Agents of Diffusion：通过多智能体强化学习增强扩散语言模型的结构化数据生成（扩展版）

Aja Khanal 西安大略大学 加拿大伦敦
akhanal3@uwo.ca

Kaushik T. Ranade 西安大略大学 加拿大伦敦
kranade@uwo.ca

Rishabh Agrawal 西安大略大学 加拿大伦敦
ragrawa9@uwo.ca

Kalyan S. Basu
ICASSSD 加拿大剑桥
ks.basu
@gmail.com

Apurva Narayan 西安大略大学 加拿大伦敦
apurva.narayan@uwo.ca

arXiv:2601.07152v1 [cs.MA] 12 Jan 2026

摘要

生成高质量的结构化数据（如 JSON 记录）仍是大语言模型（LLM）面临的一项根本挑战，尤其是在语义丰富性必须与严格的模式遵循并存时。自回归大语言模型具有很强的结构一致性，但往往难以实现语义变化和输出多样性。相比之下，扩散语言模型（DLM）为丰富语义和双向解码提供了强大的机制，却缺乏可靠保持结构所需的归纳偏置。我们提出了**Agents of Diffusion（AoD）**，这是一个新颖的框架，通过语言介导的强化学习，将扩散语言模型的生成灵活性与自回归模型的推理能力结合起来。AoD 将结构化文本生成构建为一个多智能体对齐过程，其中提示优化智能体与评判智能体协作，利用自然语言反馈迭代引导扩散语言模型。该方法无需修改模型参数，也不依赖手工设计的约束，即可实现可控且符合模式的生成。AoD 推进了可控生成的发展，证明在协作智能体的监督下，扩散模型能够兼具较高的语义新颖性和结构保真度。在多个结构化数据基准上，AoD 的表现始终优于扩散和自回归基线，为兼顾结构感知与多样性的文本合成开辟了新路径。代码：https://github.com/Idsl-group/AgentsOfDiffusion。

关键词

结构化数据生成、合成数据、多智能体系统、强化学习、扩散语言模型

ACM 参考文献格式：

Aja Khanal、Kaushik T. Ranade、Rishabh Agrawal、Kalyan S. Basu 和 Apurva Narayan。2026。Agents of Diffusion：利用多智能体强化学习增强扩散语言模型，实现结构化数据生成（扩展版）。载于第*25*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2026*），塞浦路斯帕福斯，*2026*年*5*月*25*日至*29*日，IFAAMAS，20页。

第*25*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2026*），*C. Amato*、*L. Dennis*、*V. Mascardi*、*J. Thangarajah*（编），*2026*年*5*月*25*日至*29*日，塞浦路斯帕福斯。© 2026 国际自主智能体与多智能体系统基金会（www.ifaamas.org）。本作品根据知识共享署名4.0国际版（CC-BY 4.0）许可协议授权。

1 引言

扩散智能体（AoD）是一种用于可控数据生成的多智能体强化学习框架，将扩散语言模型（DLM）的语义丰富性与自回归大语言模型（LLM）智能体的结构精确性结合起来。AoD 探索了一种独特思路：无需微调、手工制定规则或标量奖励塑形，即可利用自然语言反馈监督扩散语言模型。两个大语言模型智能体（提示优化器和评判器）通过语言反馈进行交流，以迭代方式优化提示，引导扩散语言模型生成符合模式且多样化的输出。

自回归大语言模型广泛用于合成数据流水线，因为其归纳偏置有利于结构和词元顺序[17, 31], ，但这些相同的偏置也可能限制多样性，并导致重复或幻觉 [18, 49]。扩散语言模型则通过以非因果、双向的方式迭代去噪序列来生成文本 [16, 26], ，从而促进更广泛的语义变化。然而，它们缺乏用于格式保持的位置先验，因此不太适合嵌套 JSON 合成等结构敏感型任务 [22]。AoD 旨在结合两者的优势，同时弥补其不足。

