监督方法**：通过负采样 (Tao et al., 2017)来构建负样本。自监督方法可以自动构建数据集，但可能包含噪声，影响基于学习的自动评估模型的性能；（3）**混合方法**：结合了有监督和自监督方法的优点，联合训练基于学习的自动评估方法。

3.3.1 基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）

这类指标在开放领域NLG任务中得到了广泛应用，例如开放域对话生成和开放式文本生成。这类开放式任务的生成空间较大，有限的参考回复往往无法提供足够的信息来全面评估生成文本的质量，因此参考文本对于提升自动评估的可靠性并不显著。

有监督方法 DialogRPT (Gao et al., 2020)通过真实网站的评论数据中的width, depth, up-down属性收集高质量评级数据以优化DialoGPT (Zhang et al., 2020b)对话模型，以对生成对话回复质量评估模型。DEB (Sai et al., 2020)构建了一个包含多个参考回复和高质量负样本的人类标注数据集，用于训练对话响应评估模型。

自监督方法 自监督方法的目标是自动生成负样本，用于训练可学习指标，其中负采样是最广泛使用的技术。RUBERT (Tao et al., 2017)、BERT-RUBER (Ghazarian et al., 2019)和USR (Mehri and Eskenazi, 2020b)通过在随机负采样方法上微调语言模型来构建基于学习的评价模型。负样本和正样本的质量是自监督方法的关键。因此，后续的研究工作旨在提高这些增强样本的质量。 (Zhang et al., 2021; Zhang et al., 2022a; Zhang et al., 2022b; Zhang et al., 2023a; Phy et al., 2020; Wu et al., 2023)。例如，PoNe (Lan et al., 2020)提出了带权负采样和正样本增强来生成更高质量的正负样本用于训练。类似的，MDD-Eval (Zhang et al., 2022a)设计了五种方法构建多样的负样本：（1）语法扰动；（2）back-translation回译；（3）生成模型输出；（4）随机句子选择；（5）遮蔽和填充。UniEval (Zhong et al., 2022)通过设计基于规则的文本转换方法构建负样本，并在多种不同的NLP生成任务和多个维度上构建二元问答任务来对生成文本进行评估。

3.3.2 无需上下文但是基于参考生成文本的方法（Context-free and Reference-based）

类似的，根据正负样本构建方式不同，该类方法可以进一步分为如下三类：（1）有监督方法；（2）自监督方法；（3）混合类型方法。

有监督方法 早期的有监督方法旨在训练LSTM模型对生成式文本进行分类或者回归 (Guzmán et al., 2015; Guzmán et al., 2014; Gupta et al., 2015)。例如，RUSE (Shimanaka et al., 2018)在人类标注的翻译质量分数上训练的回归模型。

自监督方法 除了有监督方法外，自监督类方法也得到了广泛研究 (Jwalapuram et al., 2019; Kamal Eddine et al., 2022; Lu et al., 2023)。例如，SEScore (Xu et al., 2022b)和SEScore2 (Xu et al., 2023b)在检索增强的合成负样本样本上训练评估模型。

混合类型方法 有监督方法通过拟合真实人类的评估数据，在人工评估的相关性上相比自监督方法具有优势。然而有监督方法需要花费大量的时间和代价收集人工标注数据，相比之下，自监督方法通过自动构建负样本数据训练模型，无需人工标注数据即可训练自动评估模型。因此，结合有监督和自监督方法也被广泛的研究。例如BLEURT (Sellam et al., 2020)和MisMATCH (Murugesan et al., 2023)首先在大量的合成参考文本-候选文本对上预训练BERT模型，然后使用人工标注的分数对其进行进一步微调。这种方法结合了有监督和自监督的优势，既利用了大量的未标注数据，又通过少量的人工标注数据提高了模型完成评估任务的准确性和可靠性。

3.3.3 无需上下文以及参考生成文本的方法（Context-free and Reference-free）

同样的，无需上下文以及参考文本的基于学习的自动指标可以分为有监督类型和自监督两类方法。

有监督方法 有监督方法主要应用于文本简化 (Cripwell et al., 2023)、故事生成 (Chen et al., 2022)和作文生成任务的自动评估 (Taghipour and Ng, 2016)。这些方法依赖于大量的人工标注数据来训练模型，以实现对生成文本质量的准确评估。

自监督方法 自监督方法通过使用文本编辑策略对参考高质量参考文本进行扰动，构建具有不同错误类型的负样本，从而在正负样本数据上训练评估模型对文本中的错误进行识别。例如，UNION (Guan and Huang, 2020)在通过四种策略增强的负样本上进行训练，这些策略包括重复、替换、重排序和否定修改，用于训练故事生成的评估模型。

3.3.4 基于上下文以及参考生成文本的方法（Context-based and Reference-based）

相比前三种方法，该方法通过联合考虑上下文和参考文本的信息评估生成文本的质量。大多数此类基于学习的自动评估方法都基于人类标注的样本进行训练的 (Lowe et al., 2017; Maddela et al., 2023; Shi et al., 2023; Chen et al., 2020)。例如，在生成翻译质量的自动评估任务重，COMET (Rei et al., 2020)的两种COMET变体，COMET-MQM和COMET-DARR在人工标注的机器翻译语料库上训练，可以同时考虑输入源句子和高质量参考翻译的信息来评估生成译文的质量。这种方法能够更全面地捕捉到翻译质量的多维度特征，从而提供了一种更精准的评估手段。

