文本、视觉和语音生成的自动评估方法综述

TIAN LAN*, Beijing Institute of Technology, China
YANG-HAO ZHOU*, Beijing Institute of Technology, China
ZI-AO MA*, Beijing Institute of Technology, China
FANSHU SUN*, Beijing Institute of Technology, China
RUI-QING SUN*, Beijing Institute of Technology, China
JUNYU LUO, Peking University, China
RONG-CHENG TU, Nanyang Technological University, Singapore
HEYAN HUANG, Beijing Institute of Technology, China
CHEN XU, Beijing Institute of Technology, China
ZHIJING WU, Beijing Institute of Technology, China
XIAN-LING MAO†, Beijing Institute of Technology, China

深度学习的最新进展显著增强了生成式人工智能在文本、图像和音频方面的能力。然而，自动评估这些生成输出的质量仍然是一个持续的挑战。尽管存在许多自动评估方法，但是当前的研究缺乏一个系统的框架来全面组织这些在文本、视觉和音频模态中的方法。为了解决这个问题，我们对所有三种模态下生成内容的自动评估方法进行了全面回顾并提出了一个统一的分类法；我们确定了五个基本范式，描述了这些领域中现有评估方法的特征。我们的分析首先从文本生成的评估方法开始，在这方面技术最为成熟。然后，我们将这一框架扩展到图像和音频生成，以展示其广泛的适用性。最后，我们讨论了跨模态评估方法研究的未来发展方向。

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for,Your, Paper

ACM Reference Format:

Tian Lan, Yang-Hao Zhou, Zi-Ao Ma, Fanshu Sun, Rui-Qing Sun, Junyu Luo, Rong-Cheng Tu, Heyan Huang, Chen Xu, Zhijing Wu, and Xian-Ling Mao. 2018. 文本、视觉和语音生成的自动评估方法综述. In *Proceedings of Make sure to enter the correct conference*

*Equal Contribution
†Xian-Ling Mao is the corresponding author of this paper.

Authors' Contact Information: Tian Lan, lantiangmftby@gmail.com, Beijing Institute of Technology, Beijing, Beijing, China; Yang-Hao Zhou, zhouyh77@bit.edu.cn, Beijing Institute of Technology, Beijing, Beijing, China; Zi-Ao Ma, maziaoylwt@gmail.com, Beijing Institute of Technology, Beijing, Beijing, China; Fanshu Sun, sunfs@bit.edu.cn, Beijing Institute of Technology, Beijing, Beijing, China; Rui-Qing Sun, 2325557558@qq.com, Beijing Institute of Technology, Beijing, Beijing, China; Junyu Luo, luojunyu@stu.pku.edu.cn, Peking University, Beijing, Beijing, China; Rong-Cheng Tu, rongcheng.tu@ntu.edu.sg, Nanyang Technological University, Singapore, Singapore, Singapore; Heyan Huang, Beijing Institute of Technology, Beijing, Beijing, China, hhy63@bit.edu.cn; Chen Xu, Beijing Institute of Technology, Beijing, China, chenxu05037@bit.edu.cn; Zhijing Wu, Beijing Institute of Technology, Beijing, China, zhijingwu@bit.edu.cn; Xian-Ling Mao, Beijing Institute of Technology, Beijing, China, maoxl@bit.edu.cn.

Manuscript submitted to ACM

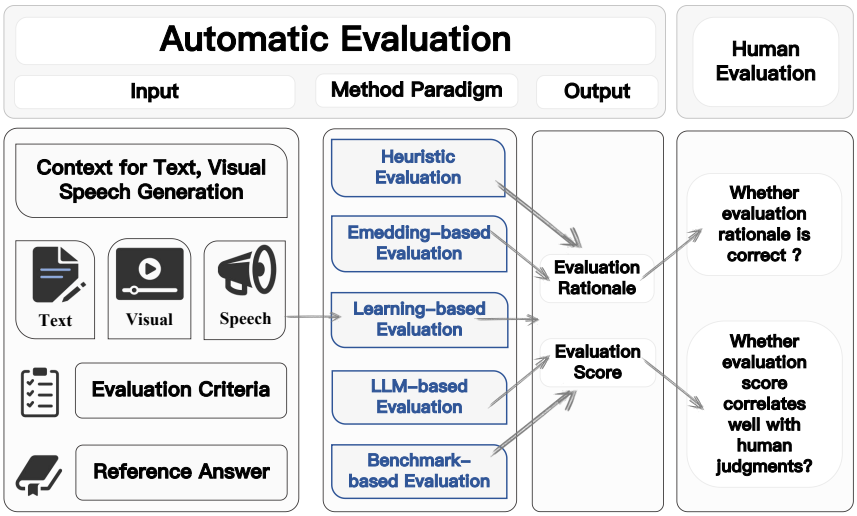

Fig. 1. 文本、视觉和语音生成的自动评估示例。

title from your rights confirmation email (Conference acronym 'XX). ACM, New York, NY, USA, 66 pages. https://doi.org/XXXXXXX.XXXXXXX

1 介绍

近年来，深度学习技术取得了显著进展，推动了各个领域和任务中生成模型的重大进步 [227, 277, 317, 377, 381, 479] 。诸如 GPT-4 和 Claude 等大型语言模型（LLMs）如今能够生成非常类似人类的对话 [277] ，而像 DALL-E 和 Stable Diffusion [317] 之类的扩散模型 [105, 295, 317, 355–357] 则改变了图像和视频的合成。这一快速发展引发了一个关键的研究问题：我们如何才能实现对模型生成内容的可靠和准确的自动评估?

尽管人工评估仍然是评估内容质量的黄金标准，但其高成本和固有的不可重复性限制了其在大规模应用中的可扩展性 [167, 170] 。这促使研究人员开发能够与人类判断高度相关的自动评估指标 [180, 194] 。如图 1 所示，自动评估旨在基于特定的评估标准、参考答案和语境，使用适当的评估方法来评估模型生成内容的质量。

尽管取得了这些进展，该领域仍然缺乏对不同任务和模态下自动评估当前发展的系统性综述。为了解决这一差距，本文提供了对文本、视觉和语音模态自动评估方法的全面回顾和统一分类法，提供了对这一发展领域的见解。我们首先深入分析了自然语言生成（NLG）中自动评估技术，此领域已有显著进展 [180, 194, 370] 。具体而言，我们描述并分析了现有自动评估方法的性质，并对现有自动评估方法进行了系统的元评估。在此分析基础上，我们将评述扩展到另外两个重要的生成式 AI 任务：视觉生成和音频生成。对于每个任务，我们总结了自动评估方法的当前发展，并概述了未来研究的有前景方向。

与相关调查的区别. 现有关于自动评估技术的调查主要集中在 NLG 任务中的特定方法论方法，如基于大模型的评估方法 [180, 194, 370] 。与之相反，我们的工作为跨越三个关键模态的自动评估提供了一个统一的框架：文本、视觉和语音。我们涵盖了评估方法的完整演变过程，从传统的启发式方法到现代的基于大模型的技术。这种跨模态的视角为生成式人工智能系统的评估方法学提供了更全面的理解。

2 预备知识

在本节中，我们介绍自动评估方法的基本概念（第 2.1 节），然后在第 2.2 节中描述三种主流评估协议：单项评估、配对评估和语料库评估。

2.1 评估概念

如图 1 所示，自动评估需要四个关键组件：（1）上下文；（2）评估标准；（3）参考答案；以及（4）待评估的模型生成内容。

上下文（$\boldsymbol{c}$）. 指的是模型用来生成内容的输入信息。例子包括对话生成任务中的对话历史 [166] 和文本到图像或文本到视频应用中的文本提示 [378, 441] 。

评价标准（$\boldsymbol{cri.}$）. 包含专为评估设计的任务特定维度，例如用于开放域文本生成的流利度 [71, 349] 和用于文本摘要的连贯性 [64] 。

参考答案（$\boldsymbol{r}$）. 广泛用于典型自然语言生成（NLG）任务的稳健评估，例如翻译和总结 [202, 283] 。然而，依赖于单一或有限数量的参考答案使得在开放式生成任务中难以有效覆盖可能输出的广阔空间 [167] ，导致次优的评估。因此，无参考评估方法得到了广泛关注 [71] ，因为它们在评估过程中不需要参考答案。相反，依赖参考答案的方法被称为基于参考的评估方法。

生成（$\boldsymbol{g}$）. 指的是正在评估的模型生成内容，包括文本、图像、视频和音频。

2.2 评估协议

与之前的工作不同 [194] ，我们将现有的自动评估方法分为三种主流评估协议：单个级别、成对级别和语料库级别的评估。下面描述这些协议。

单项测评. 直接评估一个特定生成结果（g）的质量：

$$(r^*, s) = M_{\text{AE}}(c^*, cri.^*, r^*, g) \tag{1}$$

，其中 M_{AE} 代表任何自动评估方法。r^* 是文本推理，分析并描述 g 的质量，通常由基于 LLM 的评估方法生成 [497] （见第 3.4 节）。s 是反映生成质量的质量分数，通常以限制范围内的李克特分数表示，其中更高的分数表示更高质量。上下文 c 、标准 $cri.$ 以及参考答案 r 是可选的，这取决于评估设置和任务，因此用 $*$ 标出。

成对评估. 是另一个重要且广泛使用的协议 [186] ，用于确定两代之间的偏好（g_A, g_B）。这里，p 是偏好标签，指示哪一代更好。相比单一评估，成对评估更为稳健和客观 [188] 。

语料库级别评价. 在语料库级别对测试集中的所有生成的质量进行评估：

$$s = M_{\text{AE}}(c^*, cri.^*, r^*, R, G) \tag{2}$$

其中 $R = \{r_i\}_{i=1}^N$ 和 $G = \{g_i\}_{i=1}^N$ 代表语料库中的所有 N 参考文献和模型生成的样本。与单项和成对协议不同，大多数语料库级别的评估方法不生成解释性理由 [294] 。由于语料库级别的评估不评估单个样本，因此比单项和成对评估协议更加粗粒度。

如图 1 的右部所示，一个好的自动评估方法应与人工判断有很高的相关性，并且生成的评估依据应是有效的、准确的和有帮助的。

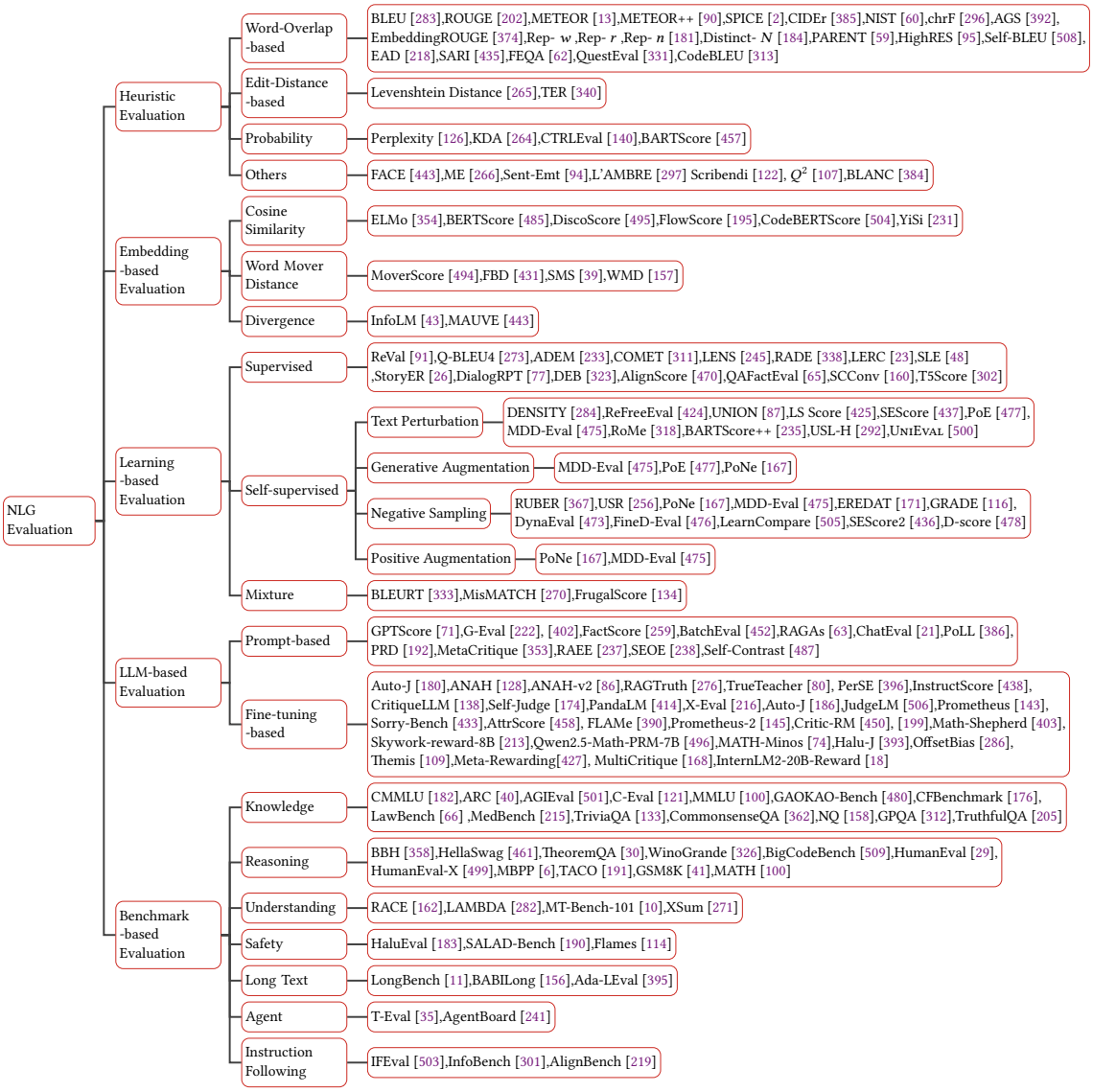

Fig. 2. **自然语言生成（NLG）中代表性自动评价方法的分类。在第 3 节中可以找到基于大型语言模型（LLM）的评价方法的详细分类。**

3 神经文本生成 (NLG) 的自动评估

在本节中，我们全面回顾了用于神经语言生成（NLG）任务的自动评估技术的进展，包括开放式文本生成、摘要、翻译等。图 2 展示了 NLG 任务自动评估方法的全面分类，主要分为五大类：（1）启发式评估使用基于规则和启发式的自然语言特征进行评估，例如词重叠度量或生成概率；（2）基于嵌入的评估测量参考文本和生成文本之间的语义相似性；（3）基于学习的评估通过在人类标注的数据上训练神经网络来评估文本质量；（4）基于大语言模型（LLM）的评估使用精心设计的提示和连锁思维推理来执行评估，即所谓的 LLM-as-a-judge；（5）

基于基准的评估使用人工标注的基准来测试特定能力，如数学推理和代码生成。以下小节详细探讨了每个类别及其代表性工作。最后，在第 3.5 节中，我们系统地比较和分析了这些自动评估范式的特征。

3.1 启发式评估

启发式评估依赖于基于规则或启发式特征的评估。现有的启发式方法可以分为三类：(1) 单词重叠；(2) 编辑距离；(3) 生成概率。表 1 全面概述了这些方法。

基于词重叠的启发式指标通过三个主要标准来评估文本生成：

(1) 文本相似度：. 这些指标假设高质量生成文本应与真实文本紧密匹配 [59, 95, 313] 。常见的例子包括 BLEU [283] 、ROUGE [202] 和 METEOR [13] ，这些用于通过 n -gram 重叠（精准度和召回率）来衡量机器翻译和摘要的质量。此外，chrF [296] 和 METEOR [13] 在 n -gram 匹配中结合了精准度和召回率，以提供更准确的相似度评估。NIST [60] 和 CIDEr [385] 为不同的 n -gram 分配权重，重点关注更重要的词汇和短语。

(2) 文本多样性：. 多样性对于开放域文本生成至关重要。指标如 Rep- n [348] 和 Distinct- N [184] 通过计算输出中唯一 n -grams 的比例来衡量生成质量。Self-BLEU [508] 计算每个生成句子的 BLEU 分数，并使用平均分数来衡量多样性。为了解决 Distinct- N 对较长文本的偏向，Expectation-Adjusted Distinct (EAD) [218] 根据统计预期调整不同词元的数量。

（3）事实一致性：. 事实一致性或忠实性对于文本摘要来说是至关重要的。大多数自动评估方法使用问答 (QA) 模型 [307] 来生成摘要和源文档的答案。然后，词重叠指标用于衡量这些答案之间的相似性 [62, 331, 392] 。

大多数基于词重叠的指标需要参考（基于参考），因为它们测量生成文本和参考文本之间的差异。尽管 HighRES [95] 是无需参考的，但它仍然需要来自文件的人类标注的源句子作为参考代理来评估生成的摘要。

3.1.1 **编辑距离** . 不同于基于词重叠的方法，这些方法计算 n -gram 的重叠，基于编辑距离的方法通过计算将生成文本转换为参考文本所需的字符或词级别的转换次数来评估文本质量 [265] 。TER [340] 和 WER [272] 是两个代表性的方法，它们通过计算与真实值的编辑距离来评估模型生成的摘要和翻译的质量。

3.1.2 **基于概率** . 虽然词重叠和编辑距离方法使用 n -gram 词汇特征，但许多方法利用概率特征来进行评价 [140, 264, 348] ，例如困惑度（PPL）[126] 。这些方法假定高概率的生成具有更好的质量。预训练语言模型（PLMs）如 BERT [58] 和 BART [178] 常用于计算生成概率。例如，BARTScore [457] 使用预训练语言模型给出的输入或参考来衡量生成概率。此外，CTRLEval [140] 通过在设计的文本填充任务中计算生成概率，根据上下文（前缀和属性标签）来评估文本生成的多个维度——连贯性、一致性和属性相关性。现在越来越多的研究使用大型语言模型（LLMs）评估生成文本概率 [194] ，如 GPTScore [71] 。尽管这些方法与其他基于概率的方法有相似之处，但由于 LLMs 的特定使用，我们在 LLM 基础的评估范式下对它们进行单独讨论（部分 3.4 ）。

除了词重叠、编辑距离和生成概率之外，一些方法利用其他启发式特征进行评估。例如，FACE [443] 基于文本的熵估计进行傅里叶分析来衡量文本相似性。一些方法专注于语料库级别的评估。Mark-Evaluate [266] 引入了受生态学中广泛使用的人口规模估算启发的评估指标。Zipf [106] 分析文字排名和频率之间的指数关系，研究生成文本如何遵循自然语言分布模式。对于可信度评估，Q^2 [107] 使用自然语言推理（NLI）模型，将从知识来源生成的答案与对话生成的答案进行比较，从而量化来源与生成之间的信息一致性。

3.2 基于嵌入的评估

启发式评价方法常常忽视文本生成中的语义特征，从而导致自动化评价与人工评价结果之间的显著差异。随着分布式词嵌入和表示学习技术的快速发展 [291, 374] ，基于嵌入的评价已经成为一种替代方法，可以测量生

Heuristic Methods	NLG Task	Category	Need Reference	Need Context	Evaluation Protocols
BLEU [283]	Machine Translation	Word-overlap	Yes	No	Single
ROUGE [202]	Text Summarization	Word-overlap	Yes	No	Single
METEOR [13]	Machine Translation	Word-overlap	Yes	No	Single
METEOR++ [90]	Machine Translation	Word-overlap	Yes	No	Single
SPICE [3]	Image Caption	Word-overlap	Yes	No	Single
CIDEr [385]	Image Caption	Word-overlap	Yes	No	Single
NIST [60]	Machine Translation	Word-overlap	Yes	No	Single
chrF [296]	Machine Translation	Word-overlap	Yes	No	Single
EmbeddingROUGE [374]	Text Summarization	Word-overlap	Yes	No	Single
MEANT2.0 [230]	Machine Translation	Word-overlap	Yes	No	Single
CodeBLEU [313]	Code Generation	Word-overlap	Yes	No	Single
PARENT [59]	Data-to-Text	Word-overlap	Yes	No	Single
HighRES [95]	Text Summarization	Word-overlap	No	Yes	Single
EAD [218]	General Text Generation	Word-overlap	Yes	No	Single
Rep-n [348]	General Text Generation	Word-overlap	Yes	No	Single
Rep-w [181]	General Text Generation	Word-overlap	Yes	No	Single
Rep-r [181]	General Text Generation	Word-overlap	Yes	No	Single
Distinct- N [184]	General Text Generation	Word-overlap	Yes	No	Single
Self-BLEU [508]	General Text Generation	Word-overlap	Yes	No	Single
QAGS [392]	Text Summarization	Word-overlap	No	Yes	Single
FEQA [62]	Abstractive Summarization	Word-overlap	Yes	Yes	Single
QuestEval [331]	Text Summarization	Word-overlap	No	Yes	Single
SARI [435]	Text Simplification	Word-overlap	Yes	Yes	Single
TER [340]	Machine Translation	Edit-Distance	Yes	No	Single
Levenshtein [265]	Consultation Note Generation	Edit-Distance	Yes	No	Single
BARTScore [457]	General Text Generation	Probability	Yes	No	Single
Perpleity [126]	General Text Generation	Probability	Yes	No	Single
KDA [264]	Multiple Choice Questions	Probability	Yes	No	Single
CRTLEval [140]	Controlled Text Generation	Probability	No	Yes	Single
FACE [443]	General Text Generation	Other	Yes	No	Single
ME [266]	General Text Generation	Other	Yes	No	Corpus
Zipf [106]	General Text Generation	Other	No	No	Corpus
Sent-Emt [94]	Dialogue Generation	Other	No	Yes	Corpus
L'AMBRE [297]	Machine Translation	Other	No	No	Single
Scibendi [122]	Grammatical Error Correction	Other	No	Yes	Single
Q^2 [107]	Knowledge-Grounded Dialogue	Other	No	Yes	Single
BLANC [384]	Text Summarization	Other	No	Yes	Single

Table 1. **启发式方法的代表性评估方法。根据是否需要参考资料，这些方法可以进一步分为基于参考的和无参考的。同样的原则也适用于上下文维度（无上下文和基于上下文）。**

成文本 [485] 与上下文或参考之间的语义相似性。基于用于测量嵌入相似性的方法，现有的基于嵌入的评价方法可以分为三大类型：（1）余弦相似度；（2）嵌入距离；以及（3）散度。表 2 总结了当前的基于嵌入的评价指标，我们将在下面详细讨论。

余弦相似度是测量两个句子嵌入之间语义相似性的最常见方法。在像 BERT 这样的预训练语言模型出现之前，研究人员通常使用像 Word2Vec 和 GloVe 这样的预训练词嵌入来构建文本的语义向量。例如，Vector Extrema、Embedding Average 和 Greedy Matching 应用余弦相似度来基于预训练词嵌入评估对话生成的质量。随着像 BERT 和 RoBERTa 这样的预训练语言模型（PLMs）的出现，这些模型的语义嵌入展示出更强的捕捉语

Embedding-based Methods	NLG Task	Category	Need Reference	Need Context	Evaluation Protocols
ELMO [354]	Text Summarization	Cosine Similarity	Yes	No	Single
FlowScore [195]	Dialogue Generation	Cosine Similarity	No	Yes	Single
DiscoScore [495]	General Text Generation	Cosine Similarity	Yes	No	Single
YiSi [231]	Machine Translation	Cosine Similarity	Yes	No	Single
CodeBERTScore [504]	Code Generation	Cosine Similarity	Yes	Yes	Single
BERTScore [485]	General Text Generation	Cosine Similarity	Yes	No	Single
FBD [431]	Dialogue Generation	WMD	Yes	No	Corpus
SMS [39]	General Text Generation	WMD	Yes	No	Single
WMD [157]	General Text Generation	WMD	Yes	No	Single
MoverScore [494]	General Text Generation	WMD	Yes	No	Single
MAUVE [294]	General Text Generation	Divergence	Yes	No	Corpus
InforLM [44]	General Text Generation	Divergence	Yes	No	Single

Table 2. **基于嵌入的评估方法的完整列表**。

义信息的能力，从而导致更为健壮的评估方法。一个显著的例子是 BERTScore，一种代表性的基于嵌入的度量，它使用 BERT 提取语义向量并计算从参考文本和生成文本的每个词元对中最大 IDF 加权的余弦相似度。

3.2.1 **词移动距离** *(WMD)*. WMD 计算两个词嵌入集之间的最小传输成本（距离），其中欧几里得距离和 $\mathcal{L}_p$ 距离（其中 $p \in 1, 2, +\infty$ ）是最常用的度量标准 [44] 。例如，MoverScore [494] 使用欧几里得距离作为传输成本，通过使用 BERT 词嵌入计算生成文本和参考文本之间的 WMD。类似地，WMD [157] 和句子移动相似性（SMS）[39] 度量使用语义嵌入将生成的文本转化为参考文本的成本。此外，Xiang et al. [431] 引入了 Frechet Bert 距离（FBD）来计算参考文本和生成文本之间的距离。

3.2.2 **散度**. 散度代表了另一类基于嵌入的评估指标。InfoLM [43] 计算由预训练语言模型生成的词汇上的离散概率分布的三种散度。MAUVE [294] 通过分析参考文本和生成文本的混合分布来测量它们之间的 KL 散度。

3.3 基于学习的评估

基于嵌入的评估方法通常与人类判断的相关性有限，原因在于两个关键限制：（1）它们在粗粒度水平上操作，未能捕捉关键错误模式，尤其是在语法和多样性方面；（2）它们依赖有限的参考集，无法充分代表高质量响应的全部范围，可能会惩罚那些与参考不一致的有效生成。借鉴数据驱动的机器学习方法，基于学习的评估通过训练在专门构建的评估数据集上来使用深度神经网络模拟人类评估模式 [233] 。训练数据集构成了基于学习的评估方法的核心。当前的方法可以根据数据集构建的方法分为三类：（1）监督方法在人工标注的数据集上训练评估模型；（2）自监督方法在通过启发式策略生成的合成数据集上训练评估模型；（3）混合方法采用两阶段训练方法，通常是先在大规模合成数据上进行预训练，然后在人工标注的数据集上进行微调。

3.3.1 **监督**. 基于监督学习的自动评估方法利用高质量的人类标注数据集 [17, 221, 253, 298, 319, 407, 421, 474] 训练模型以模拟人类评估。表 3 提供了这些方法的完整列表，下面描述了具有代表性的方法。

（1）基于参考和上下文无关的方法：. 早期基于参考和无上下文的监督方法通常通过微调超参数或基于 RNN/LSTM 架构训练成对的分类器或回归模型 [91–93] 。例如，Q-BLEU4 [273] 在人工评估得分上调优两个超参数，将可回答性得分与自动问题生成 (AQG) 系统的 BLEU-4 分数相结合。RUSE [339] 实现了一个直接在人类评分的翻译质量得分上训练的回归模型。

Metric	NLG Task	Need Reference	Need Context	Evaluation Protocols
T5Score [302]	Text Generation	Both	Yes	Single
ReVal [91]	Machine Translation	Yes	No	Single
Q-BLEU4 [273]	Question Generation	Yes	No	Single
LENS [245]	Text Simplification	Yes	Yes	Single
RADE [338]	Dialogue Generation	Yes	Yes	Single
LERC [23]	Read Comprehensive	Yes	Yes	Single
SLE [48]	Text Simplification	No	Yes	Single
StoryER [26]	Story Generation	No	No	Single
DEB [324]	Dialogue Generation	No	Yes	Single
BEER [346]	Machine Translation	Yes	No	Single
LEIC [51]	Image Caption	No	Yes	Single
RUSE [339]	Machine Translation	No	Yes	Single
SentBLEU [481]	Machine Translation	No	Yes	Single
ADEM [233]	Dialogue Generation	Yes	Yes	Single
COMET [311]	Machine Translation	Yes	No	Single
DialogRPT [76]	Dialogue Generation	No	Yes	Single
AlignScore [471]	General Text Generation	No	Yes	Single
QAFactEval [65]	Text Summarization	No	Yes	Single
SCConv [161]	Text Summarization	No	Yes	Single

Table 3. **具有代表性的基于学习（监督学习）的评估方法**。

（*2*）*基于参考和基于上下文的方法*：. 这些方法通过考虑参考和上下文信息来评估生成文本。一个显著的例子是 COMET [311]，它提供了两种变体——COMET-MQM 和 COMET-DARR——这些变体是在包含源句和参考翻译的人类评估机器翻译语料库上训练的。同样地，Takahashi et al. [361] 基于 XLM 模型 [165] 开发了机器翻译评估方法。这种方法也扩展到了其他自然语言生成任务。对于文本简化，LENS [245] 通过在人类评估的简化文本上训练，表现优于像 BERTScore 这样强大的基准。RADE [338] 使用人类响应作为参考来评估对话生成质量。此外，LERC [23] 在它的 MOCHA 数据集中的人类判断上进行训练，用于生成阅读理解任务，而 ADEM [233] 优化 RNN 模型，以基于会话历史和参考响应评估对话话语质量。

