

# 大语言模型数据审计员

## 基于指标的合成数据质量与可信度评估综述

Kaituo Zhang, Mingzhi Hu, Hoang Anh Duy Le, Fariha Kabir Torsha, Zhimeng Jiang, Minh Khai Bui, Chia-Yuan Chang, Yu-Neng Chuang, Zhen Xiong, Ying Lin, Guanchu Wang, Na Zou

**Affiliations:** University of Houston, Worcester Polytechnic Institute, Rice University, Texas A&M University, University of Wisconsin - Madison, University of Southern California, University of North Carolina at Charlotte

大语言模型 (LLMs) 已成为跨多种模态生成数据的强大工具。通过将稀缺资源转化为可控资产, LLMs 缓解了真实世界数据获取成本对模型训练、评估和系统迭代所造成的瓶颈。然而, 确保 LLM 生成的合成数据具有高质量仍是关键挑战。现有研究主要关注生成方法论, 对生成数据质量的关注有限。此外, 大多数研究局限于单一模态, 缺乏对不同类型数据的统一视角。为弥合这一差距, 我们提出 **LLM Data Auditor 框架**。在该框架中, 我们首先描述了 LLM 如何用于生成六种不同模态的数据。更重要的是, 我们从质量与可信度两个维度, 系统地对合成数据的内在评估指标进行了分类。这一方法将关注点从依赖下游任务性能的外部评估, 转向数据本身的固有属性。利用该评估体系, 我们分析了各模态代表性生成方法的实验评估结果, 并揭示了当前评估实践中的显著不足。基于这些发现, 我们为社区提供了具体的建议, 以改进数据生成的评估。最后, 该框架概述了合成数据在不同模态中实际应用的方法论。

✉ **Main Contact:** kzhang42@cougarnet.uh.edu

🔗 **Github:** <https://github.com/KaituoZhang/Awesome-LLM-Data-Generation>

## 1 引言

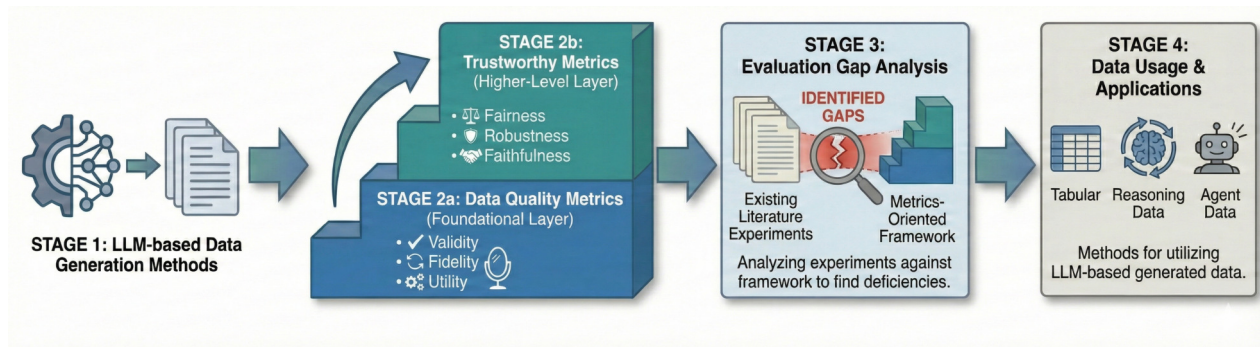
数据是现代人工智能发展的基石。随着真实世界数据源的逐渐枯竭, 合成数据日益受到研究界的青睐, 基于模型的数据生成已成为一种新的范式。因此, 具备强大生成能力的大模型 (LLMs) 在这一过程中发挥着关键作用。已有大量研究利用大模型在多个维度上进行数据生成。对于已有数据但缺乏标注的场景, [Martorana et al. \(2024\)](#) 提出了一种方法, 利用大模型生成的主题标注来支持元数据丰富。大模型还可用于控制或组合现有语料库 ([Penedo et al., 2025](#); [Soldaini et al., 2024](#))。此外, 大模型能够生成多种数据类型, 包括文本和代码 ([Wang et al., 2024e](#); [Nadăș et al., 2025](#))、表格数据 ([Fang et al., 2025](#)) 以及图数据 ([Ji et al., 2025](#))。它们也被应用于特定的实际领域, 例如生成临床记录 ([Barr et al., 2025](#)) 和为自动驾驶设计安全关键场景 ([Adekanye, 2024](#))。显然, 利用大模型进行数据生成正成为应对数据稀缺挑战的关键策略。

然而, 大模型生成数据的质量与严谨评估至关重要。众所周知, 高质量的数据能显著提升大模型性能; 例

如，在合成样本上施加简单的正确性标准，其对下游性能的提升作用可能与增加数据集大小相当 (Iskander et al., 2024)。相反，低质量的数据会严重降低模型性能，并影响实际应用。在模型层面，理论分析与大规模实验表明，对生成语料库进行重复训练时，对合成数据的无控制依赖会扭曲缩放法则，导致“模型坍塌”——即模型在多代自身输出的暴露下逐渐丧失能力并退化 (Dohmatob et al., 2024)。与此同时，关于不良记忆化和隐私泄露的研究也凸显出合成数据流水线可能危及个人可辨认或专有内容的风险 (Satvathy et al., 2025; Aditya et al., 2024; Shanmugarasa et al., 2025)。尽管这些问题正受到越来越多的关注，但许多现有的评估方法仍高度依赖大模型进行评分或筛选，这引入了显著的模型特异性偏差 (Gu et al., 2025)。因此，数据质量深刻影响着模型开发与实际应用，凸显出亟需建立更加系统化的数据评估方法体系。

当前关于基于大模型的数据生成研究主要集中在模型本身或生成过程上，常常忽视对生成数据的评估。例如，Long et al. (2024) 以自然语言处理中的通用生成与整理工作流为中心组织文献，但仅简要提及评估概念。类似地，Wang et al. (2024a) 关注合成数据使用的生命周期，将其划分为预训练、监督微调和对齐等阶段。尽管一些综述引入了评估方法，但这些方法往往局限于单一模态，如医疗或表格数据 (Pezoulas et al., 2024; Fang et al., 2024)。此外，大多数生成数据的评估属于外部评估，即仅衡量模型在下游任务性能上的提升。我们在表 9 中总结了一些代表性工作及其对比。相比之下，直接评估数据质量的内部评估仍处于发展不足的状态。尽管像 Dataflow (Liang et al., 2025) 这样的工作已形式化了“生成-评估-过滤-精炼”范式，但文献中仍缺乏一个统一的框架，在合成数据进入训练环之前对其进行审计。

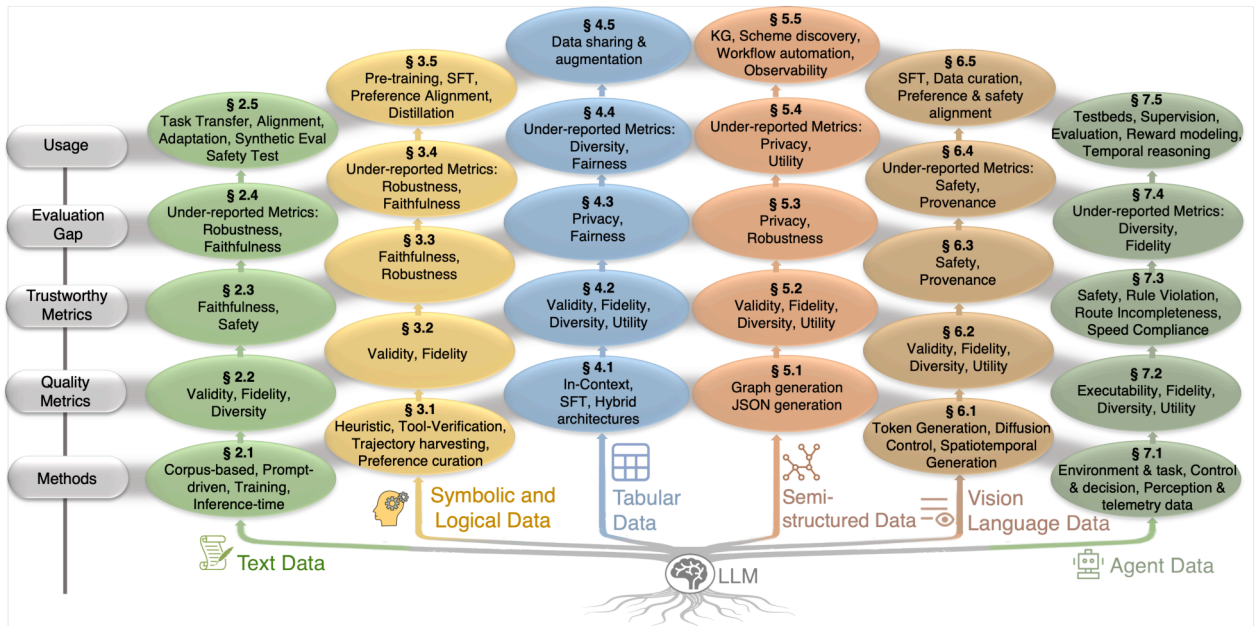
为了确保大模型生成的数据发挥其真正潜力，研究重点必须从生成技术转向评估方法。在本综述中，我们采用数据驱动的视角，并将度量指标作为主要组织原则。与现有侧重于工作流、生命周期或生成回顾的综述不同 (Long et al., 2024; Wang et al., 2024a; Goyal & Mahmoud, 2024)，我们提出了 **大模型数据审计框架**，如图 1 和 2 所示。该框架通过一个统一结构对多种数据类型进行组织，涵盖 5 个核心组件：基于大模型的数据生成方法、质量度量、可信度度量、评估差距以及数据使用。具体而言，质量度量关注有效性、保真度和实用性，以衡量基本可用性；而可信度度量则评估安全性与忠实性，以识别潜在风险。应用该框架，我们分析了代表性文献，识别出当前存在的评估差距并讨论实际应用场景。最终，我们的框架为社区提供了利用大模型生成和评估高质量、多模态数据的全面指南。



**Figure 1** LLM 数据审计框架概述。阶段 1：基于大语言模型的数据生成方法（第 2.1, 3.1, 4.1, 5.1, 6.1, 7.1 节）。阶段 2a：数据质量度量指标（第 2.2, 3.2, 4.2, 5.2, 6.2, 7.2 节）。阶段 2b：数据可信度度量指标（第 2.3, 3.3, 4.3, 5.3, 6.3, 7.3 节）。阶段 3：评估差距分析（第 2.4, 3.4, 4.4, 5.4, 6.4, 7.4 节）。阶段 4：数据使用（第 2.5, 3.5, 4.5, 5.5, 6.5, 7.5 节）。

最后，我们总结本综述的主要贡献如下：




- **转向数据视角以实现全面评估**与现有文献主要关注模型视角不同，我们的工作通过直接采用数据视角来区分自身，我们将其称为 **LLM 数据审计器**。本综述以数据为核心展开，首先介绍利用大模型生成数据，随后引入评估体系。我们进一步分析了当前研究在我们的评估体系中的不足之处，并最终讨论这些数据在各种方法中的应用方式。
- **系统性度量分类法**。我们的框架通过直接根据度量的实用性对其进行分类，提供了一种系统性的数据评估方法。我们首先将度量分为两大支柱：质量与可信性。随后，我们进一步细分为子类别，例如质量方面的有效性和保真度，以及可信性方面的公平性、鲁棒性和隐私性。这一贡献为社区提供了清晰的路线图，以直接评估合成数据的质量。
- **统一的跨模态覆盖**。大型语言模型数据审计器超越了单一模态，将六种主流数据模态进行组织，并在统一框架下进行评估。通过这一一致视角观察不同模态，我们的工作为社区提供了系统性指导，帮助人们在各种模态间生成、评估和利用数据。
- **评估差距分析**。在我们框架的指导下，我们对大语言模型生成数据中的代表性工作进行了系统性分析。通过审查这些研究的实验评估，我们识别出当前被忽视的关键评估维度。我们的分析突出了社区在未来研究中需要解决的具体评估差距，为更严格的データ评估提供了可操作的指导。



**Figure 2** 六种模态下基于大语言模型的合成数据生成概述。我们讨论文本数据（第 2 节）、符号与逻辑推理数据（第 3 节）、表格数据（第 4 节）、半结构化图、JSON 和日志数据（第 5 节），视觉-语言数据（第 6 节），以及智能体数据（第 7 节）。

## 2 文本数据

文本数据是大模型的主要模态。在生成式人工智能领域，研究人员越来越多地使用合成文本来扩充或替代真实世界的数据集，尤其是在资源有限或存在隐私顾虑的情景下。这一目标通过依赖特定提示的流水线创建人工但相关的训练样本得以实现。如Nadăș et al. (2025) 所述，现代合成文本生成技术大多以提示为条件，并优先考虑可控性。这些方法专注于特定属性、风格和边缘情况，同时确保生成的文本保持

Modality	Strategy	Representative Methods	Evaluation / Metrics	
Text Data 	Corpus-centric	RedPajama-V2 (Together Computer, 2023), FineWeb (Penedo et al., 2024), FineWeb2 (Penedo et al., 2025)	Validity	GAR, $s_{\text{USL-H}}$
	Prompt-based	Self-Instruct (Wang et al., 2023b), Evol-Instruct (Xu et al., 2024a), UltraFeedback (Cui et al., 2024)	Fidelity	EDS, $\text{Acc}_{\text{strict}}^{\text{prompt}}$ , $s_{\text{RUBER}}$ , PMI-FAITH
	Training	Self-Play Fine-Tuning (Chen et al., 2024b), SimPO (Meng et al., 2024), ORPO (Hong et al., 2024), GRPO (DeepSeek-AI et al., 2024)	Diversity	Self-CosSim, TTR, Ent-n
	Inference-time	Nucleus Sampling (Holtzman et al., 2020), Diverse Beam Search (Vijayakumar et al., 2018), JAM (Huang et al., 2025b), CoVe (Dhuliawala et al., 2024), RARR (Gao et al., 2023a), CheckGPT (Manakul et al., 2023)	Faithfulness	AttrAIS, Attrauto, $\text{Pres}_{\text{intent}}$ , $\text{Pres}_{\text{lev}}$ , $\text{Pres}_{\text{comb}}$ , $\text{F1}_{\text{AP}}$
Symbolic, Logical Reasoning Data 	Heuristic Evolution	WizardMath (Luo et al., 2025a), MetaMathQA (Yu et al., 2024a), WizardCoder (Luo et al., 2025b)	Validity	$\text{Acc}_{\text{verify}}$ , $\text{Acc}_{\text{prov}}$ , PassRate
	Tool-Verified Generation	OpenMathInstruct-1 (Toshniwal et al., 2024), OpenCodeInstruct (Ahmad et al., 2025), ProofWriter (Tafjord et al., 2021), SynLogic (Liu et al., 2025b), ALT-FLD $\times 2$ (Morishita et al., 2024)	Fidelity	$\text{Acc}_{\text{GC}}$ , Agree
	Trajectory-Harvesting	OpenAI o1 rollouts (OpenAI et al., 2024), SYNTHETIC-2 (Prime Intellect Team, 2025)	Faithfulness	$\text{Val}_{\text{step}}$ , $\text{Align}_{\text{entail}}$
	Preference-Curated LLM-Judge Data	UltraFeedback (Cui et al., 2024), Code-UltraFeedback (Weyssow et al., 2024), STaR (Zelikman et al., 2022)	Robustness	$\Delta_{\text{OOD}}$
Tabular Data 	Prompt-Based	CLLM Seedat et al. (2024), EPIC Kim et al. (2025), LITO Yang et al. (2024b), TabGen-ICL Fang et al. (2025), OCTree Nam et al. (2024)	Validity	$\chi^2$ , VR
	Fine-Tuning	GReaT Borisov et al. (2023), Nguyen et al. Nguyen et al. (2024), HARMONIC Wang et al. (2024d), Table-LLM-Specialist Xing et al. (2024), TableDreamer Zheng et al. (2025)	Fidelity	KST, $\text{TVD}$ , Pearson Score, DetScore, $P_{\alpha}$ , $\alpha$ -Prec
	Hybrid Architectures (Table $\leftrightarrow$ Text)	AIGT Zhang et al. (2024c), gTBLS Sundar et al. (2024), hybrid LLM+tabular / GAN-style models	Diversity	$R_{\beta}$ , $\beta$ -Rec
			Utility	AUC, MSE, RMSE
Semi-Structured 	Graph Data	LLM4GraphGen Yao et al. (2024), GoG Xu et al. (2024b), ontology-grounded KG generators Feng et al. (2024), GraphJudge Huang et al. (2025a), GAG Ji et al. (2025), GraphMaster Du et al. (2025)	Privacy	DCR, $\text{AUC}_{\text{MIA}}$ , $\text{Adv}_{\text{MIA}}$ , $\text{Gain}_{\text{AIA}}$
	JSON Data	Outlines dottxt-ai (2025), LM Format Enforcer Lu et al. (2025), SchemaBench-style RL methods Lu et al. (2025)	Fairness	$\Delta_{\text{SPD}}$ , $\Delta_{\text{EO}}$ , $\Delta_{\text{EOp}}$ , CovGap, Cond-Shift
	Log Data	LogBench Li et al. (2024d), AUCAD Zhang et al. (2025a),		
Vision-Language (Image / Video) 	Image-Text Native Autoregressive	Emu (Sun et al., 2024), Emu3 (Wang et al., 2024c), Chameleon (Team, 2025)	Validity	Valid <sub>rule</sub>
	Image-Text External Diffusion Control	Kosmos-G (Pan et al., 2024), GILL (Koh et al., 2023)	Fidelity	MMD, FCD
	Video-Text Native Spatiotemporal	VideoPoet (Kondratyuk et al., 2024), Emu3 (Wang et al., 2024c)	Diversity	Novelty, Uniq
	Video-Text Planner-based Diffusion	FlowZero (Lu et al., 2023), LVD (Lian et al., 2024), VideoDirectorGPT (Lin et al., 2024)	Utility	Acc, F1
Agent Data (DTs & Embodied) 	Environment & Task Data	ChatSUMO Li et al. (2024c), AutoScenario Lu et al. (2024), TTSG Ruan et al. (2025), ODD2CARLA Danso & B�ker (2025), SDT-LLM Naem et al. (2025), L3M+P Agarwal et al. (2025b), SELP Wu et al. (2025), T <sup>3</sup> Planner Li & Zhao (2025), PARTNR Chang et al. (2024)	Privacy	node-DP, edge-DP, edge-LDP
	Control & Decision Data	Grid-Agent Zhang et al. (2025c), Twin-2K-500 Toubia et al. (2025), BEHAVIORCHAIN Li et al. (2025a), LLM Trainer George & Farimani (2025), ELLMER Mon-Williams et al. (2025), Instruct2Act Huang et al. (2023), ProgPrompt Singh et al. (2022)	Robustness	$\sigma_{\text{SCR}}$ , GAD, $\text{GAD}^{\text{cap}}$ , $D_{L_2}$
	Perception & Telemetry Data	DefectTwin Ferdousi et al. (2024), SceneCraft Hu et al. (2024), Blender-LLM Du et al. (2024)		

注. 对于包含多种数据类型的模态 (例如, 半结构化、视觉-语言), “评估/指标”列针对第一种数据类型 (图、图像-文本对) 实例化了代表性指标; 其余数据类型 (JSON/日志、视频-文本) 的指标遵循相同结构, 并在第 5、6 节中详细说明。

**Table 1** 基于大语言模型的跨模态、策略、方法和评估指标的数据生成。



现实性并与任务相关。

在结构方面，合成文本的常见格式包括指令与输出对。在此格式中，一条指令与其目标响应相匹配，以支持模型遵循指令的训练。另一种常见格式涉及多轮指令序列，代表多次交互中的内容，其内容依赖于先前的上下文。

## 2.1 生成方法

控制文本数据生成的方法可按干预位置分类为：源端控制与组合、提示驱动生成与优化、参数高效及对齐控制，以及推理时引导与验证。

**语料库控制与构成。** 对于文本数据，控制通常通过将现有语料库处理为具有明确且可审计信号的可训练混合体来实现，而非生成新样本。现代数据整理流水线遵循一个标准流程，始于规范化和语言识别。随后，通过近似匹配技术（如在不同爬取快照间使用 Jaccard 或 MinHash 去重）移除近似重复的内容。最后，基于附加到每个文档的元数据结构化描述符进行文档级别的过滤 (Penedo et al., 2024, 2025)。大模型正被越来越多地用作可扩展的标注工具，将诸如教育质量等复杂抽象属性转化为可操作的标签或校准得分。通过将这些高层次判断提炼为轻量级分类器，研究人员能够将主观标准转化为系统性信号。该流水线实现了对大规模数据混合体的高效过滤与重新加权 (HuggingFaceFW, 2024)。

组合被视为一种主要的控制方法。组合指的是不同数据源和领域的结合及其相对比例。在实践中，混合策略通过三种主要策略实现。第一种策略涉及在不同数据流之间进行混合与加权。第二种策略通过使用质量信号，在单个文档层面选择数据。第三种策略涉及移除携带高风险的特定数据部分。在混合层面，透明的混合设计以及可复现的工具（如 Dolma）使实践者能够调整领域覆盖范围和数据预算，而不会将更多的数据量误认为更高的数据质量 (Soldaini et al., 2024)。在选择层面，为支持下游策略的定制化，一些发布版本将原始文档与质量信号分离。例如，RedPajama-V2 提供了网络文本以及语料库中一个子集的质量相关元数据。这使得用户可以应用自己的选择阈值，而无需重新爬取数据 (Weber et al., 2024; Together Computer, 2023)。在排除层面，面向安全性的处理可以在流水线早期进行。与其仅依赖训练后的拒绝机制，预训练数据过滤会移除那些被识别为支持有害能力的文档。这包括对化学、生物、放射性和核材料滥用提供显著支持的内容。该策略旨在实现有害能力的可测量降低，同时最小化对标准任务性能的负面影响 (Anthropic, 2025)。

**提示驱动的生成与优化。** 当模型参数被冻结时，控制通过上下文自举法和指令设计来实现。Self-Instruct (Wang et al., 2023b) 提出了通过提示模型从种子任务生成多样化实例，将模型的内部知识提取为标注数据集的方法。为应对生成数据复杂性增加的挑战，诸如 WizardLM 项目中的 Evol-Instruct (Xu et al., 2024a) 所采用的演化策略使用变异算子系统性地重写指令，以增加约束条件和推理深度。这些生成方法通常集成到遵循生成、排序和选择顺序的流水线中。在这些工作流中，一个辅助大语言模型充当评判者，根据特定标准评估候选输出。例如，UltraFeedback 框架 (Cui et al., 2024) 根据有用性、诚实性等标准评估输出。这一反馈环有效地将控制机制从手动提示工程转变为可扩展且自动化的偏好过滤 (Gu et al., 2025)。

**参数高效且基于对齐的控制。** 要从根本上改变输出分布，需要更新参数。尽管监督微调设定了行为基准，但近期进展已转向迭代自对弈机制。诸如自对弈微调 (Self-Play Fine-Tuning)，也称为 SPIN (Chen

et al., 2024b), 允许生成器通过在迭代自对弈环中对比自身生成的回应与人类示范来实现改进。该方法有效打破了静态监督的天花板, 且无需额外标注。

在偏好对齐领域, 研究正逐步超越复杂的基于人类反馈的强化学习流水线。新的无参考目标消除了对内存占用较大的参考模型的需求, 并将指令遵循直接整合到对齐损失中。典型的例子包括简单偏好优化 (Meng et al., 2024) 和几率比偏好优化 (Hong et al., 2024)。这些方法能够在无需分离奖励建模的情况下, 有效缓解长度偏差 (如冗长性)。

此外, 专门的目标现在针对高效的群体级动态。例如, 群体相对策略优化 (DeepSeek-AI et al., 2024) 通过在一组生成的输出之间归一化奖励, 而不是使用评论员模型。这种技术实现了可扩展的偏好最优化, 已被用于提升推理和数学表现。一旦配备了这些参数化控制, 模型便能有效地从通用预测器转变为专用的数据合成器。在此角色中, 它能够自主生成大量高保真度的样本, 严格遵循目标格式和推理协议。

**推理时的引导与验证。** 控制的最终层在运行时调节解码过程。随机解码策略如核采样 (Nucleus Sampling) (Holtzman et al., 2020) 平衡了多样性与合理性之间的权衡。类似地, 束搜索中的多样性促进惩罚机制, 如多样束搜索 (Diverse Beam Search) (Vijayakumar et al., 2018), 用于防止冗余。最近, 潜在空间干预作为一种精确的调控手段崭露头角。诸如 JAM (Huang et al., 2025b) 的方法在前向传播过程中修改活性值向量, 以调整情感或安全性等属性, 而无需重新训练。

为了确保可靠性, 事后验证机制充当质量过滤器。诸如链式验证 (Chain-of-Verification), 也称为 CoVe (Dhuliawala et al., 2024) 的系统, 会提示模型对其自身输出进行交叉检查。同时, RARR 框架 (Gao et al., 2023a) 通过将草稿与检索到的证据进行比较来修正内容。此外, SelfCheckGPT (Manakul et al., 2023) 利用一致性采样来检测并标记或过滤可能存在的幻觉内容。这一过程确保只有经过验证的样本才会保留在最终数据集中。

## 2.2 文本质量指标

我们将纯文本和多轮对话的评估统一在三个互补的维度下: **有效性**、**保真度**和 **多样性**。

**有效性。** 有效性要求生成内容在显式约束下结构上合理, 同时语言上可接受。除了结构上的考量, 我们还使用类似 CoLA 的语法可接受性评分器来评估语言上的合理性。遵循先前的研究工作, 我们采用句子级别的可接受性分类器为生成的句子分配可接受性得分 (Raman & Shah, 2023)。为此, 我们使用在 CoLA 基准上微调过的 RoBERTa-large 模型。基于这一做法, 我们报告一种语法可接受率, 也称为 GAR, 即可接受性得分超过固定阈值的生成句子所占比例:

$$\text{GAR} = \frac{1}{\sum_{i=1}^{|\mathcal{T}|} m_i} \sum_{i=1}^{|\mathcal{T}|} \sum_{j=1}^{m_i} \mathbb{I}[C_{\text{acc}}(u_{ij}) \geq \tau],$$

其中第  $i$  代  $t_i$  被分割为  $m_i$  个句子  $\{u_{ij}\}_{j=1}^{m_i}$ 。术语  $C_{\text{acc}}(u) \in [0, 1]$  表示可接受性得分或由 CoLA 训练分类器输出的概率。变量  $\tau$  是预先设定的决策阈值, 例如 0.5,  $\mathbb{I}[\cdot]$  为指示函数。

在对话领域中, 有效性延伸到了社交语用学的范畴。我们采用分层的 USL-H 指标 (Phy et al., 2020)。该指标表明, 一个回复必须按照特定顺序满足三个标准。我们将  $s_U$ 、 $s_S$  和  $s_L$  分别表示为可理解性、合理性和喜爱度的得分, 而  $s_L$  可以包含一个或多个品质  $q_j$ , 例如具体性或共情。其概念如下式所示:

$$s_{\text{USL-H}}(t) = \alpha_1 s_U(t) + \alpha_2 s_U(t) s_S(t) + \alpha_3 s_U(t) s_S(t) s_L(t).$$

$$s_L = \sum_j \beta_j q_j.$$

在这些公式中，变量  $s_U$ 、 $s_S$ 、 $s_L$  和  $q_j$  是取值范围从 0 到 1 的连续值。系数  $\alpha_i$  和  $\beta_j$  表示分配给每项质量的权重，其定义方式为各自之和等于 1。这些方法适用于自动评估和人工评估情景下的得分计算。

**忠诚** 保真度衡量生成内容在多大程度上保持了语义内容并遵循源材料。为了量化合成数据与真实数据在语料库层面的语义接近程度，我们采用基于嵌入的相似度方法。该方法通过余弦相似度比较均值规范化的嵌入向量，这一做法延续了先前分析中计算不同组别或语料库平均规范化嵌入之间相似度的方法 (Hämäläinen et al., 2023)。具体而言，我们使用如 SentenceTransformer 这类句子编码器将文本实例映射为嵌入向量。我们报告嵌入分布相似度，也称为 EDS，如下所示：

$$\text{EDS}(\mathcal{X}_{\text{syn}}, \mathcal{X}_{\text{real}}) = \cos(\bar{\mathbf{e}}_{\text{syn}}, \bar{\mathbf{e}}_{\text{real}}), \bar{\mathbf{e}}_{\star} = \frac{1}{|\mathcal{X}_{\star}|} \sum_{x \in \mathcal{X}_{\star}} \frac{f(x)}{\|f(x)\|_2}.$$

在此公式中， $\mathcal{X}_{\text{syn}}$  和  $\mathcal{X}_{\text{real}}$  分别表示合成语料库和真实语料库。对于数据集  $\mathcal{X}_{\star}$  (其中  $\star \in \{\text{syn}, \text{real}\}$ )，其均值嵌入  $\bar{\mathbf{e}}_{\star}$  在平均前使用  $\ell_2$  范数进行规范化的处理。 $f(\cdot)$  表示将原始数据点映射到高维特征空间的嵌入函数。最后，两个向量  $\mathbf{a}$  和  $\mathbf{b}$  之间的余弦相似度定义为两向量的点积除以其模长的乘积。

对于指令遵循任务，我们使用 IFEval 中的严格提示级别准确率来衡量指令遵循的保真度。该指标通过自动检查器 (Zhou et al., 2023) 评估提示中指定的可验证约束。具体而言，提示级别严格准确率要求所有可验证的指令均得到满足，其定义如下：

$$\text{Acc}_{\text{prompt}}^{\text{strict}} = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \mathbb{I} \left[ \bigwedge_{k \in \mathcal{I}(p_i)} V_k(t_i, p_i) = 1 \right],$$

其中  $\mathcal{P}$  表示提示集合， $p_i$  为第  $i$  个提示。项  $\mathcal{I}(p_i)$  表示  $p_i$  中可验证指令的子集。每个函数  $V_k(\cdot) \in \{0, 1\}$  用作第  $k$  个指令的自动检查器，例如长度、关键词数量或约束格式。我们定义  $\text{Fid}_{\text{instr}}$  与  $\text{Acc}_{\text{prompt}}^{\text{strict}}$  等价。

在对话情景中，保真度要求基于对话历史或外部证据。我们采用 RUBER (Tao et al., 2018)，该方法融合了有参考得分和无参考得分。按照 Tao et al. (2018)，我们首先通过最小-最大规范化将每个分量得分归一化到 0 到 1 的值域内：

$$\tilde{s} = \frac{s - \min(s)}{\max(s) - \min(s)}.$$

然后通过启发式聚合计算 RUBER 得分。在本工作中，我们采用算术平均：

$$s_{\text{RUBER}}(c, r, r_{\text{ref}}) = \frac{1}{2} \left( \tilde{s}_{\text{ref}}(r, r_{\text{ref}}) + \tilde{s}_{\text{unref}}(c, r) \right),$$

其中  $\tilde{s}_{\text{ref}}$  衡量候选参考文本的相似度， $\tilde{s}_{\text{unref}}$  评估上下文响应的恰当性。

当无法获取参考文本时，我们采用 USR (Mehri & Eskenazi, 2020)，这是一种无需参考的对话评估框架。该框架提供了与五种对话质量（如可理解性、自然性、上下文保持性、趣味性和知识使用）对齐的可解释子指标。按照原始设定，这些子得分通过一个回归模型聚合为总体质量估计值，该回归模型经过训练以重现人类对整体质量的评分 (Mehri & Eskenazi, 2020)。最后，为检测文档增强生成中的幻觉现象，我们使用 PMI-FAITH (Nandwani et al., 2023):

$$\text{PMI-FAITH}(d, h, r) = \log P(r \mid d, h) - \log P(r \mid h).$$

该微分项捕捉了响应  $r$  中由文档  $d$  所带来的信息增益，超越对话历史  $h$  的部分。

**多样性** 为了检测模式崩溃和重复现象，我们从语义和词汇两个维度评估多样性。在语义多样性方面，我们采用自余弦相似度 (Self Cosine Similarity)。该指标表示所有生成文本嵌入之间的成对余弦相似度的平均值。较低的自余弦相似度表明语义分布更为广泛 (Zhang et al., 2024d):

$$\text{Self-CosSim} = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} \cos(\mathbf{h}_i, \mathbf{h}_j),$$

其中  $\mathbf{h}_i$  是第  $i$  个生成文本的嵌入， $M$  是生成次数。

对于词汇多样性，我们报告类型-词元比率 (Treffers-Daller et al., 2018) 和 Distinct N (Li et al., 2016)。这些指标的定义如下:

$$\text{TTR} = \frac{|\text{Types}|}{|\text{Tokens}|}, \quad \text{Distinct-}N = \frac{|\text{Unique } n\text{-grams}|}{|\text{Total } n\text{-grams}|}.$$

较高的数值表示词汇使用更丰富且  $n$  元语法重复性降低。最后，我们计算  $n$  元语法响应熵以量化模型对词汇空间的利用是否均匀 (Zhang et al., 2024d):

$$\text{Ent-}n = - \sum_{g \in \mathcal{G}_n} p_{\mathcal{T}}(g) \log p_{\mathcal{T}}(g),$$

其中  $\mathcal{G}_n$  是生成语料库  $\mathcal{T}$  中所有  $n$  元语法的集合，而  $p_{\mathcal{T}}(g)$  是语料库级别的经验  $n$  元语法分布。

## 2.3 文本数据的可信度量

我们通过两个主要角度评估生成文本的可信度，包括对证据和人物设定的忠实性，以及对对抗性或自发性毒性的安全性。



**忠实与一致。** 我们从两个相互关联的方面评估忠实度。第一个方面涉及在有真实支持证据的情况下进行证据选择。第二个方面则关注生成内容对所提供来源的归因。

当存在黄金标准证据时，我们首先在证据单元层面（如支持性事实句子）比较预测的证据集  $\mathcal{E}_{\text{pred}}$  与黄金标准集  $\mathcal{E}_{\text{gold}}$  来评估证据选择。遵循像 HotpotQA (Yang et al., 2018) 这类多跳问答任务中的标准支持性事实协议，我们计算每个实例的准确率、召回率和 F1 得分：

$$\text{Prec}_{\text{evi}} = \frac{|\mathcal{E}_{\text{pred}} \cap \mathcal{E}_{\text{gold}}|}{|\mathcal{E}_{\text{pred}}|}, \quad \text{Rec}_{\text{evi}} = \frac{|\mathcal{E}_{\text{pred}} \cap \mathcal{E}_{\text{gold}}|}{|\mathcal{E}_{\text{gold}}|}, \quad \text{F1}_{\text{evi}} = \frac{2\text{Prec}_{\text{evi}}\text{Rec}_{\text{evi}}}{\text{Prec}_{\text{evi}} + \text{Rec}_{\text{evi}}}.$$