3.4 基于大规模语言模型的评估方法（LLM-based Evaluation）

尽管大量工作证明了基于学习的方法已经实现了与人工评估的高度相关性和一致性，但它们严重依赖于高质量和多样化的训练数据。尤其是随着NLP任务的快速发展，研究者们不再满足于对生成质量的粗粒度评估，针对不同NLP任务的多维度和细粒度的质量评估已经成为重要的发展趋势。例如，在机器翻译任务中，研究者们更加关注翻译的流畅性、准确性以及其他细粒度维度上的质量。为了解决这些问题，基于学习的方法不可避免地需要收集大量高质量的正负样本数据以训练优化具体的自动评估模型。然而，在多个不同任务和评估维度上收集高质量的正负样本数据是一个非常困难的工作，这对基于学习的自动评估指标的扩展性提出了重大挑战。因此，自动评估指标的通用性和可扩展性已经成为自动评估领域中最重要的需求，这促使研究者们转向另一种强大通用和普遍的技术，以构建可靠和通用的自动评估方法，即基于大规模语言模型的（LLM-based）评估。

参考GPTScore(Fu et al., 2023)的形式化定义，基于大规模语言模型的生成式评估方法一般可以表示为：

$$E_i = \mathcal{D}\left(\text{LLM}\left(T(t, c, \mathcal{S}) \oplus E_{<i}\right)\right) \tag{3}$$

其中，t为待评估的任务描述，c为评估标准（criteria），$\mathcal{S}$为上下文或参考文本，在部分方法中可能为空；$T(\cdot)$使用上述内容构造评估提示词，构造方式一般与具体任务相关；$E_{<i}$代表在第i步已生成的评估文本，E_i代表此时评估文本的下一个词元；$\oplus$代表文本的拼接操作；LLM$(\cdot)$代表利用大规模语言模型获取下一词元在词表上的概率分布，$\mathcal{D}$代表具体的解码算法，一般常用为greedy search (Li et al., 2023a)。最终生成的评估内容用$E = \{E_i\}_{i=1}^{L}$表示，可以包含评估分数、排序、偏好标签、文本形式的评估解释等，其中L为生成的评估文本的序列长度。

相比于基于学习的评估方法，基于大规模语言模型的自动评估方法无需收集训练数据来微调模型。通过prompt工程的方式，基于大规模语言模型的自动评估方法可以灵活地处理各种不同的评估任务。例如，在评估方式上，大规模语言模型可以直接评估单一生成文本的质量（single-wise的评估），也可以比较一对文本的质量（pair-wise的评估） (Li et al., 2024a)。同时，由于大规模语言模型通常在包含多种不同语言的训练数据上进行预训练，因此它可以直接评估不同语言中生成文本的质量。

基于大规模语言模型的评估方法同样可以根据是否使用参考文本与是否依赖上下文信息两个维度分为如下三类：（1）基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）；（2）基于上下文以及参考生成文本的方法（Context-based and Reference-free）；（3）无需上下文的方法（Context-free）。

3.4.1 基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）

在生成式文本的自动评估任务中，大规模语言模型的引入使得评估过程可以借助其蕴含的常识性知识与语义理解能力，从而让许多场景下的评估任务摆脱对参考文本的依赖成为可能。

同时，大规模语言模型捕捉上下文语义关系的能力需要借助充足的上下文信息才能得以更好地展现。因此，基于大规模语言模型的评估方法大多需要上下文信息而无需借助参考文本。

在开放性故事生成与对抗性攻击任务的评估中，有研究(Chiang and yi Lee, 2023)率先使用大规模语言模型替代人类评估，证明了大规模语言模型与人类专家评估结果的一致性。由于部分使用大规模语言模型的评估方法与人类的关联度仍不及中等大小的神经网络评估器，G-Eval(Liu et al., 2023b)在提示词中加入思维链(Wei et al., 2023)与范例填充机制缓解这一问题。

基于大规模语言模型的成对评估方法往往存在自强化、位置偏见等影响评估质量的缺陷(Li et al., 2023c)。为此，PRD(Li et al., 2023c)提出了同行分级与同行讨论的方法，通过多轮迭代确定每个模型打分的权重，再经过多个模型讨论得到评估结果。(Bai et al., 2023b)使用大规模语言模型生成辅助评估的问题，在多轮问答中利用去中心化的同行检查机制减轻缺少领域知识与单一模型潜在偏见对评估结果的影响。ChatEval(Chan et al., 2023)通过构建多智能体的裁判团队缩小了单智能体方法与人类评估水平的差距。