这些方法广泛用于开放域生成任务，在这种任务中，与有限参考示例不同的输出仍可能具有高质量。在这样的情况下，只有上下文信息对于评估生成质量是有用的。例如，DialogRPT 对 DialoGPT 模型进行微调，使用人类评分的数据集来创建基于排名的评估指标。DEB 专门构建了一个高质量的人类注释数据集，用于训练对话响应评估模型，结合多重参考和对抗响应。此外，还有几种方法专注于评估生成内容相对于其上下文的忠实性。QAFactEval 和 SCConv 通过使用句子级自然语言推断（NLI）分数作为特征向量来训练模型预测忠实性分数。AlignScore 采用更全面的方法，通过在七个既定任务上进行预训练来进行评估模型：NLI、问答、释义、事实验证、信息检索、语义相似度和总结。

（*4*）*无参考和上下文无关的监督方法*：. 该类别代表一种专门的基于监督学习的评估，主要在不依赖参考或上下文的情况下评估生成文本的内在质量。实例包括 StoryER [26] 和 ASE-Eval [360]，分别用于文本简化、故事生成和自动化作文评分（ASE）。这些应用通常在最小的上下文约束下运行，而是专注于生成内容的内在质量测量。

3.3.2 **自监督学习方法**. 由于创建人工标注数据集需要大量资源，研究人员开发了多种自监督策略来自动构建评估数据集。这些自监督方法可以分为四大类：（1）文本扰动技术，它对现有文本进行受控修改；（2）负采样

方法，从现有文本语料库中采样负样本；（3）生成性增强方法，利用语言模型生成样本；以及（4）正样本增强策略，生成正样本以平衡训练数据集的分布。对这些基于自监督学习的评估方法的全面概述在表 4 中展示。

Metric	NLG Task	Category	Self-Supervised Strategy	Need Reference	Need Context	Evaluation Protocols
DENSITY [284]	Dialogue Generation	Text Perturbation	Repetition Contradiction Sensitive Concatenation	No	Yes	Single
ReFreeEval [424]	Machine Translation	Text Perturbation	Reorder	No	Yes	Single
UNION [87]	Story Generation	Text Perturbation	Repetition Substitution Reorder Negation Alteration	No	Yes	Single
LS Score [425]	Text Summarization	Text Perturbation	Delete,Add Reorder	No	Yes	Single
SEScore [437]	General Text Generation	Text Perturbation	Add,Delete,Reorder Mask-and-Fill Substitution	Yes	No	Single
PoE [477]	Dialogue Generation	Text Perturbation Negative Sampling Generation	Back-Translation Delete,Reorder Repetition Mask-and-Fill	No	Yes	Single
MDD-Eval [475]	Dialogue Generation	Text Perturbation Negative Sampling Generation	Delete,Reorder Repetition Back-Translation Mask-and-Fill	No	Yes	Single
RoMe [318]	General Text Generation	Text Perturbation	Text Attack [267]	Yes	No	Single
BARTScore++ [235]	General Text Generation	Text Perturbation	Mask-and-Fill	Yes	No	Single
USL-H [292]	Dialogue Generation	Text Perturbation Negative Sampling	Delete,Reorder Substitution Repetition Mask-and-Fill	No	Yes	Single
UniEval [500]	General Text Generation	Text Perturbation	Repetition Delete,Reorder Substitution	Yes	Yes	Single
FineD-Eval [476]	Dialogue Generation	Text Perturbation	Utterance Shuffle	No	Yes	Single
DynaEval [473]	Dialogue Generation	Text Perturbation Negative Sampling	Utterance Replacement Utterance Shuffle	No	Yes	Single
SEScore2 [436]	General Text Generation	Text Perturbation	Delete,Add,Substitution	Yes	No	Single
BCR [454]	Dialogue Generation	Text Perturbation	Utterance Shuffle	No	Yes	Single
RUBER [367]	Dialogue Generation	Negative Sampling	Random Sampling	Yes	Yes	Single
PoNe [167]	Dialogue Generation	Negative Sampling Generation,PA	EDA [419] Generation	Yes	No	Single
USR [256]	Dialogue Generation	Negative Sampling	Random Sampling	No	Yes	Single
BERT-RUBER [81]	Dialogue Generation	Negative Sampling	Random Sampling	Yes	Yes	Single
EREDAT [171]	Data-to-Text	Negative Sampling	In-batch Negative Sampling	No	Yes	Single
GRADE [116]	Dialogue Generation	Negative Sampling	Lexical Sampling Embedding Sampling	No	Yes	Single
D-Score [478]	Dialogue Generation	Text Perturbation Negative Sampling	Utterance Swap Reorder Random Sampling	No	Yes	Single
LearnCompare [505]	Dialogue Generation	Generation	Generation of Past Checkpoint	No	Yes	Pair

Table 4. **完整的自监督学习评估方法列表**。

这种广泛采用的策略通过各种文本修改操作生成负样本，包括重复 [284] 、插入 [424] 、删除 [87, 500] 、替换 [87] 、重排 [425] 、反向翻译 [475] 和掩码填充 [475] 。例如，DENSITY [284] 、ReFreeEval [424] 、LS Score [425] 和 UniEval [500] 应用随机范围重复、插入、删除、替换和重排操作来生成具有故意错误的低质量文本。UNION [87] 专门对故事生成任务引入了否定变更。此外，MDD-Eval [475] 和 PoE [477] 使用反向翻译和掩码填充操作来生成看似流利但含有微妙错误的文本。最近，BCR [454] 通过打乱对话中一个说话者的发言构造了中等连贯性的负回应。

这种策略从大规模语料库中采样负例 [81, 255] 。例如，RUBER [367] 随机从语料库中采样对话话语来为每个用户查询创建负例。由于随机负例通常易于区分，许多研究人员已经开发了困难的负采样策略 [436, 473, 476, 478, 505] 。例如，PoNe [167] 和 GRADE [116] 选择与上下文相似和与参考文献相似的生成作为具有挑战性的负例。

（*3*）生成增强：. 这种方法使用预训练语言模型（PLMs）生成语言上流畅但语境上不合适的文本。MDD-Eval 通过使用 DialoGPT [491] 创建这样的样本来展示这一点。

（四）正样本增强：. 与之前主要关注收集负样本的方法不同，这种方法通过增加正样本来平衡数据集 [475] 。PoNe [167] 和 PoE [477] 都实现了这一策略。

尽管自监督方法在生成训练数据方面很高效，它们通常会在错误的负样本形式中遇到噪声。例如，生成增强和困难负样本采样方法有时会无意中将合适的生成包含在负样本集中。为了应对这一挑战，一些工作 [167, 475, 477] 发展了伪标签技术来识别和去除这些错误的负样本。例如，PoNe [167] 实现了一种专门设计用于检测此类问题样本的迭代优化算法。

3.3.3 **混合方法** . 监督学习方法通常由于高质量的人类标注而优于自监督方法。然而，自监督技术可以在没有昂贵的人工标注的情况下提供可扩展的数据集构建。混合方法结合了这两种范式的优点，提供了一种实用的折衷方案。该类别的代表性作品包括 BLEURT [333] ，FrugalScore [134] 和 MisMATCH [270] 。这些方法遵循一个包含两个阶段的训练过程：首先在通过自监督技术生成的大规模合成参考-候选对上预训练 BERT 模型，然后在较小的人类标注分数集上进行微调。这种方法利用了自监督数据的规模，同时从人类标注的质量中受益，最终得出既稳健又具成本效益的评价指标。

3.3.4 **基准测试评估** . 除自动评估方法外，基于基准的评估代表了一种独特的评估方法。该方法通过衡量 NLG 模型与人工标注的问答对的一致性来评估模型的整体能力。由于人工标注是一劳永逸地完成的，后续的模型评估仅需验证模型生成的答案与人工标注之间的匹配。这种验证可以通过简单的答案检查、精确匹配或各种自动评估方法来进行。

表格 5 提供了用于评估模型能力的广泛使用的基准的完整列表。可以发现，现有基准可以分为六种主要类型 [45] ： (1) 知识：评估一般和特定领域的知识（例如，用于 STEM 和人文学科的 C-Eval [121] ，医学领域的 MedBench）(2) 推理：测试数学 [41] 和编程任务 [29] 中的逻辑能力 (3) 理解：通过 MT-Bench101 [10] 和 XSum [271] 等基准评估上下文和查询的理解能力 (4) 长文本：通过 LongBench [11] 和 Ada-LEval [395] 测量处理扩展上下文的能力 (5) 代理：测试自动任务解决和功能调用的能力 [35, 241] (6) 指令遵循：使用 IFEval [503] 等指标评估对指令的遵循和与人类的对齐性

3.4 基于 LLM 的评估

基于学习的评估方法极度依赖于高质量和多样化的训练样本，这限制了它们在域外任务中的有效性，以及跨评估标准的泛化能力。随着展示出强大的指令遵循、上下文理解和零样本泛化能力的大型语言模型（LLMs）的

Benchmarks	Category	Metrics	Benchmarks	Category	Metrics
CMMLU [182]	Knowledge	Acc.	HumanEval-X [499]	Reasoning	Pass@K
ARC [40]	Knowledge	Acc.	MBPP [6]	Reasoning	Pass@K
AGIEval [501]	Knowledge	Acc.	TACO [191]	Reasoning	Pass@K
C-Eval [121]	Knowledge	Acc.	GSM8K [41]	Reasoning	Acc.
MMLU [99]	Knowledge	Acc.	MATH [100]	Reasoning	Acc.
GAOKAO-Bench [488]	Knowledge	Acc.	RACE [162]	Understanding	Acc.
CFBenchmark [176]	Knowledge	F1, Embedding-based	LAMBDA [282]	Understanding	Acc.
LawBench [66]	Knowledge	Acc,F1,ROUGE	MT-Bench-101 [10]	Understanding	LLM-based
MedBench [215]	Knowledge	BLEU,ROUGE, Acc,F1	XSum [271]	Understanding	ROUGE
TriviaQA [133]	Knowledge	EM,F1	HaluEval [183]	Safety	Acc.
CommonsenseQA [362]	Knowledge	Acc.	SaftyBench [492]	Safety	Acc.
NQ [158]	Knowledge	F1	SALAD-Bench [190]	Safety	F1,Acc, LLM-based
GPQA [312]	Knowledge	Acc.	Flames [114]	Safety	Acc.
TruthfulQA [206]	Knowledge	LLM-based	LongBench [11]	LongText	ROUGE-L,F1
BBH [358]	Reasoning	EM	BABILong [156]	LongText	Acc.
HellaSwag [461]	Reasoning	Acc.	Ada-LEval [395]	LongText	Acc.
TheoremQA [30]	Reasoning	Acc.	T-Eval [35]	Agent	F1,Acc,EM, Embedding-based
WinoGrande [326]	Reasoning	Acc.	AgentBoard [241]	Agent	PassRate
BigCodeBench [509]	Reasoning	Pass@K	IFEval [503]	Instruction	Acc.
HumanEval [27]	Reasoning	Pass@K	InfoBench [301]	Instruction	Decomposed Requirements Following Ratio
AlignBench [219]	Instruction	LLM-based	-	-	-

Table 5. **用于评估生成模型的基准列表**。

发展，研究人员越来越多地实施基于这些模型的自动评估系统。这种方法被称为基于 LLM 的评估或 LLM 作为裁判 [180] 。

与之前的启发式、基于嵌入和基于学习的自动评估方法相比，基于大型语言模型（LLM）的评估不仅提供评分结果，还提供分析生成内容缺陷的文本理由，并提供有价值的修订建议。这使得评估具有更精细的粒度和可解释性。目前的 LLM 评估方法主要分为两大类 [194]：（1）基于提示的方法，它在提示中加入评估指南，引导 LLM 的作用类似于标注者；（2）基于微调的方法，通过增强较小、更高效的 LLM 的评估能力，以应对像 GPT-4 这样高级模型所需的高计算成本。

3.4.1 **基于提示的方法**. 通过设计良好的提示，LLMs 可以有效地根据不同的标准评估各种自然语言生成（NLG）任务 [63, 71, 207, 380] 。例如，Wang et al. [402] 和 Liu et al. [222] 证明 GPT-3.5 和 GPT-4 作为多样化 NLG 任务的零样本评估者表现良好。基于这些发现，研究人员提出了几种技术来增强 LLM 评估能力的稳健性：

（*1*）参考：. 、Zheng et al. [498] 和 Badshah and Sajjad [8] 通过与参考响应比较来评估响应质量。最近，BatchEval [452] 通过使用批内示例作为参考来改善基于 LLM 的评估。

(2) 标准：. Lu et al. [237], Qian et al. [299] 提供详细的评估标准和评分细则以指导 LLM 评估过程。

（*3*）演示：. GPTScore [71] , ICE [123] , MSoR [343] , Little Giants [148] , 和 ALLURE [96] 通过结合少样本示例增强基于大型语言模型的评估。

（四）交换：. 换位操作广泛用于减少成对评估中的位置偏差 [170, 172, 498] 。

（5）自洽性：. Cohen et al. [42], Manakul et al. [250], Zhang et al. [487], Zheng et al. [498] 表明，随着自洽提示策略的应用，LLM-as-a-judge 的表现有所提升，该策略采样多种评估理由并采用多数投票作为最终结果。

为了应对单一模型偏见问题，研究人员开发了用于评估任务的多代理辩论框架，包括 ChatEval [22] 、PoLL [386] 和 PRD [192] 。

(7) 索赔分解：. 由于评估的回答可能包含多个独立的论断，研究人员提出将回答分解为原子信息单元 [259, 353] ，并分别验证每个单元的质量，然后汇总结果以确定整体回答质量。

总之，上述大多数基于提示的方法提出了提示策略，以解决大语言模型（LLM）输出中的不一致问题 [198, 412, 487] 。当遵循演示中提供的指导和线索、详细标准和参考响应时，LLM 在评估响应质量时能够实现更准确的判断。此外，自我一致性技术和多代理框架被广泛用于减少单一模型评估中固有的不一致性和偏见。

尽管大规模语言模型（LLM）可以达到与人类判断相媲美的性能，其高昂的推理成本限制了在大规模评估场景中的应用。最近的研究集中在降低这些成本上。例如，UniCBE [453] 引入了一个统一的以一致性为驱动的 CBE 框架，它优化了成对评估协议中的元组采样和偏好聚合策略。同样，TailoredBench [455] 提出了一种针对每个目标模型量身定制的评价方法，在相同的推理预算约束下，实现了准确性估计均绝对误差（MAE）的平均降低 31.4 %。

3.4.2 **基于微调的方法** . 降低基于 LLM 的评估成本的最直接方法是提高较小、更高效模型的评估能力。许多研究人员开发了微调方法，以增强小规模 LLM 的评估性能。这些高效评估模型解决了大型评估的计算需求，并在最近的 RLHF 训练程序中被广泛使用 [54, 172, 246, 279] 。到目前为止，当前基于微调的 LLM 评估方法可以从四个维度进行分类 [194, 463]：(1) 评价可解释性；(2) 评价细粒度；(3) 优化方法；和 (4) 数据来源。表 6 提供了这些方法的全面列表，我们在下面描述每个维度的代表性方法 [1] 。

基于可解释性，目前微调的基于 LLM 的评估模型分为两类：奖励模型和批判模型。奖励模型 [164, 279] 被训练用于模拟人类偏好，并作为人类反馈强化学习（RLHF）[279, 347] 中的关键组成部分。虽然这些模型帮助将 LLM 与人类偏好对齐，但它们只提供数字评分而没有解释性反馈，这限制了它们的可靠性和可解释性。相反，批判模型对生成内容进行文本分析，提供更详细和可解释的反馈 [219, 370] 。最近的进展引入了生成性奖励模型 [246, 447] ，例如 Critic-RM [451] ，在评分之前产生连贯的思维分析。这种方法增强了可解释性和训练数据的效率。

(2) 评估粒度：. 早期的评估方法侧重于对整个回复的评估，称为基于结果的奖励模型（ORMs）[200] 。最近的研究 [463] 表明，细粒度的过程奖励模型（PRMs）提供了更好的透明度和更有效的反馈，尽管它们需要大量人力来进行注释 [200] 。大多数基于微调的评估模型仍是基于结果的，评估整个回复的正确性 [49, 146] 。为提高透明度和解释性，一些方法识别回复中的特定缺陷并提供详细分析，包括 InstructScore [438] 、TIGERScore [129] 和 MultiCritique [169] 。然而，由于回复通常只包含少量缺陷，这些评估信号仍然稀疏。

一种日益增长的趋势是通过基于过程的奖励模型（PRMs）为所有中间步骤提供更密集的评估信号。例如，Lightman et al. [200] 注释数学解决方案中每个中间推理步骤的正确性。尽管 PRMs 提供详细的反馈，但注释中间步骤的资源消耗比评估完整的响应要多得多。为了解决这个挑战，一些研究人员采用蒙特卡罗树搜索（MCTS）自动生成已知真实答案的推理任务的过程信号，如 Math-Shepherd [403] 和 PSRLM [483] 所示。然而，大多数现有的 PRMs 仅限于具有确定性答案的推理任务，并且为更广泛的领域开发基于过程的评价模型仍然是一个巨大的挑战 [463] 。

[1] 表 6 包括选定的代表性奖励模型。

（3）优化方法：. 监督微调（SFT）是增强大语言模型（LLMs）评价能力的最常见方法 [49, 187] 。值得注意的例子包括 Auto-J [187] 、UltraCM [49] 和 Prometheus [145] ，这些方法使用由 GPT-4 生成的合成数据集来优化 Llama 模型。最近，研究人员开始探索偏好学习技术，如直接偏好优化（DPO）[304] 和强化学习（RL）[279] ，以克服行为克隆方法的局限性。行为克隆常常在分布转移时遇到困难，并且无法捕捉人类偏好背后的细微推理。像 Critic-RM [410] 和 SFR-Judge [404] 这样的模型通过在偏好评估数据集上优化 Llama 模型来展示这种转变，从而在更强大的评估能力上与人类判断模式更好地保持一致。

（四）数据来源：. 类似于 Section 3.3 中描述的基于学习的评估方法，数据集是基于微调的 LLM 评估方法的核心。截至目前，用于优化基于微调评估方法的数据集来源于三个主要渠道：（1）人工标注：这代表了构建评估数据集的最直接方法。例如，Shepherd [411] 使用领域专家标注文本评估样本以训练 Llama-7B，而 CriticGPT [254] 开发了一个人工标注的偏好评论数据集用于训练 GPT-4 模型。虽然人工标注的数据集提供了可靠的质量，但其巨大成本显著限制了可扩展性，特别是在构建详细的文本评论数据集时。（2）高级教师模型：鉴于文本评估的人工标注需要大量资源，许多研究者利用诸如 GPT-4 之类的高级教师模型来构建用于训练的合成数据集。实例包括 Auto-J [180] 和 UltraCM [49] 。然而，这些模型生成的数据集通常包含显著的噪声，可能会影响所得评估模型的鲁棒性。最近，MultiCritique [169] 提出通过聚合多代理的多样化评论意见来解决单一模型的噪声问题。（3）人机协作标注：这种混合方法结合了前述两种方法的优点。具体来说，大型语言模型生成草稿评估样本，然后使用人工标注的分数进行验证。例如，Themis [110] 和 SFR [404] 通过测量 GPT-4 判断与人工评估的一致性来筛选高质量的 GPT-4 生成样本，从而在保持成本效益的同时获得更可靠的训练数据。这种方法可以利用简单的人工标注的评估分数或偏好作为真实数据，来收集可靠的详细文本批评，从而避免从头标注详细批评所需的大量成本。

3.5 比较自动评估

在本节中，我们对现有文本生成的自动评估模式进行了定性和定量分析。

3.5.1 **定性分析** . 表 7 总结了我们对现有自动评估方法在四个关键维度上的比较分析：(1) 评估的灵活性和泛化性，它考察了方法在不同任务和评估标准上的适应性；(2) 训练数据来源，它识别出开发这些方法所使用的数据的来源和特性；(3) 评估的细粒度性，它考量了评估过程的细节程度；(4) 评估成本，它考虑了计算资源和时间需求。这些维度为理解当前自动评估方法的优势和局限性提供了一个全面的框架。

评估的灵活性和泛化 . NLG 评估技术的发展显示出一种从任务特定方法向更具普遍性方法的明显转变。早期的技术（启发式、基于嵌入和基于学习的方法）仅限于特定任务和标准 [71] 。相比之下，现代基于 LLM 的方法通过简单的提示工程实现了显著的灵活性，能够在不同任务和定制标准下进行通用评估 [225] 。

最近，基于微调的 LLM 自动评估方法已经出现，作为资源密集型提示方法的替代方案。然而，这些训练出的评估方法往往牺牲了泛化能力，特别是对于其训练数据中未涵盖的任务和评估标准 [169] 。显式学习评估标准代表着未来发展的一个关键方向。最近的工作已经开始探索这一领域。例如，HD-Eval [225] 学习使用人工标注标签来构建一个分层标准树，并证明这种标准结构改进了基于 LLM 的评估方法，如 G-Eval [222] 。同样，MultiCritique [169] 在多任务学习框架中共同学习定制的标准和评估生成，显著增强了评估的泛化能力。

训练数据来源 . 基于学习或微调（基于 LLM）的自动评估方法的可靠性在很大程度上依赖于训练数据的质量。如表 7 所示，训练数据来源要么是人工标注，要么是合成数据生成 [110, 411, 447] 。人工标注提供了精确的标签，但会产生相当高的成本，阻碍了规模化。这一限制对于最近的可解释微调自动评估方法尤为显著。合成数据提供了一种具有成本效益的替代方式，但受到生成模型能力固有的质量限制。为应对这一挑战，最近的方法开发了人机协作标注方法。诸如 SCRIT [365] 和 Critic-RM [451] 之类的方法通过用人工标注的标签验证合

Metric	Interpretability	Granularity	Optimization	Data Source	Reference	Protocols	Base Model
CriticGPT [254]	Yes	Outcome	RL	Human	No	Single	GPT-4
PRM[200]	No	Process	SFT	Human	No	Single	GPT-4
Math-Shepherd [403]	No	Process	SFT	Human-Model	No	Single	Llama2-70B
Qwen2.5-Math-PRM-7B	No	Process	SFT	Human-Model	No	Single	Qwen2.5-7B
InternLM2-20B-Reward [18]	No	Outcome	SFT	-	No	Single	InternLM2-20B
Skyword-Reward-8B [213]	No	Outcome	SFT	-	No	Single	Llama-3.1-8B
Skyword-Critic-8B [213]	No	Outcome	SFT	-	No	Pair	Llama-3.1-8B
Auto-J [187]	Yes	Outcome	SFT	GPT-4	No	Single/Pair	Llama-2-13B
UltraCM [49]	Yes	Outcome	SFT	GPT-4	No	Single	Llama-2-13B
Shepherd [411]	Yes	Outcome	SFT	Human	No	Single	Llama-7B
Prometheus [144]	Yes	Outcome	SFT	GPT-4	Yes	Single	Llama2-13B
Prometheus2 [146]	Yes	Outcome	SFT	GPT-4	Yes	Single/Pair	Llama2,Mistral
Self-Taught [410]	Yes	Outcome	SFT	LLM	No	Pair	Llama3-70B
Meta-Rewarding [428]	Yes	Outcome	PL	LLM	No	Single	Llama3-8B
InstructScore [438]	Yes	Process	SFT	GPT-4	No	Single	Llama-7B
TIGERScore [129]	Yes	Process	SFT	GPT-4	Yes	Single	Llama2-7B/13B
Themis [110]	Yes	Outcome	SFT/PL	Human-Model	No	Single	Llama3-8B
SFR [404]	Yes	Outcome	SFT/PL	Llama3-70B	No	Pair	Llama3.1-8B/70B
Critic-RM [451]	Yes	Outcome	SFT/PL	Human-Model	No	Single	Llama3.1-70B
MultiCritique [169]	Yes	Process	SFT/PL	Multi-Agent	Yes	Single	InternLM2-7B
PandaLM [415]	Yes	Outcome	SFT	GPT-3.5	Yes	Pair	Llama-7B
JudgeLM [506]	Yes	Outcome	SFT	GPT-4	Yes	Single/Pair	Llama2
CritiqueLLM [139]	Yes	Outcome	SFT	GPT-4	Yes	Single/Pair	ChatGLM3-6B
CompassJudger [19]	Yes	Outcome	SFT	Human,GPT-4	No	Single/Pair	InternLM2.5-7B
X-Eval [217]	Yes	Outcome	SFT	Human	No	Single/Pair	Llama-7B
FLAMe [391]	Yes	Outcome	SFT	Human	No	Single/Pair	PaLM-2-24B
AttrScore [459]	Yes	Outcome	SFT	Human	No	Single	Llama
Self-Judge [175]	Yes	Outcome	SFT	Human	No	Pair	Llama2-7B
Self-Rationalize [373]	Yes	Outcome	PL	LLM	No	Single	Llama3.1-8B
SCRIT [365]	Yes	Process	SFT	Human-Model	Yes	Single	Qwen2.5-72B
ANAH [128]	Yes	Outcome	SFT	Human	No	Single	InternLM2-20B
ANAH-v2 [86]	Yes	Outcome	SFT	LLM	No	Single	InternLM2-20B
RAGTruth [276]	No	Outcome	SFT	Human	No	Single	Llama2-13B
TrueTeacher [80]	Yes	Outcome	SFT	PaLM 540B	No	Single	T5-11B
PerSE [396]	Yes	Outcome	SFT	Human	No	Single/Pair	Llama-7B
SorryBench [433]	No	Outcome	SFT	Human	No	Single	Llama3-8B
MATHMinos [74]	Yes	Process	SFT	Human-Model	No	Single	MetaMATH
Halu-J [393]	Yes	Outcome	SFT/PL	Human-Model	No	Single	Mistral-7B
Offsetbias [286]	No	Outcome	SFT	GPT-4	No	Pair	Llama3-8B
DeepSeek-GRM [228]	Yes	Outcome	PL	Human	No	Single	DeepSeek-V3

Table 6. **基于微调的大型语言模型（LLM）评估模型列表。对于数据源，人-模型表示训练数据是使用 LLM 生成并由人工标注验证的。这些人工标注可以是推理任务中的参考答案 [365] 或人工标注的偏好标签 [110, 404, 451]。**

成数据，创造了更有效且具扩展性的解决方案，以提高评估能力。随着任务复杂性的增加，一个重大的挑战出现了：仅靠教师模型或人类专家都无法提供足够可靠的监督。未来的研究应专注于开发方法，以准确监督复杂推理问题或开放性问题等具有挑战性的任务的生成。

评估粒度. 评价技术的演变推动了越来越细粒度的评估能力。传统方法（启发式、基于嵌入和基于学习的方法）采用基于结果的评估 [463]，提供整体质量分数或二元偏好标签，但没有详细的理由。相比之下，基于 LLM 的评估方法利用其先进的理解和生成能力，进行多维度的评估。这些方法不仅生成定量指标，如数值分

Evaluation Methods		Flexibility and Generalization	Data Resource	Evaluation Granularity	Output Format	Evaluation Cost	Correlation with Human
Benchmark-based Evaluation		Limited Task and Criteria	Human Annotation	Instance-level	-	Small	Strong
Heuristic Evaluation	Word-overlap	Limited Task and Criteria	-	Instance-level	Scalar	Small	Weak
	Edit-Distance		-	Corpus-level	Scalar	Small	Weak
	Probability		-		Scalar	Moderate	Weak
Embedding-based Evaluation	Cosine Similarity	Limited Task and Criteria	-	Instance-level	Scalar	Moderate	Weak
	WMD		-	Corpus-level	Scalar	Moderate	Weak
	Divergence		-		Scalar	Moderate	Weak
Learning-based Evaluation	Supervised	Limited Task and Criteria	Human-annotated Score	Instance-level	Scalar	Moderate	Moderate
	Self-Supervised		Synthetic Data		Scalar	Moderate	Moderate
	Mixture		Human-annotated Score and Synthetic Data		Scalar	Moderate	Moderate
LLM-based Evaluation	Prompt-based	General Task and Customized Criteria	-	Instance-level Process-level	Scalar and Rationale	Huge	Strong
	Fine-tuning-based		Human Annotation or Synthetic Data		Scalar and Rationale	Huge	Strong *

Table 7. **对五种评估范式在六个维度上的比较结果。标量表示基于标量的评估结果，例如分数和偏好标签。流程级别指的是最近基于流程的奖励模型** [463] **，用于评估生成中每个流程的质量。**

数或偏好排名，还提供细粒度的、可解释的评估理由 [222] 。这种解释能力在过程和步骤层面上运作——检查推理路径和个别决策点，而不仅仅是最终输出。

虽然评估技术变得更加先进和稳健，但它们的计算成本也显著增加，尤其是对于最近基于 LLM 的评估方法。像 GPT-4 这样的大型基础模型需要大量的计算资源，这使得大规模评估对许多研究团队和组织来说变得非常昂贵且耗时。

为了解决这一效率挑战，基于微调的评估方法在基于 LLM 的评估范式中已成为一种有前途的解决方案。这些方法旨在将大规模 LLM（如 GPT-4）的复杂评估能力浓缩为更小、更高效的模型。由此产生的紧凑型评估器显著减少了推理时间和计算需求，同时保持了可比的评估质量。这些较小的评估模型提供了两个关键优势：首先，它们能够高效地进行大规模评估，而使用较大模型则会令人望而却步。其次，它们可以在从人类反馈中进行强化学习（RLHF）训练流水线中作为更强健和可靠的奖励模型 [54, 447] ，在优化过程中需要反复评估。这种在评估质量和计算效率之间的平衡代表了使先进评估技术更为可及和实用的重要方向。