近期，提示调优、强化学习和基于智能体的协调方面的进展提升了对自回归模型的控制能力 [25, 29, 34], ，但由于扩散语言模型问世不久，针对它们的类似方法仍鲜有探索。AoD 通过仅依靠语言对齐性来实现对扩散语言模型的智能体监督，从而弥补了这一空白。我们的优化循环无需调整参数：冻结的扩散语言模型（LLaDA-8B） [27] 不会更新权重，智能体之间仅通过自然语言交互，这有助于提升可解释性并实现模型无关控制。为此，我们提出一种强化学习算法，将近端策略优化（PPO）与 REINFORCE 原理相结合，利用自然语言反馈作为替代奖励信号来优化提示更新。

我们在四个要求语义流畅且严格遵循 JSON 模式的结构化生成基准上评估 AoD：**Mul-tiWOZ**、**Super-NaturalInstructions**、**Self-Instruct** 和**Truth-fulQA**。这些数据集包含嵌套字段、多种模式格式和丰富多样的语言风格，为结构感知型扩散语言模型控制构建了一个颇具挑战性的测试平台。在这组基准上，AoD取得了最高的**任务成功率** (0.79)和最低的**字段重叠度**

(0.29)，表现优于扩散和自回归基线，同时表明生成的输出有效且并非记忆所得。

该架构高效、可复现且易于使用。AoD 支持 LLaMA 3.1 8B、Qwen-3 8B、DeepSeek-R1 8B 和 Gemma-2 9B 等本地开放权重模型，也支持 GPT-4.1、GPT-4.1 Mini 和 GPT-4.1 Nano 等专有 API 模型。这种灵活性使其既能在 GPU受限环境中运行，也能在高性能云端环境中运行。所有实验均结合使用消费级硬件和 API端点，表明 AoD 无需专用基础设施，也能实现高质量的可控结构化生成。

贡献。 (1) 我们提出了**扩散智能体**，这是首个使用自然语言引导扩散语言模型的多智能体强化学习框架。(2) 我们提出了一种优化循环，其中大语言模型智能体通过语言批评迭代优化提示，无需奖励建模即可实现模式对齐控制。
(3) 我们在多个结构化数据集上展示了基于 JSON 的指令合成的可复现最先进结果，为符号化、格式受限领域中的可控生成奠定了基础。

2 背景与相关工作

2.1 结构化文本合成数据

合成数据在机器学习中的重要性日益凸显，尤其是在真实数据有限、敏感或获取成本高昂的情况下 [2]。尽管大语言模型在自由形式文本生成方面已初见成效 [1, 21, 46], ，但生成高质量的结构化数据（如表格记录或 JSON 输出）仍是一项重大挑战 [20]。大语言模型生成的输出通常语法正确，但在需要生成嵌套结构数据时，往往会出现幻觉并重复输出。 [39, 47]。以往的解决方案依赖包含验证模块或潜变量建模的处理流程 [9, 45], ，但这些方法难以扩展。受扩散模型在计算机视觉领域取得成功的启发 [32, 33], ，近期研究探索了将其应用于文本 [13, 48], ，但对结构的控制仍然有限。我们的工作通过使用扩散语言模型来提升合成数据的语义多样性，同时借助多智能体强化学习框架中的持续评估，确保结构保真度，从而弥补这一不足。

2.2 自回归语言模型

自回归语言模型通过按顺序预测每个词元来生成文本，这一范式能够实现较强的上下文连贯性和结构对齐 [4, 36, 38]。其单向解码使其能够有效生成语法正确且符合模式的输出，例如 JSON 或表格格式 [19, 35]。然而，对从左到右生成方式的依赖往往会限制输出多样性，强化高概率模式，并导致通用或重复的序列 [7]。尽管采样策略能够在一定程度上缓解这一问题，但自回归语言模型固有的序列偏差限制了其生成灵活性 [8, 28]。这促使我们探索扩散模型：它们能够实现双向且更加多样化的生成，同时通过多智能体控制保留结构。