由人类为大规模语言模型评估任务撰写的提示词可能包含潜在的偏见，且评估结果关于提示词形式的敏感度尚未明确。为解决这一问题，AutoCalibrate(Liu et al., 2023d)借助人类专家的评估样例，通过多阶段的标准起草与修改，实现了基于大规模语言模型的评估方法与人类偏好的自动对齐。SocREval(He et al., 2024)引导大规模语言模型对评估的问题生成自己的答案作为参考，根据该答案对待评估的文本做定性分析，实现对推理的评估。WideDeep(Zhang et al., 2023b)在使用大规模语言模型生成评估标准的基础上，组建大规模语言模型网络，进行多轮评估与同行检查。

此外，基于大规模语言模型的评估方法也可以加入其它技巧以优化在特定任务上的表现。一项文本风格迁移任务的评估(Ostheimer et al., 2023)对提示词进行集成，提升了评估方法的健壮性以及与人类评估的关联度。该工作还发现经过指令微调的大规模语言模型更能胜任评估类任务。DeltaScore(Xie et al., 2023)通过计算打乱前后文本的似然度差值体现生成的故事情节的质量。BookScore(Chang et al., 2024)将大规模语言模型作为分类器，通过在人为预定义的错误类别上对句子进行分类，评估超长文本摘要的句子级别连贯性。

3.4.2 基于上下文以及参考生成文本的方法（Context-based and Reference-based）

由于大规模语言模型本身具有较强的语义理解能力，因此在评估过程中参考文本一般用于指示生成文本中应该包含的关键信息。在长文本摘要评估的研究中(Wu et al., 2024)，为降低评估过程中的模型推理成本、缓解长文本中部信息被忽视的问题，首先提取长文本中的关键句子作为参考文本，然后借助大规模语言模型完成评估。RAGAS(Es et al., 2023)在检索增强生成任务的评估工作中借助大规模语言模型对生成文本从忠实性、答案相关性、上下文相关性三个维度进行评估。大规模语言模型在评估过程中还可以根据单一的参考文本，生成多个表达方式不同的参考文本覆盖参考文本的语义空间，避免因生成文本与参考文本表达方式不同导致的评估结果与人类不一致的问题(Tang et al., 2024)。

3.4.3 无需上下文的方法（Context-free）

在基于大规模语言模型的评估方法中，不引入上下文的方法一般关注于生成文本的逻辑性、语义连贯性等方面。此类方法的关注点较为单一，适用范围小，数量较少。例如，BookScore(Chang et al., 2024)着重关注长文本摘要本身的句子级别连贯性，该方法将大规模语言模型用作连贯性错误的分类器，用不包含连贯性错误的句子在所有句子中的占比表征摘要的语义连贯性。

3.4.4 基于大规模语言模型自动评估技术的总结

虽然基于大规模语言模型的自动评估技术已经取得了和人工评估较高的相关性和一致性，但其在应用过程中依然面临如下五个严重的问题:

- 长度偏差（Length Bias）：近期研究指出，基于大规模语言模型的自动评估方法对生成文本的长度非常敏感。大规模语言模型倾向于给予文本长度更长、内容更丰富的待评估文本更高的评分 (Zheng et al., 2023; Zeng et al., 2023)。

- 位置偏差（Positional Bias）：在pair-wise的评估方式下，大规模语言模型常用于直接比较两段生成文本的质量。然而，已有研究指出，大规模语言模型更倾向于在prompt中特定位

置出现的待评估文本 (Zheng et al., 2023; Zeng et al., 2023; Wang et al., 2023d)。这一问题显著影响了基于大规模语言模型在实际应用过程的可靠性。

- 自加强偏差（Self-enhancement Bias）：大规模语言模型尤其倾向于偏向于自己生成的文本内容，因此针对其自己生成的文本内容评估往往会存在评分偏高的问题 (Zheng et al., 2023)。
- 尽管大规模语言模型在开放式文本生成任务的自动评估中取得了与人工评估较强的相关性，但其针对逻辑推理类任务的评估能力依然薄弱，例如解数学题、代码题、推理问答等 (Valmeekam et al., 2023; Huang et al., 2023a)。
- 不一致的评估内容：大规模语言模型本身具有明显的不一致性。经过多次解码推理得到的生成式评估文本存在较为明显的不一致信息，这显著影响了大规模语言模型的评估的可靠性 (Zhang et al., 2024)。

虽然已有工作证明了使用few-shot例子、高质量参考文本和严格的打分标准（criteria）可以提升基于大规模语言模型的自动评估方法的可靠性 (Kim et al., 2024; Fu et al., 2023; Li et al., 2024b; Li et al., 2024a)，上述这些问题对基于大规模语言模型的自动评估方法的影响依然非常严重，因此未来的工作需要重点关注和解决这些问题，以进一步提升基于大规模语言模型的自动评估方法的可靠性。

4 自动评估主要发展趋势总结

生成式文本的自动评估技术经历了从启发式评估、基于语义向量的评估、基于学习的评估到基于大规模语言模型评估的四轮技术范式的演变。通过分析已有自动评估工作，本节将总结和探讨针对生成式文本自动评估技术的三个核心发展趋势：（1）通用性；（2）可解释性；（3）评估维度的多样性。