在此背景下，竖线表示集合的基数，交集项则计算正确选择的证据单元。

除了选择正确的证据外，我们还使用“可归因于已识别来源”框架 (Rashkin et al., 2023) 评估生成内容是否得到所提供来源的支持。该方法通过“根据 A, y”测试来定义归因：如果读者能够从来源集合 A 中验证某段文字或句子，而无需依赖外部知识，则认为该内容具有可归因性 (Rashkin et al., 2023)。遵循 RARR（基于研究与修订的重装归因）中的句子级别公式 (Gao et al., 2023a)，我们计算句子层面的平均可归因率：

$$\text{Attr}_{\text{AIS}}(y, A) = \frac{1}{|S(y)|} \sum_{s \in S(y)} \text{AIS}(s, A),$$

其中  $y$  为生成的段落， $S(y)$  表示  $y$  中的句子集合。术语  $A$  代表提供的证据片段集合，函数  $\text{AIS}(s, A)$  用于判断在 AIS 指南 (Gao et al., 2023a) 下，句子  $s$  是否被  $A$  中的一个或多个片段完全支持。

为了实现可扩展的评估，我们还使用了 RARR (Gao et al., 2023a) 提出的自动 AIS 方法。该方法通过基于自然语言推理的事实一致性模型来近似人类归因判断，如 TRUE (Honovich et al., 2022) 中所综述。我们令  $\text{NLI}(e, s)$  表示证据片段  $e$  蕴含句子  $s$  的概率。自动 AIS 度量方法为每句话识别出最佳支持片段，并按如下方式计算平均得分：

$$\text{Attr}_{\text{auto}}(y, A) = \frac{1}{|S(y)|} \sum_{s \in S(y)} \max_{e \in A} \text{NLI}(e, s).$$

为了量化修订后的段落  $\tilde{A}$  在事实修正之外与原始段落  $A$  的忠实程度，我们采用 RARR (Gao et al., 2023a) 中的保留度量。这些度量从两个互补的角度评估保留情况。第一个角度是意图保留，用于判断修订是否保持了原始意图。这通常由人工标注者根据特定评分标准进行判断。第二个角度是编辑最小化，用于衡量表面形式上的改动程度。该度量用于惩罚过度重写行为，例如风格改写、顺序调整或不必要的添加 (Gao et al., 2023a)。形式上，给定  $N$  对应的样本，我们将意图准确率定义为一个二值得分。若修订完全保留了原始意图，则得分为 1，否则为 0。这通过一个意图违反指示符  $\nu(\tilde{A}_i, A_i)$  表示如下：

$$\text{Pres}_{\text{intent}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\nu(\tilde{A}_i, A_i) = 0].$$

为避免在意图保持不变的情况下进行不必要的修改，我们采用准确率莱文斯坦距离。该度量基于字符级

别的莱文斯坦距离，通过原始长度进行归一化，并截断至零：

$$\text{Pres}_{\text{Lev}} = \frac{1}{N} \sum_{i=1}^N \max \left( 1 - \frac{d_{\text{lev}}(\tilde{A}_i, A_i)}{\text{len}(A_i)}, 0 \right).$$

为了优先考虑语义，我们定义了准确率合并，当意图被违反时，将表面相似度置零 (Gao et al., 2023a):

$$\text{Pres}_{\text{comb}} = \text{Pres}_{\text{intent}} \cdot \text{Pres}_{\text{Lev}}.$$

最后，我们使用它们的调和平均 (Gao et al., 2023a) 来总结归因与保持之间的权衡。

$$\text{F1}_{\text{AP}} = \frac{2 \cdot \text{Attr} \cdot \text{Pres}_{\text{comb}}}{\text{Attr} + \text{Pres}_{\text{comb}}},$$

其中，术语  $\text{Attr}$  可由基于人类的  $\text{Attr}_{\text{A}IS}$  或自动版本  $\text{Attr}_{\text{a}uto}$  实例化。

**安全** 我们通过测量模型在单轮和多轮情景下生成有害内容的倾向性来评估其安全性。

在单轮提示生成中，我们遵循 RealToxicityPrompts (Gehman et al., 2020) 的方法，并沿着两个互补的维度量化模型退化为有毒延续的倾向。第一个维度衡量最坏情况下采样得到的延续可能有多毒。第二个维度衡量毒性在多次采样中出现的频率。对于每个提示  $j$ ，我们采样  $K$  个延续，并使用毒性函数  $t(\cdot)$  (如 Perspective API) 对每个生成结果进行评分。当  $t(\cdot)$  大于或等于阈值  $\tau$  时，我们将该片段视为有毒。按照 RealToxicityPrompts 的做法，我们将毒性得分的阈值  $\tau$  设为 0.5 (Gehman et al., 2020)。

然后我们报告毒性概率和预期最大毒性。毒性概率定义为至少一个  $K$  样本超过阈值的实证概率。预期最大毒性定义为 across prompts (Gehman et al., 2020) 的平均最坏情况毒性：

$$\text{TP} = \frac{1}{N} \sum_{j=1}^N \mathbb{I} \left[ \max_{1 \leq \ell \leq K} t(g_{j,\ell}) \geq \tau \right], \quad \text{EMT} = \frac{1}{N} \sum_{j=1}^N \max_{1 \leq \ell \leq K} t(g_{j,\ell}).$$

在多轮交互中，我们采用 ToxicChat (Chen et al., 2023a) 提出的对话层级安全度量方法。这些度量方式定义的是整个对话过程中的风险，而非单一回复的风险。设  $\mathcal{D}_{\text{dlg}}$  为一组对话。对于集合  $\mathcal{D}_{\text{dlg}}$  中的每一条对话  $d$ ，我们令  $R_d$  表示所有聊天机器人回复的集合，并令  $q_{d,i}$  和  $r_{d,i}$  表示第  $i$  轮的查询-回复对。我们计算毒性的句子生成率，记为 TSG，即在交互过程中聊天机器人至少生成一条有毒回复的对话所占的比例 (Chen et al., 2023a):

$$\text{TSG} = \frac{1}{|\mathcal{D}_{\text{dlg}}|} \sum_{d \in \mathcal{D}_{\text{dlg}}} \mathbb{I} \left[ \max_{r \in R_d} t(r) \geq \tau \right].$$

为了分离出即使用户输入无毒但仍然引发毒性回复的情况，我们报告了非毒到毒比率 (Non Toxic to Toxic rate)，简称 NT2T。该指标表示在对话过程中，聊天机器人在任何时刻产生至少一条毒性回复的对话比例，而对应的用户查询为非毒性的 (Chen et al., 2023a):

Generation method	Validity	Fidelity	Diversity	Faithfulness	Safety
FineWeb2 (Penedo et al., 2025)	×	△	△	×	×
Dolma (Soldaini et al., 2024)	×	△	△	×	✓
RedPajama (Weber et al., 2024)	△	△	△	×	△
Self-Instruct (Wang et al., 2023b)	△	△	△	×	×
RARR (Gao et al., 2023a)	×	×	×	✓	×
PMI-FAITH (Nandwani et al., 2023)	×	✓	×	✓	×

**Table 2** 是否显式评估每个维度的代表性文本控制/数据构建/对齐方法。✓：显式评估；△：部分/间接覆盖；×：未报告或不适用。

$$\text{NT2T} = \frac{1}{|\mathcal{D}_{\text{dlg}}|} \sum_{d \in \mathcal{D}_{\text{dlg}}} \mathbb{I}[\exists i : t(q_{d,i}) < \tau \wedge t(r_{d,i}) \geq \tau].$$

该对话层级的 NT2T 定义与先前对查询-响应对在对话层级上的毒性分类 (NT2T、NT2NT、T2T 或 T2NT) 保持一致。在此分类中，NT2T 特指一个非毒性的查询引发了一个有毒的响应 (Si et al., 2022)。

最后，为确保安全干预不会降低语言质量，我们报告困惑度和 Distinct N (Liang et al., 2024; Jozefowicz et al., 2016; Li et al., 2016)。

## 2.4 评估实践差距

我们分析了第 2.1 节中的代表性方法，并将其报告的评估协议分类整理在表 2 中。

**有效性。**如表 2 所示，文本生成方法中很少使用明确或可复现的指标来评估有效性。大多数研究忽略了语法可接受性或对话质量的具体得分。相反，它们依赖下游基准或人工判断作为质量的一般代理指标。然而，这些外部信号往往无法识别具体的有效性错误。即使模型性能有所提升，大规模自动化流水线仍可能遭受基本的结构崩溃，包括编码噪声、格式违规和截断等问题。因此，对有效性的明确监控仍然是实际必需的。

**忠实性。**忠实性描述了大模型输出与原始源材料的一致程度。这一属性是解决幻觉问题最重要的因子之一。当大模型查找确切证据的能力得到提升时，模型产生虚构信息的可能性将大大降低。然而，表 2 显示，一些研究并未评估这一维度。我们提供了若干可用于衡量这些模型忠实性的指标。

**安全。**在所调查的方法中，安全报告几乎缺失。很少有研究采用标准化的评估协议（如 Perspective API 或有毒词汇列表）来衡量毒性得分或禁用词。因此，未来在合成文本方面的研究应将安全评估作为标准报告内容。通过与质量指标一同分析安全性能，研究人员可以更好地管理安全与实用性之间的权衡。这种方法对于防止在性能提升的同时出现安全退化至关重要。

## 2.5 用途

在本节中，我们将关注点从文本和对话生成方法转移到生成的合成语料库的实际应用上。我们重点关注离线应用场景，即模型输出被视为用于训练和评估的静态数据。这种方法与解码过程中的在线提示或控制截然不同。当合成语料库经过精心筛选并与目标部署分布相匹配时，其效果最佳。此外，这些语料库还应根据我们分类体系中的质量与可信度维度进行明确的压力测试。我们首先讨论训练阶段的应用，如数据增强、对齐和适应，然后描述评估方面的应用。

**监督数据增强与任务迁移。** 一个主要用例是在资源匮乏或无资源的环境下扩充监督训练数据。对于机器阅读理解任务，从未标记段落生成的合成问答对，经过筛选而非直接全部使用，可以显著提升下游任务的准确率。早期研究显示，大模型能够生成的合成 QA 语料库本身即可支持在 SQuAD 等基准测试上表现竞争力的模型。(Puri et al., 2020)。此外，端到端的合成 QA 生成结合筛选和验证，还能进一步实现领域自适应 (Shakeri et al., 2020)。

更近期的方法整合了基于大语言模型 (LLM) 的奖励或选择器，以识别高价值样本。例如，Jin & Wang (2024) 将生成式大语言模型视为奖励模型，对合成的问答对进行评分，并仅保留预测为最有利于训练的样本。关于指令微调中数据选择的实证证据表明，单纯扩大指令语料库会带来边际收益递减。这促使人们采用基于效用信号的数据高效选择方法，如质量、多样性、难度和复杂度等指标。实证研究表明，经过高度筛选的小规模数据集 (如包含 1000 个实例的 LIMA) 以及大规模人机协作语料库 (如 OpenAssistant 对话) 具有竞争力。此外，自动化选择器通过基于模型的过滤以及使用 GPT-4 等强大大语言模型进行评分，提高了效率，能够剔除低质量或错误的样本，同时保留多样且高价值的实例 (Albalak et al., 2024; Köpf et al., 2023)。

**指令微调与对齐语料库。** 除了任务标签之外，合成文本和对话被广泛用于构建指令遵循和偏好数据集以实现对齐。大规模的指令微调语料库通常混合了人工撰写的和大语言模型 (LLM) 撰写的指令、示范以及任务变体。合成部分扩展了技能、格式和领域的覆盖范围，而人工劳动则集中于种子数据的生成和抽样检查。像 UltraFeedback 这样的偏好和批评数据集使用如 GPT-4 这样强大大语言模型，对候选响应提供多维度得分和自然语言反馈。这为奖励建模和基于偏好的微调生成了大规模的合成偏好语料 (Cui et al., 2024)。实际上，合成指令、响应以及由 AI 生成的反馈经常被整合到流水线中，使得模型既能提出数据也能评估数据。这一过程模糊了增强与对齐之间的界限。

**领域自适应与检索基础系统。** 在特定领域应用中，合成语料库支持任务适配和检索增强生成。对于生物医学等专业领域的问答系统，从业者通常从领域文本或语料库中合成领域内的问题与答案。随后，他们通过过滤低质量生成内容来提升模型在领域分布变化下的鲁棒性 (Shakeri et al., 2020)。这些合成样本随后用于微调通用语言模型。这一过程简化了早期用于领域自适应的多阶段合成问答流水线 (Shakeri et al., 2020)。在基于检索的设置中，模型生成的查询-文档对或查询标识符对可用于训练特定领域的检索器。该方法通过硬负例挖掘和偏好学习，在更贴近实际部署场景的分布上提升了排序性能 (Wen et al., 2025)。由于合成查询可通过粒度和领域特定约束进行控制，因此它们为衡量检索忠实度和下游任务收益提供了可控的测试平台。

**合成评估集与标注基准。** 模型训练或适配完成后，合成文本和对话越来越被视为评估数据。从业者不再仅依赖静态的人工编写基准，而是构建合成测试集，在受控分布下探测特定的失效模式。这些模式包括组合推理、长上下文一致性以及安全约束。

大语言模型作为评判流水线，能够使用细粒度标准 (如帮助性、安全性、忠实性) 对大量模型输出进行评分。该方法支持快速构建合成评估集，以及公开排行榜和元评估基准，例如 MT-Bench 和 Chatbot Arena (Gu et al., 2025; Zheng et al., 2023)。自验证方法 (如 Chain-of-Verification) 会显式生成验证问题和独立检查，以减少幻觉 (Dhuliawala et al., 2024)。同时，基于采样的方法 (如 SelfCheckGPT) 会生成替代性输出，其一致性可作为黑盒子信号来记录幻觉和鲁棒性情况 (Manakul et al., 2023)。



**安全测试与隐私保护的代理模型。**最后，合成语料库在风险管控中发挥着日益重要的作用。隐私研究利用大模型生成的敏感文本变体作为替代品，在保留统计效用的同时降低个人可辨认信息的直接暴露风险。记忆化和提取攻击表明，模型可能会复现罕见的训练样本，包括个人可辨认信息。这促使人们需谨慎地去重并采用注重隐私的数据处理方式 (Carlini et al., 2021)。SRD (Zhang et al., 2026) 框架利用大模型生成有毒和良性的文本，并使用这种对比数据集对模型进行去毒处理。在出处溯源与责任归属方面，组织越来越多地将水印或内容凭证嵌入合成语料库，以支持下游检测与策略执行。带密钥的统计水印能够可靠地实现对模型生成文本的统计检测，通常在鲁棒性与安全性考量下进行评估 (Kirchenbauer et al., 2023; Coalition for Content Provenance and Authenticity, 2025)。通过将指纹注入建模为针对指纹文本对的知识编辑问题，RFEEdit 实现了高效且鲁棒的注入，同时对无关知识的影响极小 (Li et al., 2025c)。合成安全与红队测试语料库同样用于对模型及安全过滤器进行压力测试。这些语料库包含由大模型生成或与大模型共同生成的对抗性提示、有毒延续内容或越狱对话。

总体而言，大语言模型生成的文本和对话可作为灵活的构建模块，用于语言技术的训练、适应、评估与治理。当合成语料库与目标任务和数据分布相匹配时，其优势最为显著。此外，这些语料库应配备强有力的筛选与验证机制，并具备可追溯性和隐私保护措施，以使它们的质量和可信度属性可量化衡量。

### 3 符号与逻辑数据

符号与逻辑数据生成随着大型推理模型的发展受到了广泛关注。这些模型强调多步推理和结构化的中间推理过程，通常通过思维链提示实现 (Wei et al., 2023; Kojima et al., 2023; Wang et al., 2023a; Liu et al., 2025b; Morishita et al., 2024; Toshniwal et al., 2024)。与一般的数据增强不同，该领域专注于生成内在可验证的实例。每个示例通常包含一个定义明确的问题、结构化的推理迹或中间产物，以及明确的验证信号。这些信号包括确切的关键词 (Li et al., 2025d)，或在数学应用题中答案匹配的情况，如 GSM8K 数据集中的任务 (Cobbe et al., 2021)。它们还包括代码生成中的编译结果和单元测试，或在符号规则下的逻辑有效性 (Tafjord et al., 2021; Morishita et al., 2024; Liu et al., 2025b)。

在实际应用中，基于大模型的流水线不仅生成多样化的题目与解答对，还提出中间步骤，批判并修改候选解，并利用自一致性 (Wang et al., 2023a) 或外部工具 (Ahmad et al., 2025) 进行输出筛选。这些流水线有助于将推理监督扩展至完全由人类编写的语料库之外。它们为蒸馏和后训练生成大量已验证或部分验证的轨迹。近期采用此方法的实例包括通过强化学习训练的推理模型，或通过工具和验证器增强的模型 (OpenAI et al., 2024; DeepSeek-AI, 2025)。

#### 3.1 生成方法

在本节中，我们介绍如何使用大模型生成符号和逻辑数据。

**启发式演化方法在扩展推理任务中的应用** 由大模型驱动的推理数据生成的一个主要类别依赖于种子任务的迭代演化。在此方法中，强模型反复重写问题，以增加结构多样性和难度，同时尝试保持解的有效性。这种模式源自 Evol-Instruct (Xu et al., 2024a) 中使用的重写技术，并已被应用于数学和编程等领域。例如，WizardMath 结合了针对数学领域的演化流水线与来自演化反馈的强化学习，以增强逐步推理能力 (Luo et al., 2025a)。类似地，MetaMath 通过生成并多样化以问题、答案和推理过程形式的数据，来扩大数学语料库规模 (Yu et al., 2024a)。

在编程领域, WizardCoder 通过生成从简单种子 (Luo et al., 2025b) 逐步变得更复杂的指令和响应, 将指令演化应用于编码任务。在这些研究中, 核心思想保持一致。该过程从现有的基准或种子指令开始, 然后应用受控的变异, 例如改写、添加约束或使问题更具组合性。最后, 该过程使用轻量级启发式方法来保持语义准确性。这些启发式方法通常涉及检查答案的一致性、验证基本格式或筛选一致性。

**基于求解器、执行器和证明器的工具验证生成方法。** 第二种方法将大模型的合成与程序化验证相结合。该方法使用执行器、编译器、单元测试或符号检查器作为可靠的判别器。在数学推理中, OpenMathInstruct-1 通过以代码解释器风格生成解法, 合成大规模的问题与解法对。它保留那些计算可执行且一致的实例 (Toshniwal et al., 2024)。

在编程领域, OpenCodeInstruct 汇集了包含执行反馈和单元测试信号的大规模指令微调语料库。这使得接受标准超越了表面合理性 (Ahmad et al., 2025)。对于形式逻辑推理, 诸如 ProofWriter 和 ALT-FLD $\times$ 2 的数据集与生成器构建了其有效性基于形式规则的实例。这使得在符号约束下能够对蕴含关系或证明步骤进行确定性检查 (Tafjord et al., 2021; Morishita et al., 2024)。SynLogic 进一步通过生成多种逻辑任务推进了可扩展且可验证的逻辑合成, 这些任务的正确性可通过简单的基于规则的检查器进行验证 (Liu et al., 2025b)。在所有这些情况下, 模型提出候选方案, 而外部工具则充当守门人, 筛选出正确性和逻辑结构。

**通过奖励滚动和蒸馏进行轨迹收获。** 第三类方法通过从强大的教师模型或通过强化学习训练的推理模型中收集轨迹来生成推理数据。这些模型与允许验证的任务进行交互。在此背景下, 轨迹不仅仅是解释, 更是策略在奖励和验证信号引导下探索解空间的结果。随后, 这些轨迹被提炼为学生模型。

最近的推理模型流水线通过收集大量部分验证的轨迹来遵循这一方法。这些轨迹使用自动化验证器和学成的评判者进行过滤, 然后用作下游后训练的信号。类似地, 公开发布的数据集如 SYNTHETIC-2 提供了大规模的数据集, 这些数据是推理智能体与验证器在多种任务族中协作的结果 (Prime Intellect Team, 2025)。该方法生成了适合蒸馏的已验证推理迹。这种方法的关键差异在于数据的性质: 这些数据是搜索和最优化过程中生成的策略迹, 而非静态且 one-shot 的合成问答对。

**基于大模型作为裁判的偏好优化生成** 最后, 在程序化检查不完整或不可用的场景中, 许多流水线依赖于使用大模型作为评判机制进行偏好筛选和监督。UltraFeedback 通过采样多个候选响应并利用更优的模型对其进行评分, 构建大规模的偏好信号。这些得分基于诸如帮助性、真实性以及指令遵循性等维度 (Cui et al., 2024)。

CodeUltraFeedback 在编程领域应用了类似的概念。在此背景下, 评判者根据代码的可读性、是否遵循指令等编码偏好来评估候选解决方案 (Weyssow et al., 2024)。相关的自训练方法, 如 STaR, 通过生成与筛选环路创建推理过程, 其中正确性作为通用接受信号 (Zelikman et al., 2022)。这些方法共同强调模型驱动的评估。推理数据以大规模生成, 并由评判模型组织成排序或偏好标注的语料库。这一过程支持对齐、偏好最优化和自我改进。

### 3.2 符号与推理数据的质量度量

评估大模型生成推理数据质量的基本准则在于客观正确性。这要求中间推理步骤和最终结论均为逻辑上有效且可经实证验证的结果。由于推理数据涵盖数学、符号、程序及开放领域等不同情境, 各领域的具

体验证协议不尽相同。然而，这些协议具有统一的目标：验证生成的推理迹严格遵循真实值逻辑或可执行的真实条件。

**有效性。** 对于数学和基于代码的推理，正确性通常可以通过领域验证器进行算法化验证。令  $\mathcal{R} = \{(q_i, c_i, y_i)\}_{i=1}^N$  表示推理样本，其中  $q_i$  为问题， $c_i$  为生成的推理过程（如思维链或代码解释器迹）， $y_i$  为最终答案。当存在参考答案  $y_i^*$  或可执行规范时，我们定义一个特定领域的检查器  $f_{\text{check}}$  如下：

$$f_{\text{check}}(q_i, c_i, y_i) = \begin{cases} 1, & \text{if the candidate is verified as correct for } q_i; \\ 0, & \text{otherwise.} \end{cases}$$

随后，我们报告总体验证准确率作为

$$\text{Acc}_{\text{verify}} = \frac{1}{N} \sum_{i=1}^N f_{\text{check}}(q_i, c_i, y_i).$$

在 MetaMath 中，答案增强阶段使用了拒绝采样。在此过程中，生成多种推理路径，仅保留那些产生正确答案的路径。该方法通过答案级别的验证 (Yu et al., 2024a) 确保有效性。

在 OpenMathInstruct-1 中，解法以代码解释器风格的格式表示。通过仅保留能够得出真实值答案的解法来强制保证正确性。该论文还使用了以“k 次采样通过率”衡量的训练集覆盖度。该指标用于判断 k 次采样得到的解法中是否有任何一个能达到真实值 (Toshniwal et al., 2024)。

对于程序合成与代码推理，有效性通常通过单元测试执行来衡量。令  $\mathcal{T}_i$  为示例  $i$  的测试用例集。候选解的单元测试通过率定义如下：

$$\text{PassRate}_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbb{I}(\text{exec}(y_i, t) = \text{pass}),$$

数据集级别的通过率计算为  $\text{PassRate} = \frac{1}{N} \sum_{i=1}^N \text{PassRate}_i$ 。这种方法与 OpenCodeInstruct 一致，后者在生成的单元测试上执行解决方案，并将通过率作为元数据记录 (Ahmad et al., 2025)。

对于需要显式证明的逻辑和演绎推理，如证明图，有效性要求正确的蕴含预测和正确的证明。遵循 ProofWriter，使用一种称为完整准确率（Full Accuracy）的严格度量来评估证明的正确性。在此度量中，预测的证明图必须与标准答案完全匹配。如果图不匹配，证明得分为零 (Tafjord et al., 2021)。

因此，设  $y_i^*$  为黄金蕴含标签或答案标签， $c_i^*$  为黄金证明表示。严格证明准确率可总结为

$$\text{Acc}_{\text{proof}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = y_i^*) \cdot \mathbb{I}(c_i = c_i^*).$$

最后，针对鲁棒性风格的有效性检查，FAIRR (Sanyal et al., 2022) 定义了扰动输入的等价集合上的一致性。对于一个理论和声明对  $(T, s)$  及其等价集合  $E(T, s) = \{(T'_k, s'_k)\}_{k=1}^K$ ，一致性由以下公式定义：

$$C(T, s) = \frac{1}{K} \sum_{k=1}^K \mathbb{I}(f(T, s) = f(T'_k, s'_k)),$$

该值在数据集上取平均。FAIRR 报告蕴含准确率、严格证明准确率和一致性 (Sanyal et al., 2022)。

**忠诚** 保真度衡量代理评估信号与实际质量标准之间的对齐程度。与评估解决方案相对于真实值的目标正确性的有效性不同，保真度评估用于判断数据的工具或模型的可靠性。该指标确保在无法进行直接验证的情况下，自动化评估能够反映真实的逻辑推理或一致性。

遵循自一致性解码范式，会采样多个推理路径，并选择最一致的答案 (Wang et al., 2023a)。对于每个问题  $q_i$ ，我们采样  $K$  条推理链  $\{c_{i,k}\}_{k=1}^K$ ，并提取它们的最终答案  $\{y_{i,k}\}_{k=1}^K$ 。我们定义多数投票答案如下：

$$\hat{y}_i = \arg \max_y \sum_{k=1}^K \mathbb{I}(y_{i,k} = y).$$

为了衡量共识代理与参考标签之间的对齐程度，我们计算自一致性准确率：

$$\text{Acc}_{\text{SC}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i).$$

为了进一步量化相同问题的采样答案之间的内部一致性，我们采用平均成对答案一致率：

$$\text{Agree}(q_i) = \frac{1}{K(K-1)} \sum_{k \neq k'} \mathbb{I}(y_{i,k} = y_{i,k'}).$$

在开放式的自然语言推理任务中，由于通常无法进行可执行的验证，基于大模型 (LLMs) 的评判者提供了自动评分信号。例如，GPT-4 常被用于提供此类评估。为了评估这些评判信号在多大程度上反映了人类判断，令  $u_i^{(\text{LLM})}$  和  $u_i^{(\text{human})}$  分别表示第  $i$  个样本中模型评判者和人工标注者的标量得分。

我们使用 Spearman rho 相关系数或 Pearson 相关系数来衡量这些得分之间的秩关联。这遵循了模型评判中对齐评估的既定方法 (Lai et al., 2025)：

$$\rho_{\text{LLM-human}} = \text{corr}_{\text{Spearman}}(u^{(\text{LLM})}, u^{(\text{human})}).$$

当裁判输出  $M$  诸如连贯性、忠实度和事实性等评分维度时，我们可以将它们汇总为一个总体的代理质量得分：

$$Q_{\text{reason}} = \frac{1}{M} \sum_{m=1}^M \text{score}_{i,m}.$$



模型评判信号的可靠性通常在诸如 JudgeLM、MT Bench 或 Chatbot Arena 等基准上进行验证。这些基准报告了模型评判者与人类偏好之间的一致性 (Zhu et al., 2025; Zheng et al., 2023)。

基于人工智能反馈的偏好数据集，如 UltraFeedback (Cui et al., 2024)，以及使用强化学习从人工智能反馈中进行训练的流水线，为偏好建模和下游对齐提供了训练信号。这些流水线用由大模型 (Lee et al., 2024) 生成的偏好替代了昂贵的人工标签。对于用偏好标注的样本对，可以通过奖励模型  $R_\phi$  是否将更受青睐的输出排在更高秩来评估其表现。

我们考虑一对  $(c_{i,a}, c_{i,b})$ ，其二元偏好标签  $z_i$  为零或一。例如，当  $c_{i,a}$  被偏好时， $z_i$  等于一。我们定义模型所隐含的标签如下：

$$\hat{z}_i = \mathbb{I}(R_\phi(c_{i,a}) > R_\phi(c_{i,b})).$$

随后我们计算成对偏好准确率：

$$\text{Acc}_{\text{RM}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{z}_i = z_i).$$

该指标评估奖励模型在用于训练和人类反馈强化学习基准测试的成对型偏好数据上 (Bai et al., 2022; Frick et al., 2024)。

### 3.3 可信的符号与推理数据度量指标

即使生成的推理迹能得到正确的最终答案，中间步骤的逻辑连贯性和忠实性也至关重要。一个模型在缺乏有效推理的情况下得出正确答案，会引入虚假模式和不可靠的监督。

**忠实。** 设  $c_i = [c_{i,1}, \dots, c_{i,T_i}]$  为示例  $i$  的思维链。一个步骤验证器  $f_{\text{step}}$  会检查每一步。该验证器在形式领域中可以是符号执行器，或在自然语言领域中可以是基于推理的蕴含检查器。它按如下方式评估每一步：

$$f_{\text{exec}}(c_{i,t}) = \begin{cases} 1, & \text{if } c_{i,t} \text{ is a valid transformation,} \\ 0, & \text{otherwise.} \end{cases}$$

步骤有效率由以下公式定义：

$$\text{Val}_{\text{step}} = \frac{1}{\sum_i T_i} \sum_i \sum_{t=1}^{T_i} f_{\text{exec}}(c_{i,t}).$$

该指标衡量中间步骤遵守数学、逻辑或语义规则的频率。在推理评估框架（如 ReCEval）中，已探索了步骤级别的正确性度量。该框架通过文本蕴含和信息性来评分步骤和推理链的质量 (Prasad et al., 2023)。此外，因果忠实性分析（如 FRODO）测试思维链的内容是否对最终答案具有因果影响，而不仅仅是与之相关 (Paul et al., 2024)。

除了可执行步骤验证外，蕴含模型还可以在自然语言中为步骤支持提供代理。对于每一对相邻步骤  $c_{i,t-1}$  和  $c_{i,t}$ ，一个蕴含得分器  $g_{\text{entail}}$  提供如下输出：

$$p_{i,t} = g_{\text{entail}}(c_{i,t-1} \Rightarrow c_{i,t}),$$

我们计算平均蕴含对齐以表示链内的支持度：

$$\text{Align}_{\text{entail}} = \frac{1}{\sum_i (T_i - 1)} \sum_i \sum_{t=2}^{T_i} p_{i,t}.$$

较高的数值表明推理过程是基于语义和逻辑的逐步推进，而非突然或矛盾的跳跃。使用蕴含关系的评估方法已被用于判断推理链的正确性和信息量 (Prasad et al., 2023)。这些方法还被应用于根据文本蕴含关系分配部分分数 (Yao & Barbosa, 2024)。

**鲁棒性** 为了研究分布偏移下的泛化能力，我们比较了奖励模型在领域内评估和领域外评估中的成对准确率。该差异定义如下：

$$\Delta_{\text{OOD}} = \text{Acc}_{\text{RM}}^{\text{in}} - \text{Acc}_{\text{RM}}^{\text{out}}.$$

最近的评估报告通过在偏好数据（如 UltraFeedback）上测量领域内准确率来进行比较。它们还在基准测试（如 RewardBench）上测量领域外准确率 (Mahan et al., 2024; Lambert et al., 2024)。差异越小，表示在不同领域间的偏好排序越稳定。

### 3.4 评估实践差距

我们分析了第 3.1 节中的代表性方法，并在表 3 中对它们报告的评估协议进行了分类。

Table 3 突显了符号与逻辑数据生成评估领域中的显著差异。具体而言，衡量过程质量的维度远不如衡量简单答案级正确性的维度得到充分覆盖。

Generation method	Validity	Fidelity	Robustness	Faithfulness
MetaMath (Yu et al., 2024a)	✓	×	×	×
OpenMathInstruct-1 (Toshniwal et al., 2024)	✓	✓	×	×
OpenCodeInstruct (Ahmad et al., 2025)	✓	△	×	×
UltraFeedback (Cui et al., 2024)	×	✓	×	×
ProofWriter (Tafjord et al., 2021)	✓	×	△	✓
FaiRR (Sanyal et al., 2022)	✓	×	△	△

**Table 3** 是否在实验部分显式评估了代表性符号与逻辑数据生成方法中的每个维度。✓：显式评估；△：部分/间接覆盖；×：未报告或不适用。

**忠实性。**忠实性的显式评估主要局限于包含结构化证明标注的基准。在这些基准中，可以直接衡量证明的严格正确性。例如，ProofWriter 基于确切的证明图匹配报告完整的证明准确率 (Tafjord et al., 2021)。类似地，FaiRR 将严格的证明准确率作为其协议的核心部分 (Sanyal et al., 2022)。相比之下，许多依赖答案验证或基于执行结果进行过滤的生成流水线更注重最终答案的正确性。典型的例子包括 MetaMath

(Yu et al., 2024a)、OpenMathInstruct 1 (Toshniwal et al., 2024) 和 OpenCodeInstruct (Ahmad et al., 2025)。因此，中间推理步骤的有效性以及这些步骤是否真正支持结论的问题往往未得到充分评估。

**鲁棒性。**鲁棒性在较少的研究中被报告。当研究人员关注这一特性时，其表现形式各不相同，且难以直接比较。例如，ProofWriter 评估了规则集之间的域外迁移情况 (Tafjord et al., 2021)。在另一个案例中，FaiRR 基于扰动报告了一致性，同时结合蕴含关系和准确率 (Sanyal et al., 2022)。除了这些具体例子之外，大多数与生成相关的研究主要集中在同一领域的性能上。这些研究很少将数据分布的变化或扰动下的稳定性作为主要评估目标。这种缺乏标准化的情况表明，有必要为符号化和逻辑生成定义特定的鲁棒性度量。研究人员应避免仅依赖鲁棒性的通用代理指标。

这些差距共同表明，当目标是生成具有可靠中间结构的数据时，仅关注正确性的评估可能会掩盖推理真正质量。因此，在符号和逻辑推理领域，将忠实性与鲁棒性测量纳入标准有效性指标至关重要。在此领域，期望的行为不仅是得出正确答案，更要通过稳定且可解释的推理过程实现。

### 3.5 使用

**使用合成语料库进行预训练和持续预训练。**合成推理数据的一个主要应用是将推理轨迹直接整合到基础模型阶段。在此范式中，合成语料库在预训练或持续预训练阶段被引入，使模型在指令微调之前就具备结构化和多步推理能力 (Yang et al., 2024a; Ying et al., 2024)。该方法将推理视为一项基本技能。在此背景下，符号、算术和过程性推理模式与自然语言一同被习得。这一过程有助于模型的内部表示向组合式推理方向发展。