3.5.2 **定量分析** . 本节通过比较实验系统地检查各种自动评估方法中代表性作品的性能差异，不仅限于定性分析。在展示我们的定量结果之前，我们介绍元评估的概念——即评估自动评估方法的过程。

元评价 . 元评估旨在评估自动评估的可靠性。目前，元评估可分为两类：客观评估和主观评估。

元评估评估自动评估方法的可靠性，即自动评估是否与人工判断相关。目前的元评估方法主要分为两大类：客观评估和主观评估 [170] 。

(1) 目标评估：目标元评估方法分为以下几类：

- 单项评估协议衡量自动评估方法与人工判断之间的相关性。Spearman、Pearson 和 Kendall 相关系数广泛用于建立的元评估基准测试中，如 SummEval [64] 、FED [255] 和 OpenMEVA [88] 。
- 成对评估协议衡量模型生成偏好 p 与人工标注偏好（如 RewardBench [164] 和 Auto-J [187] ）之间的一致性（准确性）。
- 整体评估协议计算自动评估方法与人工标注者的质量判断之间的相关性，其中质量判断意味着选择一个特定的（生成模型和解码器）设置 [294] 。

Evaluation Methods			Summ -Eval	Topical -Chat	FED	WMT -22	Open MEVA	BAGEL	Web -NLG
Heuristic Evaluation	Word-overlap	BLEU [283]	12.0	21.6	-	19.7	-1.17	16.8	20.7
		ROUGE-L [202]	14.5	23.7	24.4	17.8	2.34	14.2	35.5
	Edit-Distance	TER [340]	-12.0	1.07	-	21.95	6.20	-0.09	-0.08
	Probability	BARTScore [457]	17.2	39.0	12.8	33.7	17.4	20.7	56.8
Embedding-based Evaluation	Cosine Similarity	BERTScore [486]	23.7	32.3	27.3	42.4	2.9	28.2	50.4
	WMD	MoverScore [494]	19.1	31.0	-	27.1	8.53	20.8	36.5
Learning-based Evaluation	Supervised	COMET22 [311]	33.8	11.6	-	56.4	39.2	13.8	40.9
	Self-Supervised	UniEval [500]	47.5	53.5	21.5	21.9	44.5	30.3	38.4
		SEScore2 [436]	39.9	-37.9	-	44.9	30.6	32.5	48.4
	Mixture	BLEURT [333]	17.3	38.8	-	48.4	27.5	22.9	16.8
		MisMatch [270]	41.0	-	-	-	-	-	49.0
LLM-based Evaluation	Fine-tuning-based Critique Models	TIGERScore [129]	36.8	34.6	-	45.0	46.4	-	42.4
		InstructScore [438]	26.3	24.1	-	51.9	16.1	34.2	59.0
		Auto-J [187]	4.8	42.8	37.6	0.4	30.1	20.4	22.2
	Fine-tuning-based Reward Models	InternLM2-20B-Reward	48.5	65.0	43.9	45.4	43.7	27.5	20.1
		Skywork-Reward-8B	44.3	43.3	42.3	30.1	39.1	25.9	25.5
	Prompt-based	GPTScore [71]	41.7	53.5	39.2	28.8	23.9	41.3	28.8
		G-Eval (GPT-4) [222]	51.4	73.2	45.5	-	47.5	27.8	43.1
		DeepSeek-V3 [54]	57.6	66.4	53.6	-	44.9	39.4	42.2
		DeepSeek-R1 [53]	52.1	64.9	54.4	-	42.9	38.0	41.5

Table 8. **代表性自动评估方法在 7 个 NLG 元评估基准上的表现（斯皮尔曼相关性得分）。由于先前研究中的评估不完整，部分结果仍为空。由于基于嵌入的散度方法的工作有限，结果不可行。**

（2）主观评价：主观元评价方法主要评估评价理由的质量。最近的研究使用 GPT-4 作为评判标准来打分评价理由质量 [49, 143, 186, 411] 。然而，由于自动评估任务的复杂性，即使是 GPT-4 也无法始终如一地提供可靠的主观元评价 [254, 411] 。CriticEval [170] 和 MetaCritique [353] 最近的工作表明，当提供人工标注的评价理由作为参考评价理由（评论）时，GPT-4 可以实现可靠的元评价。

已经提出了许多元评估基准来衡量自动评估的可靠性。如表 **??** 所示，早期的元评估基准主要集中在特定的 NLG 任务上，包括机器翻译 [70, 71] 、文本摘要 [14, 64, 85] 、数据到文本生成 [201, 248, 420, 502] 、对话生成 [255, 256] 和故事生成 [88, 407] 。随着大型语言模型（LLMs）的高级生成和泛化能力，研究人员开发了系统的通用领域元评估基准，以评估其评估能力，如 MT-Bench、Chat-Arena [497] 、RewardBench [164] 和 RM-Bench [226] 。此外，专门的元评估基准衡量 LLMs 在具有挑战性的任务中的评估性能，包括推理 [209, 239, 363] 、安全对齐 [456] 和信息检索 [237] 。

系统性元评价结果．我们对来自四种自动评价方法（启发式、基于嵌入、基于学习和基于 LLM 的方法 [2] ）的代表性工作进行了系统的比较和评估，这些方法覆盖了 12 个关键的元评估基准：(1) 用于文本摘要的 SummEval [64]；(2) 用于个性化对话生成的 Topical-Chat [256]；(3) 用于一般开放域对话生成的 FED [255]；(4) 用于机器翻译的 WMT22 [70]；(5) 用于故事生成的 OpenMEVA [88]；(6) 用于数据到文本生成的 BAGEL [248]；(7) 用于数据到文本生成的 WebNLG [502]；(8) 用于推理任务的 CriticBench [209]；(9) 用于单项（CriticEval-single）和对项（CriticEval-pair）评价协议上的 9 种多样的 NLG 任务的 CriticEval [170]；(10) 用于 58 种多样 NLG 任务的 Auto-J [180]；(11) 用于多样 NLG 任务的 PreferenceBench [145]；以及 (12) 用于评估跨聊天、安全和推理任务的奖励模型的 RewardBench [164] 。请参阅表格 **??** 了解这些元评估基准的更多细节。

[2]基于基准的评估方法不需要元评估。

Evaluation Methods			Critic -Bench	CriticEval -Single	CriticEval -Pair	Auto-J	Preference -Bench	Reward Bench
LLM-based Evaluation	Fine-tuning-based Critique Models	Auto-J [187]	67.4	36.1	49.3	75.6	74.0	42.8
		UltraCM [49]	59.4	21.5	38.0	-	78.2	-
	Fine-tuning-based Reward Models	InternLM2-20B-Reward	-	58.3	61.6	85.6	51.0	90.8
		Skywork-Reward-8B	-	52.7	51.2	76.8	51.0	92.5
	Prompt-based	G-Eval (GPT-4o) [222]	78.8	68.2	59.1	80.8	89.0	86.7
		DeepSeek-V3 [54]	73.6	57.6	71.2	74.9	87.8	91.9
		DeepSeek-R1 [53]	85.9	65.0	71.5	77.1	88.8	91.1

Table 9. **在通用领域的 5 个元评估基准上基于 LLM 的评估方法的表现。对于 CriticEval-single，度量标准是 Spearman 相关系数，而对于其他五个元基准，度量标准是偏好准确性。**

由于过去五个元评估基准涵盖不同的领域，我们仅测试了代表性的基于 LLM 的评估方法，因为启发式、基于嵌入的和基于学习的评估方法无法在这些领域中有效评估。根据表格 8 和表格 9 中的结果，我们可以得出几个重要的结论：

- 基于 LLM 的自动评估方法目前明显优于其他方法。在七个基准测试中，基于 LLM 的方法在机器翻译任务的 WMT-22 基准测试中仅略微不如 COMET-22 [311] 。
- 经过微调的 LLM（大型语言模型）评估方法也显著优于启发式、基于嵌入和基于学习的方法。这表明，从强大的 LLM 中提炼评估能力可以产生高效、高质量的自动化评估模型。
- 表 9 显示，在人类标注的偏好数据集上微调的奖励模型优于可解释的批判模型，甚至在多个基准测试（例如，Auto-J 和 RewardBench）上超越了基于提示的方法。这突显了人类标注训练数据在评估任务中的有效性。然而，表 8 显示，奖励模型在针对特定 NLG 任务的七个元评估基准测试中明显不如基于提示的方法。这一发现表明泛化能力仍然是奖励模型的一大限制。未来的工作应优先考虑提高这些模型的泛化能力。
- 最近，一些专为数学和编码问题优化的推理模型，如 OpenAI o1 和 DeepSeek-R1，在解决复杂问题时表现出强大的批判能力。这引发了一个重要问题：推理模型是否比大型语言模型（LLMs）更适合评估？如表 8 和 9 所示，最先进的推理模型 DeepSeek-R1 仅在特定的基准测试中（如 CriticBench 和 CriticEval-Pair）优于其基础模型 DeepSeek-V3 和 GPT-4。这表明推理模型并不是传统 LLM 评估方法的全面优越替代品。DeepSeek-R1 在 CriticBench 上表现尤为出色，该测试包括多样化的元评估任务，重点关注复杂推理，表明它可能特别适合在复杂推理任务中评估生成质量。

4 视觉生成的自动评估

在这一节之前，我们已经系统地调研了文本生成的自动评估方法。在此基础上，我们现在将分类扩展到视觉生成任务的自动评估技术，如文本到图像和文本到视频的生成任务。如图 3 所示，现有的视觉生成自动评估经过了五个范式的演变：（1）启发式评估：依赖于手工制作的规则或特征（例如，像素级别的差异）来量化视觉内容质量的简单方面；（2）基于嵌入的评估：利用从深度神经网络学习到的视觉特征嵌入来评估生成内容与参考内容之间的感知或语义相似性；（3）基于学习的评估：训练模型来预测人工注释的质量分数，使度量结果更贴近人类判断；（4）基于 LLM/MLLM 的评估：利用 LLM 和 MLLM 结合定制的提示在不同的视觉评估标准上进行细致的评估；（5）基准评估：使用精心策划的数据集和黄金标准的参考直接比较系统输出与已建立的性能基准。

在描述了这些工作之后，我们接着介绍用于评估视觉内容的最广泛使用的基准套件（第 4.1 - 4.5 节）。最后，我们讨论当前的挑战并概述未来研究的有前景方向（第 4.6 节）。

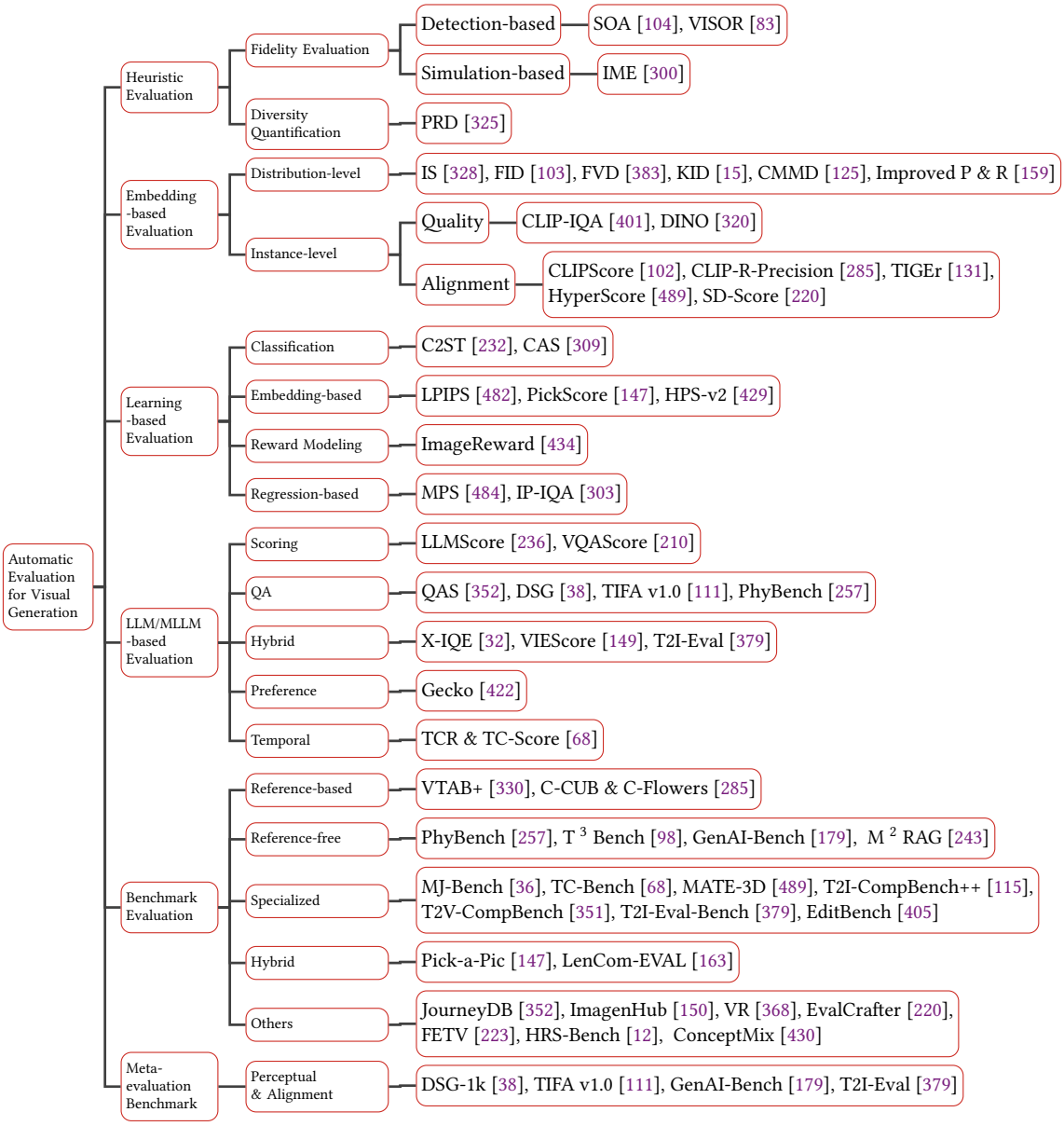

Fig. 3. **视觉生成中代表性自动评估方法的分类**

4.1 启发式评估

视觉生成的启发式评估依赖于预定义的度量和标准化程序，这些度量和程序可以通过算法计算或经过结构化的人为检查来评估模型输出。在我们的分类法中，每种方法沿以下三个轴进行特征化：(1) 参考依赖性，(2) 评估协议，以及 (3) 关注点（真实性与多样性）。表格 10 根据这些标准和其基本度量类型总结了代表性的方法。

4.1.1 **保真度评估**. 保真度评估衡量生成内容与给定条件（例如，文本提示）的语义对齐程度。大多数方法利用预训练的检测器或模拟模型来验证对象的存在、属性和跨模态一致性：

- 基于检测的度量。语义对象准确性（SOA）[104] 使用对象检测器来确认目标实体，而 VISOR [83] 通过验证检测对象之间的空间关系来扩展这一功能。
- 基于模拟的度量。隐含操控评估 (IME) [300] 通过使用视频转动作模型来模拟代理动作，测试模型的世界模拟能力，从而评估视频生成质量。

4.1.2 **多样性量化**. 多样性指标用于评估生成输出的多样性和覆盖范围，通常通过比较分布或估计样本之间的熵来实现：

- 分布比较。分布的精确-召回（PRD）[325] 分析真实数据分布和生成数据分布之间的精确-召回曲线，提供了比单一数值评分更细致的视角。
- 基于熵的度量。基于熵的启发式方法通过量化图像的信息含量来推测其感知质量，而无需参考。例如，ENIQA [31] 在空间和频域中计算熵，以预测一个连续的质量得分。尽管其方法简单，ENIQA 与人类判断具有很强的一致性，表明仅仅依靠熵就能够有效抓住细节丢失和失真多样性。

Method	Task	Objective	Reference	Protocols	Metric Type
SOA [104]	Text-to-Image Generation	Fidelity	No	Single	Detection
TIAM [84]	Text-to-Image Generation	Fidelity	No	Corpus	Hybrid
VISOR [83]	Text-to-Image Generation	Fidelity	No	Corpus	Detection
PRD [325]	Text-to-Image Generation	Diversity	Yes	Single	Distribution
IME [300]	Text-to-Video Generation	Fidelity	No	Single	Simulation

Table 10. **视觉生成任务的启发式评估方法，按关注点（真实度与 多样性）、参考依赖性、评估协议和度量类型进行分类。**

4.1.3 **局限性和未来方向**. 尽管简便且计算效率高，启发式度量具有显著缺点：

- 偏差和覆盖率。基于检测的方法继承了预训练模型的偏差，并可能忽视新的或不在词汇表中的概念。
- 人类关联。许多度量标准与人类对美学与现实感的判断相关性较弱。
- 单轴聚焦。现有的方法通常只强调保真度或多样性，而不能同时兼顾。

最近的工作如 TIAM 的多阶段验证和 VISOR 的关系检查，指出了更丰富的、多方面的评估框架。未来的工作应旨在整合无参照感知度量、联合保真度-多样性度量以及更好地反映人类偏好的学习评估器。

4.2 基于嵌入的评估

基于嵌入的评估方法利用深度神经网络学习到的语义表示来评估生成的视觉内容，从而具备更大的灵活性和语义深度。这些方法沿着两个互补的研究轨道发展：

- 分布水平度量，通过比较真实和生成输出的整体特征分布，从早期的基于 CNN 的偏差到复杂的、无偏的时空度量。
- 实例级别的指标，每个样本单独评估，分为质量指标——评估视觉保真度——和对齐指标——通过利用视觉-语言和自监督嵌入的进步来衡量与文本或时间参考的一致性。

4.2.1 **分布层面的方法**. 这些指标通过比较真实和生成特征分布的统计属性，提供了一个全局评估。随着时间的推移，研究推出了更为稳健的偏差测度，减少了偏倚，并将覆盖范围扩展到新领域，例如视频。首次在图像评估任务中引入嵌入模型的是 Inception Score (IS) [328] 。它计算了条件标签分布与其在生成样本上的边际分布之间的 KL 散度

$$\mathrm{IS}(G) = \exp\left[\mathbb{E}_{x\sim p_g} D_{KL}(p(y \mid x) \parallel p(y))\right]. \tag{3}$$

。这一开创性的指标不仅捕捉了样本的真实性（低熵 $p(y \mid x)$ ）和多样性（高熵 $p(y)$ ），虽然在不同模式共享标签时，它可能对模式崩溃不敏感 [328] 。基于 IS 的双重关注点，Fréchet Inception Distance (FID) [103] 将真实和生成的 Inception 嵌入建模为高斯分布 $\mathcal{N}(\mu, \Sigma)$ 和 $\mathcal{N}(\mu', \Sigma')$ ，然后计算它们之间的 Fréchet（2-Wasserstein）距离：

$$\mathrm{FID} = \|\mu - \mu'\|^2 + \mathrm{Tr}(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{1/2}). \tag{4}$$

这种方法直接比较特征分布，通过减轻模式崩溃展现出与人类判断更强的相关性 [103] 。为了放宽 FID 中的高斯假设，核式 Inception 距离（KID）[15] 使用了一个不偏的最大平均差异（MMD）估计器，该估计器在 Inception 特征上应用了多项式核。这种替代消除了有限样本偏差，同时保持了可靠的分布相似性估计 [15] 。为了将准确性与多样性分离开，改进的精度 & Recall [159] 无参数地估计了精度（样本质量）和召回率（覆盖率）流形。通过分别评估这些指标，可以更清楚地了解生成性能 [159] 。将 FID 框架扩展到视频上，Fréchet 视频距离（FVD）计算帧间 3D 卷积特征上的 Fréchet 距离。这个扩展捕捉了时空方面，并与人类对视频质量的评价更紧密地对齐 [383] 。CMMD [125] 通过将丰富的 CLIP 嵌入与基于核的双样本检验配对，超越了简单的高斯假设。它不仅对均值和协方差建模，而是使用高斯 RBF 核在 CLIP 特征上计算最大平均差异（MMD），使其能够在没有任何正态性假设的情况下检测真实和生成图像之间的复杂分布差异，结果是一个更健壮、不偏且样本高效的质量度量。

4.2.2 **实例级方法**. 虽然分布级别的指标捕捉了跨多个样本的全局趋势，实例级别的方法则专注于通过利用丰富的嵌入表示来单独评估每个生成的图像（或视频）。这些方法分为两个互补的类别：（1）质量指标，评估每个样本的感知和结构保真度，以及（2）对齐指标，衡量图像（或帧）与其文本或时间参考的匹配程度。

质量指标。. 质量指标旨在根据每张图像预测人类的感知判断。CLIP-IQA [401] 在 CLIP 的共享图像-文本嵌入空间中操作，使用精心设计的文本提示（例如，反义词对）来生成美学和技术质量的零样本分数；这些分数已显示出与人类对图像保真度的评估密切吻合。在由特定主题驱动生成的场景中（例如，“将此人插入新场景”），DINO 分数 [320] 使用自监督的 DINO 特征量化模型在多大程度上忠实地保持了该主题的外观，将嵌入空间中的结构一致性与感知的视觉保真度联系起来。

对齐度量. 对齐度量评估生成的视觉内容与其提示（或在视频中，其时间上下文）之间的语义一致性。CLIPScore [102] 计算图像的 CLIP 嵌入与其提示文本嵌入之间的余弦相似度，无需任何人工提供的真实值，即可对文本到图像的对齐进行稳健的、无参考的评估。扩展这一检索视角，CLIP-R-Precision [285] 通过其与给定图像的相似性评分的排名来对一组候选标题进行排序，为文本到图像任务提供了一种基于检索的对齐度量。TIGEr [131] 采用跨模态检索技术更广泛地测量语义对齐，在各种场景中比较图像和描述，以保持提示到图像的一致性。对于 3D 资产生成，HyperScore [489] 将嵌入投射到双曲空间——捕捉层次结构和几何关系——以评估生成几何图形的空间一致性。最后，对于文本到视频任务中的动态内容，SD-Score [220] 结合了帧级的图像-文本对齐和时间一致性惩罚，因此反映了随时间的视觉质量和运动保真度。

4.2.3 **总结与未来方向**. 基于嵌入的视觉指标分为分布级和实例级方法。分布级方法（例如，FID、KID、FVD）比较来自真实和生成数据的深度特征的总结统计，提供快速的整体质量检查，但有时会忽略细节。

实例级方法使用每个样本的嵌入来评估视觉保真度（例如，CLIP-IQA，DINO 评分）或提示对齐（例如，CLIPScore，CLIP-R-Precision）。这些零次度量不需要真实参考与人类对单个图像和视频的判断有很好的相关性。

综合来看，这些技术提供了一个灵活且无需参照的工具包：分布测试可以标记大规模的模式崩溃或漂移，而实例级评分则诊断每个样本的质量和相关性。未来的工作可能会将这两种观点统一为单一的、高效的指标，以捕捉整体的逼真度、语义对齐和感知微妙之处。

Method	Task	Category	Reference	Protocol
Distribution-level				
IS [328]	Text-to-Image Generation	Distribution	Yes	Corpus
FID [103]	Text-to-Image Generation	Distribution	Yes	Corpus
KID [15]	Text-to-Image Generation	Distribution	Yes	Corpus
Improved P & R [159]	Text-to-Image Generation	Distribution	Yes	Corpus
CMMD [125]	Text-to-Image Generation	Distribution	Yes	Corpus
FVD [383]	Video Generation	Distribution	Yes	Corpus
Instance-level: Quality Metrics				
CLIP-IQA [401]	Text-to-Image Generation	Quality-specific	No	Single
DINO [320]	Subject-Driven Generation	Quality-specific	Yes	Single
AC-T2I [12]	Text-to-Image Generation	Fidelity	Yes	Single
Instance-level: Alignment Metrics				
CLIPScore [102]	Text-to-Image Generation	Similarity	No	Single
CLIP-R-Precision [285]	Text-to-Image Generation	Retrieval	No	Single
TIGEr [131]	Image Captioning	Semantic	Yes	Single
HyperScore [489]	Text-to-3D Generation	Geometric	No	Single
SD-Score [220]	Text-to-Video Generation	Temporal	Yes	Single

Table 11. **基于嵌入的视觉生成任务评估指标，按分布级和实例级方法分组，并提供参考要求和评估协议的详细信息。**

4.3 基于学习的评价

基于学习的评估通过在人类标注的数据上训练神经预测模型来近似感知判断。下面，我们整理了四个主要范式——分类、嵌入、奖励建模和回归——每个范式都有简要的概述，接着是代表性方法的简要描述。

基于分类的方法。. 这些方法将质量评估视为一个辨别任务：通过训练一个分类器来区分真实图像和生成图像或识别高质量与低质量的输出，其准确性作为评估分数。C2ST [232] 在混合真实和合成样本上训练一个二元分类器；在零假设下，即两个分布相匹配时，测试集准确率应约为 50 %，因此任何偏差量化感知上的差异。CAS [309] 仅使用生成的图像-标签对构建分类器，并在真实数据上测量其 Top-1/Top-5 的准确性；准确性下降揭示了条件生成器在语义或类条件忠诚度上的差距。

基于嵌入的方法。. 嵌入式方法重用预训练的视觉–语言编码器，将图像（及其提示）映射到共享特征空间，在该空间中，利用相似性或距离来估计质量。PickScore [147] 在 50 万个用户偏好判断上微调了 CLIP 骨干网络，实现了超过人类的与人类排名的相关性，并支持绝对评分和成对比较。HPS-v2 [429] 在 HPD v2 语料库的 79.8 万人类偏好对上改进了 CLIP，预测用户在不同模型和数据中更喜欢哪一对的图像；它具有很强的泛化能力，并且完全不依赖参考。

奖励建模方法。. 奖励模型通过成对比较直接学习预测人类偏好，得到反映微妙质量区别的标量奖励。ImageReward [434] 在 137K 专业策划比较上进行训练，在匹配人类选择方面超过了 CLIP 和美学预测器超过 30 %，并且还可以作为强化学习信号来微调扩散模型，以实现更好的对齐。

基于回归的方法。. 基于回归的评估器通过学习平均意见评分，将图像（及其提示）映射到连续质量评分，提供可解释的绝对评估。IP-IQA [303] 加强了 CLIP，通过一个 Image2Prompt 预训练任务和交叉注意力融合，将图像和提示注入到一个特殊标记中；它在 AGIQA-1K/3K 基准测试中实现了最先进的绝对评分。MPS [484] 将人类偏好分解为四个轴——美学、语义对齐、细节质量和整体——并在 918 K 次比较中训练专门的回归器，实现了与用户判断紧密匹配的多维绝对评分。

Method	Task	Category	Reference	Protocol
LPIPS [482]	Image Generation	Embedding	Yes	Single
C2ST [232]	Text-to-Image	Classification	No	Pair
CAS [309]	Text-to-Image	Classification	No	Single
DreamSim [72]	Text-to-Image	Embedding	Yes	Single
ImageReward [434]	Text-to-Image	Reward	No	Single & Pair
PickScore [147]	Text-to-Image	Embedding	No	Single & Pair
HPS-v2 [429]	Text-to-Image	Embedding	No	Pair
MPS [484]	Text-to-Image	Regression	No	Single
IP-IQA [303]	Text-to-Image	Regression	No	Single

Table 12. **针对视觉生成任务的基于学习的评估方法的分类。方法按照技术方法（类别）、对参考输入的需求（参考）和评估协议（协议）进行分类。混合协议支持绝对评分和成对比较。**

利用近年来大型语言模型（LLMs）和多模态大型语言模型（MLLMs）的推理和多模态能力，我们提出了一种统一的分类法，以捕捉视觉生成任务的五大核心评估范式。如表 13 所示，我们的分类法按照机制、可解释性和协议区分方法，包括直接评分、QA 框架、可解释/混合指标、偏好建模和时间一致性评估。

直接评分方法. 这些方法利用遵循指令的 LLMs 在一次计算中得出对齐分数。该领域的早期工作将对齐视为一个语义相似性任务，LLMScore 通过利用文本描述来估计生成的图像与其提示语的匹配程度 [236]，而 VQAScore 解释了用于组合分数的概率 QA 输出 [210]。虽然这两种方法都是无参考且轻量级的，但 VQAScore 的黑箱概率聚合为了效率牺牲了可解释性，而 LLMScore 的结构化提示则提供了更透明的推理。

问答框架. 基于 QA 的方法将评估分解为离散的问答交互，以探测图像保真度的不同方面。诸如 QAS 的多模态 QA 方法应用开放式视觉问答来验证对齐和感知细节 [352]，而诊断 QA 技术如 DSG 生成结构化场景图问题以确保彻底的语义覆盖 [38]。以 TIFA v1.0 为例的分层 QA，通过将问题组织成层级（对象、计数、关系）来进一步优化这种方法，以定位具体的失败模式 [111]。以 PhyBench 为代表的评分标准 QA，提出了详细的物理常识推理标准——通过显式评估基于物理的合理性来弥补早期 QA 方法的不足 [257]。从通用 QA 向这些专业化评分标准的转变解决了早期框架的粗粒度限制，并更好地量化推理深度。