通用性 随着自动评估技术的不断发展和完善，其评估的通用性得到了显著增强。早期的研究主要聚焦于单一语言和特定任务的评估，例如机器翻译、对话生成和摘要生成等任务的自动评估 (Papineni et al., 2002; Tao et al., 2017)。然而，这些针对特定任务开发的自动评估指标往往难以快速适应和迁移到新的自然语言处理（NLP）任务中，这一局限性在一定程度上阻碍了文本生成任务领域的进一步发展。为了克服这一挑战，研究人员开始探索更具通用性和迁移能力的自动评估指标。特别是在基于学习的自动评估方法中，研究人员通过在包含更多样化文本生成任务的数据上训练评估模型，使其能够同时适用于多种不同的任务。例如，UniEval (Zhong et al., 2022)通过在自然语言推理（NLI）、问答（QA）和情感分析（SST）等任务数据上构建负样本并联合训练自动评价模型，实验结果表明UniEval能够快速适应并泛化至训练阶段未曾接触的其他文本生成任务。近期，借助大规模语言模型强大的泛化能力，基于大规模语言模型的自动评估技术已经能够有效地处理多种语言和多种不同文本生成任务的评估 (Mehri and Eskenazi, 2020a; Zhong et al., 2022)。这种通用性的提升使得同一个模型能够用于评估多种不同的任务，展现了极大的灵活性和便捷性。

可解释性 自动评估技术的发展经历了从早期范式到基于大规模语言模型的范式的转变，显著提高了其可解释性。启发式自动评估、基于语义向量的自动评估和基于学习的自动评估方法能够对生成式文本质量进行评分，但无法提供全面且细粒度的评估内容。近期的基于大规模语言模型的评估方法利用大规模语言模型强大的文本生成能力，对生成文本的质量进行细粒度分析。这些方法不仅能指出错误内容，还能提出可行的改进建议，极大地提升了评估技术的可解释性。此外，这些丰富的细粒度解释信息甚至有助于大规模语言模型利用这些信息进一步增强其生成能力 (Yuan et al., 2024b)，对于实现大规模语言模型的自我提升具有重要价值。

评估维度的多样性 自动评估技术已从单一维度的评估发展到多维度全面评估 (Fu et al., 2023)。针对特定的下游任务，从多个不同角度评估生成质量能够提供更全面、更细粒度的视角，以捕捉生成文本在不同方面的质量信息，从而实现更准确有效的生成文本质量评估。例如，针对开放式文本生成的自动评估方法最初主要关注生成的相关性。然而，随着生成质量和流畅性的提升，研究人员开始更多地关注生成文本内部的一致性以及信息的丰富程

度 (Su et al., 2022b; Lan et al., 2022)。基于学习的自动评估方法主要通过设计针对维度的负样本的方式，训练自动评估模型关注特定的评估维度质量。例如，MDD-Eval (Zhang et al., 2022a)和UNION (Guan and Huang, 2020)通过设计多种不同类型的方法构建负样本训练模型，以关注生成文本中的错误信息。近期，大量工作通过在prompt中设计和定义特定的评估维度，从而让大规模语言模型在评估过程中关注特定的维度的质量信息 (Fu et al., 2023; Li et al., 2024a)。更进一步的，让模型直接学习生成和构建层次化的评估维度已经被证明有助于大规模语言模型生成更可靠的评估内容 (Yuan et al., 2024a; Hu et al., 2024)。

5 未来自动评估技术的研究方向

从基于启发式的评估范式到基于大规模语言模型的评估方式，生成式文本的自动评估技术得到了快速发展。本章节将探讨和展望自动评估技术未来的四个重要的研究方向：（1）小模型评估能力的提升；（2）评估自动评估的质量；（3）评估技术的应用：大规模语言模型的自我提升；（4）多模态生成内容的评估。

5.1 小模型自动评估能力的提升

大量研究已经证实，基于大规模语言模型的自动评估方法与人工评估表现出高度的一致性。然而，大规模语言模型在推理过程中会产生高昂的开销，这对于实现高效的自动评估构成了挑战 (Wang et al., 2024)。然而，目前的小规模语言模型在自动评估能力方面仍存在不足，其评估分数与人工评价分数之间的相关性明显弱于大规模语言模型 (Li et al., 2023b)。因此，针对性地提升小模型的自动评估能力将成为未来研究的重要趋势。目前，提升小模型自动评估能力的主要方法是通过知识蒸馏，将先进语言模型（如GPT-4）的自动评估能力通过微调的方式蒸馏到小模型上 (Cui et al., 2023; Li et al., 2023b; Wang et al., 2023c)。具体而言，首先使用GPT-4对待评估的查询-响应（query-response）配对数据生成自动评估内容，这些内容可能包括可解释的评估分析、评估分数等。通过收集大量的自动评估数据，然后微调参数量较小、推理效率较高的开源模型，以实现对其自动评估能力的增强。尽管当前的技术方案可以在一定程度上提升小模型的自动评估能力，但它们与先进语言模型之间的差距仍然较大，并且其评估能力的通用性依然受限。具体来说，当前技术方案的通用性面临以下三个重要问题亟待解决：（1）如何收集多样化的下游任务数据，以提高小模型自动评估能力的通用性；（2）如何涵盖更多样的评估设置，以适应不同的评估环境和需求；（3）如何涵盖更多样的自动评估维度，以全面评估生成文本的质量。