2.3 扩散语言模型

扩散语言模型（扩散语言模型）通过迭代去噪生成文本，相较于自回归方法，能够实现更加灵活且语义更加多样的生成。Diffusion-LM [23]引入了连续潜在框架，用于可控的文本合成；LLaDA [27] 则扩展了掩码序列扩散，在推理和语言任务上达到或超越了自回归基线。DiffLM [50] 将离散扩散应用于结构化表格数据。尽管这些模型展现出良好潜力，但仍面临一些关键局限：在 JSON 等受约束格式中，结构保留仍然较为脆弱；提示调优对扩散语言模型行为的影响尚未得到充分理解；此外，它们在多智能体协调中的作用也基本未被探索。为弥补这些不足，我们将一个冻结的扩散语言模型嵌入由自回归大语言模型智能体构成的多智能体强化学习循环中，并利用自然语言反馈和提示空间优化，实现符合模式且语义对齐的生成。

2.4 提示调优与优化

提示优化被广泛用于在不进行微调的情况下调整大语言模型，相关方法包括思维链 [43] 等人工设计的推理策略，以及 Promptbreeder [12] 和 EvoPrompt [14]等自动化方法。此外，也存在用于增强提示效果的自评分标准策略，例如 CodecLM [42]。尽管这些方法有效，但它们将提示设计视为静态搜索问题，在离线优化提示时没有考虑实时反馈或生成动态 [3]。这限制了它们的适应能力，尤其是在需要结构精确性或迭代优化的任务中 [10]。我们的工作将提示优化重新定义为一个强化学习问题，其中提示智能体通过自然语言反馈学习优化指令。这种方法与扩散语言模型尤为契合，因为其迭代解码允许提示在多个去噪步骤中引导生成，从而实现响应迅速且具备结构感知能力的控制。

2.5 多智能体语言协调

使用大语言模型的多智能体系统（MAS）通过将推理分配给专业化的角色定义智能体，并让它们通过自然语言交流，在复杂任务上展现出强大的性能 [15, 37]。批评、协作或投票等去中心化协议，能够在代码合成和科学发现等领域产出稳健且多样化的解决方案 [11, 44]。然而，基于语言的协调容易出现交互漂移、智能体行为不一致，以及多轮交互中的记忆保持能力较弱等问题 [6, 30]。我们通过以扩散语言模型为中心锚定智能体交互来应对这些挑战。该模型采用迭代式双向解码，从而建立稳定的共享参考点。这种结构能够减少漂移、确保语义一致性，并为跨对话轮次的提示优化和评估提供依据，从而改善协调效果，且无需外部记忆模块或监督。