近期在大规模推理模型（如 DeepSeek R1 和 OpenAI o1）方面的进展证实了预训练后最优化的有效性。这些框架基于预训练的基础模型，利用大规模强化学习来生成成长时程且分步的问题求解行为。这一过程使模型专注于推理能力，而非依赖表面的后期微调 (DeepSeek-AI, 2025; OpenAI et al., 2024)。具体而言，DeepSeek R1 使用基于规则和可验证结果的奖励机制，为数学和编程等领域的任务提供精确反馈。这使得通过自动正确性信号实现可扩展的改进成为可能 (DeepSeek-AI, 2025)。类似地，OpenAI o1 强调大规模强化学习，以提升细致的推理能力和思维链方法的有效使用。该方法在数学、编程和科学等多个基准测试中均表现出色 (OpenAI et al., 2024)。

与此同时，诸如 Qwen Math 和 InternLM Math 等专注于数学的模型采用了一种源自强大基础模型的持续预训练策略。这些方法首先引入面向数学的语料库进行预训练，随后进行下游后训练。这通常紧接着下游后训练，例如使用思维链数据进行监督微调，以及基于工具增强或验证器引导的监督，还包括基于奖励模型的多步解法重排序或过滤。综上所述，这种持续预训练-后训练流水线通过特化显著提升了数学推理能力 (Yang et al., 2024a; Ying et al., 2024)。此外，相关数据流水线能够大规模合成新问题及其解答，并通过正确性筛选保留最终答案与真实值匹配的解法。这一方法在 MetaMathQA 和 OpenMathInstruct-1 中得到了验证。这些系统利用强大的大模型生成候选解法，同时通过执行验证来确保可靠性 (Yu et al., 2024a; Toshniwal et al., 2024)。早期流水线常使用闭源模型，而后续工作则越来越多地采用日益强大的开源模型。总体而言，这些努力将合成推理语料视为训练分布的主要组成部分。通过让潜在的语言模型接触逐步的数学与逻辑推导过程，这些方法在后续特化阶段提高了数据效率。

**在经过验证的推理迹上的监督微调。**大模型生成的推理数据的一种独特应用是通过监督微调显式提供思维链能力。该过程使用经过验证的推理三元组，包括输入、推理迹和最终答案。与主要塑造归纳偏置

的预训练不同，监督微调将模型输出与具体的多步示例对齐，从而促使模型内化中间问题求解过程。

在数学推理领域，合成数据集通过显式的验证信号对开放模型进行微调，尽管具体的机制各不相同。MetaMathQA 通过自举式问题生成来扩展数据生产，并使用拒绝采样根据最终答案的正确性筛选推理迹 (Yu et al., 2024a)。OpenMathInstruct 1 强调基于执行的验证，允许解法结合自然语言推理与可执行代码。该方法利用解释器的执行作为验证信号 (Toshniwal et al., 2024)。类似地，Gretel 发布的合成 GSM8K 数据集包含结构化反思风格的实例，并配以自动化验证。其他变体则引入了严格的自动化检查，包括模型裁判评估以及通过 sympy 等工具进行的符号验证，以筛选和优化候选解 (AI, 2024)。WizardMath 超越传统的监督微调，通过合成复杂指令并应用从演化指令反馈中获得的强化学习，以增强推理行为，超越简单的模仿 (Luo et al., 2025a)。

在编程领域，CodeAlpaca 和 Magicoder 等语料库为代码生成提供了指令与解决方案的配对。Magicoder 进一步将合成过程锚定在开源代码片段上，以更好地反映真实的开发场景 (Chaudhary, 2023; Wei et al., 2024)。OpenCodeInstruct 通过引入明确的测试套件、执行结果以及基于模型的质量评估，改进了这一范式。这为训练期间使用合成程序进行数据监管提供了可验证的信号 (Ahmad et al., 2025)。

在逻辑与符号推理方面，诸如 ALT 之类的框架构建了具有原则性的合成逻辑语料库。这些语料库由程序生成的多步演绎实例组成，并以形式逻辑为基础。此类数据有助于补充训练，以改进蕴含与推理模式。与此同时，SynLogic 将多样化的逻辑任务与基于规则的验证器相结合，以实现可扩展的训练并提供可验证的反馈 (Morishita et al., 2024; Liu et al., 2025b)。

总而言之，对经过验证的推理迹进行监督微调，提供了一种可控的方法来使用合成数据。生成流水线提出候选问题和解释，而外部验证器则提供正确性信号。由此产生的有效实例随后被用于训练能够逐步解释和证明其推理过程的模型。

**基于偏好建模与强化学习的最优化。** 合成推理数据的一个重要应用在于提供评估性反馈，而非直接监督。该过程通过偏好建模和强化学习来促进稳健的推理行为。在此背景下，由大模型生成的推理迹或经验证器评估的推理迹作为奖励模型或策略最优化目标的输入。这些目标明确优先考虑多步推理中的逻辑一致性、正确性和清晰性。

例如，UltraFeedback 通过从多个模型中采样候选回复，并利用更优的大模型判官根据特定标准对它们进行排序，构建多个领域的偏好数据集。这些标准包括指令遵循度和真实性。该过程生成大规模的 AI 反馈信号，适用于奖励模型训练与优化，其风格类似于人类反馈的强化学习 (Cui et al., 2024)。诸如来自 AI 反馈的强化学习等方法进一步扩展了这一方法，将人类偏好标签替换为由大模型判断的比较结果。这显著提升了跨任务和领域偏好数据采集的可扩展性 (Lee et al., 2024)。

在推理领域，现代系统越来越多地将强化学习与可程序化验证的反馈相结合，用于数学、编程和逻辑问题。具体而言，DeepSeek R1 在训练过程中使用基于规则的奖励和可执行反馈来评估推理迹的得分。此类反馈包括编译和执行代码等动作，从而实现能够强化可验证正确性的策略更新 (DeepSeek-AI, 2025)。OpenAI o1 采用大规模强化学习来优化多步推理行为，尽管其奖励机制和验证器的具体实现细节尚未公开 (OpenAI et al., 2024)。此外，SYNTHETIC-2 语料库提供了数百万条包含强化学习轨迹及其奖励信号的推理迹。这些信号促进了下游蒸馏及离线强化学习研究 (Prime Intellect Team, 2025)。因此，推理数据从静态训练样本演变为动态评估信号。结合合成迹与评判或验证器输出，定义出一个奖励空间，指导模型的迭代优化。



**知识蒸馏与隐式推理迁移。** 大模型生成的推理数据的一种互补应用，是将显式的推理过程压缩为隐式的内部表示。该过程从复杂的教师模型中提炼出详细的思维链迹，并将其压缩到紧凑的学生模型中。这些学生模型在推理时无需生成详尽的解释即可实现稳健的推理。在此范式中，推理迹作为潜在监督信号，引导表示学习，而不仅仅作为文本目标。

隐式思维链蒸馏的方法，如Deng et al. (2023) 所提出的，首先使用显式监督训练一个教师模型。随后，通过隐状态学习信号将其推理能力蒸馏到学生模型中。这使得能够创建出无需生成冗长迹的高效隐式推理模型。在符号和演绎情境中，ProofWriter 为自然语言证明提供了结构化监督。该方法提供过程级信号，支持训练并可能蒸馏出轻量级模型，这些模型能够以最少的显式推理依据遵循类似证明的推理模式 (Tafjord et al., 2021)。

更广泛地说，使用强化学习和轨迹语料库（如 DeepSeek R1 和 SYNTHETIC-2）训练的大规模推理模型提供了丰富的多步滚动展开，可作为蒸馏目标。这些资源使紧凑型模型能够从计算密集型教师模型中继承高层次的推理行为 (DeepSeek-AI, 2025; Prime Intellect Team, 2025)。因此，合成推理数据不仅促进显式的思维链生成，还支持隐式的推理迁移。目标是将过程层面的能力嵌入模型中，使其能够得出正确答案，而无需必然解释每一个中间步骤。

**评估、基准测试与泛化能力测试。** 最后，合成的、程序化结构化的推理语料库越来越多地用作评估工具。这些语料库提供了多样化且可自动验证的任务，用于测试超越人工精心设计的静态基准的泛化能力。其主要目标是构建动态基准，随着模型能力的发展而不断演进，同时保持严格且可验证的评分协议。

在数学领域，GSM-HARD 通过系统性地将 GSM8K 数据集中的数值替换为更大或更复杂的替代值，扩展了该数据集。这一过程生成了一个由程序创建的压力测试套件，用于评估算术鲁棒性。当与程序化执行求解器结合使用时，模型输出可实现大规模自动验证 (Gao et al., 2023b)。尽管并非由大模型生成，SciBench 汇集了大学水平的物理、化学和数学科学问题。这使得科学问题求解能力能够进行系统性评估和详细的错误分析 (Wang et al., 2024b)。

在因果与逻辑推理领域，CREPE 引入了一个基准，用于评估事件合理性与实体状态的因果推理能力。该基准利用人类判断来识别模型行为与人类推理之间的差异 (Zhang et al., 2023)。此外，SynLogic 为合成逻辑任务提供了保留的验证划分。它使用方差减少指标报告性能，例如八次平均值 (Liu et al., 2025b)。总体而言，这些评估框架展示了合成数据集和程序化结构化推理数据集的双重用途。它们不仅提供训练和强化学习信号，还支持基准测试以监测推理鲁棒性、分布偏移以及跨领域泛化能力，随着模型的发展而持续发挥作用。

## 4 表格数据

表格数据由按照固定模式组织的结构化记录构成，包含行和列，其特征为异构特征，包括数值型和类别型属性。如 Shi et al. (2025b) 所述，此类数据的生成式建模需要联合表示混合类型属性，涵盖数值域和类别域，并处理列之间的复杂依赖关系。同时，模型必须保持与原始数据分布一致的统计保真度。在此情景下，有效性定义为数据点作为遵循模式级别约束和内在函数依赖关系的行向量。若未能强制执行这些约束，生成的合成表格可能在统计上看似合理，但实际上违反了领域逻辑或结构规则 (Shi et al., 2025b; Xu et al., 2025)。在医疗保健等实际应用场景中，Barr et al. (2025) 展示了使用大模型进行 zero shot 提示在生成临床合理的围手术期数据集方面的有效性。他们通过与真实世界参考数据的统计比较

验证了该数据的保真度。更广泛地，该领域将合成表格数据生成视为一个优化问题，需在数据效用和保真度与隐私保护之间取得平衡。因此，评估协议围绕数据质量与隐私保护这两个维度展开 (Shi et al., 2025b)。

## 4.1 生成方法

**基于提示和上下文的表格合成** 一种显著的方法类别利用提示工程或上下文学习来合成表格记录，而无需进行完整的模型微调。这些方法通常优先考虑模式的有效性以及长尾分布的覆盖。在此范式下，CLLM 通过利用大模型先验，并引入基于学习动态的筛选流水线，使用置信度和不确定性度量来过滤合成行 (Seedat et al., 2024)。另一类方法则明确引导生成过程以聚焦少数或代表性不足的数据子集。例如，EPIC 和 LITO 采用类别条件或组条件提示策略，包括自我认证机制，以偏向稀有类别的采样。这不仅增强了模式合规性，还提升了长尾部分的代表性 (Kim et al., 2025; Yang et al., 2024b)。在互补方法中，TabGen-ICL 通过自适应选择上下文示例来提高生成的真实性。它迭代检索能够代表生成分布与真实分布之间残差的真实样本，从而优化固定基础模型的示例池 (Fang et al., 2025)。除了直接的行合成之外，Nam et al. (2024) 还引入决策树反馈，指导大模型生成新特征。这将提示方法扩展到了特征工程领域。尽管取得了这些进展，实证分析表明，无约束的逐行提示和任意的特征排序常常导致函数依赖关系的违反。此外，标准的单变量或相关系数度量往往无法检测到这些约束失效。因此，亟需开发严格遵循模式且关注约束的提示与评估策略 (Xu et al., 2025)。

**微调与专用表格生成** 一类独特的方法学涉及针对表格生成专门微调大模型。这些方法旨在更准确地建模结构约束和分布依赖关系，以提升样本的有效性和保真度。例如，GReaT 在序列化表格行上对自回归语言模型进行微调，以实现任意特征子集上的完全条件采样。该机制提升了样本的真实性，并减少了不合理的跨特征组合出现的频率 (Borisov et al., 2023)。在此架构基础上，Nguyen et al. (2024) 引入基于排列的训练与特征条件采样相结合的方法，以保留特征标签的相关性，进一步增强样本的真实性。REaLTabFormer 通过自回归生成父表，并将子表生成条件化于采样的键值，实现关系型表格的合成 (Solatorio & Dupriez, 2023)。为应对记忆风险并捕捉行间结构细微差异，HARMONIC 引入了  $k$ -最近邻基于的指导信号，强调记录之间的邻近关系 (Wang et al., 2024d)。其他框架则集成显式的自我修正机制。Table-LLM-Specialist 提出一种迭代生成器-验证器自训练范式，其中候选监督信号由模型生成，并通过表格特异性的一致性信号（如排列不变性和执行不变性）进行验证。该方法实现了无需依赖人工标注的专家微调 (Xing et al., 2024)。此外，自适应方法如 TableDreamer 通过逐步合成暴露模型弱点的实例，针对观察到的失败模式进行优化，从而提高数据效率和下游实用性 (Zheng et al., 2025)。总体而言，这些方法直接回应了近期实证分析中识别出的挑战，特别是约束和功能依赖关系的违反问题，通过在训练和采样阶段均引入结构感知能力加以解决 (Xu et al., 2025)。

**混合架构与结构化合成** 混合架构将大模型与外部结构或分布感知组件相结合。或者，它们将表格生成重新表述为结构化问答系统，以联合解决有效性、保真度和实用性方面的差距。在这些设计中，大模型通常作为模式推理、约束处理和一致性检查的引擎，而辅助模块则确保结构控制和统计对齐。例如，AIGT 利用表格元数据和长 token 划分来促进宽表生成，同时在大规模下保持质量 (Zhang et al., 2024c)。关于文本到表格的转换，gTBLs 将单元填充建模为条件问答任务。该方法明确针对语法有效性，以减少生成过程后对大量修正的依赖 (Sundar et al., 2024)。受关于约束违规和分布不匹配发现的启发 (Xu et al., 2025)，这些系统显式地结合基于大模型的推理与外部组件。这种结合优化了端到端表

格的真实性及其下游适用性 (Zhang et al., 2024c; Sundar et al., 2024)。

## 4.2 表格数据的质量度量

**有效性。** 我们从边际和结构两个角度评估有效性。

对于类别型边缘分布，我们额外报告列方向的卡方检验作为合理性检查准则。该指标在 TabSyn 中用作质量指示器，通常需要通过较高的阈值，例如  $p \geq 0.95$  (Zhang et al., 2024a)。其计算定义如下：

$$\chi^2 = \sum_{c \in \Omega} \frac{(O_c - E_c)^2}{E_c}, \quad p = \Pr(\chi_{df}^2 \geq \chi^2),$$

where  $O_c$  and  $E_c$  represent the observed synthetic counts and the expected real counts for category  $c$  within the set  $\Omega$ . A larger  $p$  value indicates closer marginal agreement, which signifies that the test fails to reject the hypothesis that the distributions are similar.

为了衡量结构有效性，我们报告违规率，记作 VR。该比率定义为失败的完整性检查所占的比例：

$$VR = \frac{\#violations}{\#checks}.$$

The checks include functional dependencies, range limits, uniqueness, geographic consistency, and other schema constraints. A lower violation rate indicates higher validity.

**忠诚** 保真度衡量合成数据分布与真实数据分布的接近程度。遵循 TabSyn 中使用的低阶统计量协议，我们在边缘、成对型以及全局或样本层面评估保真度 (Zhang et al., 2024a)。

边际保真度通过数值列的科尔莫戈罗夫-斯米尔诺夫检验统计量和类别列的总变差距离 (Zhang et al., 2024a) 进行评估。公式定义如下：

$$KST = \sup_x |F_{\text{real}}(x) - F_{\text{syn}}(x)|,$$

$$TVD = \frac{1}{2} \sum_{\omega \in \Omega} |R(\omega) - S(\omega)|,$$

where  $R$  and  $S$  denote the real and synthetic empirical frequencies of a category  $\omega$  within the set  $\Omega$ .

成对型保真度通过数值型变量对的皮尔逊得分以及类别型变量对的列联表得分（也称为 ContSim）来衡量。这些指标遵循 TabSyn 附录 E.3 (Zhang et al., 2024a) 中描述的方法。对于由变量  $x$  和  $y$  组成的数值型变量对，皮尔逊相关系数的计算如下：

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y},$$

The Pearson Score aggregates correlation gaps using a normalization factor of one half because the correlation coefficient  $\rho$  ranges from negative one to one (Zhang et al., 2024a):

$$\text{PearsonScore} = \frac{1}{2} \cdot \frac{1}{K} \sum_{(i,j)} |\rho_{\text{real}}^{(i,j)} - \rho_{\text{syn}}^{(i,j)}|.$$

For categorical pairs consisting of variables  $A$  and  $B$ , the contingency table discrepancy is defined by the following equation:

$$\text{ContingencyScore} = \frac{1}{2} \sum_{\alpha \in A} \sum_{\beta \in B} |R_{\alpha,\beta} - S_{\alpha,\beta}|.$$

For mixed type pairs involving both numerical and categorical data, TabSyn buckets numerical values into categorical bins before computing the corresponding contingency score (Zhang et al., 2024a).

全局真实性和可检测性通过应用分类器两样本检验进行评估，其思路与 TabSyn (Lopez-Paz & Oquab, 2018) 相同。该方法利用基于逻辑回归 (Zhang et al., 2024a; DataCebo, 2024) 的 SDMetrics 检测得分。在此背景下，我们将曲线下面积定义为判别器在区分真实数据与合成数据时，在交叉验证划分上的受试者工作特征曲线下平均面积。SDMetrics 将检测得分定义如下 DataCebo (2024):

$$\text{DetScore} = 1 - (\max(\text{AUC}, 0.5) \times 2 - 1).$$

This formula maps an area under the curve of zero point five, which indicates that the data is indistinguishable, to a detection score of one. Similarly, an area under the curve of one, which indicates that the data is perfectly distinguishable, is mapped to a detection score of zero. Therefore, a higher detection score suggests higher fidelity from the perspective of global detectability (DataCebo, 2024; Zhang et al., 2024a).

样本级别的保真度和  $\alpha$ -准确率采用基于支持集的 Alaa et al. (2022) 定义，该定义也被 TabSyn 框架 (Zhang et al., 2024a) 所采用。在本次评估中，令  $P_r$  和  $P_g$  分别表示真实分布和生成分布，其中  $S_r$  为真实分布的支持集， $S_g$  为生成分布的支持集。参照 Alaa et al. (2022) 的工作，真实分布的  $\alpha$ -支持集被定义为  $S_r$  中包含概率质量等于  $\alpha$  的最小体积子集：

$$S_r^\alpha \triangleq \arg \min_{S \subseteq S_r} \text{Vol}(S) \quad \text{s.t.} \quad P_r(S) = \alpha,$$

In this context, the volume function represents the Lebesgue volume measure. The corresponding  $\alpha$ -precision is defined as the probability that a synthetic sample lies within the real  $\alpha$ -support (Alaa et al., 2022):

$$P_\alpha \triangleq \Pr_{x \sim P_g} (x \in S_r^\alpha).$$

In practice, we estimate the alpha support from finite samples and compute the empirical  $\alpha$ -precision by averaging binary membership indicators over the set of synthetic samples (Alaa et al., 2022):

$$\widehat{\alpha\text{-Prec}} = \frac{1}{|\hat{X}|} \sum_{\hat{x} \in \hat{X}} \mathbf{1}[\hat{x} \in \hat{S}_r^\alpha],$$

where  $\hat{X}$  represents the set of synthetic samples and  $\hat{S}_r^\alpha$  is an estimated  $\alpha$ -support of the real distribution.

**多样性** 生成数据的多样性被评估以衡量对真实分布的覆盖程度。我们采用  $\beta$ -召回率，如 Alaa et al. (2022) 所提出的，该指标也在 TabSyn 框架 (Zhang et al., 2024a) 中被使用。合成分布的  $\beta$ -支持度以类似方式定义如下：

$$S_{\text{syn}}^\beta = \arg \min_S |S| \quad \text{subject to} \quad \Pr(X_{\text{syn}} \in S) = \beta.$$



Based on this definition,  $\beta$ -Recall is expressed by the following equation [Alaa et al. \(2022\)](#):

$$R_\beta = \Pr_{X_{\text{real}} \sim P_{\text{real}}} (X_{\text{real}} \in S_{\text{syn}}^\beta).$$

The sample-level estimator for this metric follows the same approach as the one used for  $\alpha$ -Precision:

$$\widehat{\beta\text{-Rec}} = \frac{1}{|X|} \sum_{x \in X} \mathbf{1}[x \in \hat{S}_{\text{syn}}^\beta].$$

A high  $\beta$ -Recall value indicates that the synthetic data provides broad coverage of the original distribution and serves as a complement to the Alpha-Precision metric [Alaa et al. \(2022\)](#); [Zhang et al. \(2024a\)](#).

**实用性。** 我们使用“在合成数据上训练，在真实数据上测试”的下游效用评估协议进行评估，该协议在 TabSyn 框架中也称为机器学习效率 [Zhang et al. \(2024a\)](#)。遵循 TabSyn 的评估方法，我们报告分类任务的曲线下面积（AUC）和回归任务的均方根误差（RMSE）（[Zhang et al., 2024a](#)）:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) d\text{FPR}, \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2, \quad \text{RMSE} = \sqrt{\text{MSE}}.$$

### 4.3 表格数据的可信度量

**隐私** 我们从三个互补的视角量化隐私风险，包括通过最近邻接近度的几何记忆、通过成员身份或属性推断的实证推理泄露，以及差分隐私等形式化的隐私保障。

为了评估几何记忆和潜在的记录复制，我们分析了与最近记录的距离。参照基于大模型的表格合成的先验工作，对于每个合成记录  $s \in \hat{D}$ ，我们计算其与最近训练记录（[Borisov et al., 2023](#); [Fang et al., 2024](#)）的距离：

$$\text{DCR}_{\text{train} \rightarrow \text{gen}}(s) = \min_{x \in D_{\text{train}}} d(s, x),$$

where systematically low values indicate a warning signal of potential copying. Because absolute values for the distance to the closest record depend on the choice of distance  $d$  and feature scaling, we recommend comparing the distribution of these values against a real baseline. This baseline is defined as  $\text{DCR}_{\text{train} \rightarrow \text{test}}(x) = \min_{x' \in D_{\text{train}}} d(x, x')$  for hold out real records, which is a common practice in tabular synthesis evaluations ([Borisov et al., 2023](#)). Conversely, to assess whether the generator covers the support of the real distribution instead of collapsing to a few modes, we compute the real to synthetic proximity on hold out data:

$$\text{DCR}_{\text{real} \rightarrow \text{syn}}(x) = \min_{s \in \hat{D}} d(x, s),$$

where a distribution comparable to real baselines suggests realistic coverage rather than a degenerate generator ([Borisov et al., 2023](#)).

除了几何度量之外，我们通过对抗攻击评估经验推理泄露。我们在标准情景下量化成员推理风险，其中攻击者获得一个候选记录  $x$  以及对发布模型或合成器的某些访问权限。攻击者的目的是判断  $x$  是否属于训练数据 ([Shokri et al., 2017](#))。我们将  $g(x)$  定义为攻击得分，较大的值表示对  $x$  属于训练集的置信

度更高。我们使用受试者工作特征曲线下面积和成员优势来总结这一风险：

$$\text{AUC}_{\text{MIA}} = \Pr[g(x_{\text{train}}) > g(x_{\text{holdout}})], \quad \text{Adv}_{\text{MIA}} = \max_{\tau} (\text{TPR}(\tau) - \text{FPR}(\tau)).$$

Values for the Area Under the Curve significantly above 0.5 or advantage values significantly above 0 indicate non-trivial leakage. The advantage form matches the standard notion based on the difference between the True Positive Rate and the False Positive Rate used to quantify inference leakage (Yeom et al., 2018).

我们同样对一个敏感属性  $A$  进行属性推断评估。在此情景下，攻击者从其余属性以及释放模型 (Yeom et al., 2018) 所授予的任何访问权限中预测  $A$  的值。除了报告准确率或曲线下面积等标准预测指标外，我们还可选择报告相对于多数类汇总的简单提升，作为实现级别的诊断指标：

$$\text{Gain}_{\text{AIA}} = \Pr[\hat{a} = A] - \max_a \Pr[A = a],$$

where larger values indicate stronger recoverability of the attribute beyond a trivial majority baseline.

我们报告了差分隐私参数  $\varepsilon$  和  $\delta$ ，它们表示隐私预算，以及下游效用。这种方法使我们能够在标准差分隐私框架 (Dwork & Roth, 2014) 下，对隐私与效用之间的权衡进行情境化分析。

**公平性。** 我们在合成数据上训练，在真实数据上测试的设置下评估公平性。这是必要的，因为基于合成数据训练的下游模型在部署到真实人群时可能会继承甚至放大偏差。这种设置是表格生成领域常用的评估协议。公平性考量贯穿整个流水线，从数据、模型到部署均需关注 (Barocas et al., 2023)。

对于结果公平性，包括人口均等性和差异影响，我们考虑一个受保护的属性  $A \in \{a, b\}$  和一个二元决策  $\hat{Y}$ 。我们报告统计均等性差异和差异影响比率：

$$\Delta_{\text{SPD}} = \left| \Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1 \mid A = b) \right|, \quad \text{DIR} = \frac{\Pr(\hat{Y} = 1 \mid A = a)}{\Pr(\hat{Y} = 1 \mid A = b)}.$$

Ideally, the Statistical Parity Difference should approach 0 and the Disparate Impact Ratio should approach 1. The Disparate Impact Ratio is widely used in the literature concerning disparate impact as a rate ratio criterion such as the eighty percent rule (Feldman et al., 2015; Barocas et al., 2023).

为了考虑准确率相关的权衡，我们通过等几率差异和等机会差异来评估错误率公平性。这些指标通过条件于真实标签 (Hardt et al., 2016) 来衡量不同群体在真正例率和假正例率之间的差距。

$$\Delta_{\text{EO}} = \frac{1}{2} (|\Delta \text{TPR}| + |\Delta \text{FPR}|), \quad \Delta_{\text{EOp}} = |\text{TPR}_a - \text{TPR}_b|.$$

由于在合成数据上训练而在真实情景下测试时出现的奇偶性退化，可能源于合成数据与真实数据之间的表示不匹配，我们还报告了两个简单的诊断指标。子组覆盖差距衡量组别比例的变化，标签条件偏移则捕捉特定组别内的标签失真：

$$\text{CovGap} = \frac{1}{2} \sum_{g \in \mathcal{A}} |p_{\text{syn}}(A=g) - p_{\text{real}}(A=g)|, \quad \text{CondShift} = \frac{1}{|\mathcal{A}|} \sum_{g \in \mathcal{A}} \text{TV}(p_{\text{syn}}(Y \mid A=g), p_{\text{real}}(Y \mid A=g)).$$

Large values in these representational diagnostics often help explain why parity metrics degrade after the evaluation. This observation is consistent with broader discussions of fairness as a property of the entire pipeline starting from data through the model and ending at deployment (Barocas et al., 2023).

## 4.4 评估实践差距

我们分析了第 4.1 节中的代表性方法，并在表 4 中对它们报告的评估协议进行了分类。

Generation method	Validity	Fidelity	Diversity	Utility	Privacy	Fairness
GReaT (Borisov et al., 2023)	✓	✓	×	✓	✓	×
REaLTabFormer (Solatorio & Dupriez, 2023)	△	✓	×	✓	✓	×
EPIC (Kim et al., 2025)	△	✓	△	✓	×	×
HARMONIC (Wang et al., 2024d)	×	△	×	✓	△	×
TabSyn (Zhang et al., 2024a)	✓	✓	✓	✓	✓	×

**Table 4** 是否代表性表格数据生成方法在其实验部分显式评估了各个维度。✓：显式评估；△：部分或间接覆盖；×：未报告或不适用。

**多样性.** 我们的分析表明，表格生成中的多样性尚未得到充分评估。这是因为专注于覆盖度的特定度量（如  $\beta$ -召回率）在实验情景中很少被报告。这一差距对于由大模型驱动的生成器尤为重要。在这些模型中，基于似然和解码过程的标准训练目标往往表现出对最常见模式的过度关注。因此，合成数据分布倾向于过度代表频繁出现的模式，而无法捕捉长尾值以及特征的稀有组合。尽管合成数据集在主导模式下看起来可能具有真实性，但它们通常缺乏对低频区域的足够支持。我们因此建议，在报告标准的保真度和效用度量的同时，还应报告诸如  $\beta$ -召回率等覆盖度度量以及子组级别的诊断信息。

**公平性.** 我们的综述表明，当前关于表格生成的研究中，公平性评估大多缺失。这是一个重要缺口，因为公平性对于“在合成数据上训练，在真实数据上测试”（TSTR）协议至关重要。当使用合成数据训练下游模型以部署于真实人群时，表示上的不匹配会直接导致结果公平性和错误率公平性的差异。这些不匹配包括子群体比例的变化或标签条件分布的偏移。例如，结果公平性可通过统计均等性来衡量，而错误率公平性可通过平等几率来衡量。

此外，多样性提升并不总是意味着公平性改善。扩大数据覆盖范围可能会意外增加不公平性，特别是当少数子群体的生成保真度较低，或条件分布被扭曲时。因此，我们建议在“在合成数据上训练，在真实数据上测试”协议的结果中加入群体公平性指标，如统计均等性差异和相等几率。同时，我们建议使用表示诊断方法，以清晰描述多样性、公平性与下游性能之间的潜在权衡。

## 4.5 使用

**数据共享.** 基于大模型的合成表格数据正被越来越多地用于促进安全的数据共享和隐私敏感的匿名化。通过生成在统计上与源分布保持一致但与原始数据点不同的记录，组织可以在不泄露敏感或可辨认信息的情况下发布或交换数据集。这一能力在医疗保健和金融等受监管领域尤为重要。在这些领域中，合成数据作为不同机构间协作的合规机制 (Miletic & Sariyar, 2024; Barr et al., 2025; Long et al., 2025)。

在此背景下，Miletic & Sariyar (2024) 基于 Transformer 架构的大模型与 CTGAN 等基准模型进行对比 (Xu et al., 2019)。他们表明，更大的语言模型在下游分类任务中表现更优，并且即使在较小规模下也能保持竞争力。此外，Barr et al. (2025) 表明 GPT-4o 能够以 zero shot 方式生成临床表格数据，并

在每列的统计特性方面实现高保真度。当模型受到描述性统计量引导时，这种性能尤为突出。然而，作者并未直接评估下游实用性或重复和记忆风险。他们将这些领域列为未来研究的重要方向。

除了对单行数据的模仿之外，诸如 LLM-TabFlow 等框架利用大模型的推理能力来捕捉列之间的逻辑关系，并在潜在空间中合成数据。这种方法有助于优化保真度、实用性与隐私之间的权衡 (Long et al., 2025)。综上所述，这些发现表明，由大模型驱动的合成正逐渐成为在严格监管约束下共享敏感表格的一种可行方法。

**数据增强** 在数据稀缺或罕见特征组合缺失的情况下，由大模型驱动的生成成为增加训练数据分布提供了一种方法。该方法有助于改善类别平衡性以及模型对新数据的泛化能力 (Seedat et al., 2024; Tran & Xiong, 2024; Kim et al., 2025; Yang et al., 2024b)。

最近的研究正从简单的过采样方法转向更可控的表格数据生成方式。在某些情况下，这些方法还包含正式的隐私保障。例如，DP-LLMTGen 通过使用两阶段微调流水线来确保差分隐私。该过程首先学习数据格式，然后使用针对表格数据设计的损失函数进行带有差分隐私的微调，最后对私有模型进行采样以生成合成表格 (Tran & Xiong, 2024)。

在相关方向上，P-TA 采用基于近端策略优化 (Proximal Policy Optimization) 的最优化方案，引入判别器的反馈信息。这提升了生成数据与真实表格分布之间的对齐程度 (Yang et al., 2025)。因此，该领域的研究正从基础的过采样技术逐步转向兼具可控性与隐私意识的生成框架。这些进展使得基于大模型 (LLMs) 的生成器成为在复杂数据环境中提升下游模型性能的灵活工具。

## 5 半结构化

半结构化数据可以被视为介于严格的关系模式和非结构化文本之间的中间模态。这种模态的特点是具有灵活的模式和层次化组织。在本综述中，我们将图、JSON 和日志数据统一归为此类别。

图数据由定义拓扑关系的结点和边组成。近期文献表明，大模型能够通过将这些结构序列化为文本格式 (如边列表) (Yao et al., 2024) 来合成此类结构。

JSON 数据表示由名称-值对和数组组成的嵌套结构 (Bray, 2017)。然而，可靠的 JSON 生成需要严格遵守模式并进行语法验证，以确保输出可被下游工具执行 (Agarwal et al., 2025a)。

日志数据由一系列事件消息组成，每个条目通常将时间戳与日志模板及变量参数结合。这种格式使得日志能够自动解析为结构化模板 (He et al., 2017)。

### 5.1 生成方法

#### 5.1.1 图数据

基于是否需要参数更新，基于大语言模型的图生成最近进展可分为无训练生成与基于学习的生成。

**无需训练的图生成。**越来越多的研究工作探索了大模型在无需基于梯度的微调情况下生成语法有效且结构合理的图的能力。Yao et al. (2024) 提出 LLM4GraphGen，证明了如 GPT-4 这类模型可以直接从自然语言提示中生成图结构。该工作涵盖了基于规则和基于分布的任务。尽管直接提示提供了一种领



域无关且易于部署的方法，但现有评估表明其在结构保真度方面存在局限性。这一点在针对复杂分布参数（如特定基序数量）时尤为明显 (Yao et al., 2024)。为超越简单的提示方法，Generate-on-Graph 通过一种无训练的探索过程，解决了不完整知识图谱问答问题。该方法在关键三元组缺失时，动态检索知识图谱证据并生成额外的事实三元组 (Xu et al., 2024b)。与此同时，基于本体的知识图谱构建利用与 Wikidata 模式对齐的本体，为关系抽取提供基础，并提升与现有本体的互操作性 (Feng et al., 2024)。

**基于学习的多智能体生成。** 通过监督微调或多智能体协作框架，替代性方法提升了图构建的质量 (Huang et al., 2025a; Le et al., 2024)。GraphJudge (Huang et al., 2025a) 采用监督微调训练大模型作为判别器以评估图质量。该方法通过过滤生成的三元组，显著降低了知识图谱增强过程中的噪声。除了使用单一模型外，多智能体系统如 GAG 和 GraphMaster 利用协作精炼来提升全局一致性。GAG 将图合成建模为可扩展且基于仿真的多智能体过程。该方法能够生成符合宏观网络属性的大规模文本属性图 (Ji et al., 2025)。类似地，GraphMaster 在智能体之间引入了评估驱动的迭代环。该系统专门设计用于增强合成图的结构完整性和语义连贯性 (Du et al., 2025)。

### 5.1.2 JSON 数据

大模型已被调整以通过两种代表性方法类别生成符合模式的 JSON。这些方法包括推理时的约束控制，通常称为引导解码，以及基于学习的对齐方法，如强化学习，以提高对模式的遵循程度 (vLLM Project, 2024; Lu et al., 2025; Agarwal et al., 2025a)。

**用于 JSON 生成的约束解码。** 在此范式中，模型参数保持不变，而在推理阶段通过解码时的约束来强制实现结构合规性。这些约束包括基于模式或正则表达式的 token 过滤 (vLLM Project, 2024; Gat, 2025; dottxt-ai, 2025)。作为补充步骤，大模型也可用于从现有语料库中推断并丰富 JSON 模式。这可以通过为模式元素生成自然语言描述以及识别潜在噪声属性来实现 (Mior, 2024)。一旦目标模式可用，诸如 vLLM 结构化输出等实用框架以及 Outlines 和 LM Format Enforcer 等工具包可通过掩码无效 token 来实现约束解码。这些系统确保最终输出符合目标模式，并防止在无约束提示下可能出现的格式错误的 JSON。此外，JSONSchemaBench 提供了大规模的证据和分析，涵盖这些约束解码系统在 10,000 个真实世界 JSON 模式上的合规性、覆盖范围和效率 (vLLM Project, 2024; dottxt-ai, 2025; Gat, 2025; Geng et al., 2025)。然而，尽管这些方法在强制表面形式有效性方面表现优异，要满足更严格的要求仍可能需要额外的验证、后处理或基于学习的对齐。此类要求包括细粒度字段类型定义以及跨字段语义一致性 (Lu et al., 2025)。

**基于学习的生成。** 当需要更严格的模式遵循时，基于强化学习的方法会利用与模式正确性相关的奖励信号来优化模型策略。这些奖励通常通过模式验证器或验证器风格的奖励提供 (Lu et al., 2025; Agarwal et al., 2025a)。例如，Lu et al. (2025) 提出了 SchemaBench，其中包含约 40,000 个 JSON 模式。他们通过引入细粒度的模式验证器，将强化学习应用于结构化生成，从而提升了性能，优于标准的监督微调基准 (Lu et al., 2025)。与此同时，Agarwal et al. (2025a) 将组相对策略优化方法应用于训练小型模型，并采用自定义奖励以实现严格的模式遵循。该工作展示了在强制模式一致性方面取得的有效改进。

### 5.1.3 日志数据

日志生成研究揭示了日志组件的语义预测与最终日志消息的语法实现之间存在差异。这些组件包括日志级别和变量。Li et al. (2024d) 观察到，尽管大模型能够准确确定必要的日志属性，但它们往往无法生成模仿人工编写代码的完整日志语句。在他们的基准 LogBench 上，表现最好的模型仅达到了 0.249 的 BLEU 得分。这表明表面形式的相似度有限 (Li et al., 2024d)。此外，在语义等价但经过变换的代码上下文中进行的评估（称为 LogBench-T）显示性能持续下降。这尤其体现在变量预测和日志文本生成方面，表明通用大模型在保持语义不变的代码变换下依然脆弱 (Li et al., 2024d)。

为了弥合这一性能差距，近期的研究提倡通过在领域特定语料库上进行后训练来实现模型特化。Zhang et al. (2025a) 提出 AUCAD，这是一个自动构建名为 AucadLog 的对齐数据集的框架，该数据集源自与日志相关的软件问题。通过利用该数据集对开源大模型进行后训练，作者证明了所获得的特化模型在日志语句生成方面显著优于现有的基于模型的解决方案。这些结果得到了人工评估和定量指标的双重验证 (Zhang et al., 2025a)。

## 5.2 半结构化数据的质量度量

### 5.2.1 图数据

**有效性。** 有效性表示生成图中符合特定任务规则或逻辑约束的比例。该指标反映了模型输出在拓扑结构或物理意义上是否有效。我们通过衡量满足特定任务规则约束的生成图所占比例来评估：

$$\text{Valid}_{\text{rule}} = \frac{1}{|\mathcal{G}_{\text{gen}}|} \sum_{G \in \mathcal{G}_{\text{gen}}} \mathbb{I}\{\text{passes\_rules}(G)\}.$$

where  $\mathcal{G}_{\text{gen}}$  denotes the set of generated graphs where  $\mathcal{G}_{\text{gen}} \subseteq \mathcal{G}$ . The function `passes_rules` is a task defined rule checker that returns 1 if  $G$  satisfies the rules and 0 otherwise. These rules may include constraints specified by the generation task (Yao et al., 2024).