可解释和混合方法. 这些方法利用多模态大模型（MLLM）生成自然语言解释，并结合无参考和有参考的线索。X-IQE 生成思维链推理，同时为对齐、美学和真实性打分 [32]，而 VIEScore 则通过视觉指令调整将对齐和感知判断融合到一个统一的分数中 [149]。最新的 T2I-Eval 工具包通过在基于参考的比较和基于提示的问答之间动态选择，增强了这种融合，提高了在多样化生成风格上的稳健性 [378]。

偏好建模．基于偏好的方法将质量评估重新定义为成对排序。Gecko 使用大型语言模型驱动的偏好判断，根据语义保真度对候选输出进行排序，从而减轻单实例方法固有的绝对评分偏差，并实现更细致地对高质量图像进行比较 [422] 。

时间的．时间方法通过验证时间上的断言扩展了视频生成中的对齐评估。TCR 度量通过检查帧间的过渡完成情况来确保叙事一致性，而 TC-Score 将帧级别的检查汇总为语料库级别的测量，为动态内容提供可扩展的评估 [68] 。

Method	Task	Interpretability	Category	Dimension	Reference	Protocol
LLMScore [236]	Text-to-Image	High	Direct Scoring	Alignment	No	Single
VQAScore [210]	Text-to-Image	Low	Direct Scoring	Alignment	No	Single
QAS [352]	Text-to-Image	Medium	Multimodal QA	Alignment & Perceptual	No	Single
X-IQE [32]	Text-to-Image	High	Hybrid Interpretable Metric	Alignment & Perceptual	No	Single
DSG [38]	Text-to-Image	High	Diagnostic QA	Alignment	No	Single
TIFA v1.0 [111]	Text-to-Image	High	Hierarchical QA	Alignment	No	Single
PhyBench [257]	Text-to-Image	High	Rubric-based QA	Alignment & Commonsense	No	Single
VIEScore [149]	Text-to-Image	Medium	Hybrid Interpretable Metric	Alignment & Perceptual	No	Single
Gecko [422]	Text-to-Image	Medium	Preference Modeling	Alignment	No	Pair
T2I-Eval [378]	Text-to-Image	High	Hybrid QA	Alignment & Perceptual	Both	Single
TCR [68]	Video Gen	High	Temporal Verification	Alignment	No	Corpus
TC-Score [68]	Video Gen	High	Frame-level Checking	Alignment	No	Single

Table 13. **基于 LLM/MLLM 的评估方法的分类，按评估协议（单实例、成对、语料库级别）和主要评估维度进行分类。参考要求表示是否需要对比真实值。**

值得注意的方法创新包括 T2I-Eval 的 [378] 混合方法，通过动态提示选择结合基于参考和无参考的评估，以及 TCR 的 [68] 将时间断言验证应用于视频生成的新方法。然而，目前在评估开放领域创意生成和量化微妙感知质量方面仍存在局限性，这为基于 MLLM 的评估研究指明了未来的方向。

4.4 基准测试评估

基准测试的评估方法使用系统构建的数据集，结合人工标注或合成标准来建立标准化的评估框架。这些基准通过三个主要维度实现可重复且可量化的生成模型比较：(1) 参考要求（基于参考 vs. 无参考），(2) 评估范式（自动指标 vs. 人类判断），和 (3) 评估标准（任务特定能力 vs. 跨模态对齐）。如表 14 所示，现代基准测试展示了三个进化趋势：在任务特定评估方面的专业化程度增加（例如，视频生成中的时间一致性），集成多模态大型语言模型（MLLMs）进行语义对齐评估，以及结合自动指标与人类偏好建模的混合方法。

基于参考的方法需要真实样本进行比较评估。例如，VTAB+ [330] 使用准确性指标建立了涵盖 35 个视觉任务的多任务基准，而 C-CUB & C-Flowers [285] 则专注于通过细粒度属性匹配进行组合图像生成。这些方法在受控比较中表现出色，但在开放式生成任务中面临可扩展性挑战。

无参考方法利用模型驱动评估而不需要目标输出。PhyBench [257] 基准通过基于 LLM 的推理链评估生成图像的物理常识，而 T ^{3} Bench [98] 则结合神经评分器与 LLM 评估器进行 3D 资产对齐。近期的进展如 M ^{2} RAG [243] 证明了用于多模态一致性评估的检索增强评估的潜力。

专业评估分类法在不同模态中涌现：

- 图像生成：MJ-Bench 中的分层评估 [36] 涵盖了安全检查（NSFW 检测）、语义对齐（CLIPScore）和感知质量（FID）。
- 视频生成：TC-Bench [68] 引入了使用 LLM 基于轨迹分析的时间组合性指标
- 3D 生成：MATE-3D [489] 结合了几何一致性指标与常识推理评估

新兴的混合方法融合了多种评估范式。Pick-a-Pic [147] 结合了人类偏好建模与成对比较指标，而 LenCom-EVAL [163] 则集成了 OCR 验证与神经语义评分，用于复杂文本渲染。该领域正在逐渐从单一指标评估转向多维评估框架，以解决低层次感知质量和高层次语义保真度。

Benchmark	Task	Evaluation Criteria	Metrics
Pick-a-Pic [147]	Text-to-Image	Human preference modeling	Pairwise comparison
JourneyDB [352]	Text-to-Image	Prompt-image comprehension	Heuristic scoring
TC-Bench [68]	Text/Image-to-Video	Temporal consistency	LLM-based trajectory analysis
PhyBench [257]	Text-to-Image	Physical commonsense	LLM reasoning chains
VTAB+ [330]	Multi-task	Cross-task generalization	Accuracy
T2I-CompBench++ [115]	Text-to-Image	Compositional alignment	Attribute matching
T2V-CompBench [351]	Text-to-Video	Cross-frame coherence	MLLM/Detection/Tracking
MJ-Bench [36]	Text-to-Image	Safety & Quality	NSFW detection, CLIPScore
C-CUB/Flowers [285]	Text-to-Image	Fine-grained alignment	CLIPScore, Human eval
ImagenHub [150]	Image Edit	Semantic preservation	Human evaluation
VR [368]	Image-to-Text	Visual relation capture	Heuristic scoring
EditBench [405]	Image Inpainting	Context consistency	Human eval
LenCom-EVAL [163]	Text-to-Image	Complex text rendering	CLIPScore, OCR, NLD
MATE-3D [489]	Text-to-3D	Geometric consistency	Point cloud analysis
EvalCrafter [220]	Text-to-Video	Multi-aspect quality	Dover, IS, CLIPScore
FETV [223]	Text-to-Video	Temporal alignment	BLIPScore, CLIPScore
HRS-Bench [12]	Text-to-Image	Human resemblance	Face detection metrics
GenAI-Bench [179]	Multi-modal	Cross-modal alignment	VQAScore
T^3 Bench [98]	Text-to-3D	Semantic fidelity	LLM-based scoring
T2I-Eval-Bench [378]	Text-to-Image	Perceptual & Alignment	MLLM-based
ConceptMix [430]	Text-to-Image	Concept integration	LLM-based QA
M^2 RAG [243]	Multi-modal	Contextual coherence	Retrieval accuracy

Table 14. **基于基准的视觉生成任务评价方法的分类，按模态、评价重点和度量范式分类。表格突出了三个关键维度：参考要求（真实值依赖性）、评价范式（自动 vs. 人工）、以及评估重点（特定模态能力 vs. 跨模态对齐）。**

4.5 视觉生成的元评估基准

Benchmark	Task	Type	Focus	Protocol
TIFA v1.0 [111]	Text-to-Image	Objective	Alignment	Single
DSG-1k [38]	Text-to-Image	Objective	Alignment	Single & Corpus
GenAI-Bench [179]	Text-to-Image/Video	Objective	Overall Quality	Pair
T2I-Eval [378]	Text-to-Image	Objective & Subjective	Perceptual & Alignment	Single

Table 15. **视觉生成任务的元评估基准分类，按模式、评估重点和度量范式进行分类。表格突出显示了三个关键维度：参考需求（对真实值的依赖性）、评估范式（自动 vs. 人工）和评估重点（针对特定模式的能力 vs. 跨模式的一致性）。**

如表 15 所示，目前只有少数公开可用的基准测试设计用于评估自动评估指标在文本到视觉生成任务中与人类判断的相关程度。这些基准测试在以下三个主要方面有所不同：(1) 任务模式：四个基准测试都主要集中在文本到图像合成上，但 GenAI-Bench 进一步扩展到视频生成评估。(2) 评估类型：TIFA v1.0 和 DSG-1k 仅采用纯客观的基于参考的测量，而 T2I-Eval 结合了客观和人类（主观）的评级以捕捉感知质量。(3) 协议：TIFA v1.0

和 T2I-Eval 遵循单实例协议（每个候选实例独立评判），而 DSG-1k 和 GenAI-Bench 还支持语料库级聚合，允许在一组示例中评估度量的一致性。

由于视觉生成模型的元评价研究处于早期阶段，目前的基准仅提供了对当前指标可靠性的有限视角。例如，TIFA v1.0 构建在相对较小的高质量图像标题集之上，强调对基于 CLIP 的分数的校准；相比之下，DSG-1k 提供了更大且更多样化的数据集，但仍然是纯自动化的。GenAI-Bench 将视频数据纳入其中，标志着朝向多模态评估的重要一步，但其仍然依赖于标准的客观协议。最后，T2I-Eval 的混合设计展示了通过人类感知判断补充基于参考的评分的价值，但其范围仅限于单图像。

根据目前研究进展的观察，我们可以总结出一些关键见解和未来方向：

- 纳入更丰富的主观评价。正如 T2I-Eval 所显示的，仅靠客观评分可能会遗漏视觉一致性和创造力的细微差别；通过可扩展的、低成本的方法（如众包的成对比较）扩展大规模人工注释，将增强基准的有效性。
- 拓宽协议多样性。语料库级别的协议捕捉图像间的一致性以及对数据集偏差的鲁棒性；未来的基准测试应系统地比较单一与语料库级别的相关性，涵盖更广泛类型的内容（例如，卡通片，医学图像）。
- 扩展模态覆盖范围。随着视频、3D 和交互式图形成为主流，元评估框架必须发展以处理时间动态和空间复杂性，也许可以通过整合时空对齐分数和人体运动判断来实现。
- 促进开放、模块化的基准套件。一个由社区驱动的基准，可以将客观库（例如 CLIP、基于 ViT 的评分）与即插即用的人 - 在 - 环模块统一起来，这将通过允许研究人员在一致、可扩展的协议下评估新指标来加速进展。

通过填补这些空白——更丰富的人类信号、多样化的协议、扩展的模态和模块化设计——未来的元评估基准可以更全面地描述文本到视觉测量的真实优缺点，最终指导开发出更可靠且符合人类的评估方法。

4.6 当前方法的比较与未来有前景的方向

Table 16. **文本到图像生成任务的评价方法比较。**

Category	Method	T2I-Eval		Tifa v1.0	
		ρ	τ	ρ	τ
Embedding-based	FID [103]	-0.1231	-0.0862	-	-
	CLIPScore [102]	0.1505	0.1016	0.3382	0.2456
	BLIPv2Score [185]	0.2152	0.1423	0.4049	0.2944
Learning-based	LPIPS [482]	-0.1244	-0.0856	-	-
	DreamSim [72]	-0.1382	-0.0968	-	-
	PickScore [147]	0.3944	0.2803	0.4279	0.3137
	ImageReward [434]	0.4046	0.2839	0.6211	0.4659
LLM-based & MLLM-based	LLMScore $_{G}PT-4$ [236]	0.3096	0.2228	0.4969	0.3753
	TIFA $_{m}PLUG$ [111]	0.3252	0.2455	0.5922	0.4717
	DSG $_{D}ependent$ [38]	0.4582	0.3512	0.6046	0.4893
	DSG $_{I}ndependent$	0.4704	0.3655	0.6108	0.4954
	VQAScore $_{C}LIP-FlanT5$ [210]	0.5116	0.3712	0.6950	0.5321
	VIEScore $_{G}PT-4o$ [149]	0.5545	0.4170	0.5388	0.4065
	T2I-Eval $_{M}iniCPM-V-2.6$ [379]	0.5802	0.4409	0.6061	0.4692
	T2I-Eval-R1 [244]	0.5874	0.4380	0.7043	0.5510

对两个人工标注的文本到图像基准（T2I-Eval 和 TIFA v1.0）的定量比较揭示了三个明显趋势（表 16 ）：

随着骨干模型的进步（从 Inception-v3（FID）到 CLIP（CLIPScore），再到 BLIP-v2（BLIPv2Score）），基于嵌入的方法稳步提高，但它们与人工评判的相关性仍然有限（例如，BLIPv2Score 在 T2I-Eval 上达到 ρ =0.4049，在 TIFA v1.0 上达到 ρ =0.2944）。

基于学习的度量方法直接以人类偏好进行训练，例如 PickScore 和 ImageReward，提供更强的对齐（ImageReward 在 T2I-Eval 中达到 ρ =0.6211，在 TIFA v1.0 中达到 ρ =0.4659）。这些模型能够捕捉到超越静态嵌入的细致感知和语义判断。

基于 LLM/MLLM 的评估器提供了最高的相关性。例如，使用 MiniCPM-V-2.6 的 T2I-Eval 和使用 CLIP-FlanT5 的 VQAScore 在所有其他范式中表现突出。它们能够将视觉内容分解为丰富且具有指导性的推理，这对于类人评估至关重要。

尽管这些进展，视觉评估仍处于初期阶段。目前的基准测试在规模、模式覆盖和评估协议方面都有限。为了推动进步，我们确定了四个有前途的方向：

- 扩展和多样化元评估数据集：未来的基准测试应超越静态图像对，加入视频、3D 资产和交互式图形，并注重时间一致性、几何结构和用户体验。规模更大、来源广泛的收集将提高统计可靠性，并涵盖罕见或创造性的失败模式。
- 开发多维的、混合的指标：单一得分的输出会掩盖忠实性、多样性和美感之间的权衡。结合分布性测量（例如，PRD）、嵌入对齐和由 LLM 衍生的评论在一个模块化框架中，将使研究人员能够根据特定应用需求定制评估。
- 集成人机协作：利用大型语言模型（LLMs）提出候选批评，然后由人工标注者进行完善或验证，可以以更低的成本获得高质量的标签。随着模型的演化，这一流程可以启动基于学习的评估器，并持续更新基准测试。
- 标准化可扩展的开放评估平台：一个由社区驱动的即插即用模块库——涵盖基于参考的度量、无参考评分器、LLM/MLLM 评估器以及人机交互界面——将使得在一致的协议下进行同类对比和新方法的快速测试成为可能。

通过解决这些缺口——更丰富的注释、混合评估策略、人机协作和开放工具——我们可以向更加稳健、具有普遍性以及人性化的视觉生成评估迈进。

5 音频和语音相关生成的自动评估

在本节中，我们回顾了音频和语音生成任务的自动评估方法的发展，包括文本转语音（TTS）、语音转换（VC）、文本到音频生成（TTA）、及生成性音频-语言模型（ALMs）。随着这些领域生成系统能力的增强，评估技术也相应演变。这些方法的发展可以分为五个不同的范例，如图 4 所示：（1）启发式评估。早期方法是基于简单的规则度量，如信噪比、对数频谱距离和韵律统计，用于量化失真或保真度。虽然计算上负担轻，但这些方法在复杂韵律或对说话者敏感的环境中通常与人的感知判断不一致。（2）基于嵌入的评估。随着自监督学习的兴起，基于嵌入的方法利用从模型中学习到的音频表示，例如 Wav2vec [9, 329] 、HuBERT [108] 、和 VGGish [101] ，用于测量声学相似性、说话者身份和时间动态。这些嵌入捕捉了音频信号的细小特征，并能够比手工度量方法进行更具感知相关性的比较。（3）基于学习的评估。这一类别的方法依赖于在诸如语音、一般音频或音乐等音频模态数据上训练的监督或自监督模型，以及人工标注或自动生成的质量评分。从语音生成的角度来看，这些模型旨在近似通用和专家级的主观评估指标，包括平均意见得分（MOS）。通过直接学习预测这些评分，模型旨在估计自然性和可理解性等关键感知属性。通过从反映人类偏好的标注数据中学习，这些模型通常表现出与主观判断更好的对齐，并在任务中表现出强大的泛化能力。在音频生成的更广泛背景下，这种范式通常使用在包含与质量评分配对的音频样本的数据集上训练的神经网络。这些注释可以通过人工标注获得，

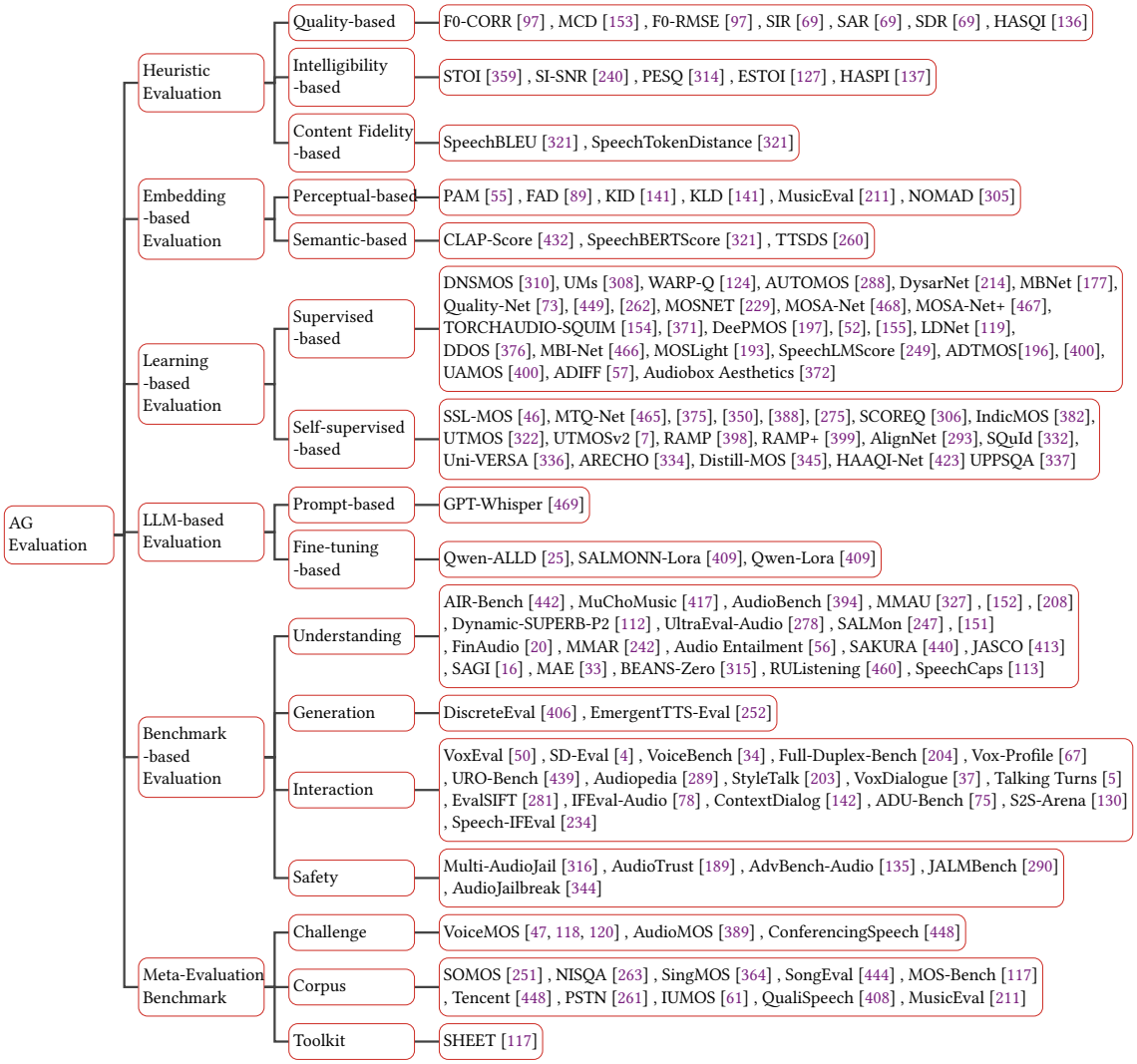

Fig. 4. **音频和语音生成（AG）自动评估指标研究的分类法。有关自动评估方法的详细分类，可参见第 5 节。**

也可以使用大型语言模型进行丰富。目标是实现对多种感知维度的自动评估，如差异、美学质量和内容一致性。例如，类似 ADIFF [57] 的方法结合由大型语言模型生成的标签增强来增加训练数据的多样性和丰富性。尽管这些技术利用大型模型来改进标注过程，但其基础评价方法仍然根植于监督学习。因此，它们被归类为基于学习的评价范式。(4) 基于 LLM/ALM 的评价。大语言模型和音频-语言模型的最新进展使得音频和语音生成的整体评价机制成为可能。这些模型不仅能够产生可与平均意见评分（MOS）评级相比较的标量预测，还能基于声学特征和语义内容提供解释性反馈和描述性评估。在大规模的音频文本语料库上训练的音频语言模型表现出很强的泛化能力，其性能可以通过在语音评估数据集上进行微调来进一步提升。它们跨越语音和文本对齐多模态表示的能力可以进行更全面和可解释的评估。因此，这种范式代表了一种从传统的基于学习的评估迈向新一代以 LLM 和 ALM 驱动的评估方法的转变。在这篇调查中，我们系统性地整合了该领域的最新工作，

并在第 5.7 节中提供了基于学习的方法与基于 LLM/ALM 的方法之间的比较分析，重点放在评估性能和泛化能力上。(5) 基准测试的评估。为了支持强大的和标准化的系统级比较，近年来出现了几个基于基准测试的对于 ALMs 和 MLLMs 的评估套件。这些基准测试旨在系统地评估模型在多个维度的能力，包括在语音、音频和音乐等模态中的理解、交互、推理和生成。与文本到图像或文本到视频生成中已经成熟的基准相比，文本到语音、音频或音乐生成的评估协议相对尚未被充分探讨。虽然像 LibriSpeech [280] 和 LibriTTS [462] 这样的高质量数据集被广泛用于训练语音生成模型，以及像 TTS-Arena [268] 这样的努力通过人机交互评分提供严格的排行榜型评估，而 Seed-TTS [1] 则提供了标准化和可靠的测试集，但这些努力不属于我们定义的用于评估 ALMs 和 MLLMs 能力的自动化、通用性和多任务可扩展性基准的范围。因此，我们认为在 ALM 层面上开发一个统一的、可扩展的语音和音频生成评估框架代表了一个重要但尚未充分研究的研究方向。在此背景下，我们将基于基准的评估定义为一种自动评估模式，其中使用一套标准化任务来评估 ALM 在语音和音频相关场景中的多模态能力，这包括理解、交互、推理和生成。值得注意的是，最近的大多数基于基准的评估套件越来越多地采用由语音和音频生成模型生成的语音或音频作为 ALM 或 MLLM 的输入，以评估其在理解、交互、推理和相关任务中的能力。尽管这些基准最初并非为生成评估而设计，但它们仍然为评估生成质量提供了有意义的信号。尤其是在理解任务中评估语义一致性、事实对齐和音频文本连贯性等方面，为模型改进提供了关键的诊断见解。因此，这些新兴基准便成为我们所认为的 ALM 基于基准评估的核心基础。

在这个概念概述之后，我们介绍了用于元评估语音和音频生成模型的代表性基准套件（第 5.6 节）。最后，我们讨论在 ALM 和 MLLM 时代推进自动评估的持续挑战和潜在方向（第 5.8 节）。

5.1 启发式评估

启发式评价指的是使用预定的、通常是基于规则或统计的指标来评估合成语音。这些方法轻量、可解释，并且广泛用于评估音频和语音信号的核心属性。根据它们的评估目标，我们将这些方法分为三大类型，如图 4 所示，并在表 17 中总结：(1) 基于质量的评估，衡量生成语音的声学保真度和信号级相似性；(2) 基于可理解性的评估，评估在各种条件下语音被清晰感知和理解的程度；(3) 基于内容保真的评估，重点关注生成内容的符号或语义准确性，通常通过基于转录的比较。

5.1.1 . 基于质量的评估 基于质量的评估根据声学特性评估合成语音与自然录音的一致性。已经开发出一系列的指标来捕捉如频谱相似性、音高准确性和信号失真等方面的特征。在频谱指标中，梅尔倒谱失真（MCD）[153] 被广泛用于量化生成语音与参考语音在帧级别频谱差异，从而实现短期频率对齐的精细分析。为了评估韵律准确性，F0-CORR 和 F0-RMSE [97] 用于测量音高轮廓的对齐和偏差，反映出语调和节奏的自然性。除了帧级别特征外，信号级别指标提供全面评估。具体而言，信号失真比（SDR）、信号干扰比（SIR）和信号伪影比（SAR）[69] 在增强和分离任务中是标准指标，分别捕捉整体失真、源干扰和伪影存在。感知驱动的指标也已被引入，特别是在助听器情境下。例如，助听器语音质量指标（HASQI）[136] 整合了听觉模型来估计感知质量，并在听力损伤场景中有效。

5.1.2 . 基于可理解性的评估 可懂度评估用于衡量合成语音在噪声或退化环境中能够被感知和理解的清晰程度。这些指标对于评估语音在实际条件下的可用性至关重要。短时目标可懂度（STOI）[359] 通过关联干净语音和退化语音的时间包络来估计可懂度，在静态噪声中表现出强大的鲁棒性。它的扩展形式，扩展 STOI（ESTOI）[127]，考虑了频带间的依赖性，提高了在调制和非稳定噪声中的敏感度。除了包络关联，感知模型提供了更直接的可懂度估计。语音质量的感知评估（PESQ）[314] 使用掩蔽和响度模型模拟听觉感知，以评估噪声和失真造成的退化。尺度不变信噪比（SI SNR）[240] 通过比较目标和估计信号并进行尺度归一化来衡量可懂度，这在语音分离和增强中特别有效。助听器语音感知指数（HASPI）[137] 通过模拟听觉外围响应来预测

各种退化条件下的可懂度。这些方法结合在一起，从声学细节到感知相关性，在不同抽象层次上提供互补的见解，并被广泛用于评估语音修复系统。

5.1.3 . 基于内容保真度的评估 内容保真度评估度量了合成语音与其参考之间的符号和语义一致性，通常通过比较它们的转录或标记化形式来实现。这对于以文本为条件的生成任务至关重要，因为保持语言内容是首要的。SpeechBLEU [321] 将传统的 BLEU 指标扩展到语音领域，通过计算来自自监督模型的离散语音标记上的 n 元组重叠，有效地捕捉词汇相似性，同时对声学细节保持不变。与之并行，SpeechTokenDistance [321] 使用编辑距离度量（如 Levenshtein 或 Jaro Winkler）评估序列级别的一致性，提供更加精细的标记对齐和顺序度量。这些方法在评估语义准确性方面尤为有用，如文本到语音合成和音频描述。总体而言，内容保真度度量提供了一种符号视角，补充了基于声学和可懂度的评估，并对于验证以文本为条件的语音生成系统的语义完整性至关重要。

启发式度量因其简便性、高效性和可解释性而被广泛用于评估合成语音。然而，每种类别都有其固有的局限性。基于质量的度量强调声学相似性，但常常未能捕捉感知自然性和语义连贯性，导致在表达性生成任务中与人类判断的契合度有限。基于可理解性的度量在受控降级环境中可靠地评估语音清晰度，但对失真敏感并且通常忽略内容忠实度。相比之下，基于内容忠实度的度量评估符号和语义准确性，通常通过转录级别的比较进行，但可能受到声学–韵律不匹配的影响。在这些度量中，编辑距离度量已展示出与人类感知更强的相关性，特别是在捕捉跨发言者和不同长度输入的内容一致性方面。总体而言，这些度量提供了互补的见解，并且在语言条件语音生成基准测试中仍然必不可少，同时为开发与感知性和语义性更一致的评估框架提供参考。

Method	Task	Objective	Dimension	Function Type
MCD [153]	Speech Generation	Signal Quality	Frame	Spectral Distance
F0-RMSE [97]	Signal Restoration	Signal Quality	Frame	Pitch Error
SDR [69]	Signal Restoration	Signal Quality	Utterance	Signal-to-Noise Ratio
SIR [69]	Signal Restoration	Signal Quality	Utterance	Interference Ratio
SAR [69]	Signal Restoration	Signal Quality	Utterance	Artifact Ratio
HASQI [136]	Speech Generation	Signal Quality	Utterance	Perceptual Quality Model
HASPI [137]	Speech Generation	Intelligibility	Utterance	Perceptual Intelligibility Model
STOI [359]	Signal Restoration	Intelligibility	Utterance	Temporal Envelope Correlation
ESTOI [127]	Signal Restoration	Intelligibility	Utterance	Inter-band Temporal Correlation
PESQ [314]	Signal Restoration	Intelligibility	Utterance	Perceptual Evaluation Model
SI-SNR [240]	Signal Restoration	Intelligibility	Utterance	Scale-Invariant SNR
SpeechBLEU [321]	Speech Generation	Content Consistency	Utterance	n-gram Token Overlap
SpeechTokenDistance [321]	Speech Generation	Content Consistency	Utterance	Token Sequence Edit Distance