3 方法论

3.1 预备知识

令 X 表示结构化文本字符串（JSON对象）的空间，令 q_{real} 表示 X 上未知的概率分布。我们观测到 n 个独立同分布（i.i.d.）样本

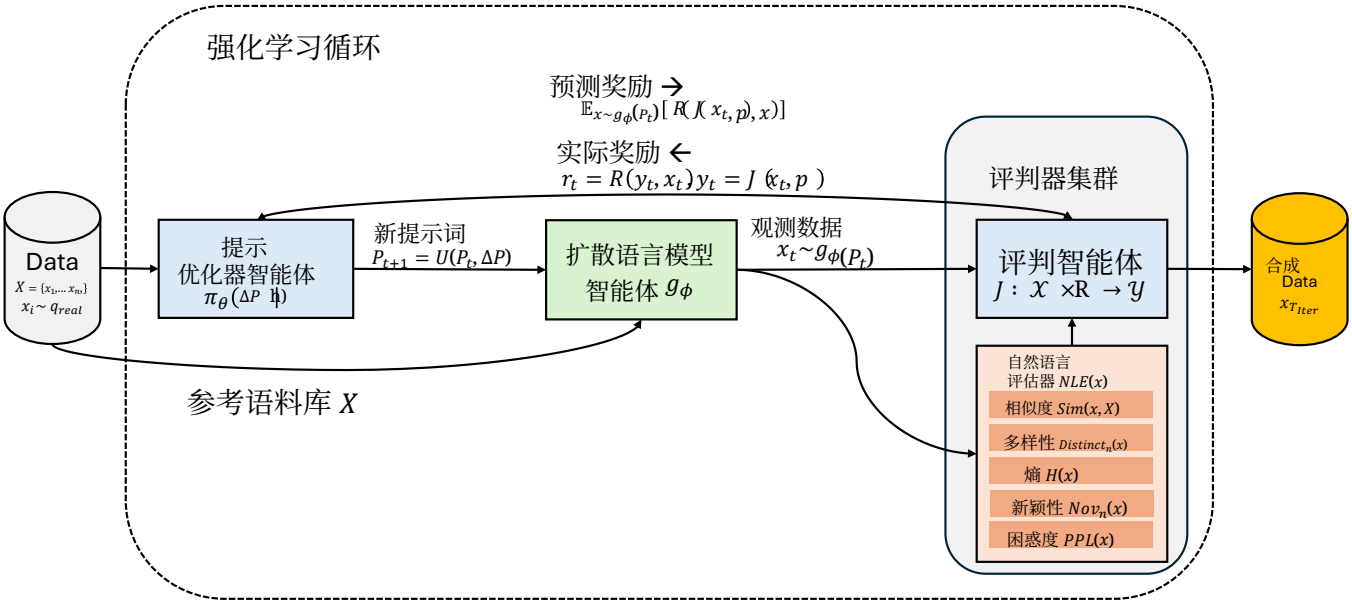

图1：Agents of Diffusion：多智能体训练框架概览。

$X=\{x_1、\ldots\ldots、x_n\}$，其中 $x_i \sim q_{\text{real}}$。模式 S定义了有效子集 $\mathrm{L}(S) \subseteq \mathrm{X}$，二元验证器 $V_S: \mathrm{X} \to \{0、1\}$在 $x \in \mathrm{L}(S)$ 时返回 $V_S(x) = 1$，否则返回 0。接下来，令 P 为提示空间，P_Δ 为允许的提示编辑集合。参数为 ϕ $g_{\phi\,\tau}\ z_{\tau-1}$ |的扩散语 z_τ P $\tau = 1 \ldots T$ v_T言模型（DLM）定义了反向时间马尔可夫核、(、)（适用于、……、），以及潜在词元空间 Z 上的噪声先验（解码后与 X 对应）。要从 $x \sim g_\phi(P)$ 采样，先抽取 $z_T \sim v_T$，然后通过 $z_{\tau-1} \sim g_\phi、\ _\tau(z_{\tau-1} \mid z_\tau、\ P)$ 迭代去噪，直到 $x = z_0$。由核 $k_\tau(z_\tau \mid z_{\tau-1})$ 定义的前向破坏过程是标准的，此处省略。

该框架包含四个相互作用的组件 $\mathrm{A} = \{$扩散语言模型 J，R，$\pi\}$。扩散语言模型 g_ϕ生成文本样本。评判器$J: \mathrm{X} \times \mathrm{R} \to \mathrm{Y}$ 将候选项 x和评分标准 $\rho \in \mathrm{R}$ 映射为反馈 $y = J(x,\ \rho) \in \mathrm{Y}$。评分器 $R: \mathrm{Y} \times \mathrm{X} \to \mathbb{R} \times \mathbb{R}^k$返回一个标量奖励 $r = R(y,\ x)$ 和一个子奖励向量 $\mathbf{s} = R_{\text{vec}}(y,\ x) = (s_1,\ \ldots,\ s_k)$，其中每个 s_i都表示一种可测量的质量，例如语义相似度、多样性、新颖性、困惑度或熵。

提示优化器是一种随机策略 $\pi_\theta(\Delta P \mid h)$，由 $\theta \in \Theta \subset \mathbb{R}^d$参数化，其中 Θ 是可接受的参数向量集合（例如神经网络权重）。给定总结了先前提示、样本、反馈和评分的历史记录$h \in \mathrm{H}$，π_θ 会输出编辑操作的分布 $\Delta P \in \mathrm{P}$。编辑算子 $U: \mathrm{P} \times \mathrm{P}_\Delta \to \mathrm{P}$ 随后确定性地更新提示，形式为 $P^+ = U(P,\ \Delta P)$。外层优化迭代的索引为 $t = 1,\ \ldots,\ T_{\text{iter}}$，而扩散步骤的索引为 $\tau = 1,\ \ldots,\ T$。