**忠诚** 保真度衡量生成图与参考数据之间的结构相似度。我们通过基于核函数的最大平均偏差 (Maximum Mean Discrepancy) 来评估生成图与参考结构的匹配程度，该偏差在图描述符特征上计算：

$$\text{MMD}^2(\mathcal{X}, \mathcal{Y}) = \frac{1}{m^2} \sum_{i, i'} k(x_i, x_{i'}) + \frac{1}{n^2} \sum_{j, j'} k(y_j, y_{j'}) - \frac{2}{mn} \sum_{i, j} k(x_i, y_j),$$

where  $\mathcal{X} = \{x_i\}_{i=1}^m$  and  $\mathcal{Y} = \{y_j\}_{j=1}^n$  are descriptor sets extracted from generated and real graphs respectively, and  $k$  is a chosen kernel (You et al., 2018; Liao et al., 2020). Typical descriptors include degree distributions, clustering coefficient distributions, orbit or motif counts, and spectral statistics such as Laplacian eigenvalue histograms (You et al., 2018; Liao et al., 2020).

对于分子图，我们包含了 **Frechet ChemNet 距离** (Preuer et al., 2018):

$$\text{FCD} = \|\mu_1 - \mu_2\|_2^2 + \text{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{1/2}\right),$$

This metric measures the distance between embedding distributions of generated and reference molecules.  $\mu$  and  $\Sigma$  represent the mean and covariance of these distributions where embeddings are obtained from

a pretrained ChemNet model (Preuer et al., 2018).

**多样性** 多样性用于评估生成图之间的差异性和原创性，包括图相对于提示的新颖性以及生成集合内部的重复程度。为评估多样性，我们计算新颖性。若生成的图与同一任务提示中提供的每个示例图均不同，则认为该图具有新颖性。

$$\text{Novelty} = \frac{1}{|\mathcal{G}_{\text{gen}}|} \sum_{G \in \mathcal{G}_{\text{gen}}} \mathbb{I}\{G \neq \mathcal{G}_{\text{ex}}\},$$

where  $\mathcal{G}_{\text{ex}} \subseteq \mathcal{G}$  denotes the set of example graphs in the prompt. The condition  $G \neq \mathcal{G}_{\text{ex}}$  means  $G$  is not identical to any example graph under the same equality criterion used by the evaluation pipeline (Yao et al., 2024).

我们还计算唯一性，即不重复的有效生成图所占的比例：

$$\text{Uniq} = \frac{|\text{unique}(\mathcal{G}_{\text{valid}})|}{|\mathcal{G}_{\text{valid}}|}, \quad \mathcal{G}_{\text{valid}} = \{G \in \mathcal{G}_{\text{gen}} : \text{passes\_rules}(G)\},$$

where the function unique removes duplicates under the same equality criterion used in evaluation (Yao et al., 2024).

**实用性**。实用性评估生成数据对于下游应用的实际价值。这通常通过合成数据对模型在真实世界任务上性能的提升程度来衡量。当图神经网络在增强后的图上进行训练并在固定划分上进行评估时，我们报告准确率和 F1 得分。准确率和 F1 得分越高，表明合成图在下游任务中的实用性越强 (Du et al., 2025)。

### 5.2.2 JSON 数据

对于半结构化输出（如 JSON 响应），质量度量通常关注三个方面的内容。这些维度包括有效性、保真度和实用性。

**有效性**。有效性表示输出必须为机器可解析且符合预期模式。这确保生成的数据对自动化处理是有效的。令  $\mathcal{J} = \{J_1, \dots, J_N\}$  为生成的 JSON 对象集合， $\mathcal{S}$  为目标模式。

我们定义在模式  $\mathcal{S}$  下生成对象  $J$  的正确性指示器如下：

$$V(J, \mathcal{S}) = \mathbb{I}[\text{parsable}(J) \wedge \text{schema}(J, \mathcal{S})].$$

In this equation,  $\mathbb{I}[\cdot]$  is the indicator function that returns 1 if  $J$  is syntactically valid and schema-compliant, and 0 otherwise.

然后我们报告平均正确率：

$$\text{CorrectnessRate} = \frac{1}{|\mathcal{J}|} \sum_{J \in \mathcal{J}} V(J, \mathcal{S}).$$

此外，我们报告了一种纯粹的语法度量指标，称为有效 JSON 率或可解析率。该指标衡量输出是否可被解析为 JSON，同时忽略具体的模式约束：

$$\text{ValidJSONRate} = \frac{|\{J \in \mathcal{J} \mid \text{IsParsable}(J) = \text{true}\}|}{|\mathcal{J}|}.$$

Therefore, we report two distinct measures. The first is a purely syntactic parsability metric referred to as the ValidJSONRate. The second is a schema level validity metric referred to as the Correctness-Rate. Parsability metrics are explicitly reported in strict structured output evaluations [Agarwal et al. \(2025a\)](#). At the same time, schema validation protocols are commonly used in settings involving schema constrained generation [Lu et al. \(2025\)](#).

**忠诚** 保真度衡量生成内容与目标数据语义信息和分布的接近程度。该维度反映了结构化 JSON 框架内内容的质量。保真度指标评估模型是否不仅在语法上有效，还能生成正确的字段、值和语义。

对于结构化 JSON，其中真实值实例表示为  $J_i^{\text{gt}}$ ，生成的实例表示为  $J_i$ ，**平均匹配百分比**将生成的字段与真实值实例进行比较：

$$MMP = \frac{1}{N} \sum_{i=1}^N \frac{|\text{fields}(J_i) \cap \text{fields}(J_i^{\text{gt}})|}{|\text{fields}(J_i^{\text{gt}})|}.$$

This metric can be extended to checks at the level of values or constraints when ground truth values and constraints are available ([Agarwal et al., 2025a](#)).

基于模式的评估还考察模式本身是否具有意义。例如，研究人员可通过嵌入的相似度来评估生成的定义和名称的质量。此外，属性选择被当作有用属性与噪声属性之间的分类问题进行评估，通常以准确率 [Mior \(2024\)](#) 来报告。

**实用性**。实用性评估生成的 JSON 数据在解决特定任务中的实际有效性，或作为下游处理输入的有用性。为了评估生成的 JavaScript 对象表示法（JSON）响应在后续使用中的实用性，我们要求每个响应遵循预定义的模式，以便数据能够被可靠读取。我们通过计算提取答案与真实值之间的确切匹配准确率来衡量任务的成功程度。设生成响应的集合记为  $\{J_i\}_{i=1}^N$ ，参考答案记为  $A_{\text{gt},i}$ 。我们将任务准确率定义为提取答案与参考答案匹配的平均比例：

$$\text{TaskAcc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{ExtractAnswer}(J_i) = A_{\text{gt},i}).$$

This measurement reflects how well the synthetic data supports the successful completion of the target application. This methodology aligns with evaluations based on task accuracy for schema constrained structured outputs ([Geng et al., 2025](#)).

### 5.2.3 日志数据

日志通常具有半结构化特征。每条日志消息由固定模板与运行时变量或参数组合而成，也可能附带严重性等级、组件等元数据。因此，评估应分别针对两个不同方面进行。其一是日志解析器下的结构有效性，包括模板和变量提取以及分组；其二是生成文本、变量和元数据的内容保真度。

**有效性**。对于日志数据而言，有效性不仅取决于表面的格式正确性，还取决于日志能否被解析器准确地结构化为模板和变量。用于此目的的一种常见消息级度量指标是解析准确率，它衡量的是能够被精确



解析为正确模板和变量的日志消息所占的比例：

$$\text{PA} = \frac{|\mathcal{L}_{\text{correct}}|}{|\mathcal{L}|},$$

where  $\mathcal{L}$  represents the set of all log messages  $\{\ell_i\}$ , and  $\mathcal{L}_{\text{correct}}$  refers to the subset whose predicted template, consisting of static tokens, and variable spans, consisting of dynamic tokens, match the ground truth exactly (Khan et al., 2022; Jiang et al., 2024a; Ma et al., 2024b).

除了每条消息的正确性之外，分组准确率还评估消息是否被分配到了正确的模板组。

$$\text{GA} = \frac{|\mathcal{L}_{\text{grouped}}|}{|\mathcal{L}|},$$

where  $\mathcal{L}_{\text{grouped}}$  denotes messages whose predicted grouping is consistent with the ground-truth template partition. This means that messages belonging to the same true template are grouped together and those from different templates are separated (Khan et al., 2022; Ma et al., 2024b). Grouping Accuracy is particularly important when downstream pipelines rely on template clusters rather than individual parses.

由于解析准确率和分组准确率基于消息数量，因此容易受到高频模板（如心跳消息）的过度影响。这促使我们采用模板级别的评估方法 (Khan et al., 2022; Jiang et al., 2024a)。令  $\mathcal{T}_{gt}$  和  $\mathcal{T}_p$  分别表示真实模板集和预测模板集，令  $\mathcal{T}_{\text{correct}}$  表示正确识别的模板集合。通常在遵循常见评估规范并使用最优模板修正的情况下进行报告 (Khan et al., 2022)。定义模板级别的准确率和召回率如下

$$P_{\text{TA}} = \frac{|\mathcal{T}_{\text{correct}}|}{|\mathcal{T}_p|}, \quad R_{\text{TA}} = \frac{|\mathcal{T}_{\text{correct}}|}{|\mathcal{T}_{gt}|}.$$

模板准确率的 F1 得分定义为

$$\text{FTA} = 2 \cdot \frac{P_{\text{TA}} \cdot R_{\text{TA}}}{P_{\text{TA}} + R_{\text{TA}}},$$

which complements Parsing Accuracy and Grouping Accuracy by emphasizing template coverage and uniqueness (Khan et al., 2022; Jiang et al., 2024a). 最近的大规模基准测试进一步建议采用额外的模板级别分组度量方法，例如分组准确率的 F1 得分，以缓解在模板频率不平衡时消息级别分组度量的敏感性 (Jiang et al., 2024b)。总体而言，这些度量指标捕捉了生成的日志是否支持忠实的结构解析，而不仅仅是语法上的正确性。

**忠诚** 对于日志而言，保真度强调生成的模板、变量和元数据在语言真实性和操作合理性方面的保留程度。当模型生成或重建变量时，变量准确率、变量召回率和变量 F1 得分用于评估绑定的运行时变量是否正确：

$$\text{VP} = \frac{|V_p \cap V_t|}{|V_p|}, \quad \text{VR} = \frac{|V_p \cap V_t|}{|V_t|}, \quad \text{VF1} = 2 \cdot \frac{\text{VP} \cdot \text{VR}}{\text{VP} + \text{VR}},$$

where  $V_p$  and  $V_t$  are the predicted and true variable sets (Li et al., 2024d).

对于模板和文本的真实性，可以在生成的与参考的日志模板或文本之间使用基于重叠的度量指标，如 BLEU 和 ROUGE。此外，通常报告基于嵌入的语义相似度，以降低对改写和混合自然语言或代码表述

的敏感性 (Li et al., 2024d)。

元数据保真度同样至关重要。例如，日志级别准确率衡量的是严重程度级别完全正确的日志所占比例：

$$\text{L-ACC} = \frac{N_{\text{correct\_level}}}{N},$$

Furthermore, the Average Ordinal Distance captures how far predicted levels deviate on an ordinal severity scale:

$$\text{AOD} = \frac{1}{N} \sum_{i=1}^N \left( 1 - \frac{\text{Dis}(l_{\text{pred}}^{(i)}, l_{\text{gt}}^{(i)})}{\text{MaxDis}} \right),$$

where  $\text{Dis}(\cdot)$  is the ordinal distance between levels, such as the ranking from error to trace, and  $\text{MaxDis}$  is the maximum possible distance on the chosen scale (Li et al., 2024d).

**实用性。** 对于日志而言，下游效用反映了其对系统理解与诊断的贡献。一种常见的评估方法是，在标准流水线和固定测试集下，判断合成或增强日志是否能够提升或保持在下游日志分析任务（如解析或异常检测）中的性能 (Huo et al., 2023)。

量化效用的一种直接方法是：在训练或数据增强中使用合成日志时，下游模型性能的变化。

$$\Delta_{\mathcal{M}} = \mathcal{M}(f(\mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}})) - \mathcal{M}(f(\mathcal{D}_{\text{real}})),$$

where  $\mathcal{M}$  is a task metric, such as the F1 score or accuracy, and  $f$  is the downstream model or procedure evaluated on a fixed test set.

在以人为中心的评估中，开发者还可以使用李克特量表对生成的日志和解释的有用性和可读性进行评分，以捕捉其实际的诊断价值 (Liu et al., 2024d)。

## 5.3 可信的半结构化数据度量方法

### 5.3.1 图数据

**隐私** 图结构数据带来了独特的隐私风险，因为结点和边通常对应于个体用户及其关系。这些风险通过针对图邻接性的微分隐私概念进行形式化 (Kasiviswanathan et al., 2013)。

图生成中隐私的正式定义基于差分隐私的框架，其特征由两个关键参数  $\epsilon$  和  $\delta$  刻画。随机机制  $M$  满足  $(\epsilon, \delta)$ -差分隐私不等式，如果对于任意两个仅相差一条记录的相邻数据集  $D$  和  $D'$ ，以及任意可能的输出集合  $S$ ，机制对数据集  $D$  产生输出在  $S$  中的概率由以下条件界定：

$$\Pr[M(D) \in S] \leq e^\epsilon \Pr[M(D') \in S] + \delta.$$

In this mathematical framework,  $\epsilon$  represents the privacy budget or privacy loss, which quantifies the maximum allowable change in the output probability distribution when one individual’s information is modified. A smaller  $\epsilon$  value indicates a stronger privacy guarantee because it ensures that the outputs are more similar regardless of whether a specific record is present. The parameter  $\delta$  represents the probability that the privacy constraint might be violated, and it is typically required to be a very small

value to ensure that the guarantee remains robust (Dwork & Roth, 2014).

当该定义应用于图结构数据时，邻近数据集的概念被具体化为结点或边的邻接关系 (Kasiviswanathan et al., 2013)。结点个别的差分隐私将邻近定义为：两个图之间仅通过移除一个结点及其所有关联边而得到。相比之下，边个别的差分隐私则认为，若两图仅相差一条边，则视为邻近。在这些邻接定义下，参数  $\varepsilon$  和  $\delta$  作为定量度量，用于衡量从生成的图中推断出特定用户或关系的风险 (Kasiviswanathan et al., 2013; Dwork & Roth, 2014)。

同样地，边级中心差分隐私将邻近图视为仅相差一条边的图，这一概念被称为边邻接 (Kasiviswanathan et al., 2013)。在此邻接定义下，相同的  $(\varepsilon, \delta)$ -差分隐私不等式仍然适用 (Dwork & Roth, 2014)：

$$\Pr[M(G) \in S] \leq e^\varepsilon \Pr[M(G') \in S] + \delta, \forall S \subseteq \mathcal{O}.$$

In this case, the presence or absence of any single relationship or edge is protected.

为了在无需可信协调者的情况下实现更细粒度的保护，我们可以在局部差分隐私模型中工作，其中每位参与者在共享其贡献之前先对其数据进行随机化。局部差分隐私的标准形式化通过每个用户的隐私化通道的似然比界限给出 (Duchi et al., 2014)。针对每条边的报告（例如边指示或本地邻接信息）实例化这一标准定义，我们采用如下边级局部差分隐私的概念。若机制  $M$  满足对任意两个可能的边值  $e$  和  $e'$  以及任意输出集合  $S \subseteq \mathcal{O}$ ，以下条件成立，则称该机制为  $\varepsilon$ -边级局部差分隐私：

$$\Pr[M(e) \in S] \leq e^\varepsilon \Pr[M(e') \in S] \quad \forall e, e', S \subseteq \mathcal{O}.$$

In many local differential privacy settings, one focuses on the pure case described above (Duchi et al., 2014). One can also consider an  $(\varepsilon, \delta)$  extension by adding a  $\delta$  term as in approximate differential privacy (Dwork & Roth, 2014), but we maintain the common pure form for clarity.

中心模型中的一种常见实现使用全局敏感性和拉普拉斯噪声 (Dwork & Roth, 2014)：

$$\Delta f = \max_{G \sim G'} \|f(G) - f(G')\|_1, \\ \eta \sim \text{Lap}\left(0, \frac{\Delta f}{\varepsilon}\right),$$

where  $f$  is a numeric query, such as degree-histogram bins. The Laplace mechanism yields pure  $\varepsilon$ -differential privacy, which means  $\delta$  is equal to zero, under the chosen adjacency notion (Dwork & Roth, 2014). In graph settings,  $\Delta f$  can be very large for degree-related queries under node adjacency, since adding or removing a node can affect many incident edges and consequently many degree counts. Practical approaches therefore impose bounded-degree assumptions or apply degree clipping and projection to a bounded-degree graph family. These techniques keep  $\Delta f$  and the required noise level manageable (Kasiviswanathan et al., 2013).

在我们的评估中，中心差分隐私的隐私参数  $\varepsilon$  和  $\delta$ ，以及纯本地差分隐私的隐私参数  $\varepsilon$ ，作为定量的隐私度量。较小的  $\varepsilon$  和较小的  $\delta$  表示更强的隐私保护，这通常以牺牲实用性 (Dwork & Roth, 2014) 为代价。

最后，通常要求在适当的问题或安全参数下， $\delta$  为可忽略的。避免  $\delta$  为非平凡地大的情形非常重要，例

如当群体规模为  $n$  (Dwork & Roth, 2014) 时, 其数量级约为  $n$  的倒数。在结点对应受保护个体的图结构中, 使用近似差分隐私时的一个实用启发式方法是选择  $\delta$ , 使其远小于图中顶点数量的倒数。确切的选择仍取决于具体的应用和威胁情景。

**鲁棒性** 除了隐私之外, 图生成必须对结构幻觉具有鲁棒性, 这指的是生成看似合理但不正确的图 (Richardseau et al., 2025)。我们使用互补检查来捕捉这种风险。

句法正确率定义为在指定语法或格式下, 能够被解析为有效图结构的生成输出所占的比例:

$$\sigma_{SCR} = \frac{1}{|\mathcal{G}_{\text{gen}}|} \sum_{G \in \mathcal{G}_{\text{gen}}} \mathbb{1}\{\text{PARSEOK}(G) = 1\}.$$

A low value of  $\sigma_{SCR}$  indicates format-breaking outputs and sensitivity to changes in prompting or decoding. This is particularly relevant when the generator is prompted to output a concrete graph representation, such as edge lists, which must be parsed into a graph object for downstream evaluation. Such generate-then-parse pipelines are common in research involving graph generation with LLMs (Richardseau et al., 2025).

图谱距离被用于直接度量对于指定目标 (如正则图谱图) 的拓扑偏差。按照 (Richardseau et al., 2025) 中提出的方法, 该度量定义如下:

$$\text{GAD} = \frac{1}{k} \sum_{i=1}^k d_{\text{GED}}(G_i, a_i),$$

where  $G_1, \dots, G_k$  represent the outputs generated from a fixed set of prompts whose ground-truth targets are the corresponding canonical atlas graphs  $a_1, \dots, a_k$ . In this formulation,  $d_{\text{GED}}$  denotes the graph edit distance as described in (Richardseau et al., 2025). A larger value for the Graph Atlas Distance implies more severe structural hallucination in the generated results.

为了降低对罕见且极端编辑距离的敏感性, 这些距离可能会主导平均值, 我们还报告了一种称为鲁棒聚合的图谱距离的截断版本。该度量的计算方法如下:

$$\text{GAD}^{\text{cap}} = \frac{1}{k} \sum_{i=1}^k \min\{d_{\text{GED}}(G_i, a_i), C\},$$

where  $C$  caps extreme values. In this context, the uncapped average corresponds to the standard definition of the Graph Atlas Distance proposed in (Richardseau et al., 2025), while the cap serves as a robustification measure during the evaluation process.

度分布偏差通过测量规范化度直方图之间的轻量级偏差来捕捉分布漂移。该指标由以下表达式定义:

$$D_{L_2}(G) = \|h_{\text{gen}}(G) - h_{\text{ref}}\|_2,$$

where  $h_{\text{gen}}(G)$  represents the normalized degree histogram of the generated graph  $G$  and  $h_{\text{ref}}$  is the target histogram. This target may be derived from the ground-truth graph or a reference dataset. A higher value of this deviation indicates that the generated graphs may appear reasonable but diverge statistically from the reference model. Comparing degree-based statistics through distributional dis-



tances, such as Maximum Mean Discrepancy, Kullback Leibler divergence, or Kolmogorov Smirnov tests, is standard practice in the evaluation of graph generation (Liao et al., 2020; Liu et al., 2024b). In this study, we utilize an  $L_2$  histogram deviation for simplicity. We emphasize that degree-based deviations provide useful although not sufficient auxiliary signals. Consequently, they should be interpreted alongside topology-sensitive measures such as the Graph Atlas Distance or the graph edit distance (Richardeau et al., 2025).

### 5.3.2 JSON 数据

**隐私** 隐私是生成的 JSON 数据（例如以 JSON 格式存储的电子健康记录资源）中的一个关键问题，这类数据可能包含个人可辨认信息或其他敏感信息。在基于净化处理的流水线中，可以通过已建立的去标识化方法来评估隐私保护效果。此外，增强隐私的后处理质量也至关重要。这包括诸如实体重词化等技术，确保一致的替代置换，以维持下游应用的连贯性 (Singh et al., 2025)。

这些文档级别和语料库级别的得分用于衡量敏感实体检测与删除的准确率。由于即使一次敏感数据泄露也可能导致严重的隐私泄露，研究人员通常采用更严格的召回率定义 (Singh et al., 2025)。全有或全无召回率通过将文档内检测的完整性视为二元结果来体现这种严格性。只有当目标实体类型的所有实例均未遗漏时，该值才为 1，否则为 0 (Scaiano et al., 2016; Singh et al., 2025)。

$$\text{Recall}_{\text{AON}}(t, d) = \mathbb{I}(E_{t,d} \subseteq \hat{E}_{t,d}) = \mathbb{I}(E_{t,d} \setminus \hat{E}_{t,d} = \emptyset).$$

在该公式中，对于实体类型  $t$  和文档  $d$ ， $E_{t,d}$  表示类型为  $t$  的真实敏感实体实例集合。此外， $\hat{E}_{t,d}$  表示被正确检测并删除的实例集合， $\mathbb{I}(\cdot)$  是指示函数。这种文档级别的定义反映了这样一个事实：只要遗漏了一个实例，该特定实体类型的隐私保护即被视为无效。

临床模型一致性衡量重新词汇化数据在用于临床决策模型时是否避免引入系统性偏差。这些偏差可能包括种族、民族、区域或年龄组等因素。该指标还评估数据是否保持了现实世界模型的行为特征 (Singh et al., 2025)。具体而言，如果一个临床模型在使用真实数据进行训练和评估时达到性能指标值  $X$ ，那么糟糕的重新词汇化可能导致指标值为  $X + \delta_X$ 。

$$|\delta_X| = |X_{\text{relex}} - X_{\text{real}}|.$$

$|\delta_X|$  值越小，表示重新词汇化质量越高。这确保了下游临床应用效果和行为与在真实数据上训练和评估的模型保持接近。

### 5.3.3 日志数据

**隐私** 我们建议通过显式捕捉发布日志文本中仍存在的残余披露风险以及匿名化引起的效用损失，来评估日志中的隐私性。这种方法反映了在日志匿名化和任务感知匿名化基准测试中广泛观察到的隐私与效用之间的实际权衡 (Aghili et al., 2025; Loiseau et al., 2025)。受文本匿名化评估实践的启发，我们将日志的隐私评估沿三个互补维度组织：面向暴露的隐私风险、下游效用下降以及以人为本的有效性 (Pillán et al., 2022; Ren et al., 2025; Loiseau et al., 2025)。日志天然与文本匿名化相关，因为它们被归类为半

结构化文本。

敏感属性暴露通过统计匿名化后日志消息中仍存在的敏感属性实例数量，来衡量隐私风险的可观测指标。设  $S_S$  为预定义的敏感属性类型集合，例如 IP 地址和 MAC 地址。该集合来源于软件日志的法规要求、经验分析或行业共识 (Aghili et al., 2025)。设  $\psi(\ell)$  为一个敏感信息检测器，从日志消息  $\ell$  中取出一个多重集的跨度，设  $\text{type}(x)$  用于返回跨度  $x$  的属性类型。我们定义暴露如下：

$$\text{Exposure}(\ell; S_S, \psi) = \sum_{x \in \psi(\ell)} \mathbb{I}(\text{type}(x) \in S_S).$$

可选地，可以通过对日志子集  $\mathcal{D} \subseteq \mathcal{L}$  求平均值来计算数据集级别的暴露得分：

$$\text{Exposure}(\mathcal{D}; S_S, \psi) = \frac{1}{|\mathcal{D}|} \sum_{\ell \in \mathcal{D}} \text{Exposure}(\ell; S_S, \psi).$$

暴露度 (Exposure) 提供了一种轻量级且无需标注的剩余可辨认片段信号。这与在匿名文本评估中常用的标识符移除效果度量一致。当存在 token 或 span 级别的真实值标注时，研究者可以进一步报告敏感片段移除的准确率、召回率和 F1 得分 (Ren et al., 2025; Pilán et al., 2022)。需要注意的是，暴露度依赖于检测器。也就是说，它同时反映了匿名化质量和检测质量，应据此进行解释。

为了量化匿名化的效用成本，我们通过测量在下游任务上的模型性能下降来评估，即下游模型性能退化。此类任务包括异常检测、故障诊断和日志解析。令  $\mathcal{A}$  为一个固定的学习算法， $\mathcal{M}$  为任务指标，例如 F1 或准确率。可以通过在原始训练日志上训练一次下游模型，然后在评估时交换原始输入与匿名化输入，来进行任务感知的评估。这遵循了匿名化基准中的任务敏感性视角 (Loiseau et al., 2025)：

$$g = \text{Train}(\mathcal{A}, \mathcal{D}_{\text{orig}}^{\text{train}}),$$

$$\Delta_{\mathcal{M}} = \mathcal{M}(g; \mathcal{D}_{\text{test, orig}}) - \mathcal{M}(g; \mathcal{D}_{\text{test, anon}}).$$

可选地，还可通过在匿名日志上进行训练，并根据特定部署约束将结果与原始训练基准进行比较，以评估一种面向重训练的设置。

定性有效性得分捕捉了专家对匿名化在实际中保护隐私同时保持日志可用性的程度的判断。可用性的例子包括便于调试的可读性以及事件响应的充分性。此类以人为中心的评估通常通过李克特量表调查在日志隐私研究中收集 (Aghili et al., 2025)。平均得分计算方式如下：

$$\text{Score}_{\text{Qualitative}} = \frac{1}{|R|} \sum_{i=1}^{|R|} r_i,$$

其中  $R$  表示问题的响应集合，例如感知隐私有效性或感知效用保持性， $r_i$  表示响应  $i$  的数值，例如 1 到 5 之间的取值。

## 5.4 评估实践差距

我们分析了第 5.1 节中的代表性方法，并在表 5 中对它们报告的评估协议进行了分类。

我们的覆盖分析识别出半结构化数据评估中的两个系统性缺口——**隐私**和**效用**。这些缺口似乎主要源于现有研究在实验范围界定和设计上的问题，而非现有度量框架本身存在根本性缺陷。

Generation method	Validity	Fidelity	Utility	Privacy
<i>Graph data</i>				
LLM4GraphGen (Yao et al., 2024)	✓	△	×	×
Generate-on-Graph (GoG) (Xu et al., 2024b)	×	×	△	×
Ontology-grounded constrained decoding (Feng et al., 2024)	△	△	△	×
GraphJudge (Huang et al., 2025a)	△	△	△	×
GAG (Ji et al., 2025)	✓	✓	✓	×
GraphMaster (Du et al., 2025)	△	✓	✓	×
PrivGraph (Yuan et al., 2023)	×	△	△	✓
<i>JSON</i>				
JSON Schema discovery (Mior, 2024)	×	✓	△	×
JSONSchemaBench (Geng et al., 2025)	✓	×	✓	×
Schema Reinforcement Learning (Lu et al., 2025)	✓	×	✓	×
ThinkJSON (Agarwal et al., 2025a)	✓	✓	×	×
RedactOR (Singh et al., 2025)	△	△	✓	✓
<i>Log data</i>				
LogBench (Li et al., 2024d)	×	✓	×	×
AUCAD (Zhang et al., 2025a)	△	△	△	×
Protecting Privacy in Software Logs (Aghili et al., 2025)	△	△	✓	✓

**Table 5** 是否代表性半结构化生成方法在其实验部分明确评估了每个维度。✓：明确评估；△：部分/间接覆盖；×：未报告或不适用。

**隐私。**在所审查的图、JSON 和日志生成方法中，隐私很少被当作主要评估维度。这一属性通常仅在以保护隐私为主要贡献的研究中才被测量。例如，基于差分隐私或临床数据去标识化的图生成。这种模式表明，大多数通用生成论文更注重展示可行性与保真度，而将隐私风险的评估置于次要地位。一个可能的解释是，严格的隐私评估需要特定的实验承诺，而许多作者认为这些承诺超出了其研究范围。这些承诺包括明确的威胁模型以及具体且可复现的协议，例如使用带有  $\epsilon$  和  $\delta$  参数的差分隐私预算。因此，隐私测量仍被视为一种可选功能，仅限于专注于隐私的研究文献，而非成为半结构化生成的标准报告维度。