Table 17. **语音生成启发式评估方法的分类，按照评估目标（信号质量、可理解性或内容一致性）、评估粒度（维度）和功能类型进行组织。维度列表示度量应用的时间分辨率。帧级度量在语音波形的短、固定长度段上操作，允许对局部声学特性如音高或频谱形状进行细粒度分析。相反，语句级度量是在整个语音段上计算的，捕捉整体特征如可理解性、失真或语义一致性。功能类型列反映了每种度量的方法学基础，指示其在评估过程中如何实现相似性、失真或语义对齐。**

5.2 基于嵌入的评价

基于嵌入的评估方法通过使用从深度神经网络中学习的高级感知和语义嵌入来解决启发式指标的局限性，已成为一种有前途的替代方案。这些方法促进了更具内容意识和感知对齐的评估框架，特别是在低级别声学指标无法捕捉与人类感知对齐的质量方面的情况下。表 18 中提供了代表性嵌入评估方法的总结，以及它们的评估目标、参考要求和协议。根据评估重点，这些方法可以分为两个主要类别：(1) 基于感知的指标，旨在近似人类的听觉感知；以及 (2) 基于语义的指标，评估音频与相应的语言或背景信息之间的对齐。

5.2.1 . 基于感知的方法

基于感知的指标通过在嵌入空间中的感知相似性建模来评估音频质量，而不是依赖于直接比较声学特征。一个代表性的例子是 PAM [55]，其使用提示音频-语言模型计算音频样本之间的成对距离，以逼近人类判断。基于分布的指标如 Fréchet Audio Distance (FAD) [141]、Kernel Inception Distance (KID) 和 Kullback–Leibler Divergence (KLD) 则作用于从真实和生成的音频信号中提取的嵌入的统计分布。这些方法通常被应用于通用音频生成任务，特别适用于没有参考信号或需要评估长形式、高维音频内容的情况。MusicEval-Score [211] 通过利用基于 CLAP 的嵌入来捕捉语料库级别的感知特征，将这种分布框架扩展到了音乐领域。此外，NOMAD [305] 引入了一种无监督的方法，通过在嵌入空间中将退化的语音信号与不相关的干净参考进行比较，估计感知音频质量，从而消除了对真实对齐或人为注释的需求。

5.2.2 . 基于语义的方法

基于语义的指标侧重于评估生成的音频内容是否与关联的文本或上下文信息保持语义一致性。CLAP-Score [432] 应用跨模态嵌入来评估音频与其参考文本在单一参考设置下的语义一致性。SpeechBERTScore [321] 通过比较从预训练语音编码器生成的嵌入以捕获配对话语的上下文一致性。最近，TTSDS [260] 引入了令牌级别的距离指标，以量化文本到语音（TTS）系统中的语义保真度。这些方法在语言调节生成场景（如 TTS 和音频字幕生成）中特别有效，在这些场景中，保持输入语义内容是至关重要的。此外，它们还能够检测语义错误，包括遗漏、幻觉内容和主题漂移，而这些错误通常被传统的基于声学的评估方法忽视。

基于嵌入的评估提供了一个统一且可扩展的框架，能够同时捕捉感知相似性和语义保真度，有效解决了传统启发式指标的主要限制。具体而言，基于感知的方法利用学习的嵌入来逼近人类的听觉感知，使得在各种声学条件下进行稳健且参考高效的评估成为可能。与此相辅相成，基于语义的方法专注于保持语言意义并确保上下文的一致性，这对于评估如 TTS 和 TTA 等语言条件生成任务至关重要。展望未来，大规模多模态嵌入架构，例如 Gemini [173] 和相关的视觉-语言-语音模型 [82, 426] 的发展，预计将扩大基于嵌入的评估的能力。这些进展可能有助于为涉及跨模态对齐和音频基础语义理解的复杂生成系统提供更全面和更具普遍性的评估策略。因此，基于嵌入的方法预计将在感知上符合实际、在语义上连贯且可扩展的评估框架的未来中发挥核心作用。

Method	Task	Category	Reference	Protocol
PAM [55]	Audio Generation	Perceptual-based	No	Pairwise
FAD [141]	Audio Generation	Perceptual-based	Yes	Corpus
KID [141]	Audio Generation	Perceptual-based	Yes	Corpus
KLD [141]	Audio Generation	Perceptual-based	Yes	Corpus
MusicEval-Score [211]	Music Generation	Perceptual-based	Yes	Corpus
NOMAD [305]	Audio/Speech Enhancement	Perceptual-based	Yes	Pairwise
CLAP-Score [432]	Audio Generation	Semantic-based	Yes	Single
SpeechBERTScore [321]	Speech Generation	Semantic-based	Yes	Pairwise
TTSDS [260]	Speech Generation	Semantic-based	Yes	Corpus

Table 18. **嵌入式评估方法的分类用于语音、音乐和音频生成任务，按照评估目标（感知或语义）、参考需求和评估协议进行分类。**

5.3 基于学习的评价

基于学习的评估通过在主观质量标注上训练模型来预测人类感知质量。这个范式已经成为评估语音和音频生成系统的基石，特别是用于估计感知属性，如平均意见分（MOS）。总体而言，现有的方法可以大致分为两大类，如表 19 所示：（1）基于监督的方法，直接在人工标注的 MOS 数据集上训练神经网络以模拟主观评估行

为；（2）基于自监督的方法，利用从大规模未标注语料库中获取的声学表示来提高预测性能，尤其在数据稀缺的情况下。

Method	Task	Category	Reference	Protocol
DNSMOS [310]	Text-to-Speech	Supervised-based	No	Single
UMs [308]	Text-to-Speech	Supervised-based	No	Single
WARP-Q [124]	Text-to-Speech	Supervised-based	Yes	Single
AUTOMOS [288]	Text-to-Speech	Supervised-based	No	Single
DysarNet [214]	Text-to-Speech	Supervised-based	No	Single
MBNet [177]	Text-to-Speech	Supervised-based	No	Single
Quality-Net [73]	Text-to-Speech	Supervised-based	No	Single
[449]	Text-to-Speech	Supervised-based	No	Single
[262]	Text-to-Speech	Supervised-based	No	Single
MOSNET [229]	Text-to-Speech	Supervised-based	No	Single
MOSA-Net [468]	Text-to-Speech	Supervised-based	No	Single
MOSA-Net+ [467]	Text-to-Speech	Supervised-based	No	Single
TORCHAUDIO-SQUIM [154]	Text-to-Speech / Text-to-Audio	Supervised-based	No	Single
[371]	Text-to-Speech	Supervised-based	No	Single
DeePMOS [197]	Text-to-Speech	Supervised-based	No	Single
[52]	Text-to-Speech	Supervised-based	No	Single
[155]	Text-to-Speech	Supervised-based	No	Single
LDNet [119]	Text-to-Speech	Supervised-based	No	Single
DDOS [376]	Text-to-Speech	Supervised-based	No	Single
MBI-Net [466]	Text-to-Speech	Supervised-based	No	Single
MOSLight [193]	Text-to-Speech	Supervised-based	No	Single
SpeechLMScore [249]	Text-to-Speech	Supervised-based	No	Single
ADT-MOS [196]	Text-to-Speech	Supervised-based	No	Single
UAMOS [400]	Text-to-Speech	Supervised-based	No	Single
ADIFF [57]	Text-to-Audio	Supervised-based	No	Pairwise
Audiobox Aesthetics [372]	Text-to-Audio	Supervised-based	No	Single
SSL-MOS [46]	Text-to-Speech	SSL-based	No	Single
MTQ-Net [465]	Text-to-Speech	SSL-based	No	Single
[375]	Text-to-Speech	SSL-based	No	Single
[350]	Text-to-Speech	SSL-based	No	Single
[388]	Text-to-Speech	SSL-based	No	Single
[275]	Text-to-Speech	SSL-based	No	Single
SCOREQ [306]	Text-to-Speech	SSL-based	Optional	Single
IndicMOS [382]	Text-to-Speech	SSL-based	No	Single
UTMOS [322]	Text-to-Speech	SSL-based	No	Single
UTMOSv2 [7]	Text-to-Speech	SSL-based	No	Single
RAMP [398]	Text-to-Speech	SSL-based	No	Single
RAMP+ [399]	Text-to-Speech	SSL-based	No	Single
AlignNet [293]	Text-to-Speech	SSL-based	No	Single
SQuId [332]	Text-to-Speech	SSL-based	No	Single
Uni-VERSA [336]	Text-to-Speech	SSL-based	Yes	Single
ARECHO [334]	Text-to-Speech	SSL-based	Optional	Single
Distill-MOS [345]	Text-to-Speech	SSL-based	No	Single
HAAQI-Net [423]	Text-to-Audio	SSL-based	No	Single
UPPSQA [337]	Text-to-Speech	SSL-based	No	Pairwise

Table 19. **学习基础的语音和音频生成评价方法的分类。方法根据生成任务（文本到语音或文本到音频）、学习策略（有监督或自监督）、参考依赖性（是否需要参考语音或信号）以及评价协议（单一、成对或语料库级别）进行分类。此分类法反映了当前主流学习型方法的多样性。**

5.3.1 . 基于监督的方法 监督模型在标注数据集上进行训练，以学习从声学信号到感知质量的直接映射。早期的工作如 DNSMOS [310] 、Quality-Net [73] 和 MBNet [177] 采用光谱或感知特征作为输入（例如 STFT 或 log-mel），并使用卷积或循环网络来执行 MOS 回归。为了增强时间建模和容量，最近的模型集成了先进的架构设计。例如，MOSNET [229] 和 MOSA-Net [468] 利用注意机制来捕捉长距离依赖性，而 MOSA-Net+ [467] 则通过残差连接来加深网络。轻量模型如 LDNet [119] 、MOSLight [193] 和 DeePMOS [197] 强调通过使用压缩友好的组件提高推断效率。几个监督方法也处理域泛化和不确定性。SpeechLMScore [249] 使用语音语言预训练来实现更具迁移性的表示。UAMOS [400] 和 ADT-MOS [196] 融合了不确定性建模和域适应，以增强在不匹配条件下的鲁棒性。ADIFF [57] 聚焦于对齐敏感场景，引入了交叉投影模块及多阶段训练策略，以捕捉音频录制之间的细粒度语义和情感降级。Audiobox Aesthetics [372] 提出了一种无参考美学评估模型，将感知判断分解为四个可人类解释的维度，从而在生成设置中实现音频愉悦度的细致预测。与完全盲目的方法相比，像 WARP-Q [124] 这样的方法保留了基于参考的估计，以确保更强的感知锚定。因此，监督模型在可解释性、可扩展性和感知保真度之间提供了广泛的权衡，使其非常适合在各种条件下进行 MOS 估计。

为了克服对标注数据的依赖，自监督学习（SSL）方法利用在大型未标注语料库上预训练的语音表示。这些由基础模型如 Wav2vec 2.0 [9] 、HuBERT [108] 或 WavLM [28] 派生的嵌入，为下游质量预测任务提供丰富的上下文和语音线索。基于 SSL 的基础模型包括 SSL-MOS [46] 和 MTQ-Net [465] ，它们将冻结或微调的语音编码器与简单回归器结合在一起。更高级的系统如 RAMP [398] ，通过从大量标注数据集中检索相关实例，采用融合网络动态调整检索范围并进行置信度感知加权，从而增强了解码器性能。RAMP+ [399] 进一步结合先验知识和自适应检索机制以提高领域鲁棒性。相比之下，UTMOSv2 [7] 通过单独的预测器融合了基于频谱图和基于 SSL 的特征，并通过微调阶段改进校准和泛化。SQuId [332] 应用知识蒸馏，将多种 SSL 变体整合为统一表示。SCOREQ [306] 引入了对比三元组损失目标，解决了基于 L2 的回归的泛化失败问题，并增强了跨域的预测一致性。IndicMOS [382] 通过为印度语言使用通用编码器，将基于 SSL 的建模扩展到多语言环境。ARECHO [334] ，在 Uni-VERSA [336] 的基础上，提出了动态分类器链和置信度感知解码，以联合估计如 PESQ、STOI 和 MOS 之类的相关语音指标。Uni-VERSA 本身提供了预测自然度、可懂度、说话者相似性和韵律的一体化架构，从而能够在多个感知轴上实现全面评估。为了支持轻量级部署，Distill-MOS [345] 采用模型剪枝和蒸馏策略，显著减少模型尺寸，同时保持对人类评分的高度保真度。HAAQI-Net [423] 将 SSL 应用于助听器音乐评估领域，结合 BLSTM 注意力和基于 BEATs 的特征，展示了在不同声压水平下的鲁棒性。UPPSQA [337] 通过利用一个语义-声学驱动的 MOS 预测模型并在多种配对语音数据类型上进行训练，解决了现有偏好评分基础的 SQA 方法在内容匹配场景中的泛化限制。这些发展总体上展示了基于 SSL 评估的可扩展性、灵活性和标签效率，使其特别适合低资源、多语言和无参考的质量评估场景。

随着基于学习的评估模型的逐渐成熟，最近的研究趋势显示出逐渐向统一的多指标架构和与基础模型的集成方向发展。一方面，传统的 MOS 专用预测器正被扩展为在单一框架内同时估计多个感知指标，如可懂度、自然性和说话者相似性。代表性例子包括 Uni-VERSA [336] 和 ARECHO [334] ，它们展示了分类器链化、联合优化和置信度感知推理如何提供一致的多维预测。这种统一建模策略不仅提高了计算效率，还增强了评估轴上的可解释性和一致性。另一方面，有监督的 MOS 预测的可靠性和可扩展性仍然依赖于大规模、高质量的人类注释数据的可用性。该领域的未来改进可能依赖于更有结构和多样化的众包管道，以及注释协议和感知维度之间更好的对齐。此外，自监督学习继续超越传统的 SSL 编码器进行演进。最近的发展开始探索使用大型语言模型（LLM）或音频语言模型（ALM）进行质量预测，将评估视为基于音频输入的生成或推理任务。虽然基于 LLM 或 ALM 的评估目前处于初期阶段，但其在整体建模上下文、语义和感知方面的潜力表明了一个充满希望的未来方向。展望未来，我们预计下一代基于学习的评估方法将越来越多地整合统一的多指标建模、改进的数据注释协议和基础模型推理，最终引向更稳健、可解释和可迁移的质量评估系统。

5.4 基于 LLM/ALM 的评估

大型语言模型（LLM）和音频语言模型（ALM）的出现为合成语音的自动评估开辟了新的可能性。在这种背景下，如表格 20 所示，目前的方法可以大致分为两类：（1）基于提示的评估，通过精心构建的提示来利用通用 LLM 的零样本能力，以及（2）基于微调的评估，其中具备音频处理能力的 LLM 通过参数高效的微调适应语音评估任务。

5.4.1 . 基于提示的方法 基于提示的方法为语音质量评估提供了一种无需参考和训练的范式。一个显著的例子是 GPT-Whisper [469]，它将 OpenAI 的 Whisper ASR 系统与多模态的 GPT-4o 集成在一起。具体来说，该系统使用 Whisper 转录输入语音，然后提示 GPT-4o 评估转录输出的自然性。与此相反，可懂度的评估是通过直接使用原始音频输入查询 GPT-4o，而不依赖中间转录。尽管设计简单，GPT-Whisper 在与人类 MOS 评分的相关性上表现出中等水平，并且在字符错误率（CER）等 ASR 指标上表现出高度一致性。这些结果表明，基于提示的大语言模型评估为低资源或参考稀缺的场景提供了可行的解决方案。

5.4.2 . 基于微调的方法 微调方法使预训练的音频语言模型适应不同的语音质量评估任务。这种方法特别受到用于感知评估的大规模、高质量标注数据集有限性的推动。由于收集这些标签通常需要昂贵的人力标注或众包 MOS 评分，现有的公共资源仍然稀缺，主要由诸如 BVCC [118] 、NISQA [263] 、SOMOS [251] 等基准所代表。这些元评估基准将在后面的部分中详细讨论。为了在维护适应性的同时缓解数据的限制，代表性的设计利用了通过参数高效技术如 LoRA 在标注的语音质量数据集上微调的听觉 LLM。这些模型由任务特定的提示引导，并支持包括平均意见分（MOS）和说话者相似性（SIM）预测、A/B 偏好测试以及多方面自然语言描述生成在内的广泛评估功能。SALMONN-lora [409] 和 Qwen-Lora [409] 的实验结果表明，这些经过微调的听觉大语言模型可以作为多功能的语音质量评估器，在多个评估场景中相对于最先进的任务特定模型实现具有竞争力的表现。最近，提出了与大语言模型蒸馏 (ALLD) 方法 [25] 对齐，以增强听觉语言模型的感知推理能力。现有的听觉语言模型通常缺乏对输入语音质量的认识，主要是因为语音评估任务通常由于缺乏注释良好的数据集而被排除在多任务训练之外。为了解决这一限制，ALLD 引入了一个自然语言基础的评估语料库，超越了传统的主观质量得分标签，包括多维质量属性、降级分析和描述性 A/B 比较。基于这一资源，ALLD 应用令牌级蒸馏，将听觉语言模型的输出与专家注释对齐，使模型能够执行一系列评估任务。这些任务包括主观质量得分预测、成对质量判断以及基于音频内容生成自然语言的解释。通过将感知监督整合到模型训练中，ALLD 推动了听觉语言模型的发展，使其能够产生与人类对齐的评估和可解释的响应，从而促进了更加感知觉知和可靠的多模态智能体的发展。

最近基于 LLM/ALM 的评估反映了几个关键的发展：

- 从完全监督的训练到参数高效的泛化：早期的方法通常需要大量的标记数据和全模型训练来进行语音和音频质量预测。相比之下，最近的方法追求更高效的范式，例如参数高效的微调（例如，LoRA）和基于提示的零样本推理，从而在最少监督的情况下实现更好的可扩展性和领域转移。
- 从标量预测到可解释的评估：以往基于学习的系统通常侧重于预测单一的质量分数，比如 MOS。相比之下，现代基于 ALM 的方法不仅能够生成 MOS 分数，还能够生成可人类读取的理由和自然语言描述，实现更丰富和更可解释的评估输出。
- 从任务特定到统一的听觉评估：之前的方法经常为不同类型的音频（如语音、非语音声音和音乐）开发单独的评估模型。近期在音频-语言模型方面的进展使得能够在一个架构内对这些听觉领域进行统一评估，从而提升了可扩展性、一致性和可转移性。这一转变利用了先进音频-语言模型的多模态和多任务能力，以支持多样化和整体性的评估场景。

Method	Model	Finetuning	Type	Dimension
GPT-Whisper [469]	Whisper + GPT-4o	No	Prompt-based	Quality, Intelligibility
SALMONN-Lora [409]	SALMONN (vic1.0 & 1.5)	Yes (LoRA)	Fine-tuning-based	MOS, SIM, Descriptions
Qwen-1 & 2-Lora [409]	Qwen1 & 2-Audio	Yes (LoRA)	Fine-tuning-based	MOS, SIM, Descriptions
ALLD [25]	Qwen2-Audio + LLM Distillation	Yes (Token-level)	Fine-tuning-based	MOS, SIM, Explanations

Table 20. **基于 LLM/ALM 的自动合成语音评估方法的分类表。该表总结了最近利用 LLM/ALMs 进行语音质量评估和 MOS 预测的方法。方法根据模型结构、是否应用微调、评估类型（基于提示或基于微调）以及预测感知维度的范围进行分类，包括平均意见得分（MOS）、可理解性、说话者相似性（SIM）、以及自然语言解释或描述。该表突出了由生成和多模态语言模型推动的多维、可解释评估框架的日益增长趋势。**

5.5 基准测试评估

基于基准的评估已成为评估语音生成系统的必要组成部分，它提供了基于精选数据集和标准化任务的结构化协议。这些基准支持在不同评估维度上进行可重复的比较，包括理解、生成质量、安全性和交互性能。基于近期的发展，我们将基准评估分为四大类型，如表格 21 所示：(1) 理解导向的基准，评估模型在语音、音乐和环境声等领域中解释、推理和语义理解音频输入的能力；(2) 生成导向的基准，评估合成音频的质量、可懂度、说话者一致性和韵律控制，通常结合主观和客观指标；(3) 交互导向的基准，检验对话连贯性、响应性、指令遵循和其他对于实时语音交互至关重要的特性；(4) 安全导向的基准，评估音频语言模型在对抗性、误导性或敏感输入条件下的稳健性和伦理对齐性。这包括对越狱易感性、拒绝行为、多语言安全对齐，及日益重要的语音伪造和模型误用检测的评估。

5.5.1 . 理解导向的基准测试 以理解为导向的基准旨在评估模型理解、推理以及生成基于语音、音频或音乐输入响应的能力。此类别包括多个子类型：

基于问答的理解。最近几个基准采用问答格式来评估模型从原始音频输入中直接提取结构化和上下文相关知识的能力。这些任务涵盖不同领域，如音乐、语音、环境声音和动物鸣叫，旨在探讨多模态推理和听觉感知能力。MuChoMusic [417] 通过人类标注者验证的 1,187 个多项选择题，以及 644 首曲目，来针对音乐理解。它评估音乐理论和历史的事实知识，以及风格和文化的解释。然而，即使在噪音输入情况下文本模型取得高分，也表明对语言先验的高度依赖，强调了需要更基于感知的评估。为了应对这一局限性，RUListening [460] 介绍了感知指数，这一指标量化了每个问题对音频感知的依赖。通过生成对抗性干扰项和分析模型的不确定性，它构建了需要真实听觉理解的问答项目，并揭示了音频基础模型和仅语言模型之间显著的性能差距。SALMon [247] 评价了语音中细粒度的声觉意识，关注背景噪音、情感、说话者身份和房间声学等方面。它并不使用直接分类，而是采用基于模型的评分方法，比较正确和错误样本的对数似然值，提供了一种可扩展的方法来测量对声学线索的敏感性。MAE [33] 转向多音频场景，汇集了涉及言语和环境声音的 11 个任务的 20 个数据集，评估模型处理同时或顺序音频流的能力，反映了现实世界的听觉复杂性。所提出的 MALLM 模型表明合成多音频数据可以在不需要广泛手动标注的情况下提高性能。在生物声学领域，BEANS-Zero [315] 提供了一种零样本基准，用于解释动物的发声。涵盖跨不同物种的分类和检测任务，它在训练数据有限的情况下评估泛化能力，并强调特定领域的音频语言对齐。这些基于问答的基准旨在通过制定明确定义的问题回答任务，系统地探测音频语言模型在特定领域的能力。它们不仅仅测试浅显的模式匹配，而是关注模型是否能够从音频输入中提取结构化、与上下文相关的信息，并将其答案建立在感知和语义理解上。

基于推理的理解。随着音频-语言模型（ALM）的不断进步，语音处理研究的重点逐渐从发声者识别、情感分类和语音转录等低层感知任务转向更高层的语义和上下文理解。这种转变导致了面向推理的基准测试的出现，这些基准测试系统地评估模型理解和推理复杂音频输入的能力，评估的不仅是识别准确性，还有多步骤推理、时间顺序和多模态整合的表现。SAKURA [440] 提供了一个结构化任务套件，涵盖发声者性别、语言和情

感状态等属性，强调多跳推理；虽然模型在直接感知任务上表现良好，但它们在跨步整合信息时往往表现不佳。JASCO [413] 针对环境声音与人类语音的联合推理，显示出当前模型往往过度依赖某一种模态，揭示了融合方面的挑战。为了评估逻辑一致性和声音理解，[152] 引入了关于对象存在性、事件顺序和声音归因的诊断任务，利用前后音频对比进行测试。通过指导模型在回答前产生中间的听觉描述，该工作展示了预测准确性和可解释性的提升。MMAR [242] 扩展了推理评估的范围，包含了 1,000 个 QA 项目，分类为信号、感知、语义和文化层次，每个都附有思维链推理以探究多步骤推理。结果表明，即便是先进的模型在推理需要领域知识或抽象理解时也面临困难。基于这些见解，音频蕴涵基准 [56] 评估是否可以从音频内容中逻辑推断出文本描述，在从 AudioCaps 和 Clotho 生成的假设中提供蕴涵、中立和矛盾标签。在零样本和线性探针条件下，目前的 ALM 表现有限。这些基准测试共同推动了 ALM 评估超越表面识别，明确了模型在多模态整合、逻辑推理和因果关系中的能力，并指明了模型开发和任务设计的未来方向。

多任务理解。随着 ALM 向更广泛的任务泛化能力发展，评价模式也从早期以 QA 和推理为基础的基准测试转变为一个更全面的框架，涵盖多任务处理、多模态理解和指令泛化。Dynamic-SUPERB Phase-2 [112] 提出了涵盖语音、音乐和环境声音的 180 个分类、回归和生成任务。结果显示不同任务之间的性能变化相当大，目前还没有模型能表现出稳定的、全方位的能力，这突显了任务泛化和指令理解的挑战。AudioBench [394] 专注于语音理解、音频场景识别和副语言分析。它结合了多样的指令模板和音频文本的开放任务，以在长形式和指令驱动场景中测试模型的鲁棒性，揭示出一般音频推理的持久性限制。UltraEval-Audio [278] 提供了第一个完全开源的基准框架，支持语音理解和生成。它整合了 34 个权威基准，涵盖 10 种语言，具有自动化和标准化的评估工作流程，大大提高了多语言和跨任务环境下的评估效率。FinAudio [20] 针对长期被忽视的金融语音领域，定义了三个核心任务：短格式自动语音识别（ASR）、长格式自动语音识别（ASR）和概述。结果显示，模型在处理长格式的金融音频时仍面临显著挑战，而开源模型在隐私敏感的应用中显示出特别的优势。SAGI 基准 [16] 提出了一个五阶段路线图，用于推进语音理解能力，从基础的自动语音识别（ASR）进展到涉及抽象声学推理和副语言信号解释的复杂任务，从而识别当前一般语音智能的局限性。这些多任务评估基准标志着音频语言模型能力测试从碎片化向更全面和系统化方法的转变。它们促进了跨领域、多模态和语境复杂的评估，同时为开发具有更强泛化能力、指令适应性和现实鲁棒性的下一代 ALMs 提供了统一的平台和关键支持。

5.5.2 . 生成导向的基准测试 生成导向型基准测试旨在系统地评估语音合成系统的感知和统计质量，从低层次的声学特征到高层次的人类感知，涵盖广泛的方面。评估子类型包括：

基于质量的评估。DiscreteEval [406] 提出了一个定量评估框架，检验模型在可理解性、韵律、发音者一致性和自然性方面的表现。基于离散语音标记的语音生成模型在自然性和韵律变化方面优于传统的 TTS，但在可理解性和发音者一致性方面仍然不足，并且容易出现幻听或非语音伪影。实验还表明，扩大模型规模可以在稳健性上带来小幅改进。

多维度评估。EmergentTTS-Eval [252] 引入了一个全面的评估套件，针对六种复杂的合成场景，包括情感表达、副语言线索、外来词、句法复杂性、困难的发音和问题。它包含了由大型语言模型自动生成和扩展的 1,645 个测试样本。该基准采用模型作为裁判的范式，其中音频语言模型在诸如韵律、情感、语调和发音准确性等维度上评估语音输出。模型评估结果显示与人工判断有很强的一致性。应用于开源和专有的 TTS 系统，EmergentTTS-Eval 揭示了在开放式生成任务中细微的性能差异，为未来的模型改进提供了细致的见解。

在当前的语音交互评估框架中，研究人员通常根据自动语言模型的特定交互能力对评估任务进行分类。这种任务分类方法使得能够更精确地描述模型在复杂的人机语音互动场景中的表现。评估任务通常分为以下五个子类别：

基于 QA。这一类别关注模型理解和回答口头问题的能力，评估模型能否准确从音频输入中提取查询要素并生成语义上合适的回答。代表性基准包括 VoxEval [50] 和 Audiopedia [289] 。VoxEval 通过引入一个全口语化的

QA 框架解决了现有 QA 基准在支持端到端语音评估方面的局限性，该框架评估 ALM 在多样的声学条件下的知识理解和复杂推理能力。作为补充，Audiopedia 构建了三个子任务，即单音频 QA、多音频 QA 和检索增强 QA，以全面评估模型在音频理解和外部知识推理方面的能力，弥补传统 AQA 任务在知识密集场景中的不足。

基于口语对话。这一类别强调多轮语音驱动对话中模型的连贯性和互动性，需要对上下文具有强大的理解能力并保持语义一致性。代表性基准包括 Full-Duplex-Bench [204] 、ContextDialog [142] 和 Talking Turns [5] 。Full-Duplex-Bench 通过系统地纳入暂停处理、中断管理和轮流交谈等维度来解决传统半双工模型中粗粒度和有限行为评估的问题，以实时互动评估为目标。在此基础上，ContextDialog 针对模型是否能够在多轮对话中保留和利用会话历史记录进行测试，揭示了开源模型的记忆限制。进一步拓展这一研究方向，Talking Turns 引入了一种评估协议，测试模型该何时发言，是否打断或留下不当的沉默，从而检验模型的轮流预测和对话时机技巧。

基于会话行为。这一类别通过关注模型对副语言提示的敏感度和适应能力，扩展了口语对话任务，这些提示包括情感、语调和说话风格。代表性基准包括 StyleTalk [203] 、SD-Eval [4] 、VoxDialogue [37] 和 Vox-Profile [67] 。StyleTalk 评估模型感知说话风格的能力，并在控制语义下生成风格一致的应答。为了满足对更丰富的副语言理解的需求，SD-Eval 提供了一个综合基准，结合情感、口音、年龄和背景噪声，使用客观和主观评估分析模型如何适应复杂的副语言环境。同时，VoxDialogue 识别十二种非文本的声学属性，测试模型识别和利用节奏、重音和背景声音等提示以生成连贯自然对话的能力。最后，Vox-Profile 提出了一种多维评估方案，结合静态和动态声学特征，以衡量语言模型如何跨不同说话人配置文件泛化，支持个性化语音识别和生成，并在定制互动场景中保持一致性能。