最后，对于任意随机变量 Z，除非另有说明，$\mathbb{E}[Z]$表示关于来自 q_{real}、扩散语言模型采样过程g_ϕ以及策略 π_θ的随机性的期望。

3.2 问题表述

如图1所示，AoD 将文本JSON数据的合成形式化为多智能体强化学习框架中的可控结构化生成问题。其目标是将扩散语言模型的语义多样性与自回归智能体的结构精确性结合起来。在此设置中，扩散模型提出多样化的 JSON 候选项，而两个大语言模型智能体通过自然语言反馈迭代优化条件提示，以确保符合模式并保持语义一致性。这种协作反馈循环无需修改模型参数，即可实现结构化文本的可控生成。

在每次外层迭代 $t \in \{1, \ldots, T_{iter}\}$ 中，系统维护一个提示 $P_t \in \mathrm{P}$ 和一份摘要历史 $h_t \in \mathrm{H}$。扩散语言模型 g_ϕ生成一个结构化候选项 $x_t \sim g_\phi(P_t)$。评判器 J依据评分标准 $\rho \in \mathrm{R}$ 对其进行评估x_t，生成反馈 $y_t = J(x_t, \rho)$；评分器 R将其转换为数值信号 $r_t = R(y_t, x_t)$ 和 **分**$_t = R_{vec}(y_t, x_t)$。提示优化器以 θ 为参数，对编辑 $\Delta P_t \sim \pi_\theta(\Delta P \mid h_t)$ 进行采样，并通过以下方式更新提示：

$$P_{t+1} = U(P_t, \Delta P_t). \tag{1}$$

这定义了一个情节式马尔可夫决策过程（MDP），其状态为 $z_t = (P_t, h_t)$，动作是 $a_t = \Delta P_t$，转移核由 (g_ϕ, J, R, U) 诱导，奖励为 r_t。策略参数 θ经过优化，以最大化期望折扣回报：

$$\max_\theta \mathbb{E}\left[\sum_{t=1}^{T_{iter}} \gamma^{t-1} r_t\right] \tag{2}$$

受资源约束

$$Tokens \le B_{tok} \quad Calls \le ,,,)$$

B_{calls} $T_{iter} \le B_{iter}$ $\gamma \in [0, 1$.

算法 1 Agents of Diffusion（AoD）中的多智能体强化学习循环

输入： 模式 S、评分标准 ρ、扩散模型 g_ϕ、策略 π_θ、编辑算子 U
初始化： 提示 $P_1 \in \mathrm{P}$、摘要历史记录 $h_1 \in \mathrm{H}$
对于 $t = 1$ **到** T_{iter}，**执行**
  采样结构化候选项 $x_t \sim g_\phi(P_t)$
  评判器提供反馈 $y_t = J(x_t, \rho)$
  计算奖励 $r_t = R(y_t, x_t)$ 和子奖励 $\mathbf{s}_t = R_{vec}(y_t, x_t)$
  更新历史摘要 $h_t = f(P_t, x_t, y_t, \mathbf{s}_t)$
  采样提示编辑 $\Delta P_t \sim \pi_\theta(\Delta P \mid h_t)$
  应用编辑：$P_{t+1} = U(P_t, \Delta P_t)$
  更新策略参数：$\theta \leftarrow \theta + \eta \widehat{\nabla}_\theta \mathbb{E}_{\pi_\theta}[r_t]$
结束循环
返回： 最终提示 $P_{T_{iter}}$，最终候选项 $x_{T_{iter}}$

在收敛时（T_{iter}），使用能够衡量结构有效性、语义相关性和多样性的终端目标来评估性能：

$$\max_\theta \ \alpha\,\mathbb{E}[V_S(x_{T_{iter}})] + \beta\,\mathbb{E}[Sim(x_{T_{iter}}, X)] + \delta\,\mathbb{E}[Distinct_n(x_{T_{iter}})],$$

或者等价地，以约束形式表示，

$$\max_\theta \ \mathbb{E}[Distinct_n(x_{T_{iter}})] \tag{3}$$

使得 $\mathbb{E}[V_S(x_{T_{iter}})] \geq \tau_{valid}$，$\mathbb{E}[Sim(x_{T_{iter}}, X)] \geq \tau_{sim}$.