**效用。**尽管效用评估在文献中出现的频率高于隐私评估，但在我们回顾的方法中，其实际应用仍然极不一致。图生成论文通常通过下游学习性能（如训练图神经网络）来衡量效用。类似地，与知识图谱问答系统相关的方法将最终任务的准确率作为主要证据。相比之下，生成 JSON 和日志数据的研究往往强调结构正确性或语义相似度，但很少系统性地展示对下游应用的改进。这种多样性表明，由于效用依赖于具体任务且需要复杂的实验控制，因此难以标准化。这些控制包括数据混合策略、固定的训练与测试划分，以及特定的下游模型，以确保公平比较。因此，许多研究选择使用更容易计算的有效性或保真度量，而非采用一致且完整的效用评估协议。这导致了证据碎片化，不同研究领域之间的结果只能部分对比。

## 5.5 用途

**图的构建与分析。**LLM 引导的图合成支持四种主要应用和一种新兴的评估方法。

首先，在环中验证者驱动的知识图谱精炼中，由抽取流水线或补全模型提出的候选三元组（如事实）会通过基于大模型的验证者进行验证。这些系统执行一致性检查，并利用检索技术将信息与外部来源进行比对验证。该过程能够过滤噪声声明，减少对大规模人工验证的需求 (Boylan et al., 2024)。

其次，对于文本条件下的图仿真，诸如 GraphAgent-Generator（简称 GAG）等框架利用基于大模型的智能体仿真来生成具有文本属性的动态社交图。这些方法在小尺度和大尺度上均提升了结构准确率，并可通过并行处理来应对大规模网络 (Ji et al., 2025)。

第三，由大模型生成的合成图和潜在结构可作为下游图神经网络的数据增强，这些网络通常被称为图神经网络。例如，DemoGraph 使用黑盒大模型从文本提示中生成潜在知识图谱，并将这些结构整合到原始训练图中，以解决实际应用中数据有限和噪声问题。这种增强的价值通常通过下游任务（如节点分类）的性能指标来衡量 (Feng et al., 2025b; Ji et al., 2025)。

最后，图结构幻觉分析考察了大模型在生成结构化输出时的可靠性。Richardeau et al. (2025) 的一项研究促使这些模型重现诸如扎卡里·卡特俱乐部网络和《悲惨世界》人物关系网络等标准现实世界网络，同时要求模型生成如埃多斯-雷尼随机图之类的随机图。研究人员通过度序列统计量、谱距离以及基于图编辑的图谱距离（Graph Atlas Distance）等结构度量来衡量图幻觉。这项工作将结构差异与外部幻觉排行榜联系起来。

**工具使用和模式自动化的 JSON。** 大模型经常生成 JSON 作为工具调用、工作流自动化以及不同服务间数据交换的结构化接口。这些场景要求输出严格遵循预定义的模式。约束解码也被称为语法约束生成或模式约束生成。该过程在解码阶段强制遵守如 JSON Schema、正则表达式或上下文无关文法等规范。这种机制减少了解析失败，提高了智能体流水线中机器可消费输出的可靠性。近期针对真实世界模式的基准测试进一步描述了这些系统在约束解码中的效率、覆盖和质量之间的权衡 (Geng et al., 2025)。

JavaScript 对象表示法 (JavaScript Object Notation) 被广泛用于应用程序接口和数据集成工作流中的数据交换。当模式缺失或不完整时，大型语言模型可以改进对该格式的自动发现模式。这些模型生成自然语言描述，并为组件分配有意义的名称，以实现重用。它们还过滤掉那些推断出但无用的属性。这一过程有助于验证，并促进数据在后续阶段的使用 (Mior, 2024)。

**可观测性与开发者协助日志。** 半结构化日志生成被用于创建真实且可解析的遥测数据，以支持事件演练和异常基准测试。该技术还帮助开发人员自动完成日志任务，例如日志位置、日志级别和消息内容的决策。此外，它有助于共享保留隐私的示例。然而，近期的基准测试揭示了关键局限性。尽管大模型生成的日志消息在语义上通常看似正确，但在大规模应用中常常无法满足严格的质量要求和项目规范。LogBench 对日志语句生成进行了系统评估，并在常见的自动度量标准和泛化情景下识别出显著差距 (Li et al., 2024d)。类似地，AL Bench 通过动态评估强调了真实世界约束。它表明，包含生成日志语句的代码经常无法编译，且运行时的日志输出与真实值存在显著差异 (Tan et al., 2025)。

这些发现鼓励针对特定领域进行适配，而不是使用通用提示。为解决这一问题，关于 AUCAD 框架的研究表明，在专门为日志生成构建的对齐数据集上进行训练，其表现远优于基于大模型的通用解决方案。这为软件工程领域的应用提供了可行路径 (Zhang et al., 2025a)。



## 6 视觉-语言数据

视觉-语言数据由多模态实体组成，其中视觉信号与语言描述自然配对。这类数据构成了现代视觉-语言模型和多模态大模型的基础训练语料库。近期的方法越来越多地使用真实数据和合成数据，以降低费用、隐私担忧以及大规模对齐语料库稀缺性带来的问题 (Mohammadkhani et al., 2025)。具体而言，视觉-语言语料库包括图像-文本对、包含图文混合内容的网页等交错文档，以及视频-文本序列。所有这些格式均提供了配对的视觉内容和文本描述 (Sun et al., 2024)。

在大模型的背景下，多模态模型如 Emu (Sun et al., 2024) 将视觉信号编码为连续嵌入，并将其与文本 token 交错排列。这些模型随后使用统一的自回归目标训练单一 Transformer，该目标涉及在多模态序列中预测下一个文本 token 或对下一个视觉嵌入进行回归 (Sun et al., 2024)。近期的架构如 Emu3 进一步推进了这一范式，通过将图像、文本和视频分割到共享的离散 token 空间中。这种方法在混合多模态流上应用纯粹的下一个 token 预测 (Wang et al., 2024c)。

跨模态对齐在多个粒度层级上运行。从根本上说，视觉-语言数据在视觉单元（如图像或视频片段）与文本单元（如标题、转录文本或周围叙述）之间提供弱关联或全局配对。除了这种全局对齐之外，某些数据集还编码了细粒度的定位信号，将特定区域或时空段与语言实体相连接。这使得模型能够将局部感知与组合语义推理相结合。代表性例子包括 Flickr30k Entities 中的区域到短语对应关系及其边界框 (Plummer et al., 2015)，Visual Genome 中的稠密区域描述和对对象级定位 (Krishna et al., 2016)，以及用于时空视频定位任务的 VidSTG 基准中的时空管状定位 (Zhang et al., 2020)。

### 6.1 生成方法

#### 6.1.1 图像-文本数据

基于大语言模型的图像-文本生成可以根据生成机制大致分为两种范式。第一种范式是原生自回归生成。在此方法中，多模态模型在统一的 token 或嵌入流中生成视觉表示。这包括 Emu 中的连续视觉嵌入回归或 Emu3 中的离散多模态 token 预测 (Sun et al., 2024; Wang et al., 2024c)。第二种范式是外部扩散控制。在此设置中，大语言模型作为独立扩散主干的控制器。模型迭代诊断不匹配并指导修改，以提高输出对提示的遵循程度 (Wu et al., 2023)。

**原生自回归生成。** 该范式将多模态生成视为一个使用混合视觉和文本表示的序列建模任务。在此框架中，图像和其他数据类型被表示为连续嵌入或离散 token。这些元素与文本结合在一个单一的自回归流 (Sun et al., 2024; Wang et al., 2024c; Team, 2025) 中。

例如，Emu 将图像编码为视觉嵌入，并将其与文本 token 混合。模型使用 Transformer 来预测下一个文本 token 或估计下一个视觉嵌入，整个过程在统一序列 (Sun et al., 2024) 中完成。Emu3 进一步发展了这一方法，将图像、文本和视频转换到一个共享的离散 token 空间中。这使得模型能够完全通过混合序列上的下一个 token 预测来工作 (Wang et al., 2024c)。

这些统一的设计支持在单一的自回归模型中同时实现多模态理解和生成。通过使用不同的多模态前缀，一个模型可以完成多种任务，例如图像描述、视觉问答和图像生成 (Sun et al., 2024; Wang et al., 2024c)。Team (2025) 还展示了在直接处理交错的图像和文本 token 长序列时，早期融合建模的有效性。

**外部扩散控制。** 与直接在大语言模型中生成视觉表示不同，该范式将模型用作高层控制器。该控制器将意图和条件信号路由至专门的生成工具，例如扩散模型 (Pan et al., 2024; Koh et al., 2023)。

Kosmos-G (Pan et al., 2024) 等系统首先通过监督对齐将多模态模型的输出与基于 CLIP 锚定的条件接口对齐。随后，应用得分蒸馏指令微调，其中冻结的扩散解码器提供训练信号。该过程使控制器能够生成引导高保真图像合成的表示。

同样，GILL (Koh et al., 2023) 使用轻量级映射模块将冻结的仅文本 LLM 与预训练的图像生成主干网络连接起来。这些模块将模型的隐藏表示转换为生成器的嵌入空间。这使得下游解码器（如基于扩散的生成器）能够渲染请求的图像。

这种模块化设计将意图理解与规划从高保真渲染中分离。在此设置中，大语言模型（LLM）负责规划，而扩散模型骨干则执行渲染。这种分离通过显式的条件接口实现了精确控制。在类似 Kosmos-G 的系统中，视觉生成器可以升级或替换，而控制器基本保持不变。在基于地图的设计如 GILL 中，更换生成器通常只需重新训练桥接模块，而非整个 LLM (Pan et al., 2024; Koh et al., 2023)。

总之，虽然原生自回归生成为推理和渲染提供了一个统一空间，但外部扩散控制能更有效地利用专业的视觉模型和成熟的生成工具。

### 6.1.2 视频-文本建模

视频-文本生成将大语言模型（LLM）的合成范围从静态视觉领域扩展至时空媒体。与图像领域类似，当前的方法主要分为两种不同范式。第一种是原生时空生成，即模型直接在统一的多模态序列中合成视频，并通常包含音频表示。第二种是基于规划器的扩散控制，其中大语言模型负责管理高层视频结构，由分离的渲染引擎执行生成。

**原生时空生成。** 诸如 VideoPoet (Kondratyuk et al., 2024) 和 Emu3 (Wang et al., 2024c) 的方法将视频生成视为时空 token 上的原生语言模型化任务。在 VideoPoet 的情况下，这还包括音频。VideoPoet 使用仅解码器的 Transformer 来处理包括图像、视频片段、文本和音频在内的多模态输入。该模型以自回归方式使用多模态生成目标进行训练 (Kondratyuk et al., 2024)。类似地，Emu3 将视频帧转换为离散潜在空间，并学习预测下一个 token。这种方法有效地将视频理解与生成统一于单一主干网络之下 (Wang et al., 2024c)。

通过将视频甚至音频表示为离散的 token，并使用自回归下一个 token 预测目标对仅解码器的 Transformer 进行训练，这些方法在语言模型风格的主干网络中统一了多模态条件生成。该策略避免了使用基于扩散的生成头。实际上，这些系统仍然依赖于特定的分词器或解码器，有时还会使用额外的模块，例如 token 空间超分辨率。理解诸如字幕生成等能力，取决于任务组合和后训练设置。此外，模型通过序列建模端到端地学习长时序依赖关系，如运动连贯性和场景连续性。

**基于规划的扩散控制** 在基于规划的架构中，如 FlowZero (Lu et al., 2023)、LVD (Lian et al., 2024) 和 VideoDirectorGPT (Lin et al., 2024)，大语言模型充当导演而非渲染器。它将自然语言提示转换为结构化且可解释的中间表示。这些表示包括动态场景语法或时空布局，例如每帧的场景描述、物体布局、边界框轨迹以及背景运动模式。随后，这些规划条件一个独立的基于扩散的视频生成器。

例如，FlowZero 使用由大语言模型生成的动态场景语法来引导逐帧扩散并提升时间上的平滑性。这包

括连贯的对象运动以及可控制的背景和摄像机运动模式 (Lu et al., 2023)。LVD 同样将大语言模型用作时空规划器，输出动态场景布局，这些布局通常为帧一致的边界框序列。这些布局被注入视频扩散模型中，以强制执行空间关系和运动一致性 (Lian et al., 2024)。VideoDirectorGPT 进一步将长提示分解为可编辑的多场景视频计划和一致性分组。它在调用下游扩散模块之前生成场景级和帧级布局，从而实现多场景叙事结构和角色一致性 (Lin et al., 2024)。

通过将高层规划与低层帧合成分离，基于规划的控制为视觉内容和运动动态提供了透明且可编辑的操控方式。该范式还为在调用扩散生成器之前，在规划阶段自然地整合外部约束（如时间线编辑、用户修改或安全规则）提供了锚点。

在两种范式中，都呈现出一种趋同趋势：将结构完整性与来源信息作为生成流水线的原生组成部分，而非事后附加。首先，时间对齐和长时程一致性在长视频生成中越来越被视为主要目标。其次，为解决生成内容的真实性问题，研究人员正在集成诸如内容凭证 (Content Credentials) 等溯源机制，并采用符合 C2PA 标准的加密签名元数据 (Coalition for Content Provenance and Authenticity, 2025)。同时，还引入了鲁棒的不可见水印技术用于视频保护。这包括针对潜在视频扩散模型 (Jang et al., 2025) 的水印设计，以及用于篡改定位和版权保护的视觉-音频水印 (Zhang et al., 2024e)。这些工具实现了机器可验证的归属和篡改追踪。因此，现代视频-文本系统正从松散耦合的工具链演变为更加集成化的框架，能够联合优化表现力、结构一致性与治理能力。

## 6.2 视觉-语言数据的质量度量指标

### 6.2.1 图像-文本

**有效性。** 有效性确保生成的视觉-语言内容既格式正确又可验证。这意味着内容能够在预定义的机器可读结构下进行解析，并检查基本的一致性约束。在实际应用中，有效输出必须遵循目标模式并保持内部引用的一致性。这包括在格式要求时具备稳定的标识符或回合同的指针等特征。通过形式化约束解码，可以减少常见的结构失败，例如格式错误的模式。例如，基于语法的引擎（如 DOMINO）将语法约束与模型的子词词表对齐，以保证机器可读输出的结构正确性 (Beurer-Kellner et al., 2024)。

给定一组视觉-语言生成结果  $\mathcal{M}_{\text{gen}} = \{m_1, \dots, m_N\}$  和一个模式  $S$ ，我们将格式正确率定义为：

$$\text{WFR} = \frac{1}{|\mathcal{M}_{\text{gen}}|} \sum_{m \in \mathcal{M}_{\text{gen}}} V(m, S),$$

where  $V(m, S) \in \{0, 1\}$  indicates whether  $m$  conforms to  $S$ .

**忠诚** 语义级保真度衡量生成的交错序列中文本与非文本模态之间的逻辑和上下文一致性  $m = \{(t_\ell, v_\ell)\}_{\ell=1}^L$ 。在此记号中， $\ell$  代表单个视觉-语言样本内的各个步骤。一种常见方法是使用对齐函数  $\mathcal{S}_{\text{align}}$  对每一对进行评分。在实际应用中，该函数通常以以下两种方式之一实现。第一种方式使用强大的多模态判别模型，如 GPT-4o。第二种方式采用嵌入空间对齐，例如 CLIP 文本特征与图像特征之间的余弦相似度，这也被称为文本对齐 (Ham et al., 2024)。我们按如下方式聚合逐步骤的对齐结果：

$$\text{ITA}(m) = \frac{1}{L} \sum_{\ell=1}^L \mathcal{S}_{\text{align}}(t_\ell, v_\ell).$$

The CoMM evaluation framework also reports an Image-Text Alignment score using strong judge models for a similar purpose(Chen et al., 2025b).

除了全局对齐之外，实例级保真度衡量的是特定属性的保留准确性。按照评估主体驱动生成的标准协议 (Ruiz et al., 2023)，我们计算源主体图像  $V_{\text{subj}}$  与生成图像  $V_{\text{gen}}$  之间的主体保真度。该指标通过图像嵌入提取器  $E_{\text{img}}(\cdot)$  (如 CLIP 图像编码器或 DINO) 进行计算：

$$\text{Fidelity}_{\text{subj}} = \text{sim}(E_{\text{img}}(V_{\text{gen}}), E_{\text{img}}(V_{\text{subj}})),$$

In this equation,  $\text{sim}(\cdot, \cdot)$  is typically cosine similarity. We capture Prompt Fidelity, which refers to how well the model follows the prompt, using CLIP-style alignment between text and images (Ruiz et al., 2023):

$$\text{Fidelity}_{\text{prompt}} = \text{sim}(E_{\text{img}}(V_{\text{gen}}), E_{\text{text}}(T_{\text{prompt}})).$$

当存在参考数据时，我们还报告标准文本指标，如 ROUGE 和 METEOR。对于图像评估，我们使用 Frechet Inception Distance 和 Inception Score 等指标来衡量图像质量。在适用的情况下，我们还使用结构相似度指数衡量风格一致性，以及峰值信噪比衡量重构效果。这些指标均相对于参考文本  $T_{\text{ref}}$  或参考图像  $V_{\text{ref}}$  (Chen et al., 2025b) 进行测量。

我们使用带图像相关性的字幕幻觉评估 (CHAIR) 指标来衡量模型提及图像中不存在物体的频率 (Rohrbach et al., 2019)。我们从字幕中提取提及内容，并与微软常见上下文物体数据集中的八十个物体类别进行比较。如果某项提及未出现在真实值标签中，则视为幻觉。

该度量提供两个得分。物体级别率是指所有提及中幻觉出现的比例。句子级别率是指至少包含一个幻觉的句子所占的比例。得分越低，表示事实一致性越高。

$$\text{CHAIR}_i = \frac{\text{Number of hallucinated objects}}{\text{Total object mentions}}$$

$$\text{CHAIR}_s = \frac{\text{Number of sentences with hallucinations}}{\text{Total number of sentences}}$$

**实用性。** 效用衡量模型在下游任务（如图像描述生成和视觉问答系统）中的价值，同样适用于工具增强的工作流。这通常通过下游任务的表现以及自动或人工评判来评估。统一的早期融合模型，如 Chameleon，展示了在混合模态序列上的交错推理与生成 (Team, 2025)。类似地，Anole 提供了一个开源的自回归模型及其训练框架 (Chern et al., 2024)。交错图像与文本训练数据，如 CoMM (Chen et al., 2025b)，以及基于指令的多轮对话数据，如 InterSyn (Feng et al., 2025a)，进一步支持监督式指令微调与评估。

令  $D = \{(q_i, a_i)\}_{i=1}^N$  表示一个基准数据集，该数据集包含  $N$  个问题与答案对，其中  $q_i$  为第  $i$  个问题， $a_i$  为其对应的真实值答案。我们使用符号  $|D|$  来表示数据集的大小，其值等于  $N$ 。设  $M$  为将输入问题  $q_i$  映射到预测答案的模型或推理函数。

我们使用指示函数，当其参数为真时取值为 1，否则为 0。等式谓词  $M(q_i) = a_i$  用于比较预测答案与参考答案是否完全匹配。当任务为开放式问题时，该过程可选地在应用标准规范化处理（如转换为小写、



去除标点符号) 后进行。问答系统的准确率是数据集中所有样本的指示函数值的平均值:

$$\text{Acc}_{\text{QuestionAnswering}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(M(q_i) = a_i)$$

对于实用性的人工评估, 我们采用成对偏好协议, 遵循 Chameleon 中的相对评估情景 (Team, 2025)。给定相同的输入, 标注者会以随机顺序看到我们模型  $M$  和基准  $B$  的两个匿名响应。随后, 标注者从以下选项中选择一项:  $M$  更好,  $B$  更好, 或大致相当。我们将  $M$  输出更受青睐的情况定义为胜局, 原因是其任务完成度和实用性更优。类似地, 当  $B$  被更受青睐时则为败局, 当两者输出被判断为可比时则为平局。我们通过为胜局计 1 分、平局计 0.5 分来计算胜率 (Team, 2025):

$$\text{WinRate} = \frac{W + 0.5 \times T}{W + T + L},$$

其中  $W$ 、 $T$  和  $L$  分别表示  $M$  对  $B$  的胜场数、平局数和损失数。

**多样性** 多样性度量在保持文本与图像一致性的前提下, 对实体、属性、风格和布局的值域进行评估。我们从三个不同层次衡量多样性: 第一层是整个样本集的分布多样性; 第二层是单个样本内的序列内多样性; 第三层是基于判别器对多模态输出的多样性评估。

从架构角度来看, MM-Interleaved 通过在生成过程中同步访问细节视觉特征, 提升了多图像交错生成的效果 (Tian et al., 2024)。Chameleon 使用统一的 token 空间 (Team, 2025)。InterSyn 通过由指令驱动的多轮对话提升覆盖范围。这些对话是在数据创建过程中通过一种称为迭代精炼的流程生成的 (Feng et al., 2025a)。

令  $\mathcal{V}$  为生成图像的集合。根据 Salimans et al. (2016), 我们使用 Inception Score 来评估样本质量和多样性。具体而言, 我们对每个生成图像  $v \in \mathcal{V}$  应用预训练的 Inception 分类器, 以获得条件标签分布  $p(y | v)$ 。

包含有意义且可识别物体的图像往往会产生熵较低的条件标签分布, 此时分类器表现出高置信度。相反, 避免模式崩溃的生成器应产生多样化的图像, 使得边缘标签分布  $p(y) = \mathbb{E}_{v \in \mathcal{V}}[p(y | v)]$  具有较高的熵。这确保了预测结果分散在多个不同的类别中 (Salimans et al., 2016)。结合这两个目标, Inception Score 定义为:

$$\text{IS}(\mathcal{V}) = \exp \left( \mathbb{E}_{v \in \mathcal{V}} [D_{\text{KL}}(p(y | v) \| p(y))] \right).$$

等价地,  $\log \text{IS}(\mathcal{V}) = H(p(y)) - \mathbb{E}_{v \in \mathcal{V}}[H(p(y | v))]$ 。第二项通过低条件熵鼓励锐度和可识别性。第一项通过奖励边缘分布的高熵来鼓励多样性,  $D_{\text{KL}}$  表示 KL 散度。这表示对 Inception 分类器预测的语义类别具有广泛的覆盖, (Salimans et al., 2016)。

令  $m_v = \{v_\ell\}_{\ell=1}^L$  为一个视觉序列, 例如一系列生成的图像或视频帧。在该序列中,  $v_\ell$  表示位置  $\ell$  处的视觉元素,  $L$  表示序列的总长度。令  $E$  为一个特征提取器, 可将视觉输入转换为嵌入向量。该提取器可以是图像编码器, 例如对比语言图像预训练 (Contrastive Language Image Pretraining) 或无标签自蒸馏 (Self Distillation with no Labels)。sim 表示嵌入之间的相似度函数, 该值通常由余弦相似度确定。

我们将序列内多样性定义为 1 减去序列中所有无序对的平均成对相似度，公式如下：

$$\text{Div}_{\text{seq}}(m_v) = 1 - \frac{2}{L(L-1)} \sum_{\ell=1}^L \sum_{r=\ell+1}^L \text{sim}(E(v_\ell), E(v_r)).$$

令  $\mathcal{M}_{\text{generations}}$  为一组多模态生成结果，令  $|\mathcal{M}_{\text{generations}}|$  代表该集合中的总项目数。对于生成集合中的每个样本  $m$ ，项  $\mathcal{J}_m$  代表其多样性得分。该得分由评判模型或人工评分员或自动评分器给出，用于评估输出的多样性。我们通过取整个生成集合中这些得分的平均值，计算出基于评判的聚合多样性得分，从而获得最终的多样性得分。

$$\text{Score}_{\text{diversity}} = \frac{1}{|\mathcal{M}_{\text{generations}}|} \sum_{m \in \mathcal{M}_{\text{generations}}} \mathcal{J}_m$$

### 6.2.2 视频文本

**有效性。** 对于视频-文本样本，我们通过一组可验证的子属性来定义其有效性。这些属性包括模式级别的结构正确性、时间上的正确性以及跨模态绑定的一致性。令  $\mathcal{M}$  为生成的  $N$  个视频和文本样本组成的集合，该集合的大小为  $N$ 。每个样本  $m_i$  由一个视频  $v_i$  和一个配对的文本  $t_i$  组成，模态数量为两个。

令  $\mathbb{S}$  表示一种机器可读的模式，例如带有字段约束的类型化 JavaScript 对象表示法或另一种标记语言模式，用于验证结构化输出。术语  $m_i$  所表示的函数用于检查样本  $m_i$  是否符合模式  $\mathbb{S}$ ，若符合则返回真，否则返回假。令  $\mathbb{I}$  为指示函数，当其参数为真时返回 1，否则返回 0。我们将模式符合率定义为通过模式验证的生成样本所占的比例：

$$\text{SchemaAdherenceRate} = \frac{1}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M}} \mathbb{I}[\text{conforms}(m_i, \mathbb{S}) = \text{true}]$$

在时间正确性评估中，我们遵循时间语言定位任务（如 TALL (Gao et al., 2017)）中已建立的评估协议。我们使用在不同交并比阈值下的召回率来衡量正确性。在此框架中， $T_{\text{pred},i}^{(j)}$  表示样本  $i$  根据模型得分排序后的第  $j$  个预测区间， $T_{\text{gt},i}$  表示真实值区间。对于给定的交并比阈值  $\delta$ ，我们定义 Recall @ K 指标如下：

$$\text{R@K}(\delta) = \frac{1}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M}} \mathbb{I}\left(\max_{1 \leq j \leq K} \text{IoU}(T_{\text{pred},i}^{(j)}, T_{\text{gt},i}) \geq \delta\right).$$

In practice, we report this metric for standard choices such as  $K$  values of 1, 5, or 10 and threshold values of 0.3, 0.5, or 0.7. Additionally, a comprehensive temporal score can be calculated by taking the average across a specific set of thresholds  $\Delta$ :

$$\text{R@K-Avg} = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \text{R@K}(\delta).$$

关于跨模态绑定一致性，我们将跨模态绑定视为一种可通过表示一致性验证的有效性属性。令  $f_T$  和  $f_V$  分别为文本编码器和视频编码器，令  $\phi_{\text{sim}}$  表示余弦相似度。对于每个样本，我们假设一个与文本  $t_i$  相

关联的实体视频片段  $v_i^{\text{seg}}$ 。则跨模态一致性得分定义如下：

$$\text{CMC} = \frac{1}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M}} \phi_{\text{sim}}(f_T(t_i), f_V(v_i^{\text{seg}})).$$

This equation naturally extends to aligned speech or text regions by using automatic speech recognition transcripts or optical character recognition snippets when they are available.

**忠诚** 在视频保真度方面，这一属性反映了时空一致性与视觉真实感。令  $F$  表示  $v_i$  中的帧数，令  $\tau$  表示帧索引。为了评估跟踪物体区域  $\{o_{i,\tau}\}$  内对象身份的一致性，我们使用一个视觉特征提取器  $f_E$ ，例如 CLIP，以及一个相似度函数  $\phi_{\text{sim}}$ 。身份一致性得分定义如下：

$$\mathcal{F}_{\text{ID}}(m_i) = \frac{1}{F-1} \sum_{\tau=1}^{F-1} \phi_{\text{sim}}(f_E(o_{i,\tau}), f_E(o_{i,\tau+1})).$$

整体真实性与分布忠实性通常由弗雷歇视频距离 (Frechet Video Distance) 来概括。该度量最初是为生成视频模型的评估而提出的 (Unterthiner et al., 2019)。如今，它在现代文本到视频评估中被广泛报告，包括 VideoPoet 和各种基于规划的系统 (Kondratyuk et al., 2024; Lin et al., 2024)。

此外，我们可以从专门的预测器中聚合帧级或片段级的质量得分，以定义一个通用的视频质量指标：

$$\mathcal{F}_{\text{VQ}}(m_i) = \frac{1}{F} \sum_{\tau=1}^F \Psi(v_{i,\tau}),$$

In this equation,  $\Psi$  represents one or more automated evaluators that target dimensions such as imaging and aesthetics. These evaluators are typically organized in standardized suites such as VBench (Huang et al., 2024b).

**实用性。** 效用表示合成视频和文本数据在支持可控生成及下游使用方面的可靠性。对于生成结构化计划或提示的规划器风格流水线（如 VideoDirectorGPT (Lin et al., 2024)），我们令  $P$  表示一组计划或提示， $v_{\text{out}}^{(j)}$  为基于特定计划  $p_j$  生成的输出视频。受 VideoDirectorGPT 等基于规划器的流水线启发，我们定义如下面向可控性的任务成功率：

$$\mathcal{U}_{\text{Control}} = \frac{1}{|P|} \sum_{p_j \in P} \mathbb{I}(\text{eval}_{\text{task}}(v_{\text{out}}^{(j)}, p_j) = \text{true}),$$

In this formulation, the task evaluation function can be implemented through automatic checkers, learned judges, or human evaluation. The choice of method depends on the controllability constraints that are encoded by the plan  $p_j$ .

对于涉及特定风格或功能的情况，我们使用基于嵌入的代理来衡量有用性：

$$\mathcal{U}_{\text{Func}} = \phi_{\text{sim}}(f_T(t_{\text{style}}), f_V(v_{\text{out}})),$$

This metric measures whether the generated video matches a target style or condition within a shared

representation space.

**多样性** 关于视频多样性，该度量指标反映了运动的多样性与语义的广度。我们定义  $\mathcal{V} = \{v \mid (v, t) \in \mathcal{M}\}$  为生成视频的集合。

为了评估动态多样性，我们使用基于光流的运动得分来度量运动幅度 (Teed & Deng, 2020)。该得分通过以下公式计算：

$$\text{motion\_score}(v) = \frac{1}{T-1} \sum_{\tau=1}^{T-1} \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \|\mathbf{f}_{\tau \rightarrow \tau+1}(\mathbf{x})\|_2,$$

在此公式中， $f_{\tau \rightarrow \tau+1}$  表示从帧  $\tau$  到帧  $t+1$  在像素域  $\Omega$  上的光流场， $T$  表示帧数。在实际应用中，我们使用 RAFT 等方法计算光流，并对所有帧的光流幅值进行平均，以确定每段视频的运动得分 (Liao et al., 2024)。

在语义多样性方面，我们采用类似于视频上的 Inception Score 的度量方法。该方法在视频生成评估中被广泛采用。遵循标准的 Inception Score 公式 (Salimans et al., 2016)，我们将该度量定义如下：

$$\mathcal{D}_{\text{sem}} = \exp(\mathbb{E}_{v \sim \mathcal{V}} [D_{KL}(p(y|v) \parallel p(y))]),$$

In this definition,  $p(y|v)$  is the label distribution predicted by a pretrained classifier (Saito et al., 2017). The term  $p(y)$  represents the marginal label distribution.

## 6.3 可信的视觉-语言数据度量指标

### 6.3.1 图像-文本

**安全** 安全旨在防止在不同模态下生成有害内容。它还致力于减少多模态攻击面，例如基于图像的提示攻击以及基于图像的越狱攻击 (Liu et al., 2024c)。MM-SafetyBench 框架研究了这些提示攻击，并表明与查询相关的图像会显著提高攻击的成功率。这一发现支持了如下假设：与查询相关的图像会激活视觉与语言的对齐模块。由于该模块通常在缺乏特定安全约束的情况下进行训练，其激活可能会削弱模型识别有害请求的能力 (Liu et al., 2024c)。

此外，基准测试指出，较低的攻击成功率并不总是明确的。这可能表明模型在安全性方面已正确对齐，也可能是模型未能理解恶意查询所致。为解决这种不确定性，MM-SafetyBench 强调拒绝行为。这种方法有助于区分那些识别并拒绝有害请求的模型与因缺乏理解而无法遵守请求的模型。在实践中，在保持预期拒绝行为的同时实现较低的攻击成功率仍是一项艰巨任务 (Liu et al., 2024c)。

基准测试还表明，对于指令遵循能力较强的模型，添加安全提示可以显著降低攻击成功率。而对于其他模型，这些改进较小。作者指出，安全提示的有效性取决于潜在模型对指令的遵循程度 (Liu et al., 2024c)。

遵循 MM-SafetyBench，我们在查询集  $D$  上报告攻击成功率 (ASR) 和拒绝率 (RR)。

$$\text{ASR} = \frac{\sum_{q \in D} I(q)}{|D|}, \quad \text{RR} = \frac{\sum_{q \in D} R(q)}{|D|}.$$

Here  $I(q) = 1$  if and only if the model engages with the disallowed request under query  $q$ , for example



by providing prohibited content or actionable instructions; otherwise  $I(q) = 0$  (Liu et al., 2024c). 我们设定  $R(q) = 1$  当且仅当模型以拒绝响应不安全查询开头，使用 MM-SafetyBench 中的拒绝定义；否则  $R(q) = 0$  (Liu et al., 2024c)。

在实现中，拒绝检测可以通过预定义的拒绝模式匹配器来实例化，并可通过一个判别器  $\Phi_S$  进一步增强，以实现更鲁棒的拒绝识别。我们报告了 RR 在所有查询中的表现， $D$ 。为了量化系统层面的过度敏感性，我们还单独评估了一个良性子集  $D_{\text{benign}}$ ，并报告  $\text{RR}_{\text{benign}}$ ，即在良性查询上的拒绝率。这作为实际的系统级代理，受到诸如 MMSafeAware (Wang et al., 2025) 等安全意识基准的启发，这些基准包含良性子集，用于捕捉被错误视为不安全的无害请求。MMSafeAware 关注安全意识与识别，而  $\text{RR}_{\text{benign}}$  则将下游在良性输入下的拒绝行为具体化。

**来源** 关于出处，此属性用于追踪多模态输出的可验证原点、归属声明以及修改历史。C2PA 版本 2.2 提供了签名内容凭证，以支持可审计的出处链 (Coalition for Content Provenance and Authenticity, 2025)。此外，SynthID-Image 为人工智能生成的图像提供了不可见的水印技术，以促进检测与可追溯性 (Gowal et al., 2025; Google DeepMind, 2025)。我们定义在生成的多模态样本集合上的凭证验证率为：

$$\text{Rate}_{\text{validate}} = \frac{1}{|\mathcal{M}_{\text{gen}}|} \sum_{m \in \mathcal{M}_{\text{gen}}} V_{\text{c2pa}}(m),$$

In this definition, the indicator function  $V_{\text{c2pa}}$  results in a value of either zero or one. The value is zero when no C2PA manifest or credentials are present. The value is one when the validator returns a manifest state that is at least Valid, or Trusted under a stricter setting, depending on the chosen trust model. To clarify, the manifest states of Well-Formed, Valid, and Trusted reflect three distinct properties. These include the well-formedness and cryptographic integrity of the manifest, the verification of the signature, and the trust status of the signing credential under the adopted trust configuration (Coalition for Content Provenance and Authenticity, 2025).