基于指令执行。这一类别评估模型能否正确解释口头指令并生成相应的动作或语音响应。代表性基准包括 S2S-Arena [130] 、EvalSIFT [281] 和 Speech-IFEval [234] 。S2S-Arena 关注现实世界中的语音到语音任务，结合副语言特征来评估任务和领域间的指令理解及风格一致性。在此基础上，EvalSIFT 建立在大型多语言语音文本指令数据集之上，并提供标准化测试集来评估模型在任务和语言间的泛化能力。为了进一步诊断模型的局限性，Speech-IFEval 将语音感知与指令执行分离，提供了一种诊断框架，揭示了基础指令执行中的性能不稳定性和提示敏感性。

这些子类别共同反映了语音交互评估的发展，从单轮响应到多轮对话，从语言学理解到副语言适应，以及从通用任务到个性化交互。这种分类法为系统分析在真实交互环境中 ALMs 的能力界限提供了一个清晰的框架。

5.5.3 . 基于安全性的基准测试 这一类别主要关注在对抗性、欺骗性或潜在有害的音频条件下评估音频语言模型（ALMs）的稳健性、伦理对齐和拒绝行为。代表性的基准通常分为两种子类型：

多维度基于安全性。AudioTrust [189] 提出了第一个全面的 ALM 值得信赖性基准，涵盖公平性、幻觉、安全性、隐私性、鲁棒性和认证这六个核心维度，通过 18 个真实世界的实验设置和 9 个音频特定评估指标，提供了一个可扩展的道德模型部署框架。

越狱和对抗性。这一子类别主要关注于评估音频语言模型（ALMs）在面对旨在绕过安全限制或引发有害响应的恶意音频输入时的弹性。AudioJailbreak [344] 提出了 AJailBench，这是第一个通过对抗性音频提示及其扰动变体系统性评估越狱漏洞的基准，目标是生成语义上一致但违反政策的输出。在此方向上，AdvBench-Audio [135] 引入了 AdvWave，一个双阶段优化框架，利用基于梯度的攻击生成感知上自然的对抗性扰动，大幅提高了对先进音频语言模型攻击成功率。JALMBench [290] 通过提供一个涵盖超过 50,000 个对抗性音频样本的大规模统一基准，进一步扩展了这一领域，支持各种攻击和防御策略，并实现标准化的跨模型比较。补充这些工作，Multi-AudioJail [316] 发现多语言和带口音的音频输入带来额外的安全风险，显示跨语言语音变化和声学扰动可以显著提高越狱成功率——尤其是在多模态系统中。

这些基准测试共同为评估 ALMs 在现实世界和高风险音频理解与交互中的安全边界提供了一个全面的视角。

尽管最近取得了一些进展，未来基于基准的 ALM 评估将越来越强调：

- 统一和多维度评估。现有的基准测试在任务类型上是分散的。未来的工作应推动综合的、基于情景的评估，以共同评估推理、安全性和交互性。
- 开放世界和低资源泛化。目前大多数数据集是经过整理的高资源数据集。稳健的评估需要针对低资源、零样本和长尾音频情境的基准测试。
- 实时个性化和长对话上下文。当前的交互基准在模拟说话者多样性、个性化和长期对话中的记忆方面有所不足。未来的基准应该能够反映对用户资料的动态适应、不断发展的对话状态以及持续的语音驱动交互。
- 超越越狱：社会和伦理风险。安全评估仍然过于狭隘。更广泛的评估应在实际部署条件下包括公平性、偏见、深度伪造滥用和隐私问题。
- 个性化和档案感知的生成评估。未来的基准应该评估模型是否能够将语音生成适应于用户特定的档案，捕捉说话者身份、情感、意图和历史背景，以实现更加自然、一致和以用户为中心的交互。
- 可扩展且透明的评估框架。对于支持自动化、多语言任务以及在基准测试中统一评分的可重复、可扩展评估流程的需求日益增加。

5.6 语音生成的元评价基准

目前，大多数音频模态生成的元评估基准集中于文本到语音（TTS）系统的自动 MOS 预测，主要评估自动评分指标与人类主观评分之间的相关性。在音频和音乐领域，由于依赖于广泛的专家注释进行美学评估，相关的元评估基准数据仍然稀缺。表 22 中显示的现有元评估基准可以分为三种主要类型：（1）面向挑战的，（2）面向数据集的，以及（3）面向工具包的。需要注意的是，基于工具包的元评估方法与直接用于语音和音频质量评估的工具包（如 VERSA [335] 和 Aquatk [387] ）不同。它们的主要作用是作为评估 MOS 预测模型的辅助工具包，为研究人员提供便利的元评估支持。

5.6.1 . 挑战导向基准测试

面向挑战的基准通常组织为社区竞赛，旨在测试 MOS 预测模型在现实、复杂和具有挑战性的声学环境中的泛化能力和鲁棒性。代表性的例子包括 VoiceMOS，这是一个涵盖零样本预测、跨语言鲁棒性和歌声质量预测的多年评估系列。该系列基于大规模的人类主观评级语料库构建了一个动态评估框架。最近，AudioMOS 挑战通过引入创新任务如文本到音乐 MOS 预测、高保真语音质量评估和音频审美对齐，扩展了评估范围，建立了通用音频生成质量评估的新基准。此外，ConferencingSpeech 2022 专注于远程会议场景，评估多说话者复杂通信环境中的语音清晰度和通话质量。

5.6.2 . 面向数据集的基准测试

面向数据集的基准提供了带有人类和专家评级的标准化数据集，为训练和评估语音质量模型提供了坚实的数据基础。它们的优点在于涵盖了语音、唱歌和音乐的多个维度，支持在不同声学环境和任务上的综合分析。SOMOS [251] 是第一个完全由神经 TTS 合成样本组成的大规模 MOS 数据集。它包含大约两万个由两百个不同的神经声学模型生成的语音样本。所有语音样本都使用统一的声码器，以确保差异仅来自声学模型，从而便于训练自动 MOS 预测器和现代语音合成器的质量评估。NISQA [263] 针对现实世界的通信场景，包含具有模拟失真（如数据包丢失、滤波和编码伪影）、真实背景噪音，以及实况电话和 VoIP 通话的多样化语音样本。所有样本都根据 ITU-T P.808 和 P.800 标准进行了多层主观评级，以确保标注的准确性和可靠性。该数据集支持在复杂网络和多设备条件下的语音质量预测模型的评估。SingMOS [364] 通过提供中文和日文的歌唱合成质量数据集来满足跨文化音乐表达的需求。它涵盖了来自各种合成和转换系统的输出，支持多风格分析并填补了音乐

Benchmark	Category	Modality	Evaluation Subtype	Representative Metrics
AIR-Bench [442]	Understanding	Speech/Audio/Music	Reasoning-based	Task Success Rate
MuChoMusic [417]	Understanding	Music	QA-based	QA Accuracy
MMAU [327]	Understanding	Speech/Audio/Music	Multi-task	Task-specific Metrics
AudioBench [394]	Understanding	Speech/Audio	Multi-task	Task-specific Metrics
Dynamic-SUPERB-P2 [112]	Understanding	Speech/Audio/Music	Multi-task	Task-specific Metrics
MMAR [242]	Understanding	Speech/Audio/Music	Reasoning-based	Accuracy
UltraEval-Audio [278]	Understanding	Speech/Audio/Music	Multi-task	Task-specific Metrics
SALMon [247]	Understanding	Speech	QA-based	Accuracy
FinAudio [20]	Understanding	Speech	Multi-task	Task-specific Metrics
Audio Entailment [56]	Understanding	Audio	Reasoning-based	Classification metrics
SAKURA [440]	Understanding	Speech/Audio	Reasoning-based	LLM-vetted Accuracy
JASCO [413]	Understanding	Speech/Audio/Music	Reasoning-based	Relevance Score
SAGI [16]	Understanding	Speech/Audio/Music	Multi-task	Task-specific Metrics
MAE [33]	Understanding	Speech/Audio	QA-based	Task-specific Metrics
BEANS-Zero [315]	Understanding	Audio	QA-based	Task-specific Metrics
RUListening [460]	Understanding	Music	QA-based	Accuracy
SpeechCaps [113]	Understanding	Speech	QA-based	Accuracy
[152]	Understanding	Audio	Reasoning-based	Accuracy
[208]	Understanding	Speech/Audio/Music	Reasoning-based	Micro-averaged Accuracy
DiscreteEval [406]	Generation	Speech	Quality-based	Task-specific Metrics
EmergentTTS-Eval [252]	Generation	Speech	Multi-dimensional-based	LLM Judges
VoxEval [50]	Interaction	Speech	QA-based	QA Accuracy
SD-Eval [4]	Interaction	Speech/Audio	Spoken Dialogue	LLM Judges
VoiceBench [34]	Interaction	Speech/Audio	Conversational Behavior	LLM Judges
Full-Duplex-Bench [204]	Interaction	Speech	Conversational Behavior	LLM Judges, Latency
Vox-Profile [67]	Interaction	Speech	Personalization & Profiling	Accuracy
URO-Bench [439]	Interaction	Speech	Conversational Behavior	MOS, WER, Latency
Audiopedia [289]	Interaction	Audio	QA-based	QA Accuracy
StyleTalk [203]	Interaction	Speech	Conversational Behavior	Task-specific Metrics
VoxDialogue [37]	Interaction	Speech/Audio/Music	Spoken Dialogue	Task-specific Metrics
IFEval-Audio [78]	Interaction	Audio/Speech	Instruction Following	Instruction Following Rate
Talking Turns [5]	Interaction	Speech	Spoken Dialogue	Turn-Taking Metrics
EvalSIFT [281]	Interaction	Speech	Instruction Following	LLM Judges
Speech-IFEval [234]	Interaction	Speech	Instruction Following	LLM Judges
ContextDialog [142]	Interaction	Speech	Spoken Dialogue	LLM Judges
ADU-Bench [75]	Interaction	Audio	Spoken Dialogue	LLM Judges
S2S-Arena [130]	Interaction	Speech	Instruction Following	LLM Judges
AudioTrust [189]	Safety	Speech/Audio	Multi-dimensional Safety	Group Fairness Metrics
AdvBench-Audio [135]	Safety	Audio	Jailbreak & Adversarial	Jailbreak Success Rate
AudioJailbreak [344]	Safety	Audio	Jailbreak & Adversarial	Jailbreak Success Rate
JALMBench [290]	Safety	Audio	Jailbreak & Adversarial	LLM judges
Multi-AudioJail [316]	Safety	Audio	Jailbreak & Adversarial	Jailbreak Success Rate

Table 21. **用于语音、音频或音乐的基于基准的评估的精细分类。”评估子类型”列根据评估目的将基准细分为更细的类型，在理解、生成、交互和安全任务中提供更清晰的区分。**

相关语音评估的空白。MOS-Bench [117] 集成了来自文本转语音、语音转换和语音增强领域的十九个 MOS 数据集，并结合 SHEET 工具包建立了一个统一的评分流程和可解释的评估生态系统。腾讯 [448] 专注于会议语音质量，提供大规模的真实会议音频样本和详细的质量注释，特别关注网络抖动和数据包丢失对语音质量的影响。PSTN [262] 收集了来自传统公共交换电话网络的语音样本，涵盖各种编码格式和传输损伤，作为电话语音质量评估的基准。IUMOS [61] 针对噪声环境下的语音质量评估，包括来自工厂、交通和公共场所的录音，支

持开发强大的语音评估模型。此外，MusicEval [211] 作为第一个生成音乐的元评估基准，通过专家注释和严格的实验设计，为文本到音乐生成系统提供了一个可靠的验证平台。

5.6.3 . 工具包导向基准测试

面向工具包的基准测试在系统实现层面提供了标准化的基础设施，旨在支持自动化评估、结果的可重复性和对语音生成系统的可解释分析。SHEET [117] 提供了综合的评估流水线解决方案，不仅支持标准的 MOS 模型评估，还创新性地整合了评分差距分析、潜在空间可视化和泛化诊断，涵盖了从基本指标到深度分析的评估。

这三类基准测试形成了一个紧密连接的框架。数据集导向的基准测试为挑战导向的评估提供了基本的数据支持，而工具包导向的基准测试则为这两者的实现提供了公平高效的流水线支持。元评估基准测试在未来将变得越来越重要，这不仅仅是因为自动评估指标需要更紧密地与人类偏好和意图对齐，还因为评估自动评估模型的价值在提升语音生成系统的质量和发展方面起着关键作用。未来的趋势将不可避免地转向与人类判断紧密对齐的、可解释的、多维度的元评估框架。

Benchmark	Type	Focus	Data
VoiceMOS [47, 118, 120]	Challenge	MOS Prediction	Speech/Singing
SingMOS [364]	Challenge	MOS Prediction	Singing
AudioMOS [389]	Challenge	MOS & Aesthetics Prediction	Audio
ConferencingSpeech [448]	Challenge	MOS Prediction	Speech
QualiSpeech [408]	Corpus	Description & Explanation	Speech
SongEval [444]	Corpus	Aesthetics Prediction	Singing
SOMOS [251]	Corpus	MOS Prediction	Speech
NISQA [263]	Corpus	Speech Quality	Speech
MOS-Bench [117]	Corpus	MOS Prediction	Speech
Tencent [448]	Corpus	Speech Quality	Speech
PSTN [261]	Corpus	Speech Quality	Speech
IUMOS [61]	Corpus	Speech Quality	Speech
MusicEval [211]	Corpus	Music Quality	Music
SHEET [117]	Toolkit	MOS Prediction	Speech

Table 22. **自动语音评估的代表性元评估基准总结。这些基准根据其类型分类，分为挑战导向的社区竞赛或数据集导向的精选语料库；它们的重点，如平均意见得分 MOS 预测、感知质量建模或自然语言解释；以及它们支持的数据模式，包括语音、歌唱声音和通用音频。这些基准提供标准化资源，用于评估自动语音质量评估模型在不同领域、语言和声学条件下的性能和泛化能力。**

5.7 比较用于语音生成的自动 MOS 评估

自动平均意见评分（MOS）预测在评估合成语音的感知质量方面起着至关重要的作用，作为现代文本到语音（TTS）系统中传统人工评估的可扩展和成本效益高的替代方案。这些方法在标注的数据集上训练模型，以模拟主观的人类评分，从而实现大规模和一致的质量评估。我们比较了反映该领域演变的两种核心方法：基于学习的方法和基于 ALM 的方法。基于学习的方法通过全监督或自监督技术来学习 MOS 预测，而基于 ALM 的方法则利用在特定任务的 MOS 数据上微调的大规模预训练多模态模型。为了系统地评估这些模型，我们在多个基准数据集上进行实验，这些数据集涵盖了通用和特定领域的场景。为确保一致性和可比性，我们采用了统一的语句级评估协议。通用基准包括 NISQA [263] 、VoiceMOS [118] 和 SOMOS [251] 的测试集，与 MOS-Bench [117] 的设置一致。对于领域泛化评估，我们选择了 SingMOS [364] 的整个测试集和腾讯 [448] 的开发集。我们采用三种标准指标来衡量模型性能：线性相关系数（LCC）和 Spearman 等级相关系数（SRCC）用于评估预测结果与人工评分之间的线性和单调相关性，均方误差（MSE）则用于量化绝对预测误差。较高的 LCC 和 SRCC 值以及

较低的 MSE 表明模型预测和人类感知之间的更强一致性。评估的模型包括基于监督学习的方法，如 MOSNet [229] 和 LDNet [119]；自监督模型，如 UTMOSv2 [7] 、SCOREQ [306] 、RAMP [398] 和修改的 SSL-MOS [117]；以及基于 ALM 的方法，如 SALMONN-Lora 和 Qwen2-Audio-Lora [409]，它们利用了任务特定 Lora 微调的预训练音频-语言骨干。

我们在表 23 和表 24 中的比较分析评估了基于学习和基于 ALM 的方法在多个基准中对自动平均意见得分 (MOS) 预测的表现。在通用数据集如 NISQA、VoiceMOS-BVCC 和 SOMOS 上，基于 ALM 的模型，特别是 SALMONN-Lora，始终优于基于学习的方法。例如，SALMONN-Lora 在 NISQA 上的 LCC 为 0.861，SRCC 为 0.859，MSE 为 0.347，表明其与人类感知高度一致，预测误差低。同样地，在 SOMOS 上，其 LCC 达到 0.644，MSE 为 0.196，超过了许多基准模型。相反，基于学习的模型如 UTMOSv2 在净化或域内数据如 VoiceMOS-BVCC 测试中表现出竞争力，其 LCC 为 0.945，SRCC 为 0.949。然而，在领域转变的情况下，它们的性能显著下降。在特定领域的基准如 SingMOS（歌声合成）和 TencentDev（真实世界会议语音）中，许多基于学习的模型与人类评分表现出较弱甚至负的相关。例如，在 SingMOS 上，SALMONN-Lora 的相关性下降到 LCC 为 0.372，SRCC 为 0.347，MSE 增加到 1.928，而基于学习的 RAMP 则取得了更好的 LCC 为 0.505。同时，SALMONN-Lora 在 Tencent-Dev 上表现出卓越的鲁棒性，达到 LCC 为 0.758 和 MSE 为 0.394，大幅优于基于学习的替代方案。

这些发现表明，基于 ALM 的方法受益于大规模的预训练和上下文表示学习，这些方法增强了在不同声学和语义条件下的泛化能力和鲁棒性。同时，当经过精心调整时，基于学习的模型在专门领域可以表现优异。总体而言，将多模态预训练与领域感知微调相结合对于开发可靠且具有高度可推广性的 MOS 预测系统至关重要，这使得 ALM 成为未来感知语音评估的有希望的基础。基于 ALM 的方法仍处于早期阶段，但具有显著的进一步发展潜力。未来的趋势可能包括扩展到更多样化和复杂的元评估数据集，同时保持稳定性，通过如从 ALLD [25] 知识蒸馏、改进的提示工程、微调策略、以及结合可解释性驱动的训练数据等技术，专门优化 ALM 用于语音评估。这些进展可能进一步增强基于 ALM 的 MOS 预测模型的准确性、鲁棒性和可解释性。

Category	Model	$\text{NISQA}_{\text{AVG}}$			$\text{VoiceMOS-BVCC}_{\text{Test}}$			$\text{SOMOS}_{\text{Test}}$		
		LCC ↑	SRCC ↑	MSE ↓	LCC ↑	SRCC ↑	MSE ↓	LCC ↑	SRCC ↑	MSE ↓
Learning-based	MOSNET [229]	0.413	0.394	0.975	0.501	0.530	0.993	-	-	-
Learning-based	LDNET [119]	0.290	0.345	0.936	0.606	0.596	0.880	-	-	-
Learning-based	UTMOSv2 [7]	0.629	0.614	0.696	0.945	0.949	0.439	0.438	0.413	0.306
Learning-based	SCOREQ [306]	0.711	0.694	0.732	0.893	0.892	0.240	0.449	0.436	0.849
Learning-based	RAMP [398]	-	-	-	0.904	0.903	0.177	0.395	0.387	0.958
Learning-based	Modified SSL-MOS [117]	0.613	0.614	0.860	0.895	0.891	0.253	0.492	0.482	0.662
ALM-based	SALMONN(vic1.5)-Lora [409]	0.861	0.859	0.347	0.826	0.833	0.282	0.644	0.636	0.196
ALM-based	Qwen2-Audio-Lora [409]	0.768	0.780	0.643	0.681	0.678	0.493	0.583	0.572	0.216

Table 23. **对比自动 MOS 预测方法在 NISQA、VoiceMOS-BVCC 和 SOMOS 上的表现。这些基准测试侧重于语音质量评估。**

5.8 挑战和未来趋势

尽管取得了显著进展，当前用于语音合成的自动评估方法仍然面临诸如泛化能力有限、可解释性不足，以及在多样化条件下的鲁棒性较弱等挑战。此外，高质量和多样化的标注数据的稀缺，加上从听众和专家获取标注的高成本，尤其限制了在诸如广义音频、情感语音、带口音的语音和语音障碍等新兴场景中构建高质量元评估基准。这些限制进一步制约了自动评估模型的适用性和泛化性。此外，对提示设计的敏感性和对预测置信度的不充分校准也阻碍了其实际效用和稳定性。

随着大型音频-语言模型（ALMs）专用于语音理解的出现，未来自动评估的发展将集中于提高通用性、可解释性和多维覆盖。基于 ALM 的方法不仅能够生成标量质量评分，还能够以自然语言生成诊断反馈，显著增强

Category	Model	$\text{SingMOS}_{\text{Test}}$			$\text{Tencent}_{\text{Dev}}$		
		LCC ↑	SRCC ↑	MSE ↓	LCC ↑	SRCC ↑	MSE ↓
Learning-based	MOSNET [229]	0.327	0.158	1.379	0.375	0.417	0.972
Learning-based	LDNET [119]	0.424	0.200	1.158	-0.285	-0.288	1.166
Learning-based	UTMOSv2 [7]	0.506	0.452	4.449	0.435	0.395	1.575
Learning-based	SCOREQ [306]	0.493	0.480	3.348	0.442	0.311	2.111
Learning-based	RAMP [398]	0.505	0.480	3.489	0.325	0.240	1.669
Learning-based	Modified SSL-MOS [117]	0.429	0.415	5.740	0.338	0.297	2.553
ALM-based	SALMONN(vic1.5)-Lora [409]	0.372	0.347	1.928	0.758	0.767	0.394
ALM-based	Qwen2-Audio-Lora [409]	-	-	-	-	-	-

Table 24. **模型在 SingMOS 和腾讯数据集上的泛化评估，这些数据集反映了更为多样化和开放域的语音场景。**

了评估结果的透明度和可解释性。ALMs 在跨语言和跨领域泛化方面显示出强大的潜力，为建立稳健且通用的语音质量评估系统奠定了坚实的基础。

本调查讨论了与语音和音频生成评估方法相关的每个部分的未来趋势。在此，我们特别强调 ALM（多模态语言模型）的多模态感知能力为跨模态一致性评估开辟了新途径，并促进了音频评估与其他模态之间的整合。例如，在说话人脸合成中，ALM 可以评估语音与面部表情之间的情感一致性；在音乐生成中，它们可以评估歌词、旋律和视觉节奏之间的连贯性；在沉浸式 3D 音频场景中，评估必须涵盖声源定位和空间听觉一致性。这些多模态挑战突显了开发具备多模态理解和推理能力的综合评估框架的必要性。因此，针对目前有限的元评估数据集进行高质量微调和学习以优化 ALM 成为一个关键趋势。

展望未来，音频生成的自动评估有望超越传统的基于失真的指标，并发展成为统一的、多维的、可推广的框架，这些框架融合了语义理解、情感表达、跨模态对齐和可解释的推理。这样的评估系统不仅将推动生成模型技术的创新，还将为智能多模态交互和内容创作提供强有力的支持。

6 结论与未来工作

本文系统地回顾了用于生成模型的自动评估方法，涵盖文本、视觉和音频模态，并介绍了一种统一的分类法，将现有方法分为五类：基于启发、嵌入、学习和基于 LLM/VLM/ALM 的评估。通过详细的比较分析，我们阐明了每种评估范式的优点和局限性，为推进评估方法提供了基础。未来的工作将扩展此框架到其他模态和任务，同时解决如评估偏差、跨领域泛化和在日益复杂的生成系统中实现可扩展性等关键挑战。

References

[1] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. arXiv preprint arXiv:2406.02430 (2024).

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 . Springer, 382–398.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In Computer Vision – ECCV 2016 , Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 382–398.

[4] Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. arXiv preprint arXiv:2406.13340 (2024).

[5] Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025. Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics. arXiv preprint arXiv:2503.01174 (2025).

[6] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. arXiv:2108.07732 [cs.PL]

[7] Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The T05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In 2024 IEEE Spoken Language Technology Workshop (SLT) . IEEE, 818–824.

[8] Sher Badshah and Hassan Sajjad. 2024. Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text. arXiv:2408.09235 [cs.CL] https://arxiv.org/abs/2408.09235

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33 (2020), 12449–12460.

[10] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. arXiv preprint arXiv:2402.14762 (2024).

[11] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. arXiv preprint arXiv:2308.14508 (2023).

[12] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In Proceedings of the IEEE/CVF International Conference on Computer Vision . 20041–20053.

[13] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization , Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[14] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) .

[15] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018).

[16] Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. 2024. Roadmap towards superhuman speech understanding using large language models. arXiv preprint arXiv:2410.13268 (2024).

[17] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 291–305. https://doi.org/10.18653/v1/2022.emnlp-main.20

[18] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 Technical Report. arXiv:2403.17297 [cs.CL] https://arxiv.org/abs/2403.17297

[19] Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. CompassJudger-1: All-in-one Judge Model Helps Model Evaluation and Evolution. arXiv:2410.16256 [cs.CL] https://arxiv.org/abs/2410.16256

[20] Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao-yang Liu, Koduvayur Subbalakshmi, et al. 2025. FinAudio: A Benchmark for Audio Large Language Models in Financial Applications. arXiv preprint arXiv:2503.20990 (2025).

[21] 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. arXiv preprint arXiv:2308.07201 (2023).

[22] 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. arXiv:2308.07201 [cs.CL] https://arxiv.org/abs/2308.07201

[23] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A Dataset for Training and Evaluating Generative Reading Comprehension Metrics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6521–6532. https://doi.org/10.18653/v1/2020.emnlp-main.528

[24] Boxing Chen and Hongyu Guo. 2015. Representation Based Translation Evaluation Metrics. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) , Chengqing Zong and Michael Strube (Eds.). Association for Computational Linguistics, Beijing, China, 150–155. https://doi.org/10.3115/v1/P15-2025

[25] Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025. Audio Large Language Models Can Be Descriptive Speech Quality Evaluators. arXiv preprint arXiv:2501.17202 (2025).

[26] Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva,

and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1739–1753. https://doi.org/10.18653/v1/2022.emnlp-main.114

[27] 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. (2021). arXiv:2107.03374 [cs.LG]

[28] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing 16, 6 (2022), 1505–1518.

[29] Wang Chen, Piji Li, and Irwin King. 2021. A Training-free and Reference-free Summarization Evaluation Metric via Centrality-weighted Relevance and Self-referenced Redundancy. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 404–414. https://doi.org/10.18653/v1/2021.acl-long.34

[30] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In The 2023 Conference on Empirical Methods in Natural Language Processing .

[31] Xiaoqiao Chen, Qingyi Zhang, Manhui Lin, Guangyi Yang, and Chu He. 2019. No-reference color image quality assessment: From entropy to perceptual quality. EURASIP Journal on Image and Video Processing 2019 (2019), 1–14.

[32] Yixiong Chen, Li Liu, and Chris Ding. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. arXiv preprint arXiv:2305.10843 (2023).

[33] Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D ' Haro, Robby Tan, and Haizhou Li. 2024. Beyond Single-Audio: Advancing Multi-Audio Processing in Audio Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2024 . 10917–10930.

[34] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. arXiv preprint arXiv:2410.17196 (2024).

[35] Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, et al. 2023. T-Eval: Evaluating the Tool Utilization Capability Step by Step. arXiv preprint arXiv:2312.14033 (2023).

[36] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. 2024. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? arXiv preprint arXiv:2407.04842 (2024).

[37] Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, et al. 2025. VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words?. In The Thirteenth International Conference on Learning Representations .

[38] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. arXiv preprint arXiv:2310.18235 (2023).

[39] Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2748–2760. https://doi.org/10.18653/v1/P19-1264

[40] 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:1803.05457 [cs.AI]

[41] 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. arXiv preprint arXiv:2110.14168 (2021).

[42] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing . 12621–12640.

[43] Pierre Colombo, Chloe Clave, and Pablo Piantanida. 2021. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. In AAAI Conference on Artificial Intelligence . https://api.semanticscholar.org/CorpusID:244896426

[44] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10450–10466. https://doi.org/10.18653/v1/2021.emnlp-main.817

[45] OpenCompass Contributors. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

[46] 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) . IEEE, 8442–8446.

[47] 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) . IEEE, 1–7.

[48] Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12053–12059. https://doi.org/10.18653/v1/2023.emnlp-main.739

[49] 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. arXiv:2310.01377 [cs.CL]

[50] Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. VoxEval: Benchmarking the Knowledge Understanding Capabilities of End-to-End Spoken Language Models. arXiv preprint arXiv:2501.04962 (2025).

[51] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to Evaluate Image Captioning. arXiv:1806.06422 [cs.CV] https://arxiv.org/abs/1806.06422

[52] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan KA Reddy, Christian Schüldt, and Saikat Chatterjee. 2025. Impairments are Clustered in Latents of Deep Neural Network-based Speech Quality Models. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[53] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[54] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[55] 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 (2024).

[56] Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2025. Audio Entailment: Assessing deductive reasoning for audio understanding. In Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 39. 23769–23777.

[57] Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. 2025. ADIFF: Explaining audio difference using natural language. arXiv preprint arXiv:2502.04476 (2025).

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[59] 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 Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4884–4895. https://doi.org/10.18653/v1/P19-1483

[60] 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 (San Diego, California) (HLT '02) . Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 138–145.