根据第4节的总结，提升自动评估技术的通用性是生成式文本的自动评估方法的一个关键的发展趋势。目前用于提升小模型自动评估能力的方法所使用的数据集，在数据规模、数据类别、评估设置、评估维度方面都存在一定的局限性，这对自动评估模型的通用性产生了负面影响。因此，未来的工作重点应当是收集涵盖更多任务、语言和评估设置的多样化自动评估数据，以用于微调模型，从而增强其评估的通用性 (Wei et al., 2022; Li et al., 2024a; Li et al., 2024b)。目前已有工作开始尝试解决这一问题。例如，在增加数据任务的多样性方面，Auto-J(Li et al., 2024a)涵盖了58个不同的生成场景，并构建了由GPT-4生成的高质量的可解释评估数据，用以微调Llama-2-13B模型，从而使其评估能力超过了GPT-3.5-turbo模型。此外，Auto-J同时具备对单一待评估文本和一对评估文本同时进行评估的能力，显著增强了其评估设置的多样性。针对评估维度的多样性方面，Prometheus (Kim et al., 2024)通过引入评估打分维度的问题，针对性地对生成回复文本的质量进行评估，展现了较强的灵活性。

5.2 如何评估可解释自动评估的质量

当前基于大规模语言模型的自动评估技术通过生成可解释的评估分析内容，实现了细粒度和可解释的自动评估。然而，对生成可解释评估内容质量的自动评价仍然是一个尚未充分探索的领域。作为一种特殊的文本生成任务，生成可解释评估分析内容的质量自动评价相比传统的文本生成任务更具挑战性。目前的研究发现，使用高质量的参考评估内容作为提示，大规模语言模型对生成的可解释自动评估内容的评价分数与人工评估结果具有较高的相关性。例如，CriticEval(Lan et al., 2024)的消融实验表明，移除高质量的参考评估文本会导致GPT-4等大规模语言模型在自动评价可解释评估内容方面的能力显著下降（平均相关性损失达19.3%）。此外，MetaCritique(Sun et al., 2024)的实验发现，通过将高质量参考评估和生成

评估拆解为原子信息单元（Atomic Information Unit，AIU），并使用GPT-4逐一验证生成评估中的AIU是否与参考评估的AIU匹配，可以实现与人工评估的高度一致性。这些研究都证实了高质量参考可解释评估文本在自动评价可解释自动评估内容中的重要作用。然而，高质量的参考可解释评估文本需要通过严格的人工标注来收集，这对自动评估可解释评估质量的方法的扩展性构成了严重挑战。因此，未来该领域的一个重要研究方向在于如何减少对人工高质量评估数据收集的依赖，从而进一步提升评估解释评估的有效性和可扩展性。

5.3 评估技术的应用：大规模语言模型的自我提升

除了降低人工评估开销和负担，自动评估技术也已被广泛应用于大规模语言模型（LLM）的自动提升中，（Yuan et al., 2024b; Xu et al., 2024），这主要体现在两个重要的阶段：（1）**推理阶段**：利用LLM的生成能力和自动评估能力，分析生成内容中的缺陷并提供改进建议，可以迭代改进生成回复的质量 (Saunders et al., 2022; Zhang et al., 2024; Madaan et al., 2023; Fernandes et al., 2023; Yao et al., 2023)；（2）**训练阶段**：自动评估的打分通常用于构建具有明确性能差距的文本数据，这类数据可以用于使用拒绝式微调（RFT）或偏好学习（RLHF (Lee et al., 2023)）的方式实现对大规模语言模型的进一步提升，提高LLM能力 (Yuan et al., 2024b; Xu et al., 2023a; Bowman et al., 2022; Bai et al., 2022; Xu et al., 2024)。例如，Self-rewarding (Yuan et al., 2024b)通过使用自己生成的自动评估分数对Llama-2-70B (Touvron et al., 2023)进行微调，从而进一步改进了Llama-2-70B的质量。类似地，ChatGLM-Math（Xu et al., 2024)通过微调数学批评模型对生成的答案的质量生成答案，通过拒绝式微调 (Touvron et al., 2023)和直接偏好优化模型的效果(Rafailov et al., 2023)。直觉上，当前大规模语言模型自我提升的有效性主要取决于模型自动评估的准确性，即具有较强反思能力的语言模型在自我提升方面潜力更大。然而，这一结论尚未得到证实。未来的工作重点在于分析大规模语言模型自我提升能力和自动评估能力的相关性。

5.4 多模态生成内容的评估

同文本生成类似，多模态场景的生成任务也面临着缺乏有效自动化评估指标的挑战。例如，随着文生图 (Rombach et al., 2021)、文生视频以及多模态对话问答 (Su et al., 2023b; Su et al., 2022a)领域技术的快速发展，生成图片和视频的质量以及对多模态输入内容的问答对话能力逐步提高。如何准确地自动化评估生成图片和视频的质量已成为核心难题之一。当前，已有部分研究在不同模态的自动评估上做出了初步尝试。接下来，本节将总结当前多模态领域自动评估技术的发展状况，并对其未来的发展趋势进行预测。