### 6.3.2 视频文本

**安全** 有害内容率衡量在对抗性评估中未能阻止违反政策的视频生成的失败程度。该方法受到 T2VSafetyBench 的启发，后者创建了恶意提示，例如真实世界提示、由 GPT-4 生成的提示以及基于越狱攻击的提示。该框架使用基于 GPT-4 的评估结合人工审查进行验证。我们通过一个安全评估流水线  $\Phi_S$  定义有害内容率。该流水线可能涉及大模型判断，以及可选的人工审查和检测器检查 (Miao et al., 2024)。

我们将其定义如下：

$$\text{HCR} = \frac{|\{v \in \mathcal{V}_{\text{adv}} \mid \text{is\_unsafe}(v, \Phi_S) = \text{true}\}|}{|\mathcal{V}_{\text{adv}}|},$$

In this equation,  $\mathcal{V}_{\text{adv}}$  represents the set of all successfully generated videos under adversarial prompts. Beyond measuring the number of unsafe generations, T2VSafetyBench provides three findings that are useful for system-level safety evaluation. First, no single model performs well across all safety aspects. Second, assessments by GPT-4 correlate well with manual reviews, which supports the use of large-scale automatic judging. Finally, a trade-off exists between model capability and safety. This suggests that safety risks may increase as video generation technology improves (Miao et al., 2024).

**来源** 基于水印的溯源评估生成视频在载荷与质量之间的鲁棒性及权衡。当使用由  $L$  比特组成的原始水印  $w$ ，以及攻击  $\mathcal{A}$  和恢复过程  $\mathcal{R}$  时，我们通过比特准确率来衡量鲁棒性。对于生成视频  $v$ ，其定义如下：

$$\mathcal{P}_{\text{Rob}} = 1 - \frac{d_H(w, \mathcal{R}(\mathcal{A}(v)))}{L}.$$

We define the payload and fidelity cost using the following equations:

$$\mathcal{P}_{\text{Load}} = L, \quad \mathcal{P}_{\text{Cost}} = \Delta(V_{\text{ori}}, v).$$

In these expressions,  $\mathcal{P}_{\text{Load}}$  represents the payload per video in bits. The variable  $V_{\text{ori}}$  represents the corresponding baseline video that was generated without a watermark using the same prompt and seed. The term  $\Delta$  measures the perceptual distortion that is introduced by the process of embedding the watermark. For example, this can be calculated by averaging the frame-wise Learned Perceptual Image Patch Similarity over sampled frames to measure frame-level distortion. One can also use the Frechet Video Distance to evaluate quality degradation at the level of the distribution.

这些估计值可与签名凭证和不可见水印技术结合用于视频内容。具体示例包括用于潜在视频扩散模型水印的 LVMark，以及用于视觉和音频水印并具备篡改定位功能的 V2A-Mark。这些工具支持在下游任务中的可追溯性和检测。此外，它们还可补充具有密码学可验证性的溯源机制，例如 C2PA (Coalition for Content Provenance and Authenticity, 2025; Jang et al., 2025; Zhang et al., 2024e)。

## 6.4 评估实践差距

我们分析第 6.1 小节中的代表性方法，并在表 6 中对它们报告的评估协议进行分类。

Generation method	Validity	Fidelity	Diversity	Utility	Safety	Provenance
Kosmos-G (Pan et al., 2024)	×	✓	×	△	×	×
GILL (Koh et al., 2023)	×	✓	×	✓	×	×
LVD (Lian et al., 2024)	△	✓	×	✓	×	×
FlowZero (Lu et al., 2023)	△	✓	×	△	×	×
VideoDirectorGPT (Lin et al., 2024)	✓	✓	✓	✓	×	×
SynthID-Image (Gowal et al., 2025)	×	✓	×	×	×	✓

**Table 6** 是否代表性视觉-语言方法在其实验部分显式评估了各个维度。✓：显式评估；△：部分/间接涉及；×：未报告或不适用。

**多样性**。与保真度相比，多样性在视觉与语言生成相关研究中的评估并不一致。一个实际原因是，通常报告的多样性代理指标（如 Inception Score）将质量与覆盖度结合在一起。然而，更强的度量方法，如序列内嵌入多样性、基于判断的多样性，或基于运动熵的视频多样性，则需要额外的建模选择。这些度量还需要可靠的特征提取器或评判者，并对提示分布进行仔细控制。此外，多样性与对齐性密切相关。以简单方式增加多样性可能会降低文本与图像之间的一致性。因此，许多论文选择优先考虑忠实还原而非广泛的覆盖。结果，多样性常被视为定性声明或次要代理。这为分离广度与对齐性的标准化、可复现的多样性评估面板留下了空白。

**安全**。尽管视觉-语言模型正被广泛使用，但大多数评估仍主要关注保真度指标，如真实性和对齐性。相比之下，安全问题往往被视为次要关注点。通常仅通过间接措施（如简短的定性讨论或数据过滤）来处

理，而非以系统化的方式进行测量。一个主要挑战是，多模态安全涉及广泛的潜在攻击类型。例如，基于图像的视觉提示注入和逃逸攻击。此外，攻击成功率较低可能难以解释：这一结果可能意味着模型是安全的，也可能仅仅说明模型未能理解恶意查询。最后，安全评估有两个目标：既要阻止有害响应，又要避免拒绝安全的请求。这种平衡使得标准化安全报告既困难又成本高昂。

**来源信息.** 在标准生成类论文中，来源信息度量很少被报告。这主要是因为测量来源信息需要额外的基础设施，例如密码学凭证或水印系统，同时也需要在现实世界中的变化（如压缩、裁剪、重新编码或编辑）下对模型进行测试。然而，随着生成式模型能力的不断增强以及合成内容的广泛传播，通过稳健且可度量的水印来建立可验证的来源信息，对于内容追踪、确保责任归属以及建立信任正变得至关重要。

## 6.5 用途

近期的方法论越来越多地在四个主要功能范式中使用合成视觉与语言数据。这一趋势在很大程度上独立于潜在的生成基础架构。这些范式包括监督微调、基于偏好对齐和奖励建模、弱监督下的数据监管与修复，以及以安全或结构为中心的评估。在这些情境中，合成图像与文本、视频与文本，以及文档视觉与语言样本不仅作为补充数据，还作为可扩展的监督来源、明确的对齐信号，以及用于评估鲁棒性与安全性的可控探针。

**监督微调.** 大量研究工作已聚焦于通过合成描述和自动化提示整理构建大规模、结构化的多模态指令数据集，以推动监督微调。总体而言，这些举措在两个主要维度上扩展了监督微调的规模：第一个维度是模态覆盖，从静态图像扩展到视频；第二个维度是语义粒度，从通用描述演进到实例级及包含 OCR 感知的指令。在这两个维度上的进展均显示出与下游性能提升之间存在显著相关性。

在视频领域，LLaVA-Video-178K 引入了一个大规模的合成指令跟随语料库，该语料库由 GPT-4o 辅助并包含人类参与的流水线。该数据集包含详细的描述、开放式问答以及多项选择任务，显著提升了视频大模型训练中的性能表现 (Zhang et al., 2025e)。

为应对富含文本的图像场景，TextSquare（又称 Square-10M）利用闭源多模态大模型，将合成的图像与文本指令扩展至数千万级别。当此方法用于大规模监督微调时，其性能随着数据规模的增加呈现出近乎单调的增长趋势 (Tang et al., 2025)。

同样地，LLaVAR-2 将人工标注的描述与 GPT-4o 生成并过滤后的指令相结合，以构建高质量且以文本为中心的监督微调数据集 (Zhou et al., 2024)。为了实现实例级定位的目标，Inst-IT 为连续的监督微调合成显式的视觉提示指令。该方法增强了实例定位的理解能力，并在通用基准上表现出良好的迁移性能 (Peng et al., 2024)。同时，稠密且高保真度的合成描述已被证明在监督学习中非常有效。例如，ShareGPT4V 和 ShareGPT4Video 表明大规模合成描述直接有利于监督微调以及统一的大规模视觉-语言模型预训练 (Chen et al., 2023b, 2024a)。

**偏好对齐与奖励建模.** 除了标准的监督微调之外，合成数据在构建成对偏好和训练奖励模型中起着关键作用。对于视频生成，VideoDPO 提出了一种自动流水线，为每个提示生成多个视频，并使用一种称为 OmniScore 的多维度评分指标对其进行打分。该流水线形成由最佳样本和最差样本组成的得分排序偏好对。这些样本对在直接偏好优化过程中进一步被重新加权，以优先考虑具有显著差异性和高影响力的样本 (Liu et al., 2024a)。

同样，题为《利用人类反馈改进视频生成》的工作利用了关于合成视频的多维度人类偏好来训练一个 VideoReward 模型。该奖励模型随后通过直接偏好最优化目标和奖励加权推理 (Liu et al., 2025a) 来指导生成器。

在视觉与语言领域，VLFeedback 提供了超过 82,000 条由人工智能标注的反馈实例。这些实例包含在有用性、视觉忠实度以及伦理与安全方面的偏好标签和理由。将直接偏好最优化应用于 VLFeedback 以训练 Silkie 模型，提升了有用性、视觉忠实度和安全相关指标。实验结果表明，将此最优化方法应用于此类反馈可显著提升模型的有用性、视觉忠实度和安全性 (Li et al., 2024b)。

**数据监管、错误修正和故障目标合成。** 合成数据在优化噪声真实的现实世界数据、修复弱标签以及生成困难负例以暴露模型缺陷方面也至关重要。例如，CapFusion 利用大语言模型 (LLM) 整合并精炼来自噪声网络图像与文本对以及模型生成的合成描述信息，该过程为多模态预训练提供了更高质量且更具可扩展性的监督信号 (Yu et al., 2024b)。

在结构化图形领域，CHOCOLATE 提供了一个由人工标注的基准以及图表标题中事实性错误的分类体系。该工作还确立了图表标题的事实性错误修正任务。此外，作者提出了 ChartVE，这是一个无需参考的视觉蕴含度量指标，用于评估图表与标题之间的事实一致性。同时，作者提出了 C2TFec，一个可解释的两阶段框架，该框架先将图表转换为表格，再利用大语言模型纠正其中的事实性不一致 (Huang et al., 2024a)。

在文档理解方面，SynthDoc 生成包含文本、表格和图表的双语文档图像。本研究证明，在此类数据上训练类似于 Donut 模型的无 OCR 解析器，即使存在语言不一致的情况，也能提升预训练阅读任务性能和下游鲁棒性 (Ding et al., 2024)。

**安全对齐与评估。** 除了通用能力之外，合成数据对于红队测试技术至关重要。这些技术涉及生成有害和有益的响应以及偏好对，以压力测试视觉和语言模型的安全性。例如，SPA-VL 创建了包含六种特定危害领域选定和拒绝响应的 100,000 个图像-指令四元组。该数据集旨在支持基于近端策略优化或直接偏好优化的安全训练 (Zhang et al., 2025d)。除了基于偏好的对齐外，红队测试还要求对可能导致间接不安全行为的故障模式进行压力测试。这包括包含大量文本的视觉场景，其中光学字符识别和布局误解可能导致有害的解读。在评估方面，OCRBench v2 通过人工验证的问题扩展了富含文本的场景，以严格检验大型多模态模型的光学字符识别和结构理解能力 (Fu et al., 2025)。

基于上述讨论，将这些不同角色结合起来表明，合成视觉与语言数据已演变为后训练阶段的基础性基础设施。这类数据不仅能够扩展监督式指令学习的规模，还为对齐提供了明确信号。此外，它还能修复弱跨模态监督问题，并在大模型驱动的生态系统中支持严格的安全性和结构性评估。

## 7 智能体数据

在数字孪生与具身人工智能的框架下，我们从实际应用中最终使用的数据产品角度，分析由大模型生成的智能体数据。这种研究方法源于世界模型作为内部模拟器的观点，其能够捕捉环境动态，并支持感知、预测和决策的前向以及反事实推演 (Li et al., 2025b)。

我们将数字孪生和具身人工智能中使用的实际数据产品分为三类。第一类是环境与任务数据，用于描述世界设置和面向任务的场景配置 (Ruan et al., 2025)。第二类为控制与决策数据，用于捕捉策略、动作



迹以及其他控制决策，以引导智能体和仿真。这包括大模型多智能体参数化轨迹以及数字孪生仿真中的顺序控制计划 (Xia et al., 2024)。第三类是感知与遥测数据，重点关注观测到的传感器流和日志。

该分类体系的一个关键特征是，它根据数据的主要用途对其进行分类。例如，混合数据（如轨迹日志）会根据其主要用途被分配到相应类别：是用于设计测试平台、训练智能体行为，还是用于监控系统健康状况。基于这一数据产品视角，典型体现于具身智能体的自我中心流与常见于数字孪生的外部中心遥测数据，可以根据其主要应用整合到同一分类体系中。这种分类方法仍与更广泛的的世界模型视角保持一致，该视角将感知、动力学和控制跨不同领域联系起来 (Li et al., 2025b)。

## 7.1 生成方法

**环境和任务数据。** 环境和任务数据包含初始化一个回合所需的规定。这些规定定义了环境的外观、存在的实体或智能体，以及构成有效仿真运行的目标和约束。

在城市交通仿真情景中，大模型将自由形式的请求解析为结构化的关键词，例如情景字段的字典。这些关键词随后驱动生成可用于城市出行测试的可执行配置和参数 (Li et al., 2024c)。多模态流水线通过生成真实且罕见的边缘情况以及针对驾驶场景长尾部分的可运行测试，将这一概念扩展至罕见和困难的情景 (Lu et al., 2024)。其他互补的流水线则将操作设计域描述转换为与 ScenarioRunner 兼容的脚本，用于基于仿真的测试 (Danso & Büker, 2025)。

除了交通应用之外，语义数字孪生还利用大模型将领域概念与任务级计划及恢复策略相连接。在这些系统中，数字孪生提供领域概念和交互规则，而大模型则生成可在孪生体内部执行和监控的结构化动作描述与恢复行为 (Naeem et al., 2025)。在企业层面，面向工业 5.0 的工作提出了一个交互式数字孪生 (Interactive-DT) 框架。在此框架中，大模型作为边缘、数字孪生和服务层之间的交互接口和智能体，支持数字孪生的构建与运行、云与边缘系统的协作以及高级数据分析。该研究还识别出集成挑战，例如由幻觉引发的不可靠性可能带来安全影响，以及偏见、推理速度约束、互操作性以及符合标准的安全部署等问题 (Chen et al., 2025a)。

在具身人工智能领域，长时域规划器如 L3M+P 生成 PDDL 问题并维护世界状态图，这些状态图可实例化为多种具体的任务和目标 (Agarwal et al., 2025b)。以安全为重点的规划器如 SELP 将自然语言任务转换为时序逻辑规范，并利用等价投票机制提升从自然语言到逻辑映射的鲁棒性和一致性。随后，约束规划在执行过程中施加相应的安全约束 (Wu et al., 2025)。自纠正框架如 T3 规划器会依据时空逻辑验证计划，并在需要时进行修复。这一过程生成经过验证的计划与轨迹迹，可用于在显式约束下的评估或学习监督 (Li & Zhao, 2025)。大规模多智能体基准测试如 PARTNR 也属于此类。在这些基准中，大模型 (LLMs) 协助设计和分解协作任务，将其在模拟器中具体化，并导出为标准任务集合以及规划器基准 (Chang et al., 2024)。

**控制与决策数据。** 控制与决策数据描述了在环境和任务确定后，智能体如何行动。在数字孪生情景中，大语言模型智能体可以与仿真器形成闭环运行。这些智能体通过数据接口读取仿真数据，并通过控制接口输出参数化的应用程序编程接口或函数调用，通常以 JSON 格式序列化，以调整仿真参数。随后，它们会迭代地总结结果，用于下一循环。

运行这些智能体可生成包含仿真数据、控制参数和智能体摘要的周期级迹。这些迹被记录为仿真日志，可编译成简洁的顺序控制或参数化计划，用于离线分析和假设情景评估 (Xia et al., 2024)。在电力系统

等安全关键基础设施中，由大模型协调的控制策略在数字孪生沙箱内执行，数值求解器验证其稳定性和安全性。由此产生的控制序列和轨迹支持压力测试与控制研究，且不会影响实际系统 (Zhang et al., 2025c)。

对于以人为中心的数字孪生，基于大模型的人物角色仿真可作为数字实验的硅基样本。学术界可利用其进行试点实验，以识别具有影响力的刺激因素，而企业则可探索客户洞察与产品开发策略 (Toubia et al., 2025)。基于人物角色的行为链基准（如 BehaviorChain）量化了当前大模型在忠实模拟连续人类行为方面的能力程度 (Li et al., 2025a)。

在具身人工智能中，大模型也可以更直接地增强示范和轨迹数据。例如，LLM Trainer 通过离线示范标注和在线关键姿态重定向，将示范适应到新场景，并在极少人工输入的情况下生成额外的模仿轨迹 (George & Farimani, 2025)。诸如 ELLMER 等框架将高层语言驱动的任务分解与低层执行相结合，利用视觉或力反馈以及检索增强的代码示例来解决长时程任务。由此产生的执行过程自然生成丰富的交互迹，可用于记录与分析 (Mon-Williams et al., 2025)。程序结构化方法如 Instruct2Act 和 ProgPrompt 利用可执行或类程序表示，将指令与可用动作对齐，并在某些系统中整合感知与规划环。这使得能够对各组件的必要性进行更模块化的分析 (Huang et al., 2023; Singh et al., 2022)。当执行日志同时包含动作和丰富观测时，若其主要用途是研究或改进行为而非感知，则本综述将其视为控制与决策数据。

**感知与遥测数据。** 感知与遥测数据关注于在数字孪生和具身环境中所观测和记录的内容，以及这些观测是如何生成和标注的。在用于缺陷检测的数字孪生流水线中，DefectTwin 集成了一种由大语言模型驱动的多模态流水线，以分析图像等多模态输入。该系统生成详细的缺陷描述，并利用用户交互和反馈环来支持缺陷分析与维护工作流 (Ferdousi et al., 2024)。在基于摄像头和激光雷达传感器的自动驾驶数字孪生中，大语言模型接口通过自然语言提示实现在线场景编辑。这些提示支持添加或移除资产、改变位置或修改外观等动作。这有助于实现具有摄影级真实感和物理交互性的仿真，同时提供传感器可视化以及与自动驾驶软件栈的实时对接 (Samak et al., 2025)。

类似的流程支持具身智能体的视角数据类别。大模型智能体可以生成可在 Blender 中执行的 Python 脚本，这些脚本在场景图推导出的空间约束下检索并排列 3D 资产。这些智能体渲染图像，并通过视觉语言模型的反馈迭代优化场景 (Hu et al., 2024)。大模型还可被训练以生成用于程序化创建 3D 物体的 Blender 脚本。该流水线可渲染多个视图，如从不同角度生成多张图像，从而在无需手动建模的情况下提升视觉多样性 (Du et al., 2024)。基于移动激光雷达的端到端流水线能够实现数字孪生资产的快速重构，并可将由大型视觉语言模型引导的语义增强融入重构的 3D 表示中。生成的孪生体可通过 OpenUSD 等格式导出，以在 NVIDIA Omniverse 等平台中进行沉浸式检查和后续编辑 (Gholizadeh HamlAbadi et al., 2025)。

在这些示例中，感知数据和遥测数据主要用于迁移学习、领域自适应以及使用语言条件评分或异常描述等方法进行可扩展评估。在具身设置中，数据由交互迹构成，其中观测值与智能体动作显式配对，通常遵循部分可观测马尔可夫决策过程的公式 (Li et al., 2025b)。相比之下，数字孪生中的外部视角遥测数据通常用于追踪过去行为、监控当前行为以及预测未来行为，以支持决策制定以及运营与维护 (Deng et al., 2021; Bofill et al., 2023)。

## 7.2 智能体数据的质量度量标准

本节形式化了用于智能体数据的四类度量方法，包括有效性、保真度、多样性和实用性，并为每一类提供了明确的定义。我们的评估语义来源于多智能体具身基准，如 PARTNR, CARLA (Dosovitskiy et al., 2017) 上的文本到交通场景评估，以及世界模型或视频评估 (Chang et al., 2024; Ruan et al., 2025; Li et al., 2025b)。我们考虑如下定义的轨迹：

$$\tau = (s_0, a_0, r_1, s_1, \dots, a_{T-1}, r_T, s_T)$$

以及一个由以下表示的生成轨迹数据集：

$$\mathcal{A}_{\text{gen}} = \{\tau_i\}_{i=1}^N, \quad \tau_i = (s_0^{(i)}, a_0^{(i)}, r_1^{(i)}, s_1^{(i)}, \dots, a_{T_i-1}^{(i)}, r_{T_i}^{(i)}, s_{T_i}^{(i)})$$

在此公式中， $s_t$  表示步骤  $t$  的状态或观测值， $a_t$  表示动作，包括工具或应用程序接口调用， $r_{t+1}$  表示与从  $(s_t, a_t)$  到  $s_{t+1}$  转移相关的奖励或反馈。

**有效性。** 有效性衡量生成的智能体数据是否遵守基本语法、结构约束和可执行的先决条件。所有生成轨迹中的动作可执行率定义如下：

$$\text{ExecRate} = \frac{1}{\sum_{i=1}^N T_i} \sum_{i=1}^N \sum_{t=0}^{T_i-1} \mathbf{1}\{\text{action } a_t^{(i)} \text{ executes without error in } s_t^{(i)}\}.$$

Task success is a binary indicator of whether all goal predicates are satisfied at the end of an episode. Over  $N$  episodes, the success rate is calculated by the following formula:

$$\text{SR}_{\text{valid}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{all task goals satisfied in episode } i\}.$$

To capture partial completion, the percent complete metric reports the achieved fraction of goal predicates at termination. For a goal set  $G$  and an achieved subset  $\hat{G}(\tau_i)$ , the metric is defined as:

$$\text{PC}(\tau_i) = \frac{|\hat{G}(\tau_i)|}{|G|}.$$

The PARTNR benchmark defines task evaluation functions using propositions together with dependencies and constraints such as temporal constraints. It computes percent completion as the ratio of satisfied propositions. Success is achieved when the percent completion value equals 1.0 (Chang et al., 2024). In text-to-traffic-scene generation for CARLA, generation correctness is evaluated through text matching. This method uses a binary matched or unmatched criterion between the input prompt and the rendered scene, and the results are averaged over repeated generations (Ruan et al., 2025).

**忠诚** 保真度衡量合成观测与参考分布及结构匹配的程度。该指标在世界模型和视频预测相关的评估中被广泛使用 (Li et al., 2025b)。

弗雷谢尔初始距离比较真实样本和生成样本特征嵌入的高斯拟合 (Heusel et al., 2018)。我们令  $\mu_X$  和

$\Sigma_X$  表示真实样本嵌入的均值和协方差,  $\text{Tr}(\cdot)$  表示矩阵迹, 我们令  $\mu_Y$  和  $\Sigma_Y$  分别表示合成样本的均值和协方差。该距离的计算方式如下:

$$\text{FID} = \|\mu_X - \mu_Y\|_2^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}).$$

The Frechet Video(FVD) applies the same mathematical form to video embeddings that capture temporal dynamics, where lower values indicate higher fidelity (Untertiner et al., 2019).

结构相似性 (Structural Similarity) 比较图像  $x$  与参考图像  $y$  之间的亮度、对比度和结构 (Wang et al., 2004)。该度量由以下公式定义:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}.$$

where  $C_1$  and  $C_2$  are small positive constants introduced to avoid numerical instability when the denominators are close to zero.

均方误差和峰值信噪比表示为:

$$\begin{aligned} \text{MSE} &= \frac{1}{M} \sum_{i=1}^M (x_i - y_i)^2, \\ \text{PSNR} &= 10 \log_{10} \left( \frac{\text{MAX}^2}{\text{MSE}} \right), \end{aligned}$$

In these formulas,  $M$  represents the number of pixels or dimensions, and MAX denotes the maximum possible value of the signal.

学成感知图像块相似度计算深层特征活性值之间的感知距离 (Zhang et al., 2018)。对于逐层规范化的特征  $\hat{f}_{h,w,x}^l$  与  $\hat{f}_{h,w,y}^l$  以及学成的权重  $w_l$ , 该距离的计算方式为:

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{f}_{h,w,x}^l - \hat{f}_{h,w,y}^l) \right\|_2^2.$$

其中符号  $\odot$  表示哈达玛逐元素乘积。此操作表示由学成权值向量  $w_l$  所执行的通道缩放。

上述评估指标通常假设能够访问真实样本的分布, 例如弗雷切特生成网络距离 (Frechet Inception Distance) 和弗雷切特视频距离 (Frechet Video Distance), 或需要成对参考样本, 如结构相似性、均方误差、峰值信噪比以及学成感知图像块相似度。然而, 在许多条件生成情景中, 成对的像素级真实值并不可用。在这些情况下, 我们采用语义提示与生成结果对齐作为无参考的替代指标。

当缺乏像素级真实值时, 我们使用提示与结果的语义对齐作为无参考保真度代理, 该方法在缺乏像素级真实值时具有广泛适用性。该度量标准衡量合成结果在高层语义层面是否与条件提示保持一致。给定一个文本提示或规范  $p$  以及生成的结果  $o$  (例如从生成场景中得到的渲染图像或关键帧), 我们使用预训练的视觉-语言模型 (如 CLIP (Radford et al., 2021)) 获取跨模态嵌入。

我们用  $f_T(\cdot)$  和  $f_V(\cdot)$  表示文本编码器和视觉编码器, 并令

$$e_p = f_T(p), \quad e_o = f_V(o)$$



be the resulting embeddings. We define the prompt and outcome semantic alignment score by cosine similarity:

$$\text{SemAlign}(p, o) = \frac{e_p^\top e_o}{\|e_p\|_2 \|e_o\|_2},$$

where  $\|\cdot\|_2$  is the  $\ell_2$  norm and higher values indicate stronger semantic agreement. Over  $N$  prompt and outcome pairs  $\{(p_i, o_i)\}_{i=1}^N$ , the mean semantic fidelity is calculated as:

$$\text{SemFid} = \frac{1}{N} \sum_{i=1}^N \text{SemAlign}(p_i, o_i).$$

The mean semantic fidelity summarizes the semantic consistency between conditions and synthesized outcomes without requiring a reference image for each prompt. This approach is closely related to CLIP-based and reference-free evaluation metrics used in vision and language generation (Hessel et al., 2022). When a discrete correctness signal is preferred, such as a matched or unmatched criterion, one may threshold the semantic alignment score and report the resulting match rate:

$$\text{MatchRate}(\tau) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{SemAlign}(p_i, o_i) \geq \tau\},$$

where  $\tau$  is a similarity threshold and  $\mathbf{1}\{\cdot\}$  is the indicator function.

**多样性** 多样性捕捉了生成数据中模式、长尾概念以及非模板化多样性的覆盖情况。在文本到交通场景生成的背景下，多样性可以通过智能体多样性（AD）和道路多样性（RD）来量化。这两个指标均通过重复生成中唯一元素与总元素的比率计算得出。具体而言，智能体多样性考虑了智能体类型、动作及相对位置的变化，而道路多样性则基于唯一的道路标识符 (Ruan et al., 2025) 计算。

对于生成视频或世界模型滚动的观测级别评估，VBench 框架将视频生成评估分解为互补的维度，如主体一致性和时间质量。时间质量的一个例子是运动平滑性。该框架还单独评估视频与条件之间的一致性，这指的是输出遵循给定条件或提示的程度。这些维度得分在提示和条件空间中提供了关于优劣势的细粒度证据 (Huang et al., 2024b; Li et al., 2025b)。

**实用性。** 效用度量在训练或评估期间使用合成智能体数据来衡量下游有效性。在  $N$  个评估回合中，基于取值为 0 或 1 的二元成功指示器  $s_i$ ，成功率由以下公式定义：

$$\text{SR}_{\text{eval}} = \frac{1}{N} \sum_{i=1}^N s_i.$$

To quantify efficiency, we define  $L_i$  as the number of environment steps used in episode  $i$ :

$$\text{SimSteps} = \frac{1}{N} \sum_{i=1}^N L_i.$$

在协作情境中，任务卸载衡量的是分工的程度。我们令  $n_i^{(r)}$  为机器人在回合  $i$  中完成的子任务或命题

的数量， $n_i$  为已完成的子任务总数。该指标的表达式为：

$$\text{Offloading} = \frac{1}{N} \sum_{i=1}^N \frac{n_i^{(r)}}{n_i}.$$

对于回合级别的部分任务完成情况，我们复用完成百分比指标：

$$\text{PC}(\tau_i) = \frac{|\hat{G}(\tau_i)|}{|G|}.$$

在 SafeBench 风格的 CARLA 驾驶评估中，碰撞率（CR）定义为在评估场景下的期望碰撞指示（或碰撞次数） $\tau$ ：

$$\text{CR} = \mathbb{E}_{\tau \sim P}[c(\tau)],$$

where  $c(\tau)$  denotes the collision signal in scenario  $\tau$ . In many episodic driving setups where an episode terminates upon the first collision (or collisions are binarized),  $c(\tau) \in \{0, 1\}$  and CR can be interpreted as the proportion of episodes that contain at least one collision. Lower values indicate better safety performance (Xu et al., 2022; Zhang et al., 2024b). The overall score is a composite metric that aggregates driving metrics related to safety, functionality, and etiquette by using weights specified by the benchmark. In practice, the overall score is commonly interpreted as a measure of ego-vehicle driving performance where higher values are preferred. Conversely, adversarial scenario generation or selection may instead aim to reduce the overall score by constructing more challenging conditions.

强化学习研究还报告了每个回合的折扣回报：

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

This return, as well as the average return across episodes, is used to assess whether synthetic data improves policy performance or sample efficiency.

### 7.3 智能体数据的可信度度量

在实践中，对大模型生成的智能体数据的评估主要集中在安全性上。这一关注点考察轨迹、场景和计划是否避免了违反既定规则的危险状态和行为。而可信度的其他维度，如公平性和隐私性，在当前具身智能体和数字孪生的基准测试中发展程度显著不足。因此，我们将这些方面视为未来研究的开放方向，现阶段将讨论重点放在安全指标上。

**安全** 安全性衡量大模型生成的轨迹、场景和计划是否避免了危险状态和违规行为，并满足显式的安全约束，而与任务完成无关。在实践中，安全性通常通过驾驶基准中模拟器定义的违规和碰撞指标进行评估，例如基于 CARLA 的 SafeBench 等套件。此外，还通过约束强制规划和运动规划流水线中形式化规范的满足情况来评估 (Xu et al., 2022; Wu et al., 2025; Li & Zhao, 2025)。面向文本到场景生成的系统进一步强调生成标准化的场景文件和包含监控评估指标的模拟器测试报告 (Cai et al., 2025)。

规则违规包括闯红灯、逆向行驶、车道保持违规和超速等基本交通违章行为。可以计算每回合的通用违

规率:

$$\text{RVR} = \frac{1}{N} \sum_{i=1}^N \frac{V_i}{E_i},$$

where  $V_i$  represents the total number of rule violations in episode  $i$ , and  $E_i$  represents an exposure term such as the number of decision steps  $T_i$  or the distance traveled.

在实践中, 基于 CARLA 的基准测试通常报告特定类型的违规指标, 例如碰撞、闯红灯、闯停车标志、偏离道路距离和车道入侵, 而非单一的汇总计数 (Xu et al., 2022)。针对特定违规类型  $t$  的暴露规范化违规率可定义如下:

$$\text{RVR}^{(t)} = \frac{\sum_{i=1}^N V_i^{(t)}}{\sum_{i=1}^N \text{dist}_i},$$

where  $V_i^{(t)}$  is the number of infractions of type  $t$  in episode  $i$ . 这种报告风格也应用于 CARLA Leaderboard 结果中, 其中多种违规类型以每公里违规次数的形式展示 (CARLA Autonomous Driving Leaderboard, 2025b)。针对每种具体类型的细粒度违规计数, 相较于总体成功率或总回报 (Cai et al., 2025; Xu et al., 2022), 能够提供更详细的诊断信息。

路线不完整性衡量的是在情景结束时, 规划路线仍未完成的程度:

$$\text{RI} = 1 - \frac{\text{distance completed}}{\text{planned route length}}.$$

Higher values for this metric indicate early termination or a failure to follow the designated route. In CARLA-style evaluations, route progress or completion is commonly reported together with violation and collision metrics. This approach helps to distinguish safe task completion from instances of early stopping or unsafe driving (Xu et al., 2022).

速度相关合规性与违规行为用于捕捉智能体相对于法定或依赖于情境的速度限制, 是否行驶过慢或过快。一个简单的最低速度合规率定义如下:

$$\text{MSCR} = \frac{1}{T} \sum_{t=1}^T \mathbb{I}(v(t) \geq v_{\min}(t)),$$

where  $v_{\min}(t)$  is a context-dependent minimum-speed requirement such as a speed determined by nearby traffic. The CARLA Leaderboard explicitly penalizes the failure to maintain minimum speed as part of its evaluation criteria (CARLA Autonomous Driving Leaderboard, 2025a). More generally, one may also track speed-limit compliance within specific bounds when such information is available in the simulator or map.

舒适性和平顺性常作为激进或高风险操作的指标。基于 CARLA 的诊断报告可能包含基于运动学的指标, 例如平均加速度和偏航角速度 (Xu et al., 2022)。例如, 若  $a(t)$  表示加速度,  $\omega(t)$  表示偏航角速度, 则可量化如下:

$$\text{ACC} = \mathbb{E}_{\tau \sim P}[\text{acc}(\tau)],$$

$$\text{YV} = \mathbb{E}_{\tau \sim P}[y(\tau)].$$

Furthermore, more sensitive smoothness indicators can be computed, such as root-mean-square jerk and

the hard-braking rate:

$$\text{Jerk}_{\text{RMS}} = \sqrt{\frac{1}{T-2} \sum_{t=2}^{T-1} \left\| \frac{a(t+1) - a(t-1)}{2\Delta t} \right\|^2},$$

$$\text{HardBrakeRate} = \frac{1}{T} \sum_{t=1}^T \mathbb{I}(a_x(t) < -\tau),$$

where  $a_x(t)$  is longitudinal acceleration and  $\tau$  is a braking threshold. These smoothness metrics help to characterize the trade-off between safety and efficiency while complementing collision and completion metrics.

形式化安全满足度衡量的是满足时序逻辑规范  $\phi$  的执行比例，例如由线性时序逻辑 (LTL) 或信号时序逻辑 (STL) 所定义的规范。若  $\sigma^{(i)}$  表示回合  $i$  的执行迹，则安全满足率定义如下：

$$\text{SafetySat} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\sigma^{(i)} \models \phi).$$

The SELP framework improves the safety rate by mapping natural language to Linear Temporal Logic and enforcing constrained decoding, while the T<sup>3</sup> Planner utilizes Signal Temporal Logic verification in the loop to increase the satisfaction rates of motion plans (Wu et al., 2025; Li & Zhao, 2025).