[61] Xuan Dong and Donald S Williamson. 2020. A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals. (2020).

[62] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

[63] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations , Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 150–158. https://aclanthology.org/2024.eacl-demo.16/

[64] Alexander Richard 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 (2021), 391–409.

[65] Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies . 2587–2601.

[66] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. arXiv:2309.16289 [cs.CL] https://arxiv.org/abs/2309.16289

[67] Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, et al. 2025. Vox-Profile: A Speech Foundation Model Benchmark for Characterizing Diverse Speaker and Speech Traits. arXiv preprint arXiv:2505.14648 (2025).

[68] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. 2024. TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation. arXiv preprint arXiv:2406.08656 (2024).

[69] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent. 2005. BSS_EVAL Toolbox User Guide–Revision 2.0. (2005).

[70] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. 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) , Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 46–68. https://aclanthology.org/2022.wmt-1.2/

[71] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. arXiv:2302.04166 [cs.CL] https://arxiv.org/abs/2302.04166

[72] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. Advances in Neural Information Processing Systems 36 (2023), 50742–50768.

[73] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. 2018. Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. arXiv preprint arXiv:1808.05344 (2018).

[74] Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Junyang Lin, Chang Zhou, Wen Xiao, et al. 2024. LLM critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. CoRR (2024).

[75] Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2024. Benchmarking Open-ended Audio Dialogue Understanding for Large Audio-Language Models. arXiv preprint arXiv:2412.05167 (2024).

[76] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 386–395. https://doi.org/10.18653/v1/2020.emnlp-main.28

[77] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue Response RankingTraining with Large-Scale Human Feedback Data. In EMNLP .

[78] Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. 2025. IFEval-Audio: Benchmarking Instruction-Following Capability in Audio-based Large Language Models. arXiv preprint arXiv:2505.16774 (2025).

[79] Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1347–1354. https://doi.org/10.18653/v1/2020.acl-main.124

[80] Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing . 2053–2070.

[81] Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation , Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 82–89. https://doi.org/10.18653/v1/W19-2310

[82] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition . 15180–15190.

[83] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022).

[84] Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. 2024. TIAM-A metric for evaluating alignment in Text-to-Image generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision . 2890–2899.

[85] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) , Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 708–719. https://doi.org/10.18653/v1/N18-1065

[86] Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH-v2: Scaling Analytical Hallucination Annotation of Large Language Models. arXiv preprint arXiv:2407.04693 (2024).

[87] Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 9157–9166. https://doi.org/10.18653/v1/2020.emnlp-main.736

[88] 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 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) , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6394–6407. https://doi.org/10.18653/v1/2021.acl-long.500

[89] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1331–1335.

[90] Yinuo Guo, Chong Ruan, and Junfeng Hu. 2018. Meteor++: Incorporating Copy Knowledge into Machine Translation Evaluation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers , 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 (Eds.). Association for Computational Linguistics, Belgium, Brussels, 740–745. https://doi.org/10.18653/v1/W18-6454

[91] Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing , Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 1066–1072. https://doi.org/10.18653/v1/D15-1124

[92] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to Differentiate Better from Worse Translations. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 214–220. https://doi.org/10.3115/v1/D14-1027

[93] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise Neural Machine Translation Evaluation. In 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) , Chengqing Zong and Michael Strube (Eds.). Association for Computational Linguistics, Beijing, China, 805–814. https://doi.org/10.3115/v1/P15-1078

[94] Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and Improving Semantic Diversity of Dialogue Generation. In Findings of the Association for Computational Linguistics: EMNLP 2022 , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 934–950. https://doi.org/10.18653/v1/2022.findings-emnlp.66

[95] Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based Reference-less Evaluation of Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3381–3392. https://doi.org/10.18653/v1/P19-1330

[96] Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. 2023. ALLURE: Auditing and Improving LLM-based Evaluation of Text using Iterative In-Context-Learning. arXiv:2309.13701 [cs.CL] https://arxiv.org/abs/2309.13701

[97] Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe. 2021. Espnet2-tts: Extending the edge of tts research. arXiv preprint arXiv:2110.07840 (2021).

[98] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. 2023. T3 Bench: Benchmarking Current Progress in Text-to-3D Generation. arXiv preprint arXiv:2310.02977 (2023).

[99] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. Proceedings of the International Conference on Learning Representations (ICLR) (2021).

[100] 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 (2021).

[101] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In 2017 ieee international conference on acoustics, speech and signal processing (icassp) . IEEE, 131–135.

[102] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021).

[103] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017).

[104] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2020. Semantic object accuracy for generative text-to-image synthesis. IEEE transactions on pattern analysis and machine intelligence 44, 3 (2020), 1552–1565.

[105] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG] https://arxiv.org/abs/2006.11239

[106] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. arXiv:1904.09751 [cs.CL] https://arxiv.org/abs/1904.09751

[107] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2 : Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7856–7870. https://doi.org/10.18653/v1/2021.emnlp-main.619

[108] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing 29 (2021), 3451–3460.

[109] Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15924–15951. https://doi.org/10.18653/v1/2024.emnlp-main.891

[110] Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability. arXiv:2406.18365 [cs.CL] https://arxiv.org/abs/2406.18365

[111] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision . 20406–20417.

[112] Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. 2024. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. arXiv preprint arXiv:2411.05361 (2024).

[113] Chien-yu Huang, Min-Han Shih, Ke-Han Lu, Chi-Yuan Hsiao, and Hung-yi Lee. 2025. SpeechCaps: Advancing Instruction-Based Universal Speech Models with Multi-Talker Speaking Style Captioning. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[114] Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2023. Flames: Benchmarking Value Alignment of Chinese Large Language Models. arXiv:2311.06899 [cs.CL]

[115] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems 36 (2023), 78723–78747.

[116] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 9230–9240. https://doi.org/10.18653/v1/2020.emnlp-main.742

[117] Wen-Chin Huang, Erica Cooper, and Tomoki Toda. 2024. Mos-bench: Benchmarking generalization abilities of subjective speech quality assessment models. arXiv preprint arXiv:2411.03715 (2024).

[118] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The voicemos challenge 2022. arXiv preprint arXiv:2203.11389 (2022).

[119] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. 2022. 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) . IEEE, 896–900.

[120] Wen-Chin Huang, Szu-Wei Fu, Erica Cooper, Ryandhimas E Zezario, Tomoki Toda, Hsin-Min Wang, Junichi Yamagishi, and Yu Tsao. 2024. The VoiceMOS challenge 2024: Beyond speech quality prediction. In 2024 IEEE Spoken Language Technology Workshop (SLT) . IEEE, 803–810.

[121] 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. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. arXiv preprint arXiv:2305.08322 (2023).

[122] Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3009–3015. https://doi.org/10.18653/v1/2021.emnlp-main.239

[123] Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-Dimensional Evaluation of Text Summarization with In-Context Learning. In Findings of the Association for Computational Linguistics: ACL 2023 , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8487–8495. https://doi.org/10.18653/v1/2023.findings-acl.537

[124] Wissam A Jassim, Jan Skoglund, Michael Chinen, and Andrew Hines. 2021. WARP-Q: Quality prediction for generative neural speech codecs. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 401–405.

[125] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. 2024. Rethinking fid: Towards a better evaluation metric for image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . 9307–9315.

[126] Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. Journal of the Acoustical Society of America 62 (1977). https://api.semanticscholar.org/CorpusID:121680873

[127] Jesper Jensen and Cees H Taal. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 11 (2016), 2009–2022.

[128] Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical Annotation of Hallucinations in Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8135–8158. https://doi.org/10.18653/v1/2024.acl-long.442

[129] Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. arXiv:2310.00752 [cs.CL] https://arxiv.org/abs/2310.00752

[130] Feng Jiang, Zhiyu Lin, Fan Bu, Yuhao Du, Benyou Wang, and Haizhou Li. 2025. S2S-Arena, Evaluating Speech2Speech Protocols on Instruction Following with Paralinguistic Information. arXiv preprint arXiv:2503.05085 (2025).

[131] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. arXiv preprint arXiv:1909.02050 (2019).

[132] Christian Johnson. 2022. Binary Encoded Word Mover's Distance. In Proceedings of the 7th Workshop on Representation Learning for NLP , Spandana Gella, He He, Bodhisattwa Prasad Majumder, Burcu Can, Eleonora Giunchiglia, Samuel Cahyawijaya, Sewon Min, Maximilian Mozes, Xiang Lorraine Li, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Laura Rimell, and Chris Dyer (Eds.). Association for Computational Linguistics, Dublin, Ireland, 167–172. https://doi.org/10.18653/v1/2022.repl4nlp-1.17

[133] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics . Association for Computational Linguistics, Vancouver, Canada.

[134] Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning Cheaper, Lighter and Faster Evaluation Metrics for Automatic Text Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1305–1318. https://doi.org/10.18653/v1/2022.acl-long.93

[135] Mintong Kang, Chejian Xu, and Bo Li. 2024. AdvWave: Stealthy Adversarial Jailbreak Attack against Large Audio-Language Models. arXiv preprint arXiv:2412.08608 (2024).

[136] James M Kates and Kathryn H Arehart. 2010. The hearing-aid speech quality index (HASQI). Journal of the Audio Engineering Society 58, 5 (2010), 363–381.

[137] James M Kates and Kathryn H Arehart. 2014. The hearing-aid speech perception index (HASPI). Speech Communication 65 (2014), 75–93.

[138] Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In Proceedings of the 62nd

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) . 13034–13054.

[139] Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CritiqueLLM: Towards an Informative Critique Generation Model for Evaluation of Large Language Model Generation. arXiv:2311.18702 [cs.CL] https://arxiv.org/abs/2311.18702

[140] 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 Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2306–2319. https://doi.org/10.18653/v1/2022.acl-long.164

[141] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr \ 'echet audio distance: A metric for evaluating music enhancement algorithms. arXiv preprint arXiv:1812.08466 (2018).

[142] Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. 2025. Does Your Voice Assistant Remember? Analyzing Conversational Context Recall and Utilization in Voice Interaction Models. arXiv preprint arXiv:2502.19759 (2025).

[143] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations .

[144] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. arXiv:2310.08491 [cs.CL] https://arxiv.org/abs/2310.08491

[145] 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. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4334–4353. https://doi.org/10.18653/v1/2024.emnlp-main.248

[146] 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:2405.01535 [cs.CL] https://arxiv.org/abs/2405.01535

[147] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems 36 (2023), 36652–36663.

[148] Neema Kotonya, Saran Krishnasamy, Joel Tetreault, and Alejandro Jaimes. 2023. Little Giants: Exploring the Potential of Small LLMs as Evaluation Metrics in Summarization in the Eval4NLP 2023 Shared Task. In Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems , Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Bali, Indonesia, 202–218. https://doi.org/10.18653/v1/2023.eval4nlp-1.17

[149] 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 (2023).

[150] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2023. Imagenhub: Standardizing the evaluation of conditional image generation models. arXiv preprint arXiv:2310.01596 (2023).

[151] Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. 2024. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. arXiv preprint arXiv:2406.08402 (2024).

[152] Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? Tackling hallucinations with multi-task assessment and stepwise audio reasoning. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[153] Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE pacific rim conference on communications computers and signal processing , Vol. 1. IEEE, 125–128.

[154] Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[155] Marie Kunešová, Jindřich Matoušek, Jan Lehečka, Jan Švec, Josef Michálek, Daniel Tihelka, Martin Bulín, Zdeněk Hanzlíček, and Markéta Řezáčková. 2023. Ensemble of deep neural network models for MOS prediction. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[156] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. arXiv:2406.10149

[157] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37) , Francis Bach and David Blei (Eds.). PMLR, Lille, France, 957–966. https://proceedings.mlr.press/v37/kusnerb15.html

[158] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics 7 (2019), 453–466.

[159] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. Advances in neural information processing systems 32 (2019).

[160] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. Transactions of the Association for Computational Linguistics 10 (2022), 163–177.

[161] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. Transactions of the Association for Computational Linguistics 10 (2022), 163–177. https://doi.org/10.1162/tacl_a_00453

[162] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683 (2017).

[163] Sanyam Lakhanpal, Shivang Chopra, Vinija Jain, Aman Chadha, and Man Luo. 2024. Refining Text-to-Image Generation: Towards Accurate Training-Free Glyph-Enhanced Image Generation. arXiv preprint arXiv:2403.16422 (2024).

[164] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. RewardBench: Evaluating Reward Models for Language Modeling. arXiv:2403.13787 [cs.LG] https://arxiv.org/abs/2403.13787

[165] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. arXiv:1901.07291 [cs.CL]

[166] Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Heyan Huang, and Xian-Ling Mao. 2024. Exploring Dense Retrieval for Dialogue Response Selection. ACM Trans. Inf. Syst. 42, 3, Article 84 (Jan. 2024), 29 pages. https://doi.org/10.1145/3632750

[167] 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. arXiv:2004.02399 [cs.CL] https://arxiv.org/abs/2004.02399

[168] Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024. Training Language Models to Critique With Multi-agent Feedback. arXiv preprint arXiv:2410.15287 (2024).

[169] Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024. Training Language Models to Critique With Multi-agent Feedback. arXiv:2410.15287 [cs.CL] https://arxiv.org/abs/2410.15287

[170] Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian ling Mao. 2024. CriticEval: Evaluating Large Language Model as Critic. arXiv:2402.13764 [cs.CL] https://arxiv.org/abs/2402.13764

[171] Teven Le Scao and Claire Gardent. 2023. Joint Representations of Text and Knowledge Graphs for Retrieval and Evaluation. In Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings) , Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 110–122. https://doi.org/10.18653/v1/2023.findings-ijcnlp.10

[172] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267 [cs.CL] https://arxiv.org/abs/2309.00267

[173] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. arXiv preprint arXiv:2503.07891 (2025).

[174] Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024. Aligning Large Language Models by On-Policy Self-Judgment. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11442–11459. https://doi.org/10.18653/v1/2024.acl-long.617

[175] Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024. Aligning Large Language Models by On-Policy Self-Judgment. arXiv:2402.11253 [cs.LG] https://arxiv.org/abs/2402.11253

[176] Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2024. CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model. arXiv:2311.05812 [cs.CL] https://arxiv.org/abs/2311.05812

[177] 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) . IEEE, 391–395.

[178] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[179] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. GenAI-Bench: A Holistic Benchmark for Compositional Text-to-Visual Generation. In Synthetic Data for Computer Vision Workshop@ CVPR 2024 .

[180] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594 [cs.AI] https://arxiv.org/abs/2411.16594

[181] Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023. Repetition In Repetition Out: Towards Understanding Neural Text Degeneration from the Data Perspective. In Thirty-seventh Conference on Neural Information Processing Systems . https://openreview.net/forum?id=WjgCRrOgip

[182] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: Measuring massive multitask language understanding in Chinese. arXiv:2306.09212 [cs.CL]

[183] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing . 6449–6464.

[184] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 110–119. https://doi.org/10.18653/v1/N16-1014

[185] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV] https://arxiv.org/abs/2301.12597

[186] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, et al. 2024. Generative Judge for Evaluating Alignment. In The Twelfth International Conference on Learning Representations .

[187] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative Judge for Evaluating Alignment. arXiv:2310.05470 [cs.CL] https://arxiv.org/abs/2310.05470

[188] Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting Human and LLM Preferences. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1790–1811. https://doi.org/10.18653/v1/2024.acl-long.99

[189] Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Zhe Wang, Xingjian Du, Shun Zhang, Hanjun Luo, et al. 2025. AudioTrust: Benchmarking the Multifaceted Trustworthiness of Audio Large Language Models. arXiv preprint arXiv:2505.16211 (2025).

[190] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. In Findings of the Association for Computational Linguistics ACL 2024 , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 3923–3954. https://aclanthology.org/2024.findings-acl.235

[191] Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. 2023. TACO: Topics in Algorithmic COde generation dataset. arXiv preprint arXiv:2312.14852 (2023).

[192] Ruosen Li, Teerth Patel, and Xinya Du. 2024. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations. arXiv:2307.02762 [cs.CL] https://arxiv.org/abs/2307.02762

[193] Zitong Li and Wei Li. 2023. MOSLight: A Lightweight Data-Efficient System for Non-Intrusive Speech Quality Assessment. In Proc. INTERSPEECH , Vol. 2023. 5386–5390.

[194] Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging Large Language Models for NLG Evaluation: Advances and Challenges. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16028–16045. https://doi.org/10.18653/v1/2024.emnlp-main.896

[195] 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 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) , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 128–138. https://doi.org/10.18653/v1/2021.acl-long.11

[196] Qiao Liang, Ying Shen, Tiantian Chen, Lin Zhang, and Shengjie Zhao. 2025. ADTMOS–Synthesized Speech Quality Assessment Based on Audio Distortion Tokens. IEEE Transactions on Audio, Speech and Language Processing (2025).

[197] Xinyu Liang, Fredrik Cumlin, Christian Schüldt, and Saikat Chatterjee. 2023. DeePMOS: deep posterior mean-opinion-score of speech. In Proceedings of INTERSPEECH . 526–530.

[198] Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. 2024. Internal Consistency and Self-Feedback in Large Language Models: A Survey. arXiv:2407.14507 [cs.CL] https://arxiv.org/abs/2407.14507

[199] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. arXiv preprint arXiv:2305.20050 (2023).

[200] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. arXiv:2305.20050 [cs.LG] https://arxiv.org/abs/2305.20050

[201] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2020 , Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1823–1840. https://doi.org/10.18653/v1/2020.findings-emnlp.165

[202] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out . Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[203] Guan-Ting Lin, Cheng-Han Chiang, and Hung-Yi Lee. 2024. Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) . 6626–6642.

[204] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025. Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities. arXiv preprint arXiv:2503.04721 (2025).

[205] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) . 3214–3252.

[206] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958 [cs.CL] https://arxiv.org/abs/2109.07958

[207] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023) , Yun-Nung Chen and Abhinav Rastogi (Eds.). Association for Computational Linguistics, Toronto, Canada, 47–58. https://doi.org/10.18653/v1/2023.nlp4convai-1.5

[208] Yu-Xiang Lin, Chih-Kai Yang, Wei-Chih Chen, Chen-An Li, Chien-yu Huang, Xuanjun Chen, and Hung-yi Lee. 2025. A Preliminary Exploration with GPT-4o Voice Mode. arXiv preprint arXiv:2502.09940 (2025).

[209] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. CriticBench: Benchmarking LLMs for Critique-Correct Reasoning. In Findings of the Association for Computational Linguistics: ACL 2024 , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1552–1587. https://doi.org/10.18653/v1/2024.findings-acl.91

[210] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In European Conference on Computer Vision . Springer, 366–384.

[211] Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[212] 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 Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing , Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2122–2132. https://doi.org/10.18653/v1/D16-1230

[213] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. arXiv preprint arXiv:2410.18451 (2024).

[214] Ganjun Liu, Xiaohui Hou, Meng Ge, Tao Zhang, and Haizhou Li. 2024. A non-intrusive approach to assessing dysarthria severity: Advancing clinical diagnosis. In Companion Proceedings of the ACM Web Conference 2024 . 1134–1137.

[215] Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024. MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models. arXiv:2407.10990 [cs.CL] https://arxiv.org/abs/2407.10990

[216] Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024. X-Eval: Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) . 8552–8571.

[217] Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024. X-Eval: Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects. arXiv:2311.08788 [cs.CL] https://arxiv.org/abs/2311.08788

[218] Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and Refining the Distinct Metric. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 762–770. https://doi.org/10.18653/v1/2022.acl-short.86

[219] Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese Alignment of Large Language Models. arXiv:2311.18743 [cs.CL] https://arxiv.org/abs/2311.18743

[220] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . 22139–22149.

[221] Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Towards Interpretable and Efficient Automatic Reference-Based Summarization Evaluation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 16360–16368. https://doi.org/10.18653/v1/2023.emnlp-main.1018

[222] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634 [cs.CL] https://arxiv.org/abs/2303.16634

[223] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. Advances in Neural Information Processing Systems 36 (2024).

[224] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro { BERT } a: A Robustly Optimized { BERT } Pretraining Approach. https://openreview.net/forum?id=SyxS0T4tvS

[225] Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition. arXiv:2402.15754 [cs.CL] https://arxiv.org/abs/2402.15754

[226] Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024. RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style. arXiv:2410.16184 [cs.CL] https://arxiv.org/abs/2410.16184

[227] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177 [cs.CV] https://arxiv.org/abs/2402.17177

[228] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-Time Scaling for Generalist Reward Modeling. arXiv:2504.02495 [cs.CL] https://arxiv.org/abs/2504.02495

[229] 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 (2019).

[230] Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In Proceedings of the Second Conference on Machine Translation , Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 589–597. https://doi.org/10.18653/v1/W17-4767

[231] Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1) , 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 (Eds.). Association for Computational Linguistics, Florence, Italy, 507–513. https://doi.org/10.18653/v1/W19-5358

[232] David Lopez-Paz and Maxime Oquab. 2016. Revisiting classifier two-sample tests. arXiv preprint arXiv:1610.06545 (2016).

[233] 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 Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1116–1126. https://doi.org/10.18653/v1/P17-1103

[234] Ke-Han Lu, Chun-Yi Kuan, and Hung-yi Lee. 2025. Speech-ifeval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models. arXiv preprint arXiv:2505.19037 (2025).

[235] Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2022. Toward Human-Like Evaluation for Natural Language Generation with Error Analysis. arXiv:2212.10179 [cs.CL] https://arxiv.org/abs/2212.10179

[236] 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 (2024).

[237] Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Chen Xu, and Heyan Huang. 2024. Beyond Exact Match: Semantically Reassessing Event Extraction by Large Language Models. arXiv:2410.09418 [cs.CL] https://arxiv.org/abs/2410.09418

[238] Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Tong Zhang, Yu-Shi Zhu, and Heyan Huang. 2025. SEOE: A Scalable and Reliable Semantic Evaluation Framework for Open Domain Event Detection. arXiv:2503.03303 [cs.CL] https://arxiv.org/abs/2503.03303

[239] Liangchen Luo, Zi Lin, Yinxiao Liu, Lei Shu, Yun Zhu, Jingbo Shang, and Lei Meng. 2023. Critique Ability of Large Language Models. arXiv:2310.04815 [cs.LG] https://arxiv.org/abs/2310.04815

[240] Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 696–700.

[241] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. arXiv:2401.13178 [cs.CL]

[242] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. arXiv preprint arXiv:2505.13032 (2025).

[243] Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Yu-Shi Zhu, Tong Zhang, Heyan Huang, and Xian-Ling Mao. 2025. Multi-modal Retrieval Augmented Multi-modal Generation: Datasets, Evaluation Metrics and Strong Baselines. arXiv:2411.16365 [cs.CL] https://arxiv.org/abs/2411.16365

[244] Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Shu-Hang Liu, Heyan Huang, Zhijing Wu, Chen Xu, and Xian-Ling Mao. 2025. T2I-Eval-R1: Reinforcement Learning-Driven Reasoning for Interpretable Text-to-Image Evaluation. arXiv:2505.17897 [cs.AI] https://arxiv.org/abs/2505.17897

[245] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 16383–16408. https://doi.org/10.18653/v1/2023.acl-long.905

[246] Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative Reward Models. arXiv:2410.12832 [cs.LG] https://arxiv.org/abs/2410.12832

[247] Gallil Maimon, Amit Roth, and Yossi Adi. 2025. Salmon: A Suite for Acoustic Language Model Evaluation. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[248] François Mairesse, Milica Gašić, Filip Jurčíček, 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 , Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre (Eds.). Association for Computational Linguistics, Uppsala, Sweden, 1552–1561. https://aclanthology.org/P10-1157/

[249] 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) . IEEE, 1–5.

[250] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing . 9004–9017.

[251] 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 (2022).

[252] Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2025. EmergentTTS-Eval: Evaluating TTS Models on Complex Prosodic, Expressiveness, and Linguistic Challenges Using Model-as-a-Judge. arXiv preprint arXiv:2505.23009 (2025).

[253] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2799–2808. https://doi.org/10.18653/v1/P19-1269

[254] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. LLM Critics Help Catch LLM Bugs. arXiv:2407.00215 [cs.SE] https://arxiv.org/abs/2407.00215

[255] Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue , Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 1st virtual meeting, 225–235. https://doi.org/10.18653/v1/2020.sigdial-1.28

[256] Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 681–707. https://doi.org/10.18653/v1/2020.acl-main.64

[257] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. 2024. PhyBench: A Physical Commonsense Benchmark for Evaluating Text-to-Image Models. arXiv preprint arXiv:2406.11802 (2024).

[258] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL] https://arxiv.org/abs/1301.3781

[259] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing . 12076–12100.

[260] Christoph Minixhofer, Ondřej Klejch, and Peter Bell. 2024. TTSDS-Text-to-Speech Distribution Score. In 2024 IEEE Spoken Language Technology Workshop (SLT) . IEEE, 766–773.

[261] Gabriel Mittag, Ross Cutler, Yasaman Hosseinkashi, Michael Revow, Sriram Srinivasan, Naglakshmi Chande, and Robert Aichner. 2020. DNN No-Reference PSTN Speech Quality Prediction. In Proc. Interspeech 2020 . 2867–2871.

[262] Gabriel Mittag and Sebastian Möller. 2021. Deep learning based assessment of synthetic speech naturalness. arXiv preprint arXiv:2104.11673 (2021).

[263] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494 (2021).

[264] 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 Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10512–10526. https://doi.org/10.18653/v1/2022.emnlp-main.718

[265] 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 Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5739–5754. https://doi.org/10.18653/v1/2022.acl-long.394

[266] Gonçalo Mordido and Christoph Meinel. 2020. Mark-Evaluate: Assessing Language Generation using Population Estimation Methods. In Proceedings of the 28th International Conference on Computational Linguistics , Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 1963–1977. https://doi.org/10.18653/v1/2020.coling-main.178

[267] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations , Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16

[268] mrfakename, Vaibhav Srivastav, Clémentine Fourrier, Lucain Pouget, Yoach Lacombe, main, and Sanchit Gandhi. 2024. Text to Speech Arena. https://huggingface.co/spaces/TTS-AGI/TTS-Arena.

[269] Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. arXiv:2408.08781 [cs.AI] https://arxiv.org/abs/2408.08781

[270] 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 Findings of the Association for Computational Linguistics: ACL 2023 , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4485–4503. https://doi.org/10.18653/v1/2023.findings-acl.274

[271] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. arXiv:1808.08745 [cs.CL]

[272] Matteo Negri, Marco Turchi, José G. C. de Souza, and Daniele Falavigna. 2014. Quality Estimation for Automatic Speech Recognition. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers , Junichi Tsujii and Jan Hajic (Eds.). Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1813–1823. https://aclanthology.org/C14-1171/

[273] Preksha Nema and Mitesh M. Khapra. 2018. Towards a Better Metric for Evaluating Question Generation Systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing , Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3950–3959. https://doi.org/10.18653/v1/D18-1429

[274] Jun-Ping Ng and Viktoria Abrecht. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing , Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 1925–1930. https://doi.org/10.18653/v1/D15-1222

[275] Tuan Nguyen, Corinne Fredouille, Alain Ghio, Mathieu Balaguer, and Virginie Woisard. 2024. Exploring pathological speech quality assessment with ASR-powered Wav2Vec2 in data-scarce context. arXiv preprint arXiv:2403.20184 (2024).

[276] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10862–10878. https://doi.org/10.18653/v1/2024.acl-long.585

[277] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024.

GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[278] OpenBMB. 2025. UltraEval-Audio: An Easy-to-Use, Fast, and Easily Integrable Tool for Evaluating Audio LLM. https://github.com/OpenBMB/UltraEval-Audio. Accessed: 2025-04-18.

[279] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[280] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) . IEEE, 5206–5210.

[281] Prabhat Pandey, Rupak Vignesh Swaminathan, KV Girish, Arunasish Sen, Jian Xie, Grant P Strimel, and Andreas Schwarz. 2025. SIFT-50M: A Large-Scale Multilingual Dataset for Speech Instruction Fine-Tuning. arXiv preprint arXiv:2504.09081 (2025).

[282] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. arXiv preprint arXiv:1606.06031 (2016).

[283] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics , Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[284] ChaeHun Park, Seungil Lee, Daniel Rim, and Jaegul Choo. 2023. DEnsity: Open-domain Dialogue Evaluation Metric using Density Estimation. In Findings of the Association for Computational Linguistics: ACL 2023 , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14222–14236. https://doi.org/10.18653/v1/2023.findings-acl.896

[285] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) .

[286] Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. 2024. OffsetBias: Leveraging Debiased Data for Tuning Evaluators. In Findings of the Association for Computational Linguistics: EMNLP 2024 , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1043–1067. https://doi.org/10.18653/v1/2024.findings-emnlp.57

[287] Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. OffsetBias: Leveraging Debiased Data for Tuning Evaluators. arXiv:2407.06551 [cs.CL] https://arxiv.org/abs/2407.06551

[288] 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 (2016).

[289] Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. 2025. Audiopedia: Audio QA with Knowledge. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[290] Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, Shengmin Xu, and Xinyi Huang. 2025. JALMBench: Benchmarking Jailbreak Vulnerabilities in Audio Language Models. arXiv:2505.17568 [cs.CR]

[291] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[292] Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In Proceedings of the 28th International Conference on Computational Linguistics , Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 4164–4178. https://doi.org/10.18653/v1/2020.coling-main.368

[293] Jaden Pieper and Stephen Voran. 2024. AlignNet: Learning dataset score alignment functions to enable better training of speech quality estimators. In Proc. Interspeech 2024 . 82–86.