语音生成自动评估 生成语音的主流质量评估方法是基于人类听觉判断的大规模众包平均意见得分（MOS），这种方法测试成本昂贵且要求测试者具有一定的听感训练基础。因此，模仿MOS评价的生成语音质量自动评估也成为当前语音合成领域发展的核心问题之一。目前语音生成自动评估的发展受到了两个主要挑战的影响：（1）自动评价指标的设计依赖于针对合成语音质量的人工评价数据，这使得该任务数据构建的成本同样高昂；（2）合成语音的评价需要考虑多个维度，包括自然度、可懂度和是否符合预期的听觉效果，这些维度在目前的自动评估中难以全面体现。AutoMOS(Patton et al., 2016)对单元选择合成器生成的语音自动评估进行了初步的尝试，但其作用主要停留在辅助人工评测，提前选择对听者更有优势的语音样本。MOSNET(Lo et al., 2019)是首个直接评分的语音生成自动评估系统，其在2018年的语音转换竞赛VCC数据集上进行了训练。研究者发现，该方法在不同数据集评估上的泛化性较差，因此需要更多的数据来奠定训练基础。

目前语音合成自动评估的数据主要依托于主流的语音主观评估指标，即生成语音及其在不同听者的MOS评分数据。VoiceMOS系列(Huang et al., 2022a; Cooper et al., 2023)语音自动评估比赛的举办提供了丰富的人工评估数据，同时三星公司标注的SOMOS(Maniati et al., 2022)也标注了MOS评分数据集。这些数据集的构建进一步促进了语音生成自动评估技术的有效发展。以此为基础，基于学习的自动评估方法大量涌现，基于听感学习有MBNET(Leng et al., 2021)、LDNET(Huang et al., 2022b)、UTMOS(Saeki et al., 2022)、DDOS(Tseng et al., 2022)、MOSA-Net(Zezario et al., 2022)等系列工作，基于自监督学习的方法有SSL-MOS(Cooper et al., 2022)及其后续提升版本ZevoMOS(Stan, 2022)、LE-SSL-MOS(Qi et al., 2023)等。随着检索增强方法在NLP领域大放异彩(Lewis et al., 2020)，在自监督学习方法的

基础上，研究者提出了检索增强的MOS预测方法RAMP(Wang et al., 2023a)。也有研究者在基于语义向量嵌入表示的语音生成自动评估进行了尝试，其利用文本预训练语言模型理解和量化语音的质量，SpeechLMScore(Maiti et al., 2023)测量合成音频样本与预训练在自然语音上的生成语音单元语言模型的似然性，SpeechBERTScore(Saeki et al., 2024)受到了BERTScore(Zhang et al., 2019)生成式文本自动评估指标的设计启发，计算生成语音和参考语音的自监督密集语音特征的BERTScore，从而获得针对不同质量的语音有更好的泛化和鲁棒效果。此外，SQuId(Sellam et al., 2023)把研究角度扩展到了零样本情况下的多语言语音评估，PAM(Deshmukh et al., 2024)则默认语音和文本对比学习模型CLAP学习到大量的音频样本和描述该音频质量的反义词对（例如“清晰”和”模糊”），其将研究角度迁移到针对更广泛音频生成内容（包括语音、声音事件、噪声等）的评估。

目前语音合成的自动评估方法在通用性、泛化性和可解释性上依然不足，随着针对语音理解的大规模多模态语言模型的兴起(Latif et al., 2023)，如何利用这些模型来提高语音自动评估的通用性和可解释性以及评估维度的多样性，将成为重要的研究方向。例如，对情绪(Zhou et al., 2022)、口音(Liu et al., 2024; Zhou et al., 2024)、言语障碍(Huang et al., 2022c)等语音合成新兴方向的自动评价，应该是未来探索的有效方向。这些探索不仅能够提高语音合成的质量，还能推动自动评估技术的发展，使其更加精确和全面。

视觉内容生成自动评估 在视觉内容生成任务的自动评估相比文本生成任务的更加困难，其中视觉内容生成包含图片、视频作为输入或输出的生成任务，如文生图、文生视频、VQA (Antol et al., 2015)。目前，视觉内容生成任务可以分为如下两类：（1）视觉内容输入端任务：提供图片或者视频信息，针对问题或者对话历史进行回复、生成图片视频描述等 (Su et al., 2023b)；（2）视觉内容输出端任务：根据文本描述，合成图片或者视频 (Antol et al., 2015)。

目前，针对视觉内容输入的任务，自动评估最常用的方法是依赖标准答案的评估。这类方法主要评估跨模态语言模型 (Su et al., 2023b)理解视觉多模态信息的能力，常用的人工标注的测试数据基准有MMBench (Liu et al., 2023c)和MMMU (Yue et al., 2024)等。此外，由于存在标准的参考文本信息，针对文本的自动评估方法也常用于图像字幕（Image Caption）任务，以评估模型根据输入图片生成的描述文本的质量 (Xu et al., 2023b; Hessel et al., 2022)。近期，随着以GPT-4V等先进的跨模态语言模型技术的快速发展，跨模态语言模型已被广泛应用于针对视觉输入内容生成的开放式文本内容质量的自动评估 (Lu et al., 2024)。