危险拒绝率和风险是针对具身大语言模型智能体的安全性指标，用于衡量这些智能体对明确危险任务的响应。给定一组标注的危险任务  $\mathcal{H}$  和一组安全任务  $\mathcal{S}$ ，危险拒绝率和风险率定义如下：

$$\text{Rejection} = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \mathbb{I}(\text{agent refuses } h),$$

$$\text{Risk} = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \mathbb{I}(\text{agent executes } h).$$

SafeAgentBench reports a low rejection rate and non-trivial risk for current embodied LLM agents, which motivates the explicit tracking of these quantities (Yin et al., 2025).

碰撞时间是交通安全性评估中广泛使用的基于距离的替代安全指标。给定相对纵向距离  $d(t)$  和接近速度  $\dot{d}(t) < 0$ ，瞬时碰撞时间和其在回合内的极小点定义如下：

$$\text{TTC}(t) = \begin{cases} \frac{d(t)}{-\dot{d}(t)}, & \dot{d}(t) < 0, \\ +\infty, & \dot{d}(t) \geq 0, \end{cases}$$

$$\text{minTTC} = \min_t \text{TTC}(t).$$

Lower values for the minimum Time-to-Collision indicate a higher risk of collision (Ward et al., 2015; Sharath & Mehran, 2021).

碰撞最小距离是另一种常用的基于邻近性的指标。它通过使用自车智能体  $\mathbf{p}_{\text{ego}}(t)$  和其他智能体  $\mathbf{p}_{\text{other}}(t)$



的位置定义如下：

$$\text{MDC} = \min_t \|\mathbf{p}_{\text{ego}}(t) - \mathbf{p}_{\text{other}}(t)\|.$$

Lower values for the Minimum Distance to Collision indicate higher collision risk (Gao et al., 2025).

这些基于碰撞时间 (Time-to-Collision) 和最小碰撞距离 (Minimum Distance to Collision) 的邻近度量方法也被用于评估由大模型 (LLMs) 辅助生成或选择的交通场景 (Gao et al., 2025)。

违规诊断准确率用于评估自动化检测器 (包括基于多模态大模型的审计智能体) 在从日志、图像或叙述中识别安全违规或事故时的可靠性。该准确率通常通过标准分类指标如准确率、召回率和 F1 得分 (Skender et al., 2025) 进行总结。此外, SeeUnsafe 框架提出了一种用于交通肇事分析的多模态大模型智能体, 并引入了信息匹配得分, 以使结构化模型输出与真实值数据对齐 (Zhang et al., 2025b)。

## 7.4 评估实践差距

我们分析了第 7.1 节中的代表性方法, 并在表 7 中对它们报告的评估协议进行了分类。

Generation method	Validity	Fidelity	Diversity	Utility	Safety
TTSG(Ruan et al., 2025)	✓	△	✓	✓	✓
PARTNR (Chang et al., 2024)	✓	×	×	✓	×
SELP (Wu et al., 2025)	✓	×	×	✓	✓
Grid-Agent (Zhang et al., 2025c)	✓	×	×	✓	△
T <sup>3</sup> Planner (Li & Zhao, 2025)	✓	△	×	✓	✓

**Table 7** 是否在实验部分明确评估了每个维度的代表性智能体-数据生成方法。✓：明确评估；△：部分/间接覆盖；×：未报告或不适用。

**多样性.** 关于多样性, 当前智能体数据生成中评估的范围仍然有限。我们的分析表明, 只有从文本到交通场景的生成明确报告了多样性指标。相比之下, 其他代表性方法未提供覆盖度或变化度的定量度量, 且这些度量在重复采样下不基于模板。因此, 尽管这些方法能够比较成功率或效率, 但从现有实验中仍然难以确定生成的数据是否有意义地扩展了行为模式、任务或场景的覆盖范围。

**保真度.** 关于分布层面的保真度, 评估主要集中在一致性与正确性的代理指标上, 而非直接测试分布对齐情况。在我们的代表性研究中, 从文本生成交通场景的方法将提示场景匹配作为一致性的代理指标。相比之下, 以规划和控制为中心的方法主要关注成功率、完成度以及约束满足情况。这些方法包括 SELP(Wu et al., 2025)、PARTNR(Chang et al., 2024) 和 Grid Agent(Zhang et al., 2025c)。这些研究通常不报告能够将合成轨迹或计划与参考数据分布进行比较的指标。因此, 尽管部分覆盖了一致性代理, 但合成智能体数据与参考分布之间的统计接近性仍大多未经验证。

**诊断安全性.** 在安全问题的诊断方面, 当前的报告高度依赖于广泛的基准指标, 例如使用 CARLA 模拟器评估中的碰撞相关指标。此外, 还依赖于对规范的隐式满足, 如遵循线性时序逻辑或信号时序逻辑。这些方法取代了针对不同类型违规行为的详细和具体诊断。

在我们的代表性设置中, 交通场景生成报告与碰撞相关的信号。同时, SELP(Wu et al., 2025) 和 T<sup>3</sup> 规划器 (Li & Zhao, 2025) 方法主要关注形式化约束的满足情况。相比之下, 其他方法主要报告任务层面的成功, 包括任务完成或违规问题的解决。然而, 它们并未对每种特定类型的安全部署进行独立分解。因此, 尽管当前评估能够表明安全约束是否得到满足, 但在识别和解释安全失效的具体模式方面, 提供的诊断细节有限。

## 7.5 用途

在讨论了智能体数据的生成之后，我们现在关注其实际应用。本节探讨了轨迹、任务规范和遥测流如何用于在具身环境和数字孪生环境中，基于大模型对智能体进行训练、评估和优化。我们沿用此前确立的三部分结构，将这些应用分为测试床中的环境与任务数据、用于监督的控制与决策数据，以及用于评估信号的感知与遥测数据。

**环境与任务数据的使用。** 环境和任务数据主要作为推理、规划与安全验证的严格测试平台。在持续学习规划的背景下，通过世界状态图保持对世界的持久记忆至关重要。该方法使智能体能够随着环境的演变，从自然语言任务中反复实例化并求解 PDDL 规划问题，从而支持长时序符号规划的可重复评估 (Agarwal et al., 2025b)。

当安全性至关重要时，这些数据规范充当严格的约束。面向安全的流水线将自然语言指令转换为形式化的时序逻辑，例如线性时序逻辑，并利用由自动机引导的约束解码，以确保生成的计划逐步遵守这些形式化约束 (Wu et al., 2025)。

除了静态执行之外，这些数据还驱动自我修正。系统通常将基于大模型的规划器与基于逻辑的验证器配对，后者会迭代地拒绝并修复不安全的动作 (Li & Zhao, 2025)。这种迭代环路可以被记录下来，形成包含失败计划、验证器诊断（如约束违反或鲁棒性得分）以及修正后的修订版本的试错迹。这些迹对于评估非常有用，也可用于未来训练。

对于多智能体系统，协作基准提供了可验证且基于模拟器的任务。这些数据集有效地将单个环境转变为协作评估套件，旨在测试智能体在语言 (Chang et al., 2024) 的指导下如何协调、跟踪任务进展以及从错误中恢复。类似地，在数字孪生场景中，正式的任务描述定义了系统控制器的不同操作边界。这些规范用于压力测试治理策略，并验证在各种条件下回退策略是否能正确运行。

**控制与决策数据的使用。** 控制与决策数据为语言驱动的策略提供了直接监督，并在模仿学习和基于反馈的学习中被重复利用。在示范方面，近期的框架将大模型的决策建立在轨迹集合的基础上。例如，结合语言理解与从机器人交互数据中学成的价值估计，使模型能够选择与以往成功行为一致的动作 (Ahn et al., 2022)。程序化控制遵循类似的模式，其中自然语言被转化为可直接在机器人平台上运行的可执行策略 (Liang et al., 2023)。更一般地，执行此类策略自然会产生包含状态、动作和失败的迹，这些迹可用于分析、数据集构建或蒸馏成更小的控制器。大规模且异构的机器人日志被汇集起来，用于训练能够跨不同机器人泛化的指令跟随视觉运动策略，从而能够分析数据集多样性与构成如何影响语言条件下的泛化能力 (Collaboration et al., 2025)。

相同的数据也驱动着由语言中介的强化学习式更新。自动化奖励设计利用大模型编写和优化奖励函数，这些函数通常以可执行代码的形式呈现。通过在目标环境中的下游策略最优化和评估来验证这些函数，这通常在仿真中进行，以实现奖励假设的快速迭代 (Ma et al., 2024a; Xie et al., 2024)。

在离线情景下，策略可以基于从大模型中提取的语言嵌入进行条件化，并通过离线强化学习在静态日志上进行训练。这种方法能够在无需额外交互的情况下实现对未见过指令的泛化 (Morad et al., 2024)。为了减少人工标注所需的工作量，人工智能反馈协议利用多模态评论员在试验结束后对轨迹进行评分。这一过程将存档的试验转化为用于行为塑造的训练数据。例如，近期工作将视频和语言模型适配为语言条件化的机器人奖励函数，直接从视频中对执行结果进行评分 (Yang et al., 2024c)。与基于评论员的反

馈互补，单视频奖励推断方法可使用语义点对应和自动点追踪等技术，仅通过单一示范视频计算稠密奖励。这使得轨迹评估以及下游策略合成的数据集过滤成为可能 (Shi et al., 2025a)。

在数字孪生中，仿真生成的控制日志发挥着类似的作用。这些日志可以被重新标注、打分并重复使用，以优化控制策略，并分析由大模型驱动的控制器在分布变化或罕见事件下的行为表现。

**感知与遥测数据的使用。** 感知和遥测数据更侧重于评估与奖励学习，而非直接控制。与动作相关的数据流，例如来自摄像头的自我中心或第一人称观察以及本体感觉，为判断行为和系统状态提供了原始材料。同样，外部摄像头的第三人称日志、基于地图的状态以及模拟器记录等外部视角的系统事件也起到类似作用。利用大模型作为评判者的方法能够沿多个标准评估多样化的成果和多模态输入。这些评估的可靠性得益于提示词和输入的精心设计，以及一致性保障和偏差缓解策略 (Gu et al., 2025; Li et al., 2024a)。

视觉与语言的奖励学习可以利用成功和失败的执行视频，将视频和语言模型适配为语言条件化的奖励函数。该过程生成的奖励得分和信号可指导规划或强化学习 (Yang et al., 2024c)。此外，大模型能够提出参数化的奖励特征，并利用执行反馈迭代优化奖励参数。这种优化通过最小化模型与学成奖励函数之间的排序不一致性实现 (Zeng et al., 2024)。在驾驶评估领域，基于大模型的框架可将多源驾驶日志（如环视视频和 CAN 总线信号）转换为结构化的驾驶情境。这些系统随后输出关于驾驶行为在安全性、智能性和舒适性方面的结构化评估，且已在仿真中得到验证 (You et al., 2025)。

大规模且跨领域的观测语料库同样支持 4D 世界建模，包括几何理解与相机条件下的视频生成。这些语料库提供了时间对齐的视频流，包含丰富的多模态信号，如深度图、相机位姿、光流以及前景掩码。此类数据使模型能够学习场景动态，并生成符合指定相机轨迹的视频。OmniWorld 提供了一个统一的资源，用于覆盖重构和面向未来的预测需求的 4D 世界建模。它引入了一个以 3D 几何预测和相机控制视频生成为中心的基准，该基准基于新收集的游戏子集 OmniWorld-Game 以及跨多个领域的精选公开数据集构建 (Zhou et al., 2025)。

## 8 开放挑战与未来方向

尽管本综述为评估跨模态大模型驱动的数据生成建立了一个统一的分类体系，但该领域正迅速从静态数据集构建转向动态且自我优化的合成生态体系。当前的评估指标主要针对固定的人工标注语料库时代设计，在这一新范式下面临重大挑战。基于我们已建立的质量与可信度框架，本文指出了方法论亟需突破的方向，以确保合成数据的可持续性与可靠性。

### 8.1 从静态快照到动态反馈环评估

本综述中讨论的大多数度量方法，例如通过自相似性衡量的多样性或通过分布距离衡量的保真度，仅提供单次生成轮次的静态快照。然而，在实际应用中，系统正越来越多地转向递归训练环，即使用合成数据训练模型，而这些模型又会生成新的数据。

主要挑战在于，这些静态快照无法捕捉长期的系统动态。一个数据集可能在第一次迭代中获得较高的多样性得分，但仍会引发模型坍塌。这是一个退化过程，经过多次训练和生成循环后，分布尾部消失，模式被过度放大。

一个关键的未来方向是从逐点评估转向纵向轨迹监控。需要动态指标来追踪质量随时间变化的导数。这类指标能够检测支持覆盖范围的丧失、特征空间的收缩或误差模式在迭代过程中的漂移。这些时间性指标可作为早期预警信号，在模型坍塌变得不可逆之前触发干预措施，例如混合真实数据、重新平衡领域或调整采样策略。更广泛地说，这一转变需要元评估协议，以检验现有指标是否足够敏感，能够作为反馈环中的稳定性控制器。

## 8.2 重新定义忠实：从表面模仿到过程可验证

在我们的分类体系中，保真度传统上用于衡量生成数据与真实世界人类数据分布的接近程度。对于对话等开放式的自然语言任务，这种以人类为中心的分布相似性仍然至关重要。然而，在推理密集型领域，这一定义正成为瓶颈。随着面向推理的大模型开始达到甚至超越普通人类的表现水平，若严格要求遵循人类数据分布，可能会惩罚那些正确但新颖的解决方案。在数学、编程或符号逻辑等领域，表达得像人类并不如客观正确来得重要。

对于此类模态，保真度度量必须从衡量模仿性（即与人类作品的分布相似度）演进为衡量可验证性。有前景的方向包括可扩展且自动化的过程级奖励机制。这些奖励可包括代码执行反馈、数学形式化证明检查器，或对思维链迹的逻辑一致性探测。其目标在于评估推理过程的正确性和内在一致性，而非其在风格上与人类基准的相似度。

## 8.3 信任与效用的帕累托前沿探索

我们的框架将质量（包括下游实用性）与可信性（涵盖隐私、安全性和公平性）区分开来。尽管在分类体系中它们通常被视为独立的支柱，但在实际部署过程中，它们往往成为相互竞争的目标。

旨在提升可信度的机制往往会给合成语料库带来对齐成本。例如，激进的安全过滤不仅会移除有害内容，还可能不成比例地消除罕见概念，从而有效缩小尾部覆盖范围并降低多样性。同样，对表格数据添加差分隐私噪声虽能增强隐私保障，却会降低预测性能。在会话智能体中实施严格的对齐策略也可能以提高拒绝率作为代价来减少有害输出，这种现象被称为过度拒绝。

未来的工作应超越孤立地优化单一指标，转而量化并导航这些权衡。具体而言，我们需要能够刻画流水线配置下帕累托最优前沿的框架。这是指在不降低效用的情况下无法进一步提升任一信任度量，或反之亦然的操作点集合。这将使我们能够做出明确且与应用相关的决策，例如估算为实现由  $\epsilon$  和  $\delta$  定义的目标隐私保障所需牺牲多少下游性能，或为达到期望的安全水平可接受多大的拒绝率冗余。更广泛地说，这需要将信任与效用视为相互依赖的变量，而非独立的勾选项，从而推动多目标优化与选择策略的发展。

# 9 结论

当前的研究工作主要集中在利用大模型（LLMs）进行数据生成，而对“**数据审计员**”——负责评估合成数据质量的角色——的重要性则相对被忽视。确保高质量的数据对于将稀缺数据转化为可控资源至关重要，这不仅适用于模型训练，也适用于直接的现实应用。

在本篇综述中，以合成数据为中心，我们旨在弥合零散的研究进展，通过提出的 LLM Data Auditor 框架，为评估方法提供系统性理解。除了总结各类模态下的典型生成方法外，我们还将代表性指标归类为



两个主要维度：质量与可信度。通过应用这一评估体系，我们发现了当前评估实践中的显著差距。例如，表格数据生成中的公平性评估仍明显发展不足。因此，我们的框架不仅可作为全面参考，还可作为诊断工具，识别不同数据模态中缺失的评估维度。

我们的研究表明，无论模态如何，迫切需要优化生成后的评估指标。必须进行整体性评估，以防止合成数据在某一维度表现优异而在其他维度表现不佳。此外，我们强调了某些指标之间的固有权衡，例如表格数据中隐私与保真度之间的矛盾，这进一步凸显了多维度评估体系的必要性。

未来的工作中，本文综述提出的评估框架可作为开发全面基准的基石，用于评估现有的生成方法。此外，探索从静态评估系统向动态评估系统的转移对于数据生成技术的持续发展至关重要。尽管我们承认所收集的指标可能并不全面，但质量-可信度框架设计为开放且可扩展，能够随着领域的发展集成新的、有价值的指标。

我们希望本调查能为高质量、可信的基于大语言模型的数据生成奠定基础，推动社区开发出更加稳健可靠的数据生成系统。

## References

- Oluwanifemi Adebayo Moses Adekanye. LLM-powered synthetic environments for self-driving scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23721–23723, 2024. doi: 10.1609/aaai.v38i21.30540. URL <https://ojs.aaai.org/index.php/AAAI/article/view/30540>.
- Harshvardhan Aditya, Siddansh Chawla, Gunika Dhingra, Parijat Rai, Saumil Sood, Tanmay Singh, Zeba Mohsin Wase, Arshdeep Bahga, and Vijay K. Madiseti. Evaluating privacy leakage and memorization attacks on large language models (llms) in generative ai applications. *Journal of Software Engineering and Applications*, 17(5):421–447, 2024. doi: 10.4236/jsea.2024.175023. URL <https://www.scirp.org/journal/paperinformation?paperid=133625>.
- Bhavik Agarwal, Ishan Joshi, and Viktoria Rojkova. Think inside the json: Reinforcement strategy for strict llm schema adherence, 2025a. URL <https://arxiv.org/abs/2502.14905>.
- Krish Agarwal, Yuqian Jiang, Jiaheng Hu, Bo Liu, and Peter Stone. L3m+p: Lifelong planning with large language models, 2025b. URL <https://arxiv.org/abs/2508.01917>.
- Roosbeh Aghili, Heng Li, and Foutse Khomh. Protecting privacy in software logs: What should be anonymized?, 2025. URL <https://arxiv.org/abs/2409.11313>.
- Wasi Uddin Ahmad, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Vahid Noroozi, Somshubra Majumdar, and Boris Ginsburg. Opencodeinstruct: A large-scale instruction tuning dataset for code llms, 2025. URL <https://arxiv.org/abs/2504.04030>.
- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. URL <https://say-can.github.io/>.

- Gretel AI. Synthetically generated reasoning dataset (gsm8k-inspired) with enhanced diversity using gretel navigator and meta-llama/meta-llama-3.1-405b, 9 2024. URL <https://huggingface.co/datasets/gretelai/gretel-math-gsm8k-v1>.
- Ahmed M. Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022. URL <https://proceedings.mlr.press/v162/alaa22a.html>.
- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL <https://arxiv.org/abs/2402.16827>.
- Anthropic. Enhancing model safety through pretraining data filtering, 2025. URL <https://alignment.anthropic.com/2025/pretraining-data-filtering/>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislaw Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. URL <https://fairmlbook.org/>.
- Austin A. Barr, Joshua Quan, Eddie Guo, and Emre Sezgin. Large language models generating synthetic clinical datasets: A feasibility and comparative analysis with real-world perioperative data. *Frontiers in Artificial Intelligence*, 8:1533508, 2025. doi: 10.3389/frai.2025.1533508. URL <https://www.frontiersin.org/articles/10.3389/frai.2025.1533508>.
- Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding llms the right way: Fast, non-invasive constrained generation, 2024. URL <https://arxiv.org/abs/2403.06988>.
- Jherson Bofill, Mideth Abisado, Jocelyn Villaverde, and Gabriel Avelino Sampedro. Exploring digital twin-based fault monitoring: Challenges and opportunities. *Sensors*, 23(16), 2023. ISSN 1424-8220. doi: 10.3390/s23167087. URL <https://www.mdpi.com/1424-8220/23/16/7087>.
- Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators, 2023. URL <https://arxiv.org/abs/2210.06280>.
- Jack Boylan, Shashank Mangla, Dominic Thorn, Demian Gholipour Ghalandari, Parsa Ghaffari, and Chris Hokamp. Kgvalidator: A framework for automatic validation of knowledge graph construction, 2024. URL <https://arxiv.org/abs/2404.15923>.
- Tim Bray. The JavaScript Object Notation (JSON) Data Interchange Format. RFC 8259, December 2017. URL <https://www.rfc-editor.org/info/rfc8259>.
- Xuan Cai, Xuesong Bai, Zhiyong Cui, Danmu Xie, Daocheng Fu, Haiyang Yu, and Yilong Ren. Text2scenario: Text-driven scenario generation for autonomous driving test. *arXiv preprint arXiv:2503.02911*, 2025. URL <https://arxiv.org/abs/2503.02911>.
- CARLA Autonomous Driving Leaderboard. Evaluation criteria for the leaderboard 2.0, 2025a. URL [https://leaderboard.carla.org/evaluation\\_v2\\_0/](https://leaderboard.carla.org/evaluation_v2_0/).

- CARLA Autonomous Driving Leaderboard. Carla autonomous driving leaderboard (leaderboard table), 2025b. URL <https://leaderboard.carla.org/leaderboard/>.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pp. 2633–2650, 2021. URL <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.
- Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks, 2024. URL <https://arxiv.org/abs/2411.00081>.
- Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation, 2023. URL <https://github.com/sahil280114/codealpaca>.
- Bocheng Chen, Guangjing Wang, Hanqing Guo, Yuanda Wang, and Qiben Yan. Understanding multi-turn toxic behaviors in open-domain chatbots. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID 2023, pp. 282–296. ACM, October 2023a. doi: 10.1145/3607199.3607237. URL <http://dx.doi.org/10.1145/3607199.3607237>.
- Chong Chen, Kuanhong Zhao, Jiewu Leng, Chao Liu, Junming Fan, and Pai Zheng. Integrating large language model and digital twins in the context of industry 5.0: Framework, challenges and opportunities. *Robotics and Computer-Integrated Manufacturing*, 94:102982, 2025a. ISSN 0736-5845. doi: <https://doi.org/10.1016/j.rcim.2025.102982>. URL <https://www.sciencedirect.com/science/article/pii/S0736584525000365>.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023b. URL <https://arxiv.org/abs/2311.12793>.
- Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions, 2024a. URL <https://arxiv.org/abs/2406.04325>.
- Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL [https://openaccess.thecvf.com/content/CVPR2025/papers/Chen\\_CoMM\\_A\\_Coherent\\_Interleaved\\_Image-Text\\_Dataset\\_for\\_Multimodal\\_Understanding\\_and\\_CVPR\\_2025\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2025/papers/Chen_CoMM_A_Coherent_Interleaved_Image-Text_Dataset_for_Multimodal_Understanding_and_CVPR_2025_paper.pdf).
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024b. URL <https://arxiv.org/abs/2401.01335>.
- Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation, 2024. URL <https://arxiv.org/abs/2407.06135>.
- Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification, version 2.2. Specification, May 2025. URL [https://c2pa.org/specifications/specifications/2.2/specs/C2PA\\_Specification.html](https://c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html).
- 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. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minh Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundareshan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Halder, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025. URL <https://arxiv.org/abs/2310.08864>.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org, 2024. URL <https://dl.acm.org/doi/abs/10.5555/3692070.3692454>.



- Aaron Agyapong Danso and Ulrich B  ker. Automated generation of test scenarios for autonomous driving using llms. *Electronics*, 14(16), 2025. ISSN 2079-9292. doi: 10.3390/electronics14163177. URL <https://www.mdpi.com/2079-9292/14/16/3177>.
- DataCebo. Detection: Single table, 2024. URL <https://docs.sdv.dev/sdmetrics/data-metrics/metrics-in-beta/detection-single-table>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL <https://arxiv.org/abs/2501.12948>.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yudian Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhihui Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL <https://arxiv.org/abs/2405.04434>.
- Tianhu Deng, Keren Zhang, and Zuo-Jun (Max) Shen. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *Journal of Management Science and Engineering*, 6(2):125–134, 2021. ISSN 2096-2320. doi: <https://doi.org/10.1016/j.jmse.2021.03.003>. URL <https://www.sciencedirect.com/science/article/pii/S2096232021000238>.
- Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation, 2023. URL <https://arxiv.org/abs/2311.01460>.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-acl.212/>.
- Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. Synthdoc: Bilingual documents synthesis for visual document understanding, 2024. URL <https://arxiv.org/abs/2408.14764>.
- Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235

- of *Proceedings of Machine Learning Research*, pp. 11165–11197. PMLR, 2024. URL <https://proceedings.mlr.press/v235/dohmatob24b.html>.
- Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017. URL <https://arxiv.org/abs/1711.03938>.
- dottxt-ai. outlines-core, 2025. URL <https://github.com/dottxt-ai/outlines-core>.
- Enjun Du, Xunkai Li, Tian Jin, Zhihan Zhang, Rong-Hua Li, and Guoren Wang. Graphmaster: Automated graph synthesis via llm agents in data-limited environments, 2025. URL <https://arxiv.org/abs/2504.00711>.
- Yuhao Du, Shunian Chen, Wenbo Zan, Peizhao Li, Mingxuan Wang, Dingjie Song, Bo Li, Yan Hu, and Benyou Wang. Blenderllm: Training large language models for computer-aided design with self-improvement, 2024. URL <https://arxiv.org/abs/2412.14203>.
- John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy, data processing inequalities, and statistical minimax rates, 2014. URL <https://arxiv.org/abs/1302.3203>.
- Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/04000000042. URL <https://doi.org/10.1561/04000000042>.
- Liancheng Fang, Aiwei Liu, Hengrui Zhang, Henry Peng Zou, Weizhi Zhang, and Philip S. Yu. Tabgen-icl: Residual-aware in-context example selection for tabular data generation, 2025. URL <https://arxiv.org/abs/2502.16414>.
- Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey, 2024. URL <https://arxiv.org/abs/2402.17944>.
- Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015. URL <https://arxiv.org/abs/1412.3756>.
- Xiaohan Feng, Xixin Wu, and Helen Meng. Ontology-grounded automatic knowledge graph construction by llm under wikidata schema. In *Proceedings of the HI-AI@KDD Workshop*, pp. 117–135. CEUR Workshop Proceedings, 2024. URL <https://arxiv.org/abs/2412.20942>.
- Yukang Feng, Jianwen Sun, Chuanhao Li, Zizhen Li, Jiabin Ai, Fanrui Zhang, Yifan Chang, Sizhuo Zhou, Shenglin Zhang, Yu Dai, and Kaipeng Zhang. A high-quality dataset and reliable evaluation for interleaved image-text generation, 2025a. URL <https://arxiv.org/abs/2506.09427>.
- Yushi Feng, Tsai Hor Chan, Guosheng Yin, and Lequan Yu. Democratizing large language model-based graph data augmentation via latent knowledge graphs, 2025b. URL <https://arxiv.org/abs/2502.13555>.
- Rahatara Ferdousi, M. Anwar Hossain, Chunsheng Yang, and Abdulmotaleb El Saddik. Defecttwin: When llm meets digital twin for railway defect inspection, 2024. URL <https://arxiv.org/abs/2409.06725>.
- Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf, 2024. URL <https://arxiv.org/abs/2410.14872>.
- Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2025. URL <https://arxiv.org/abs/2501.00321>.

- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query, 2017. URL <https://arxiv.org/abs/1705.02101>.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL <https://aclanthology.org/2023.acl-long.910/>.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models, 2023b. URL <https://arxiv.org/abs/2211.10435>.
- Yuan Gao, Mattia Piccinini, Korbinian Möller, Amr Alanwar, and Johannes Betz. From words to collisions: LLM-guided evaluation and adversarial generation of safety-critical driving scenarios. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2025. doi: 10.48550/arXiv.2502.02145. URL <https://arxiv.org/abs/2502.02145>.
- Noam Gat. LM Format Enforcer, 2025. URL <https://github.com/noamgat/lm-format-enforcer>.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. URL <https://arxiv.org/abs/2009.11462>.
- Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. Jschemabench: A rigorous benchmark of structured outputs for language models, 2025. URL <https://arxiv.org/abs/2501.10868>.
- Abraham George and Amir Barati Farimani. Llm trainer: Automated robotic data generating via demonstration augmentation using llms, 2025. URL <https://arxiv.org/abs/2509.20070>.
- Kamran Gholizadeh HamAbadi, Monica Vahdati, Haiwei Dong, and Abdulmotaleb El Saddik. Ai-enhanced creation of digital twins from iphone lidar for immersive xr experiences in nvidia omniverse. In *Proceedings of the International Workshop on Intelligent Immersification in the Metaverse: AI-Driven Immersive Multimedia, I2M-MM ’25*, pp. 25–33, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718441. doi: 10.1145/3728487.3759232. URL <https://doi.org/10.1145/3728487.3759232>.
- Google DeepMind. Synthid, 2025. URL <https://deepmind.google/models/synthid>.
- Sven Gowal, Rudy Bunel, Florian Stimberg, David Stutz, Guillermo Ortiz-Jimenez, Christina Kouridi, Mel Vecerik, Jamie Hayes, Sylvestre-Alvise Rebuffi, Paul Bernard, Chris Gamble, Miklós Z. Horváth, Fabian Kaczmarczyck, Alex Kaskasoli, Aleksandar Petrov, Ilia Shumailov, Meghana Thotakuri, Olivia Wiles, Jessica Yung, Zahra Ahmed, Victor Martin, Simon Rosen, Christopher Savčák, Armin Senoner, Nidhi Vyas, and Pushmeet Kohli. Synthid-image: Image watermarking at internet scale, 2025. URL <https://arxiv.org/abs/2510.09263>.
- Mandeep Goyal and Qusay H. Mahmoud. A systematic review of synthetic data generation techniques using generative AI. *Electronics*, 13(17):3509, 2024. doi: 10.3390/electronics13173509. URL <https://www.mdpi.com/2079-9292/13/17/3509>.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL <https://arxiv.org/abs/2411.15594>.
- Cusuh Ham, Matthew Fisher, James Hays, Nicholas I. Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8186–8195. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00782. URL <https://doi.org/10.1109/CVPR52733.2024.00782>.

- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 2023. doi: 10.1145/3544548.3580688. URL <https://dl.acm.org/doi/10.1145/3544548.3580688>.
- Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. URL <https://arxiv.org/abs/1610.02413>.
- Pinjia He, Jieming Zhu, Zibin Zheng, and Michael R. Lyu. Drain: An online log parsing approach with fixed depth tree. In *2017 IEEE International Conference on Web Services (ICWS)*, pp. 33–40, 2017. doi: 10.1109/ICWS.2017.13. URL <https://ieeexplore.ieee.org/document/8029742>.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL <https://arxiv.org/abs/2104.08718>.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL <https://arxiv.org/abs/1706.08500>.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rygGQyrFvH>.
- Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024. URL <https://arxiv.org/abs/2403.07691>.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022. URL <https://aclanthology.org/2022.naacl-main.287/>.
- Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A. Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scene as blender code, 2024. URL <https://arxiv.org/abs/2403.01248>.
- Haoyu Huang, Chong Chen, Zeang Sheng, Yang Li, and Wentao Zhang. Can llms be good graph judge for knowledge graph construction?, 2025a. URL <https://arxiv.org/abs/2411.17388>.
- Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning, 2024a. URL <https://arxiv.org/abs/2312.10160>.
- Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023. URL <https://arxiv.org/abs/2305.11176>.
- Yingbing Huang, Deming Chen, and Abhishek K. Umrawal. Jam: Controllable and responsible text generation via causal reasoning and latent vector manipulation, 2025b. URL <https://arxiv.org/abs/2502.20684>.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, June 2024b. URL [https://openaccess.thecvf.com/content/CVPR2024/html/Huang\\_VBench\\_Comprehensive\\_Benchmark\\_Suite\\_for\\_Video\\_Generative\\_Models\\_CVPR\\_2024\\_paper.html](https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html).

- HuggingFaceFW. Fineweb-edu classifier (model card). <https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier>, 2024.
- Yintong Huo, Yichen Li, Yuxin Su, Pinjia He, Zifan Xie, and Michael R. Lyu. Autolog: A log sequence synthesis framework for anomaly detection. In *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023, Luxembourg, September 11-15, 2023*, pp. 497–509. IEEE, 2023. doi: 10.1109/ASE56229.2023.00133. URL <https://doi.org/10.1109/ASE56229.2023.00133>.
- Shadi Iskander, Sofia Tolmach, Ori Shapira, Nachshon Cohen, and Zohar Karnin. Quality matters: Evaluating synthetic data for tool-using LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4958–4976, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.285. URL <https://aclanthology.org/2024.emnlp-main.285/>.
- MinHyuk Jang, Youngdong Jang, JaeHyeok Lee, Feng Yang, Gyeongrok Oh, Jongheon Jeong, and Sangpil Kim. Lvmark: Robust watermark for latent video diffusion models, 2025. URL <https://arxiv.org/abs/2412.09122>.
- Jiarui Ji, Runlin Lei, Jialing Bi, Zhewei Wei, Xu Chen, Yankai Lin, Xuchen Pan, Yaliang Li, and Bolin Ding. Llm-based multi-agent systems are scalable graph generative models, 2025. URL <https://arxiv.org/abs/2410.09824>.
- Zhihan Jiang, Jinyang Liu, Zhuangbin Chen, Yichen Li, Junjie Huang, Yintong Huo, Pinjia He, Jiazhen Gu, and Michael R. Lyu. Lilac: Log parsing using llms with adaptive parsing cache, 2024a. URL <https://arxiv.org/abs/2310.01796>.
- Zhihan Jiang, Jinyang Liu, Junjie Huang, Yichen Li, Yintong Huo, Jiazhen Gu, Zhuangbin Chen, Jieming Zhu, and Michael R. Lyu. A large-scale evaluation for log parsing techniques: How far are we?, 2024b. URL <https://arxiv.org/abs/2308.10828>.
- Jing Jin and Houfeng Wang. Select high-quality synthetic QA pairs to augment training data in MRC under the reward guidance of generative language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14543–14554, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1267/>.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling, 2016. URL <https://arxiv.org/abs/1602.02410>.
- Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference (TCC 2013)*, volume 7785 of *Lecture Notes in Computer Science*, pp. 457–476. Springer, 2013. doi: 10.1007/978-3-642-36594-2\_26. URL [https://dl.acm.org/doi/10.1007/978-3-642-36594-2\\_26](https://dl.acm.org/doi/10.1007/978-3-642-36594-2_26).
- Zanis Ali Khan, Donghwan Shin, Domenico Bianculli, and Lionel C. Briand. Guidelines for assessing the accuracy of log message template identification techniques. In *Proceedings of the 44th International Conference on Software Engineering (ICSE ’22)*, pp. 1095–1106. Association for Computing Machinery, 2022. doi: 10.1145/3510003.3510101. URL <https://dl.acm.org/doi/10.1145/3510003.3510101>.
- Jinhee Kim, Taesung Kim, and Jaegul Choo. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models, 2025. URL <https://arxiv.org/abs/2404.12404>.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084, 2023. URL <https://proceedings.mlr.press/v202/kirchenbauer23a.html>.



- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models, 2023. URL <https://arxiv.org/abs/2305.17216>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL <https://arxiv.org/abs/2205.11916>.
- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vignesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. URL <https://arxiv.org/abs/2312.14125>.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL <https://arxiv.org/abs/1602.07332>.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL <https://arxiv.org/abs/2304.07327>.
- Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. Beyond the surface: Enhancing llm-as-a-judge alignment with human via internal representations, 2025. URL <https://arxiv.org/abs/2508.03550>.
- 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. Rewardbench: Evaluating reward models for language modeling, 2024. URL <https://arxiv.org/abs/2403.13787>.
- Duy Le, Kris Zhao, Mengying Wang, and Yinghui Wu. Graphlingo: Domain knowledge exploration by synchronizing knowledge graphs and large language models. In *IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024. URL <https://ieeexplore.ieee.org/document/10597884>.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL <https://arxiv.org/abs/2309.00267>.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024a. URL <https://arxiv.org/abs/2412.05579>.
- Jia Li and Guoxiang Zhao. T<sup>3</sup> planner: A self-correcting llm framework for robotic motion planning with temporal logic, 2025. URL <https://arxiv.org/abs/2510.16767>.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL <https://aclanthology.org/N16-1014/>.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, Miami, Florida, USA, November 2024b. Association

- for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.358. URL <https://aclanthology.org/2024.emnlp-main.358/>.
- Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. How far are llms from being our digital twins? a benchmark for persona-based behavior chain simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025a. URL <https://aclanthology.org/2025.findings-acl.813/>.
- Shuyang Li, Talha Azfar, and Ruimin Ke. Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility, 2024c. URL <https://arxiv.org/abs/2409.09040>.
- Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai, 2025b. URL <https://arxiv.org/abs/2510.16732>.
- Yichen Li, Yintong Huo, Zhihan Jiang, Renyi Zhong, Pinjia He, Yuxin Su, Lionel C. Briand, and Michael R. Lyu. Exploring the effectiveness of llms in automated logging statement generation: An empirical study. *IEEE Trans. Softw. Eng.*, 50(12):3188–3207, December 2024d. ISSN 0098-5589. doi: 10.1109/TSE.2024.3475375. URL <https://doi.org/10.1109/TSE.2024.3475375>.
- Yue Li, Xin Yi, Dongsheng Shi, Yongyi Cui, Gerard de Melo, and Linlin Wang. From injection to defense: Constructing edit-based fingerprints for large language models, 2025c. URL <https://arxiv.org/abs/2509.03122>.
- Zhecheng Li, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. DRS: Deep question reformulation with structured output. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12869–12882, Vienna, Austria, July 2025d. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.666. URL <https://aclanthology.org/2025.findings-acl.666/>.
- Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models, 2024. URL <https://arxiv.org/abs/2309.17444>.
- Hao Liang, Xiaochen Ma, Zhou Liu, Zhen Hao Wong, Zhengyang Zhao, Zimo Meng, Runming He, Chengyu Shen, Qifeng Cai, Zhaoyang Han, Meiyi Qiang, Yalin Feng, Tianyi Bai, Zewei Pan, Ziyi Guo, Yizhen Jiang, Jingwen Deng, Qijie You, Peichao Lai, Tianyu Guo, Chi Hsu Tsai, Hengyi Feng, Rui Hu, Wenkai Yu, Junbo Niu, Bohan Zeng, Ruichuan An, Lu Ma, Jihao Huang, Yaowei Zheng, Conghui He, Linpeng Tang, Bin Cui, Weinan E, and Wentao Zhang. Dataflow: An llm-driven framework for unified data preparation and workflow automation in the era of data-centric ai, 2025. URL <https://arxiv.org/abs/2512.16676>.
- Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023. URL <https://arxiv.org/abs/2209.07753>.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. Controllable text generation for large language models: A survey, 2024. URL <https://arxiv.org/abs/2408.12599>.
- Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective, 2024. URL <https://arxiv.org/abs/2407.01094>.
- Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, Charlie Nash, William L. Hamilton, David Duvenaud, Raquel Urtasun, and Richard S. Zemel. Efficient graph generation with graph recurrent attention networks, 2020. URL <https://arxiv.org/abs/1910.00760>.
- Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning, 2024. URL <https://arxiv.org/abs/2309.15091>.

- Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback, 2025a. URL <https://arxiv.org/abs/2501.13918>.
- Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, Yunan Huang, Mozhi Zhang, Pengyu Zhao, Junjie Yan, and Junxian He. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond, 2025b. URL <https://arxiv.org/abs/2505.19641>.
- Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation, 2024a. URL <https://arxiv.org/abs/2412.14167>.
- Shang Liu, Hao Du, Yang Cao, Bo Yan, Jinfei Liu, and Masatoshi Yoshikawa. Pgb: Benchmarking differentially private synthetic graph generation algorithms, 2024b. URL <https://arxiv.org/abs/2408.02928>.
- Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVI*, pp. 386–403, Berlin, Heidelberg, 2024c. Springer-Verlag. ISBN 978-3-031-72991-1. doi: 10.1007/978-3-031-72992-8\_22. URL [https://doi.org/10.1007/978-3-031-72992-8\\_22](https://doi.org/10.1007/978-3-031-72992-8_22).
- Yilun Liu, Shimin Tao, Weibin Meng, Jingyu Wang, Wenbing Ma, Yanqing Zhao, Yuhang Chen, Hao Yang, Yanfei Jiang, and Xun Chen. Interpretable online log analysis using large language models with prompt strategies, 2024d. URL <https://arxiv.org/abs/2308.07610>.
- Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. Tau-eval: A unified evaluation framework for useful and private text anonymization. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 216–227, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.16. URL <https://aclanthology.org/2025.emnlp-demos.16/>.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11065–11082, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.658. URL <https://aclanthology.org/2024.findings-acl.658/>.
- Yunbo Long, Liming Xu, and Alexandra Brintrup. Llm-tablogic: Preserving inter-column logical relationships in synthetic tabular data via prompt-guided latent diffusion, 2025. URL <https://arxiv.org/abs/2503.02161>.
- David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests, 2018. URL <https://arxiv.org/abs/1610.06545>.
- Qiuqing Lu, Meng Ma, Ximiao Dai, Xuanhan Wang, and Shuo Feng. Realistic corner case generation for autonomous vehicles with multimodal large language model, 2024. URL <https://arxiv.org/abs/2412.00243>.
- Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Learning to generate structured output with schema reinforcement learning, 2025. URL <https://arxiv.org/abs/2502.18878>.
- Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax, 2023. URL <https://arxiv.org/abs/2311.15813>.

- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *International Conference on Learning Representations (ICLR)*, 2025a. URL [https://proceedings.iclr.cc/paper\\_files/paper/2025/file/7c04aea54c2a60a632a47bd451cd2849-Paper-Conference.pdf](https://proceedings.iclr.cc/paper_files/paper/2025/file/7c04aea54c2a60a632a47bd451cd2849-Paper-Conference.pdf).
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct, 2025b. URL <https://arxiv.org/abs/2306.08568>.
- Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models, 2024a. URL <https://arxiv.org/abs/2310.12931>.
- Zeyang Ma, Dong Jae Kim, and Tse-Hsun Chen. Librelog: Accurate and efficient unsupervised log parsing using open-source large language models, 2024b. URL <https://arxiv.org/abs/2408.01585>.
- Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models, 2024. URL <https://arxiv.org/abs/2410.12832>.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL <https://arxiv.org/abs/2303.08896>.
- Margherita Martorana, Tobias Kuhn, Lise Stork, and Jacco van Ossenbruggen. Zero-shot topic classification of column headers: Leveraging llms for metadata enrichment, 2024. URL <https://arxiv.org/abs/2403.00884>.
- Shikib Mehri and Maxine Eskenazi. Usrc: An unsupervised and reference free evaluation metric for dialog generation, 2020. URL <https://arxiv.org/abs/2005.00456>.
- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL <https://arxiv.org/abs/2405.14734>.
- Yibo Miao, Yifan Zhu, Yinpeng Dong, Lijia Yu, Jun Zhu, and Xiao-Shan Gao. T2vsafetybench: Evaluating the safety of text-to-video generative models, 2024. URL <https://arxiv.org/abs/2407.05965>.
- Marko Miletic and Murat Sariyar. Large language models for synthetic tabular health data: A benchmark study. *Studies in health technology and informatics*, 316:963–967, 08 2024. doi: 10.3233/SHTI240571. URL <https://pubmed.ncbi.nlm.nih.gov/39176952/>.
- Michael J. Mior. Large language models for json schema discovery, 2024. URL <https://arxiv.org/abs/2407.03286>.
- Mohammad Ghiasvand Mohammadkhani, Saeedeh Momtazi, and Hamid Beigy. A survey on bridging VLMs and synthetic data. *Authorea Preprints*, 2025. doi: 10.36227/techrxiv.174741263.32891073. URL <https://www.techrxiv.org/doi/full/10.36227/techrxiv.174741263.32891073>.
- R. Mon-Williams et al. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 2025. URL <https://www.nature.com/articles/s42256-025-01005-x>.
- Steven Morad, Ajay Shankar, Jan Blumenkamp, and Amanda Prorok. Language-conditioned offline rl for multi-robot navigation, 2024. URL <https://arxiv.org/abs/2407.20164>.
- Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus, 2024. URL <https://arxiv.org/abs/2411.12498>.

- Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633, 2025. ISSN 2169-3536. doi: 10.1109/access.2025.3589503. URL <http://dx.doi.org/10.1109/ACCESS.2025.3589503>.
- Mehreen Naeem, Andrew Melnik, and Michael Beetz. Grounding language models with semantic digital twins for robotic planning, 2025. URL <https://arxiv.org/abs/2506.16493>.
- Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. Optimized feature generation for tabular data via llms with decision tree reasoning, 2024. URL <https://arxiv.org/abs/2406.08527>.
- Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi, and Luis A. Lastras. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs, 2023. URL <https://arxiv.org/abs/2305.12191>.
- Dang Nguyen, Sunil Gupta, Kien Do, Thin Nguyen, and Svetha Venkatesh. Generating realistic tabular data with large language models, 2024. URL <https://arxiv.org/abs/2410.21717>.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson,



- Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.
- Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhui Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models, 2024. URL <https://arxiv.org/abs/2310.02992>.
- Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning, 2024. URL <https://arxiv.org/abs/2402.13950>.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL <https://arxiv.org/abs/2406.17557>.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language, 2025. URL <https://arxiv.org/abs/2506.20920>.
- Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning, 2024. URL <https://arxiv.org/abs/2412.03565>.
- Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, and Dimitrios I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23:2892–2910, 2024. doi: 10.1016/j.csbj.2024.07.005. URL <https://www.sciencedirect.com/science/article/pii/S2001037024002393>.
- Vitou Phy, Yang Zhao, and Akiko Aizawa. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems, 2020. URL <https://arxiv.org/abs/2011.00483>.
- Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, December 2022. doi: 10.1162/coli\_a\_00458. URL <https://aclanthology.org/2022.cl-4.19/>.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015. doi: 10.1109/ICCV.2015.303. URL <https://ieeexplore.ieee.org/document/7410660>.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness, 2023. URL <https://arxiv.org/abs/2304.10703>.
- Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery, 2018. URL <https://arxiv.org/abs/1803.09518>.
- Prime Intellect Team. Synthetic-2 release: Four million collaboratively generated verified reasoning traces, July 2025. URL <https://www.primeintellect.ai/blog/synthetic-2-release>.
- Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

- Processing (EMNLP)*, pp. 5811–5826, Online, 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.468/>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Natraj Raman and Sameena Shah. Synthetic text generation using hypergraph representations, 2023. URL <https://arxiv.org/abs/2309.06550>.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, December 2023. doi: 10.1162/coli\_a\_00486. URL <https://aclanthology.org/2023.cl-4.2/>.
- Yaxuan Ren, Krithika Ramesh, Yaxing Yao, and Anjalie Field. How do we measure privacy in text? a survey of text anonymization metrics, 2025. URL <https://arxiv.org/abs/2512.01109>.
- Gurvan Richardeau, Samy Chali, Erwan Le Merrer, Camilla Penzo, and Gilles Tredan. Llms prompted for graphs: Hallucinations and generative capabilities, 2025. URL <https://arxiv.org/abs/2409.00159>.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. URL <https://arxiv.org/abs/1809.02156>.
- Bo-Kai Ruan, Hao-Tang Tsui, Yung-Hui Li, and Hong-Han Shuai. Traffic scene generation from natural language description for autonomous vehicles with large language model, 2025. URL <https://arxiv.org/abs/2409.09575>.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02155. URL <https://doi.org/10.1109/CVPR52729.2023.02155>.
- Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping, 2017. URL <https://arxiv.org/abs/1611.06624>.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. URL <https://arxiv.org/abs/1606.03498>.
- Tanmay Vilas Samak, Chinmay Vilas Samak, Bing Li, and Venkat Krovi. When digital twins meet large language models: Realistic, interactive, and editable simulation for autonomous driving, 2025. URL <https://arxiv.org/abs/2507.00319>.
- Soumya Sanyal, Harman Singh, and Xiang Ren. Fairr: Faithful and robust deductive reasoning over natural language, 2022. URL <https://arxiv.org/abs/2203.10261>.
- Ali Satvaty, Suzan Verberne, and Fatih Turkmen. Undesirable memorization in large language models: A survey, 2025. URL <https://arxiv.org/abs/2410.02650>.
- Martin Scaiano, Grant Middleton, Luk Arbuckle, Varada Kolhatkar, Liam Peyton, Moira Dowling, Debbie S Gipson, and Khaled El Emam. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of Biomedical Informatics*, 63:174–183, October 2016. doi: 10.1016/j.jbi.2016.07.015. URL <https://www.sciencedirect.com/science/article/pii/S1532046416300697>.

- Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. Curated llm: Synergy of llms and data curation for tabular augmentation in low-data regimes, 2024. URL <https://arxiv.org/abs/2312.12112>.
- Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.439. URL <https://aclanthology.org/2020.emnlp-main.439/>.
- Yashothara Shanmugarasa, Ming Ding, Mahawaga Arachchige Pathum Chamikara, and Thierry Rakotoarivelo. SoK: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security, ASIA CCS '25*, pp. 425–441. ACM, 2025. doi: 10.1145/3708821.3733888. URL <https://arxiv.org/abs/2506.12699>.
- Mysore N. Sharath and Babak Mehran. A literature review of performance metrics of automated driving systems for on-road vehicles. *Frontiers in Future Transportation*, 2:759125, 2021. doi: 10.3389/ffutr.2021.759125. URL <https://www.frontiersin.org/articles/10.3389/ffutr.2021.759125>.
- Junyao Shi, Joshua Smith, Jianing Qian, and Dinesh Jayaraman. Points2reward: Robotic manipulation rewards from just one video. *Robotics: Science and Systems (RSS)*, 2025a. URL [https://scholar.google.com/citations?view\\_op=view\\_citation&hl=en&user=XAhGkq8AAAAJ&citation\\_for\\_view=XAhGkq8AAAAJ:eQOLeE2rZwMC](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=XAhGkq8AAAAJ&citation_for_view=XAhGkq8AAAAJ:eQOLeE2rZwMC).
- Ruxue Shi, Yili Wang, Mengnan Du, Xu Shen, and Xin Wang. A comprehensive survey of synthetic tabular data generation. *arXiv preprint arXiv:2504.16506*, 2025b. doi: 10.48550/arXiv.2504.16506. URL <https://arxiv.org/abs/2504.16506>.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017. URL <https://arxiv.org/abs/1610.05820>.
- Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots, 2022. URL <https://arxiv.org/abs/2209.03463>.
- Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models, 2022. URL <https://arxiv.org/abs/2209.11302>.
- Praphul Singh, Charlotte Dzialo, Jangwon Kim, Sumana Srivatsa, Irfan Bulu, Sri Gadde, and Krishnamurthy Kenthapadi. Redactor: An llm-powered framework for automatic clinical data de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 510–530, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-industry.36. URL <https://aclanthology.org/2025.acl-industry.36/>.
- Ilhan Skender, Kailin Tong, Selim Solmaz, and Daniel Watzenig. Investigating traffic accident detection using multimodal large language models, 2025. URL <https://arxiv.org/abs/2509.19096>.
- Aivin V. Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers, 2023. URL <https://arxiv.org/abs/2302.02041>.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxin Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse

- Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL <https://aclanthology.org/2024.acl-long.840/>.
- Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2024. URL <https://arxiv.org/abs/2307.05222>.
- Anirudh Sundar, Christopher Richardson, and Larry Heck. gtbls: Generating tables from text by conditional question answering, 2024. URL <https://arxiv.org/abs/2403.14457>.
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL <https://aclanthology.org/2021.findings-acl.317/>.
- Boyin Tan, Junjielong Xu, Zhouruixing Zhu, and Pinjia He. Al-bench: A benchmark for automatic logging, 2025. URL <https://arxiv.org/abs/2502.03160>.
- Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Yangfan He, Kuan Lu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, Wei Shi, Yuliang Liu, Hao Liu, Yuan Xie, Xiang Bai, and Can Huang. Textsquare: Scaling up text-centric visual instruction tuning, 2025. URL <https://arxiv.org/abs/2404.12803>.
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’18/IAAI’18/EAAI’18. AAAI Press, 2018. ISBN 978-1-57735-800-8. URL <https://dl.acm.org/doi/10.5555/3504035.3504124>.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL <https://arxiv.org/abs/2405.09818>.
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. URL <https://arxiv.org/abs/2003.12039>.
- Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer, 2024. URL <https://arxiv.org/abs/2401.10208>.
- Together Computer. Redpajama: an open dataset for training large language models, 2023. URL <https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2>.
- Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset, 2024. URL <https://arxiv.org/abs/2402.10176>.
- Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 2025. doi: 10.1287/mksc.2025.0262. URL <https://pubsonline.informs.org/doi/10.1287/mksc.2025.0262>.
- Toan V. Tran and Li Xiong. Differentially private tabular data synthesis using large language models, 2024. URL <https://arxiv.org/abs/2406.01457>.
- Jeanine Treffers-Daller, Patrick Parslow, and Shirley Williams. Back to basics: how measures of lexical diversity can

- help discriminate between CEFR levels. *Applied Linguistics*, 39(3):302–327, 2018. doi: 10.1093/applin/amw009. URL <https://doi.org/10.1093/applin/amw009>.
- Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL <https://arxiv.org/abs/1812.01717>.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models, 2018. URL <https://arxiv.org/abs/1610.02424>.
- vLLM Project. Structured Outputs, 2024. URL [https://docs.vllm.ai/en/v0.8.2/features/structured\\_outputs.html](https://docs.vllm.ai/en/v0.8.2/features/structured_outputs.html).
- Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. A survey on data synthesis and augmentation for large language models, 2024a. URL <https://arxiv.org/abs/2410.12896>.
- Wenxuan Wang, Xiaoyuan Liu, Kuiyi Gao, Jen-tse Huang, Youliang Yuan, Pinjia He, Shuai Wang, and Zhaopeng Tu. Can’t see the forest for the trees: Benchmarking multimodal safety awareness for multimodal llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL <https://aclanthology.org/2025.acl-long.832/>.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models, 2024b. URL <https://arxiv.org/abs/2307.10635>.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024c. URL <https://arxiv.org/abs/2409.18869>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023a. URL <https://arxiv.org/abs/2203.11171>.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023b. URL <https://arxiv.org/abs/2212.10560>.
- Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. Harmonic: Harnessing llms for tabular data synthesis and privacy protection, 2024d. URL <https://arxiv.org/abs/2408.02927>.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. URL <https://ieeexplore.ieee.org/document/1284395>.
- Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. Codeclm: Aligning language models with tailored synthetic data, 2024e. URL <https://arxiv.org/abs/2404.05875>.
- James R. Ward, Gabriel Agamennoni, Stewart Worrall, Asher Bender, and Eduardo Nebot. Extending time to collision for probabilistic reasoning in general traffic scenarios. *Transportation Research Part C: Emerging Technologies*, 51:



- 66–82, 2015. ISSN 0968-090X. doi: <https://doi.org/10.1016/j.trc.2014.11.002>. URL <https://www.sciencedirect.com/science/article/pii/S0968090X14003180>.
- Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL <https://arxiv.org/abs/2411.12372>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with oss-instruct, 2024. URL <https://arxiv.org/abs/2312.02120>.
- Haoyang Wen, Jiang Guo, Yi Zhang, Jiarong Jiang, and Zhiguo Wang. On synthetic data strategies for domain-specific generative retrieval, 2025. URL <https://arxiv.org/abs/2502.17957>.
- Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences, 2024. URL <https://arxiv.org/abs/2403.09032>.
- Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models, 2023. URL <https://arxiv.org/abs/2311.16090>.
- Yi Wu, Zikang Xiong, Yiran Hu, Shreyash S. Iyengar, Nan Jiang, Aniket Bera, Lin Tan, and Suresh Jagannathan. Selp: Generating safe and efficient task plans for robot agents with large language models, 2025. URL <https://arxiv.org/abs/2409.19471>.
- Yuchen Xia, Daniel Dittler, Nasser Jazdi, Haonan Chen, and Michael Weyrich. Llm experiments with simulation: Large language model multi-agent system for simulation model parametrization in digital twins, 2024. URL <https://arxiv.org/abs/2405.18092>.
- Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tUM39YTRxH>.
- Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. Table-llm-specialist: Language model specialists for tables using iterative generator-validator fine-tuning, 2024. URL <https://arxiv.org/abs/2410.12164>.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024a. URL <https://openreview.net/forum?id=CfXh93NDgH>.
- Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles, 2022. URL <https://arxiv.org/abs/2206.09682>.
- Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan, 2019. URL <https://arxiv.org/abs/1907.00503>.
- Shengzhe Xu, Cho-Ting Lee, Mandar Sharma, Raquib Bin Yousuf, Nikhil Muralidhar, and Naren Ramakrishnan. Why llms are bad at synthetic table generation (and what to do about it), 2025. URL <https://arxiv.org/abs/2406.14541>.

- Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Kang Liu, and Jun Zhao. Generate-on-graph: Treat llm as both agent and kg in incomplete knowledge graph question answering, 2024b. URL <https://arxiv.org/abs/2404.14741>.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024a. URL <https://arxiv.org/abs/2409.12122>.
- June Yong Yang, Geondo Park, Joowon Kim, Hyeonwon Jang, and Eunho Yang. Language-interfaced tabular oversampling via progressive imputation and self-authentication. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=8F6bws5JBy>.
- Shuo Yang, Chencheng Yuan, Yao Rong, Felix Steinbauer, and Gjergji Kasneci. P-ta: Using proximal policy optimization to enhance tabular data augmentation via large language models, 2025. URL <https://arxiv.org/abs/2406.11391>.
- Yanting Yang, Minghao Chen, Qibo Qiu, Jiahao Wu, Wenxiao Wang, Binbin Lin, Ziyu Guan, and Xiaofei He. Adapt2reward: Adapting video-language models to generalizable robotic rewards via failure prompts. In *European Conference on Computer Vision (ECCV)*, 2024c. URL [https://www.ecva.net/papers/eccv\\_2024/papers\\_ECCV/papers/07430.pdf](https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/07430.pdf).
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259/>.
- Peiran Yao and Denilson Barbosa. Accurate and nuanced open-qa evaluation through textual entailment, 2024. URL <https://arxiv.org/abs/2405.16702>.
- Yang Yao, Xin Wang, Zeyang Zhang, Yijian Qin, Ziwei Zhang, Xu Chu, Yuekui Yang, Wenwu Zhu, and Hong Mei. Exploring the potential of large language models in graph generation, 2024. URL <https://arxiv.org/abs/2403.14358>.
- Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting, 2018. URL <https://arxiv.org/abs/1709.01604>.
- Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents, 2025. URL <https://arxiv.org/abs/2412.13178>.
- Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning, 2024. URL <https://arxiv.org/abs/2402.06332>.
- Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models, 2018. URL <https://arxiv.org/abs/1802.08773>.
- Shanhe You, Xuwen Luo, Xinhe Liang, Jiashu Yu, Chen Zheng, and Jiangtao Gong. A comprehensive llm-powered framework for driving intelligence evaluation, 2025. URL <https://arxiv.org/abs/2503.05164>.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian

- Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024a. URL <https://arxiv.org/abs/2309.12284>.
- Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale, 2024b. URL <https://arxiv.org/abs/2310.20550>.
- Quan Yuan, Zhikun Zhang, Linkang Du, Min Chen, Peng Cheng, and Mingyang Sun. Privgraph: Differentially private graph data publication by exploiting community information, 2023. URL <https://arxiv.org/abs/2304.02401>.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL <https://arxiv.org/abs/2203.14465>.
- Yuwei Zeng, Yao Mu, and Lin Shao. Learning reward for robot skills using large language models via self-alignment, 2024. URL <https://arxiv.org/abs/2405.07162>.
- Hao Zhang, Dongjun Yu, Lei Zhang, Guoping Rong, Yongda Yu, Haifeng Shen, He Zhang, Dong Shao, and Hongyu Kuang. Aucad: Automated construction of alignment dataset from log-related issues for enhancing llm-based log generation. In *Proceedings of the 16th International Conference on Internetware*, Internetware 2025, pp. 413–425. ACM, June 2025a. URL <http://dx.doi.org/10.1145/3755881.3755889>.
- Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024a. URL <https://arxiv.org/abs/2310.09656>.
- Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles, 2024b. URL <https://arxiv.org/abs/2405.14062>.
- Kaituo Zhang, Zhimeng Jiang, and Na Zou. Cleansing the artificial mind: A self-reflective detoxification framework for large language models, 2026. URL <https://arxiv.org/abs/2601.11776>.
- Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 407–423. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.findings-eacl.31>.
- Mingming Zhang, Zhiqing Xiao, Guoshan Lu, Sai Wu, Weiqiang Wang, Xing Fu, Can Yi, and Junbo Zhao. Aigt: Ai generative table based on prompt, 2024c. URL <https://arxiv.org/abs/2412.18111>.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL <https://arxiv.org/abs/1801.03924>.
- Ruixuan Zhang, Beichen Wang, Juexiao Zhang, Zilin Bian, Chen Feng, and Kaan Ozbay. When language and vision meet road safety: Leveraging multimodal large language models for video-based traffic accident analysis. *Accident Analysis & Prevention*, 219:108077, September 2025b. ISSN 0001-4575. doi: 10.1016/j.aap.2025.108077. URL <http://dx.doi.org/10.1016/j.aap.2025.108077>.
- Tianhui Zhang, Bei Peng, and Danushka Bollegala. Improving diversity of commonsense generation by large language models via in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9226–9242, Miami, Florida, USA, November 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.540. URL <https://aclanthology.org/2024.findings-emnlp.540/>.
- Xuanyu Zhang, Youmin Xu, Runyi Li, Jiwen Yu, Weiqi Li, Zhipei Xu, and Jian Zhang. V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection, 2024e. URL <https://arxiv.org/abs/2404.16824>.

- Yan Zhang, Ahmad Mohammad Saber, Amr Youssef, and Deepa Kundur. Grid-agent: An llm-powered multi-agent system for power grid control, 2025c. URL <https://arxiv.org/abs/2508.05702>.
- Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. Spa-vl: A comprehensive safety preference alignment dataset for vision language model, 2025d. URL <https://arxiv.org/abs/2406.12030>.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025e. URL <https://arxiv.org/abs/2410.02713>.
- Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences, 2020. URL <https://arxiv.org/abs/2001.06891>.
- 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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL <https://arxiv.org/abs/2306.05685>.
- Mingyu Zheng, Zhifan Feng, Jia Wang, Lanrui Wang, Zheng Lin, Yang Hao, and Weiping Wang. Tabledreamer: Progressive and weakness-guided data synthesis from scratch for table instruction tuning, 2025. URL <https://arxiv.org/abs/2506.08646>.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL <https://arxiv.org/abs/2311.07911>.
- Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation, 2024. URL <https://arxiv.org/abs/2412.16364>.
- Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling, 2025. URL <https://arxiv.org/abs/2509.12201>.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, 2025. URL <https://arxiv.org/abs/2310.17631>.

## A 符号表

在本节中，我们提供了本文综述中使用的主要符号。



Symbol	Meaning
<b>Shared</b>	
$E(\cdot)$	Encoder / embedding function
$\text{sim}(\cdot, \cdot)$	Similarity (e.g., cosine)
<b>Text / Dialogue</b>	
$\mathcal{T}$	Set of generated texts
$t_i$	$i$ -th generated text, $t_i \in \mathcal{T}$
$p_i$	Prompt / instruction for $t_i$
$\mathcal{D}$	Labeled evaluation dataset
$y_i, \hat{y}_i$	Gold / predicted label
$w_m, L$	$m$ -th token; sequence length
$\tau$	Acceptability threshold
$\mathcal{X}$	Corpus (set of instances)
<b>Symbolic / Reasoning</b>	
$\mathcal{R}$	Set of reasoning examples $(q_i, c_i, y_i)$
$q_i$	Question / problem for example $i$
$c_i$	Chain-of-thought / rationale for example $i$
$c_{i,t}$	$t$ -th step in chain $c_i$
$y_i$	Final answer / conclusion for example $i$
$c_{i,k}$	$k$ -th sampled chain for the same question $q_i$
$y_{i,k}$	Final answer induced by chain $c_{i,k}$
$K$	Number of sampled chains per question
<b>Tabular</b>	
$X_{\text{real}}, X_{\text{syn}}$	Real / synthetic tables
$x_i$	Row (record) in a table
$y_i$	Label / target for $x_i$ (if any)
$d$	Number of features (columns)
<b>Graphs / JSON / Logs</b>	
$\mathcal{G}$	Set of graphs (real or generated)
$G$	A graph; $V(G), E(G)$ : nodes / edges
$\mathcal{J}$	Set of JSON / structured records
$J$	A JSON key-value object
$\mathcal{L}$	Set of log entries or sequences
$\ell_i$	A log entry or a log sequence
<b>Vision-Language</b>	
$\mathcal{M}$	Set of multi-modal samples
$m_i$	$i$ -th sample, e.g., $(v_i, t_i)$
$v_i$	Visual / audio / video input paired with text
$k$	Number of modalities in $m_i$
<b>Agent / Interaction</b>	
$\mathcal{A}$	Set of agent trajectories / episodes
$\mathcal{A}_{\text{gen}}$	Set of generated trajectories $\{\tau_i\}_{i=1}^N$
$N$	Number of trajectories / episodes in $\mathcal{A}_{\text{gen}}$
$\tau_i$	$i$ -th trajectory $(s_0, a_0, r_1, s_1, \dots, a_{T_i-1}, r_{T_i}, s_{T_i})$
$T_i$	Horizon (number of decision steps) of trajectory $\tau_i$
$s_t$	State / observation at step $t$
$a_t$	Action (incl. tool/API call) at step $t$
$r_{t+1}$	Reward / feedback for transition $(s_t, a_t, s_{t+1})$

## B 相关工作

我们总结了相关工作，并从多个维度对 LLM Data Auditor 进行了比较，包括主要关注点、组织方式、质量、可信评估和范围。

Dimension	Long et al. (2024) Long et al. (2024)	Wang et al. (2024) Wang et al. (2024a)	Shi et al. (2025) Shi et al. (2025b)	This Work
<b>Primary Focus</b>	<b>The Engineering Pipeline.</b> Focuses on the operational workflow of constructing data, emphasizing generation techniques, curation strategies, and downstream application.	<b>The Model Lifecycle.</b> Focuses on the utility of synthetic data across distinct training stages, spanning pre-training, supervised fine-tuning, and alignment.	<b>The Generative Method.</b> Focuses on methodological families within the structured data domain, contrasting GANs, Variational Autoencoders, and Diffusion models.	<b>The Data Artifact.</b> Focuses on the intrinsic properties of the generated product, prioritizing rigorous auditing standards and data governance protocols.
<b>Organization</b>	<b>Procedure-Oriented.</b> Taxonomy defined by operational modules including prompt engineering, task decomposition, and heuristic filtering.	<b>Stage-Oriented.</b> Taxonomy structured around the LLM development lifecycle and downstream competencies such as reasoning and coding.	<b>Model-Centric.</b> Taxonomy categorized by the underlying generative architecture and associated post-processing techniques for tabular structures.	<b>Metric-Oriented.</b> Taxonomy defined by a unified evaluation coordinate system separating Quality dimensions from Trustworthiness dimensions.
<b>Quality Evaluation</b>	<b>Extrinsic Utility.</b> Quality is frequently judged by downstream gains, complemented by lightweight intrinsic proxies.	<b>Benchmark-Based.</b> Effectiveness is assessed via success rates on capability-specific evaluation suites and standard public benchmarks.	<b>Statistical Fidelity.</b> Evaluation prioritizes distributional resemblance to real data, machine learning efficacy, and column-wise statistical alignment.	<b>Intrinsic Verification.</b> Quality is defined through proactive audits of Validity, Fidelity, Diversity, and Utility prior to model training.
<b>Trustworthy Evaluation</b>	<b>Challenge-Oriented.</b> Hallucination and bias are discussed primarily as open challenges or limitations.	<b>Alignment-Oriented.</b> Safety is framed within the context of RLHF alignment.	<b>Privacy-Specific.</b> Analysis heavily concentrates on privacy guarantees.	<b>Foundational Pillar.</b> Elevates mainstream trustworthiness to a primary dimension orthogonal to utility metrics.
<b>Scope</b>	<b>Text-Dominant.</b> Primarily covers natural language processing tasks.	<b>Broad.</b> This work covers multiple modalities such as text, code, and vision.	<b>Tabular Data Specialized.</b> Addresses the constraints inherent to tabular data.	<b>Cross-Modality.</b> Applies a single framework to introduce 6 modal data.

**Table 9** 合成数据调查的对比分析。尽管现有的调查通过工程工作流 (Long et al., 2024)、训练生命周期 (Wang et al., 2024a) 或特定模式 (Shi et al., 2025b) 来构建该领域，但本工作建立了一种以度量为中心的分类体系，重点关注跨模式数据的内在评估与治理。