[294] 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 Advances in Neural Information Processing Systems , M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 4816–4828. https://proceedings.neurips.cc/paper_files/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf

[295] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952

[296] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation , Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 392–395. https://doi.org/10.18653/v1/W15-3049

[297] 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 Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7131–7150. https://doi.org/10.18653/v1/2021.emnlp-main.570

[298] Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning Compact Metrics for MT. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 751–762. https://doi.org/10.18653/

v1/2021.emnlp-main.58

[299] Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do Large Language Models Need for Machine Translation Evaluation?. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 3660–3674. https://doi.org/10.18653/v1/2024.emnlp-main.214

[300] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. 2024. WorldSimBench: Towards Video Generation Models as World Simulators. arXiv:2410.18072 [cs.CV] https://arxiv.org/abs/2410.18072

[301] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. (2024). arXiv:2401.03601 [cs.CL]

[302] Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5Score: Discriminative Fine-tuning of Generative Evaluation Metrics. arXiv:2212.05726 [cs.CL] https://arxiv.org/abs/2212.05726

[303] Bowen Qu, Haohui Li, and Wei Gao. 2024. Bringing Textual Prompt to AI-Generated Image Quality Assessment. arXiv preprint arXiv:2403.18714 (2024).

[304] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

[305] Alessandro Ragano, Jan Skoglund, and Andrew Hines. 2024. NOMAD: Unsupervised learning of perceptual embeddings for speech enhancement and non-matching reference audio quality assessment. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1011–1015.

[306] Alessandro Ragano, Jan Skoglund, and Andrew Hines. 2024. SCOREQ: Speech Quality Assessment with Contrastive Regression. arXiv preprint arXiv:2410.06675 (2024).

[307] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[308] Aditya Ravuri, Erica Cooper, and Junichi Yamagishi. 2024. Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction. In 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) . IEEE, 580–584.

[309] Suman Ravuri and Oriol Vinyals. 2019. Classification accuracy score for conditional generative models. Advances in neural information processing systems 32 (2019).

[310] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 6493–6497.

[311] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

[312] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q & A Benchmark. In First Conference on Language Modeling . https://openreview.net/forum?id=Ti67584b98

[313] 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. arXiv:2009.10297 [cs.SE]

[314] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221) , Vol. 2. IEEE, 749–752.

[315] David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. 2024. NatureLM-audio: an Audio-Language Foundation Model for Bioacoustics. arXiv preprint arXiv:2411.07186 (2024).

[316] Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. 2025. Multilingual and Multi-Accent Jailbreaking of Audio LLMs. arXiv preprint arXiv:2504.01094 (2025).

[317] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[318] Md Rashad Al Hasan Rony, Liubov Kovriguina, Debanjan Chaudhuri, Ricardo Usbeck, and Jens Lehmann. 2022. RoMe: A Robust Metric for Evaluating Natural Language Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5645–5657. https://doi.org/10.18653/v1/2022.acl-long.387

[319] Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error Identification for Machine Translation with Metric Embedding and Attention. In Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems , Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 146–156. https://doi.org/10.18653/v1/2021.eval4nlp-1.15

[320] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition . 22500–22510.

[321] 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 (2024).

[322] 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 (2022).

[323] 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. arXiv:2009.11321 [cs.CL] https://arxiv.org/abs/2009.11321

[324] 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 (2020), 810–827. https://doi.org/10.1162/tacl_a_00347

[325] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. Advances in neural information processing systems 31 (2018).

[326] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. arXiv preprint arXiv:1907.10641 (2019).

[327] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. arXiv preprint arXiv:2410.19168 (2024).

[328] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. Advances in neural information processing systems 29 (2016).

[329] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019).

[330] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems 35 (2022), 25278–25294.

[331] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6594–6604. https://doi.org/10.18653/v1/2021.emnlp-main.529

[332] 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) . IEEE, 1–5.

[333] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

[334] Jiatong Shi, Yifan Cheng, Bo-Hao Su, Hye jin Shim, Jinchuan Tian, Samuele Cornell, Yiwen Zhao, Siddhant Arora, and Shinji Watanabe. 2025. ARECHO: Autoregressive Evaluation via Chain-Based Hypothesis Optimization for Speech Multi-Metric Estimation. arXiv:2505.24518 [cs.SD] https://arxiv.org/abs/2505.24518

[335] Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, et al. 2024. VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music. arXiv preprint arXiv:2412.17667 (2024).

[336] Jiatong Shi, Hye-Jin Shim, and Shinji Watanabe. 2025. Uni-VERSA: Versatile Speech Assessment with a Unified Network. arXiv preprint arXiv:2505.20741 (2025).

[337] Yu-Fei Shi, Yang Ai, and Zhen-Hua Ling. 2025. Universal Preference-Score-based Pairwise Speech Quality Assessment. arXiv:2506.01455 [cs.SD] https://arxiv.org/abs/2506.01455

[338] Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren, and Zhaochun Ren. 2023. RADE: Reference-Assisted Dialogue Evaluation for Open-Domain Dialogue. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12856–12875. https://doi.org/10.18653/v1/2023.acl-long.719

[339] Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers , 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 (Eds.). Association for Computational Linguistics, Belgium, Brussels, 751–758. https://doi.org/10.18653/v1/W18-6456

[340] 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 . Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 223–231. https://aclanthology.org/2006.amta-papers.25

[341] Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. LLM-as-a-Judge & Reward Model: What They Can and Cannot Do. arXiv:2409.11239 [cs.CL] https://arxiv.org/abs/2409.11239

[342] Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models. arXiv:2410.17578 [cs.CL] https://arxiv.org/abs/2410.17578

[343] Mingyang Song, Mao Zheng, Xuan Luo, and Yue Pan. 2025. Can Many-Shot In-Context Learning Help LLMs as Evaluators? A Preliminary Empirical Study. arXiv:2406.11629 [cs.CL] https://arxiv.org/abs/2406.11629

[344] Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, et al. 2025. Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models. arXiv preprint arXiv:2505.15406 (2025).

[345] Benjamin Stahl and Hannes Gamper. 2025. Distillation and Pruning for Scalable Self-Supervised Representation-Based Speech Quality Assessment. arXiv preprint arXiv:2502.05356 (2025).

[346] Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In Proceedings of the Ninth Workshop on Statistical Machine Translation , Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia (Eds.). Association for Computational Linguistics, Baltimore, Maryland, USA, 414–419. https://doi.org/10.3115/v1/W14-3354

[347] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS '20) . Curran Associates Inc., Red Hook, NY, USA, Article 253, 14 pages.

[348] Yixuan Su and Nigel Collier. 2023. Contrastive Search Is What You Need For Neural Text Generation. arXiv:2210.14140 [cs.CL]

[349] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A Contrastive Framework for Neural Text Generation. arXiv:2202.06417 [cs.CL] https://arxiv.org/abs/2202.06417

[350] Subrina Sultana and Donald S Williamson. 2025. A Pre-training Framework that Encodes Noise Information for Speech Quality Assessment. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[351] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2024. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. arXiv preprint arXiv:2407.14505 (2024).

[352] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. 2024. Journeydb: A benchmark for generative image understanding. Advances in Neural Information Processing Systems 36 (2024).

[353] Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024. The Critique of Critique. arXiv:2401.04518 [cs.CL] https://arxiv.org/abs/2401.04518

[354] Simeng Sun and Ani Nenkova. 2019. The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1216–1221. https://doi.org/10.18653/v1/D19-1116

[355] Wenhao Sun, Rong-Cheng Tu, Yifu Ding, Zhao Jin, Jingyi Liao, Shunyu Liu, and Dacheng Tao. 2025. VORTA: Efficient Video Diffusion via Routing Sparse Attention. arXiv preprint arXiv:2505.18809 (2025).

[356] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, Zhao Jin, and Dacheng Tao. 2024. AsymRnR: Video Diffusion Transformers Acceleration with Asymmetric Reduction and Restoration. arXiv preprint arXiv:2412.11706 (2024).

[357] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. 2024. Diffusion model-based video editing: A survey. arXiv preprint arXiv:2407.07111 (2024).

[358] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv preprint arXiv:2210.09261 (2022).

[359] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on audio, speech, and language processing 19, 7 (2011), 2125–2136.

[360] Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing , Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1882–1891. https://doi.org/10.18653/v1/D16-1193

[361] Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3553–3558. https://doi.org/10.18653/v1/2020.acl-main.327

[362] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937 (2018).

[363] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. JudgeBench: A Benchmark for Evaluating LLM-based Judges. arXiv:2410.12784 [cs.AI] https://arxiv.org/abs/2410.12784

[364] Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin. 2024. Singmos: An extensive open-source singing voice dataset for mos prediction. arXiv preprint arXiv:2406.10911 (2024).

[365] Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, and Junyang Lin. 2025. Enabling Scalable Oversight via Self-Evolving Critic. arXiv:2501.05727 [cs.CL] https://arxiv.org/abs/2501.05727

[366] Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, and Junyang Lin. 2025. RealCritic: Towards Effectiveness-Driven Evaluation of Language Model Critiques. arXiv:2501.14492 [cs.CL] https://arxiv.org/abs/2501.14492

[367] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. arXiv:1701.03079 [cs.CL] https://arxiv.org/abs/1701.03079

[368] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . 17918–17928.

[369] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. arXiv:2406.12624 [cs.CL] https://arxiv.org/abs/2406.12624

[370] Lan Tian, Ma Ziao, Zhou Yanghao, Xu Chen, and Mao Xianling. 2024. A Survey of Automatic Evaluation on the Quality of Generated Text. In Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum) , Xin Zhao (Ed.). Chinese Information Processing Society of China, Taiyuan, China, 169–196. https://aclanthology.org/2024.ccl-2.10/

[371] Xiaohai Tian, Kaiqi Fu, Shaojun Gao, Yiwei Gu, Kai Wang, Wei Li, Zejun Ma, and AI Bytedance. 2022. A multi-task and transfer learning based approach for MOS prediction. (2022).

[372] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. 2025. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. arXiv preprint arXiv:2502.05139 (2025).

[373] Prapti Trivedi, Aditya Gulati, Oliver Molenschot, Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Keith Stevens, Tanveesh Singh Chaudhery, Jahnavi Jambholkar, James Zou, and Nazneen Rajani. 2024. Self-rationalization improves LLM as a fine-grained judge. arXiv:2410.05495 [cs.CL] https://arxiv.org/abs/2410.05495

[374] 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) . 01–06. https://doi.org/10.1109/ICDI57181.2022.10007404

[375] Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Yist Y Lin, and Hung-yi Lee. 2021. Utilizing self-supervised representations for MOS prediction. arXiv preprint arXiv:2104.03017 (2021).

[376] 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 (2022).

[377] Rong-Cheng Tu, Zhao Jin, Jingyi Liao, Xiao Luo, Yingjie Wang, Li Shen, and Dacheng Tao. 2025. MLLM-Guided VLM Fine-Tuning with Joint Inference for Zero-Shot Composed Image Retrieval. arXiv preprint arXiv:2505.19707 (2025).

[378] Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic Evaluation for Text-to-image Generation: Task-decomposed Framework, Distilled Training, and Meta-evaluation Benchmark. arXiv:2411.15488 [cs.CL] https://arxiv.org/abs/2411.15488

[379] Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic Evaluation for Text-to-image Generation: Task-decomposed Framework, Distilled Training, and Meta-evaluation Benchmark. arXiv preprint arXiv:2411.15488 (2024).

[380] Rong-Cheng Tu, Wenhao Sun, Zhao Jin, Jingyi Liao, Jiaxing Huang, and Dacheng Tao. 2024. SPAgent: Adaptive Task Decomposition and Model Selection for General Video Generation and Editing. arXiv preprint arXiv:2411.18983 (2024).

[381] Rong-Cheng Tu, Wenhao Sun, Hanzhe You, Yingjie Wang, Jiaxing Huang, Li Shen, and Dacheng Tao. 2025. Multimodal Reasoning Agent for Zero-Shot Composed Image Retrieval. arXiv preprint arXiv:2505.19952 (2025).

[382] Sathvik Udupa, Soumi Maiti, and Prasanta Kumar Ghosh. 2024. IndicMOS: Multilingual MOS Prediction for 7 Indian languages. In Proc. Interspeech 2024 . 2690–2694.

[383] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. [n. d.]. FVD: A new metric for video generation. ([n. d.]).

[384] Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems , Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy (Eds.). Association for Computational Linguistics, Online, 11–20. https://doi.org/10.18653/v1/2020.eval4nlp-1.2

[385] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726 [cs.CV] https://arxiv.org/abs/1411.5726

[386] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796 [cs.CL] https://arxiv.org/abs/2404.18796

[387] Ashvala Vinay and Alexander Lerch. 2023. Aquatk: An audio quality assessment toolkit. arXiv preprint arXiv:2311.10113 (2023).

[388] Alexandra Vioni, Georgia Maniati, Nikolaos Ellinas, June Sig Sung, Inchul Hwang, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2023. Investigating content-aware neural text-to-speech MOS prediction using prosodic and linguistic features. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[389] VoiceMOS Challenge Steering Committee. 2025. AudioMOS Challenge 2025. https://sites.google.com/view/voicemos-challenge/audiomos-challenge-2025. Accessed: 2025-04-18.

[390] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing . 17086–17105.

[391] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation. arXiv:2407.10817 [cs.CL] https://arxiv.org/abs/2407.10817

[392] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5008–5020. https://doi.org/10.18653/v1/2020.acl-main.450

[393] Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. 2024. Halu-J: Critique-Based Hallucination Judge. arXiv:2407.12943 [cs.CL] https://arxiv.org/abs/2407.12943

[394] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. arXiv preprint arXiv:2406.16020 (2024).

[395] Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks. arXiv:2404.06480 [cs.CL]

[396] Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024. Learning Personalized Alignment for Evaluating Open-ended Text Generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13274–13292. https://doi.org/10.18653/v1/2024.emnlp-main.737

[397] Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024. Learning Personalized Story Evaluation. https://openreview.net/forum?id=7AS7vaVU8d

[398] Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023. RAMP: Retrieval-augmented MOS prediction via confidence-based dynamic weighting. arXiv preprint arXiv:2308.16488 (2023).

[399] Hui Wang, Shiwan Zhao, Xiguang Zheng, Jiaming Zhou, Xuechen Wang, and Yong Qin. 2025. RAMP+: Retrieval-Augmented MOS Prediction With Prior Knowledge Integration. IEEE Transactions on Audio, Speech and Language Processing (2025).

[400] Hui Wang, Shiwan Zhao, Jiaming Zhou, Xiguang Zheng, Haoqin Sun, Xuechen Wang, and Yong Qin. 2024. Uncertainty-Aware Mean Opinion Score Prediction. arXiv preprint arXiv:2408.12829 (2024).

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

[402] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048 [cs.CL] https://arxiv.org/abs/2303.04048

[403] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9426–9439. https://doi.org/10.18653/v1/2024.acl-long.510

[404] Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2024. Direct Judgement Preference Optimization. arXiv:2409.14664 [cs.CL] https://arxiv.org/abs/2409.14664

[405] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition . 18359–18369.

[406] Siyang Wang and Éva Székely. 2024. Evaluating text-to-speech synthesis from a large discrete token-based speech language model. arXiv preprint arXiv:2405.09768 (2024).

[407] Shuohang Wang, Ruochen Xu, Yang Liu, Chenguang Zhu, and Michael Zeng. 2022. ParaTag: A Dataset of Paraphrase Tagging for Fine-Grained Labels, NLG Evaluation, and Data Augmentation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7111–7122. https://doi.org/10.18653/v1/2022.emnlp-main.479

[408] Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025. QualiSpeech: A Speech Quality Assessment Dataset with Natural Language Reasoning and Descriptions. arXiv preprint arXiv:2503.20290 (2025).

[409] Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, et al. 2025. Enabling auditory large language models for automatic speech quality evaluation. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[410] Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-Taught Evaluators. arXiv:2408.02666 [cs.CL] https://arxiv.org/abs/2408.02666

[411] 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. 2023. Shepherd: A Critic for Language Model Generation. arXiv:2308.04592 [cs.CL]

[412] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In The Eleventh International Conference on Learning Representations . https://openreview.net/forum?id=1PL1NIMMrw

[413] Yingzhi Wang, Pooneh Mousavi, Artem Ploujnikov, and Mirco Ravanelli. 2025. What Are They Doing? Joint Audio-Speech Co-Reasoning. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[414] Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. In The Twelfth International Conference on Learning Representations .

[415] 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. arXiv:2306.05087 [cs.CL] https://arxiv.org/abs/2306.05087

[416] Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, Mark Cusick, Param Kulkarni, Zhengping Ji, Yasser Ibrahim, and Xia Hu. 2024. DHP Benchmark: Are LLMs Good NLG Evaluators? arXiv:2408.13704 [cs.CL] https://arxiv.org/abs/2408.13704

[417] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. arXiv preprint arXiv:2408.01337 (2024).

[418] Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates. arXiv:2408.13006 [cs.CL] https://arxiv.org/abs/2408.13006

[419] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 6382–6388. https://doi.org/10.18653/v1/D19-1670

[420] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing , Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 1711–1721. https://doi.org/10.18653/v1/D15-1199

[421] Jack Weston, Raphael Lenain, Udeepa Meepegama, and Emil Fristed. 2022. Generative Pretraining for Paraphrase Evaluation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4052–4073. https://doi.org/10.18653/v1/2022.acl-long.280

[422] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. 2024. Revisiting Text-to-Image Evaluation with Gecko: On Metrics, Prompts, and Human Ratings. arXiv preprint arXiv:2404.16820 (2024).

[423] Dyah AMG Wisnu, Stefano Rini, Ryandhimas E Zezario, Hsin-Min Wang, and Yu Tsao. 2025. HAAQI-net: A non-intrusive neural music audio quality assessment model for hearing aids. IEEE Transactions on Audio, Speech and Language Processing (2025).

[424] Hanming Wu, Wenjuan Han, Hui Di, Yufeng Chen, and Jinan Xu. 2023. A Holistic Approach to Reference-Free Evaluation of Machine Translation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 623–636. https://doi.org/10.18653/v1/2023.acl-short.55

[425] Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3612–3621. https://doi.org/10.18653/v1/2020.emnlp-main.294

[426] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In Forty-first International Conference on Machine Learning .

[427] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. arXiv preprint arXiv:2407.19594 (2024).

[428] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. arXiv:2407.19594 [cs.CL] https://arxiv.org/abs/2407.19594

[429] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023).

[430] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. Conceptmix: A compositional image generation benchmark with controllable difficulty. arXiv preprint arXiv:2408.14339 (2024).

[431] Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing Dialogue Systems with Distribution Distances. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2192–2198. https://doi.org/10.18653/v1/2021.findings-acl.193

[432] Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, Shuhan Qi, Kejia Zhang, Jianyuan Sun, and Wenwu Wang. 2024. A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining. arXiv preprint arXiv:2407.04936 (2024).

[433] Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. arXiv preprint arXiv:2406.14598 (2024).

[434] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems 36 (2024).

[435] 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 (2016), 401–415. https://doi.org/10.1162/tacl_a_00107

[436] Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023. SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5166–5183. https://doi.org/10.18653/v1/2023.acl-long.283

[437] Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis. In Findings of the Association for Computational Linguistics: EMNLP 2022 , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6559–6574. https://doi.org/10.18653/v1/2022.findings-emnlp.489

[438] Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5967–5994. https://doi.org/10.18653/v1/2023.emnlp-main.365

[439] Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. URO-Bench: A Comprehensive Benchmark for End-to-End Spoken Dialogue Models. arXiv preprint arXiv:2502.17810 (2025).

[440] Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. 2025. SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information. arXiv preprint arXiv:2505.13237 (2025).

[441] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2024. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796 [cs.LG] https://arxiv.org/abs/2209.00796

[442] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. arXiv preprint arXiv:2402.07729 (2024).

[443] Zuhao Yang, Yingfang Yuan, Yang Xu, SHUO ZHAN, Huajun Bai, and Kefan Chen. 2023. FACE: Evaluating Natural Language Generation with Fourier Analysis of Cross-Entropy. In Advances in Neural Information Processing Systems , A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 17038–17056. https://proceedings.neurips.cc/paper_files/paper/2023/file/37094fdc81632915a5738293cf9b7ad4-Paper-Conference.pdf

[444] Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. 2025. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. arXiv preprint arXiv:2505.10793 (2025).

[445] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736 [cs.CL] https://arxiv.org/abs/2410.02736

[446] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv:2307.10928 [cs.CL] https://arxiv.org/abs/2307.10928

[447] Zihuiwen Ye, Fraser Greenlee-Scott, Max Bartolo, Phil Blunsom, Jon Ander Campos, and Matthias Gallé. 2024. Improving Reward Models with Synthetic Critiques. arXiv:2405.20850 [cs.CL] https://arxiv.org/abs/2405.20850

[448] Gaoxiong Yi, Wei Xiao, Yiming Xiao, Babak Naderi, Sebastian Möller, Wafaa Wardah, Gabriel Mittag, Ross Culter, Zhuohuang Zhang, Donald S Williamson, et al. 2022. ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications. In Proc. Interspeech 2022 . 3308–3312.

[449] Takenori Yoshimura, Gustav Eje Henter, Oliver Watts, Mirjam Wester, Junichi Yamagishi, and Keiichi Tokuda. 2016. A hierarchical predictor of synthetic speech naturalness using neural networks. In Interspeech 2016 . International Speech Communication Association, 342–346.

[450] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. 2024. Self-generated critiques boost reward modeling for language models. arXiv preprint arXiv:2411.16646 (2024).

[451] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. 2024. Self-Generated Critiques Boost Reward Modeling for Language Models. arXiv:2411.16646 [cs.CL] https://arxiv.org/abs/2411.16646

[452] Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024. BatchEval: Towards Human-like Text Evaluation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15940–15958. https://doi.org/10.18653/v1/2024.acl-long.846

[453] Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. UniCBE: An Uniformity-driven Comparing Based Evaluation Framework with Unified Multi-Objective Optimization. In The Thirteenth International Conference on Learning Representations . https://openreview.net/forum?id=rpwGUtTeA5

[454] Peiwen Yuan, Xinglin Wang, Jiayi Shi, Bin Sun, Yiwei Li, and Prof. Kan. 2023. Better Correlation and Robustness: A Distribution-Balanced Self-Supervised Learning Framework for Automatic Dialogue Evaluation. In Advances in Neural Information Processing Systems , A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 53857–53874. https://proceedings.neurips.cc/paper_files/paper/2023/file/a8b148559549ce33261e79b4400e0d77-Paper-Conference.pdf

[455] Peiwen Yuan, Yueqi Zhang, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. Beyond One-Size-Fits-All: Tailored Benchmarks for Efficient Evaluation. arXiv:2502.13576 [cs.LG] https://arxiv.org/abs/2502.13576

[456] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. arXiv:2401.10019 [cs.CL] https://arxiv.org/abs/2401.10019

[457] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. arXiv:2106.11520 [cs.CL] https://arxiv.org/abs/2106.11520

[458] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2023 . 4615–4635.

[459] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2023 , Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4615–4635. https://doi.org/10.18653/v1/2023.findings-emnlp.307

[460] Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. arXiv preprint arXiv:2504.00369 (2025).

[461] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? arXiv:1905.07830 [cs.CL]

[462] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In Proc. Interspeech 2019 . 1526–1530.

[463] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective. arXiv:2412.14135 [cs.AI] https://arxiv.org/abs/2412.14135

[464] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating Large Language Models at Evaluating Instruction Following. In International Conference on Learning Representations (ICLR) .

[465] Ryandhimas E Zezario, Bo-Ren Brian Bai, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao. 2024. Multi-task pseudo-label learning for non-intrusive speech quality assessment model. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 831–835.

[466] Ryandhimas E Zezario, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao. 2022. MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids. arXiv preprint arXiv:2204.03305 (2022).

[467] Ryandhimas E Zezario, Yu-Wen Chen, Szu-Wei Fu, Yu Tsao, Hsin-Min Wang, and Chiou-Shann Fuh. 2024. A study on incorporating Whisper for robust speech assessment. In 2024 IEEE International Conference on Multimedia and Expo (ICME) . IEEE, 1–6.

[468] 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 (2022), 54–70.

[469] Ryandhimas E Zezario, Sabato M Siniscalchi, Hsin-Min Wang, and Yu Tsao. 2025. A study on zero-shot non-intrusive speech assessment using large language models. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 1–5.

[470] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) . Association for Computational Linguistics, Toronto, Canada, 11328–11348. https://aclanthology.org/2023.acl-long.634

[471] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. arXiv:2305.16739 [cs.CL] https://arxiv.org/abs/2305.16739

[472] Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Difficulty-Aware Machine Translation Evaluation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 26–32. https://doi.org/10.18653/v1/2021.acl-short.5

[473] 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 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) , Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5676–5689. https://doi.org/10.18653/v1/2021.acl-long.441

[474] Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2021. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation . Springer Singapore, Singapore, 53–69. https://doi.org/10.1007/978-981-15-8395-7_5

[475] Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. MDD-Eval: Self-Training on Augmented Data for Multi-Domain Dialogue Evaluation. arXiv:2112.07194 [cs.CL] https://arxiv.org/abs/2112.07194

[476] Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022. FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3336–3355. https://doi.org/10.18653/v1/2022.emnlp-main.220

[477] Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2023. PoE: A Panel of Experts for Generalized Automatic Dialogue Assessment. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023), 1234–1250. https://doi.org/10.1109/TASLP.2023.3250825

[478] Chen Zhang, Grandee Lee, Luis Fernando D' Haro, and Haizhou Li. 2021. D-Score: Holistic Dialogue Evaluation Without Reference. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 29 (April 2021), 2502–2516. https://doi.org/10.1109/TASLP.2021.3074012

[479] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. arXiv:2303.13336 [cs.SD] https://arxiv.org/abs/2303.13336

[480] Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC ^{3} : Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. In Findings of the Association for Computational Linguistics: EMNLP 2023 . Association for Computational Linguistics, Singapore, 15445–15458. https://doi.org/10.18653/v1/2023.findings-emnlp.1032

[481] Jingyi Zhang and Josef van Genabith. 2020. Translation Quality Estimation by Jointly Learning to Score and Rank. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2592–2598. https://doi.org/10.18653/v1/2020.emnlp-main.205

[482] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition . 586–595.

[483] Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025. Process-based Self-Rewarding Language Models. arXiv:2503.03746 [cs.CL] https://arxiv.org/abs/2503.03746

[484] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. 2024. Learning Multi-dimensional Human Preference for Text-to-Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . 8018–8027.

[485] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[486] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations . https://openreview.net/forum?id=SkeHuCVFDr

[487] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. arXiv:2401.02009 [cs.CL] https://arxiv.org/abs/2401.02009

[488] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the Performance of Large Language Models on GAOKAO Benchmark.

[489] Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. 2024. Benchmarking and Learning Multi-Dimensional Quality Evaluator for Text-to-3D Generation. arXiv preprint arXiv:2412.11170 (2024).

[490] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv:1911.00536 [cs.CL] https://arxiv.org/abs/1911.00536

[491] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In ACL, system demonstration .

[492] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safety-Bench: Evaluating the Safety of Large Language Models with Multiple Choice Questions. arXiv preprint arXiv:2309.07045 (2023).

[493] Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1656–1671. https://doi.org/10.18653/v1/2020.acl-main.151

[494] 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 Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 563–578. https://doi.org/10.18653/v1/D19-1053

[495] Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics , Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3865–3883. https://doi.org/10.18653/v1/2023.eacl-main.278

[496] Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. arXiv preprint arXiv:2412.06559 (2024).

[497] 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. arXiv:2306.05685 [cs.CL]

[498] 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. arXiv:2306.05685 [cs.CL]

https://arxiv.org/abs/2306.05685

[499] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. 2023. CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X. arXiv:2303.17568 [cs.LG]

[500] 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 Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2023–2038. https://doi.org/10.18653/v1/2022.emnlp-main.131

[501] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364 [cs.CL]

[502] 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+) , Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Dublin, Ireland (Virtual), 186–191. https://aclanthology.org/2020.webnlg-1.22/

[503] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. arXiv preprint arXiv:2311.07911 (2023).

[504] Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. CodeBERTScore: Evaluating Code Generation with Pretrained Models of Code. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13921–13937. https://doi.org/10.18653/v1/2023.emnlp-main.859

[505] Wangchunshu Zhou and Ke Xu. 2020. Learning to Compare for Better Training and Evaluation of Open Domain Natural Language Generation Models. arXiv:2002.05058 [cs.CL] https://arxiv.org/abs/2002.05058

[506] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. arXiv preprint arXiv:2310.17631 (2023).

[507] Wanrong Zhu, Xin Wang, An Yan, Miguel Eckstein, and William Yang Wang. 2023. ImaginE: An Imagination-Based Automatic Evaluation Metric for Natural Language Generation. In Findings of the Association for Computational Linguistics: EACL 2023 , Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 93–105. https://doi.org/10.18653/v1/2023.findings-eacl.6

[508] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. arXiv:1802.01886 [cs.CL] https://arxiv.org/abs/1802.01886

[509] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. 2024. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. arXiv preprint arXiv:2406.15877 (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009