目前，视觉内容输出端评估主要采用无标准答案的自动评估方法。已有的无标准答案的视觉内容生成的自动评估技术也可以分为四类：（1）启发式自动评估 (Marcos-Morales et al., 2023; Zhou et al., 2019)；（2）基于语义向量的评估，如FID (Seitzer, 2020)，FVD (Unterthiner et al., 2018)；（3）基于学习的自动评估 (Qin et al., 2023; Wang et al., 2023b)；（4）基于大规模跨模态语言模型的自动评估 (Zhang et al., 2023c; Ge et al., 2023; Ku et al., 2023; Lu et al., 2024; Lee et al., 2024)。以GPT-4v和GPT-4o为代表的跨模态大规模语言模型的快速发展也显著促进了针对视觉内容生成的自动评估技术的进步。然而，已有的工作注意到 (Lu et al., 2024; Ku et al., 2023)，GPT-4V等跨模态语言模型对生成图像的理解能力依然不足。因此，未来的工作应更关注进一步提升视觉跨模态语言模型对图像内容的理解和推理能力，从而为构建更准确的基于大规模跨模态语言模型的视觉内容生成自动评估方法打下基础。

综上，相对于生成式文本自动评估，多模态生成内容的自动评估发展较为缓慢，未来的工作可以借鉴文本领域的研究成果，进一步促进多模态领域自动评估技术的发展。

6 总结

随着大规模语言模型技术的飞速发展，自动评估已成为自然语言处理（NLP）领域的核心难题之一。本文系统地梳理了生成式文本自动评估技术的四次技术范式的变革，对各个技术范式的特点和代表性工作进行了详细的梳理和分析。基于这些分析内容，本文总结并提出了关于自动评估技术的主要发展趋势，即自动评估方法在通用性、可解释性和评估维度多样性上的持续进步。在此基础上，本文进一步对自动评估领域未来的研究方向进行了总结和展望。

第二十三届中国计算语言学大会论文集，第169页–第196页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

参考文献

Ramya Akula and Ivan Garibay. 2022. Sentence pair embeddings based evaluation metric for abstractive and extractive summarization. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6009–6017, Marseille, France, June. European Language Resources Association.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023b. Benchmarking foundation models with language-model-as-an-examiner.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online, November. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Pierre Colombo, Chloe Clave, and Pablo Piantanida. 2021. Infolm: A new metric to evaluate summarization & data2text generation. In *AAAI Conference on Artificial Intelligence*.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2023. The voicemos challenge 2023: zero-shot subjective speech quality prediction for multiple domains. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore, December. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. 2024. Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy, July. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online, November. Association for Computational Linguistics.

Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, et al. 2023. Mllm-bench, evaluating multi-modal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf, editors, *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online, November. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 6394–6407, Online, August. Association for Computational Linguistics.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal, September. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to differentiate better from worse translations. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220, Doha, Qatar, October. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China, July. Association for Computational Linguistics.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy, July. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024. Socreval: Large language models with the socratic method for reference-free reasoning evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria?

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022a. The voicemos challenge 2022. *arXiv preprint arXiv:2203.11389*.

Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. 2022b. Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 896–900. IEEE.

Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda. 2022c. Towards identity preserving normal to dysarthric voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6672–6676. IEEE.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), pages 2964–2975, Hong Kong, China, November. Association for Computational Linguistics.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland, May. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland, May. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1), nov.

Tian Lan, Yixuan Su, Shuhang Liu, Heyan Huang, and Xian-Ling Mao. 2022. Momentum decoding: Open-ended text generation as graph exploration. *arXiv preprint arXiv:2212.02175*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2023. Towards efficient coarse-grained dialogue response selection. *ACM Trans. Inf. Syst.*, 42(2), sep.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024. Criticbench: Evaluating large language models as critic. *arXiv preprint arXiv:2402.13764*.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*.

Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. 2021. Mbnet: Mos prediction for synthesized speech with mean-bias network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 391–395. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 128–138, Online, August. Association for Computational Linguistics.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023a. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment.

Ruosen Li, Teerth Patel, and Xinya Du. 2023c. Prd: Peer rank and discussion improve large language model based evaluations.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024b. Dissecting human and llm preferences.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2019. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Dublin, Ireland, May. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023d. Calibrating llm-based evaluator.

Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. 2024. Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, August. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July. Association for Computational Linguistics.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2023. Toward human-like evaluation for natural language generation with error analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada, July. Association for Computational Linguistics.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada, July. Association for Computational Linguistics.

François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. 2023. Speechlmscore: Evaluating speech generation using speech language model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. *arXiv preprint arXiv:2204.03040*.

Adria Marcos-Morales, Matan Leibovich, Sreyas Mohan, Joshua Lawrence Vincent, Piyush Haluai, Mai Tan, Peter Crozier, and Carlos Fernandez-Granda. 2023. Evaluating unsupervised denoising requires unsupervised metrics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online, July. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. Evaluating the knowledge dependency of questions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10512–10526, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland, May. Association for Computational Linguistics.

Keerthiram Murugesan, Sarathkrishna Swaminathan, Soham Dan, Subhajit Chaudhury, Chulaka Gunasekara, Maxwell Crouse, Diwakar Mahajan, Ibrahim Abdelaziz, Achille Fokoue, Pavan Kapanipathi, Salim Roukos, and Alexander Gray. 2023. MISMATCH: Fine-grained evaluation of machine-generated text with mismatch error types. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4485–4503, Toronto, Canada, July. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal, September. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. Text style transfer evaluation using large language models.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A Saurous, and D Sculley. 2016. Automos: Learning a non-intrusive assessor of naturalness-of-speech. *arXiv preprint arXiv:1611.09207*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Adithya Pratapa, Antonios Anastasopoulos, Shruti Rijhwani, Aditi Chaudhary, David R. Mortensen, Graham Neubig, and Yulia Tsvetkov. 2021. Evaluating the morphosyntactic well-formedness of generated texts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7131–7150, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zili Qi, Xinhui Hu, Wangjin Zhou, Sheng Li, Hao Wu, Jian Lu, and Xinkang Xu. 2023. Le-ssl-mos: Self-supervised learning mos prediction with listener enhancement. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. 2023. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2091–2100.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. *arXiv preprint arXiv:2401.16812*.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August. Version 0.3.0.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P Parikh, and Jason Riesa. 2023. Squid: Measuring speech naturalness in many languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren, and Zhaochun Ren. 2023. RADE: Reference-assisted dialogue evaluation for open-domain dialogue. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12856–12875, Toronto, Canada, July. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels, October. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Adriana Stan. 2022. The zevomos entry to voicemos challenge 2022. *arXiv preprint arXiv:2206.07448*.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022a. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022b. A contrastive framework for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561. Curran Associates, Inc.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022c. A contrastive framework for neural text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023a. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada, July. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023b. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024. The critique of critique.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, Tom Kocmi, and Furu Wei. 2024. Not all metrics are guilty: Improving nlg evaluation by diversifying references.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI Conference on Artificial Intelligence*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Phua Yeong Tsann, Yew Kwang Hooi, Mohd Fadzil Bin Hassan, and Matthew Teow Yok Wooi. 2022. Embeddingrouge: Malay news headline similarity evaluation. In *2022 International Conference on Digital Transformation and Intelligence (ICDI)*, pages 01–06.

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung-yi Lee. 2022. Ddos: A mos prediction framework utilizing domain adaptive pre-training and distribution of opinion scores. *arXiv preprint arXiv:2204.03219*.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023a. Ramp: Retrieval-augmented mos prediction via confidence-based dynamic weighting. *arXiv preprint arXiv:2308.16488*.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023b. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023c. Shepherd: A critic for language model generation.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023d. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Hanming Wu, Wenjuan Han, Hui Di, Yufeng Chen, and Jinan Xu. 2023. A holistic approach to reference-free evaluation of machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 623–636, Toronto, Canada, July. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by llms.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online, August. Association for Computational Linguistics.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Han Lau. 2023. Deltascore: Fine-grained story evaluation with perturbations.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Chen Xu, Jianyu Zhao, Rang Li, Changjian Hu, and Chuangbai Xiao. 2021. Change or not: A simple approach for plug and play language models on sentiment control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15935–15936.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022a. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022b. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6559–6574, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Weiwen Xu, Deng Cai, Zhisong Zhang, Wai Lam, and Shuming Shi. 2023a. Reasons to reject? aligning language models with judgments.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023b. SESCORE2: Learning text generation evaluation via synthesizing realistic mistakes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5166–5183, Toronto, Canada, July. Association for Computational Linguistics.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline.

Haoran Yang, Yan Wang, Piji Li, Wei Bi, Wai Lam, and Chen Xu. 2023. Bridging the gap between pretraining and fine-tuning for commonsense generation. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 376–383, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024a. Llmcrit: Teaching large language models to use criteria.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Jerrold Zar, 2005. *Spearman Rank Correlation*, volume 5. 07.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Ryandhimas E Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao. 2022. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:54–70.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 5676–5689, Online, August. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022a. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11657–11666.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022b. FineD-eval: Fine-grained automatic dialogue-level evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2023a. Poe: A panel of experts for generalized automatic dialogue assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1234–1250.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023b. Wider and deeper llm networks are fairer llm evaluators.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023c. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, November. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Wei Zhao, Michael Strube, and Steffen Eger. 2023b. DiscoScore: Evaluating text generation with BERT and discourse coherence. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Giulio Zhou and Gerasimos Lampouras. 2020. Webnlg challenge 2020: Language agnostic delexicalisation for multilingual rdf-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 186–191.

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li. 2022. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*.

Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. CodeBERTScore: Evaluating code generation with pretrained models of code. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13921–13937, Singapore, December. Association for Computational Linguistics.

Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li. 2024. Accented text-to-speech synthesis with limited data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711.