扩散参数 ϕ 保持固定；可控性完全由作用于提示的自回归策略 π_θ 实现。

3.3 DLM 集成多智能体强化学习

提示优化智能体。提示优化智能体通过引导扩散语言模型生成符合模式且语义连贯的内容，掌控基于文本的结构化 JSON 合成中的可控性。它采用自回归大型语言模型，因为提示编辑是顺序进行且依赖词元的，因此自回归架构非常适合在文本空间中学习离散编辑轨迹。该智能体定义了一个由 θ 参数化的随机策略 $\pi_\theta(\Delta P \mid h)$，生成受上下文引导的提示更新，以调节 g_ϕ 的条件输出分布。其目标是最大化评判器–评分器组合 (J, R) 提供的预期奖励：

$$\pi_\theta^* = \arg\max_{\pi_\theta} \mathbb{E}_{x \sim g_\phi(P)}[\,R(J(x, \rho), x)\,], \tag{4}$$

其中，ρ 是用于衡量模式和语义保真度的特定任务评分标准。由于 g_ϕ 不可微，π_θ 充当替代函数优化器，通过离散的、由语言条件化的更新来近似 $\nabla_\theta J(P) = \mathbb{E}_{x \sim g_\phi(P)}[R(J(x, \rho), x)]$ 的梯度，而非通过反向传播实现。

定理 1. 令 $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)}[R(J(x, \rho), x)]$ 如式（4）所示。若 $\mathcal{T}(P)$ 是局部利普希茨连续，且从 $\pi_\theta(\Delta P \mid h_t)$ 采样的提示编辑 ΔP_t 有界，则对于足够小的步长，迭代更新 $P_{t+1} = U(P_t, \Delta P_t)$ 在期望意义下构成压缩映射。因此，序列 $\{P_t\}$ 收敛到一个不动点 P^*，该不动点满足

$$\mathcal{T}(P^*) = \max_P \mathcal{T}(P),$$

从而确保稳定收敛至与奖励一致且符合模式的提示。

该智能体将离散自回归推理作为扩散动力学之上的控制层，并将自然语言反馈转化为词元级模式对齐步骤。例如，在包含出发_城市、到达_城市和日期字段的预订模式 S 下，初始提示“生成旅行详情”可能会 eliciting 出“使用 JSON 格式并包含所有字段”之类的反馈。通过迭代更新 ΔP_t，该智能体将其细化为“生成一个包含字段 {origin, destination, date} 的 JSON对象，并使用 YYYY-MM-DD 格式”。每次细化都会逐步将扩散语言模型（DLM）的条件分布 $p_\phi(x \mid P_t)$ 与有效模式子集 $\mathrm{L}(S)$ 对齐，同时保留语义多样性。从理论上讲，该机制连接了离散符号推理与随机生成，从而实现结构化文本合成中的强化驱动自适应。

扩散语言模型智能体。扩散语言模型（DLM）$g_\phi(z_{\tau-1} \mid z_\tau, P)$ 作为生成式骨干，用于合成基于文本的结构化 JSON 数据。它定义了一个逆向时间马尔可夫过程，从逐步去噪的潜在表示中重建文本词元，从而得到条件分布 $p_\phi(x \mid P)$。不同于按顺序分解 $p(x \mid P)$ 为 $\prod_i p(x_i \mid x_{<i}, P)$ 的自回归模型，扩散语言模型通过迭代去噪隐式估计 $p_\phi(x \mid P)$。这种非因果双向形式使每个去噪步骤都能够基于全局上下文，而非局部词元依赖关系进行条件建模，从而更广泛地覆盖有效的符合模式的配置。

命题 1. 设 $p_\phi(x \mid P)$ 和 $p_{AR}(x \mid P)$ 分别表示在同一结构化数据集上、基于模式 S 训练的扩散条件分布和自回归条件分布。若二者在重构误差有界且扩散时域有限 T 的条件下，均最小化与真实数据分布 $q_{\text{real}}(x)$ 之间的散度，则

$$\mathrm{KL}(q_{\text{real}}(x) \,\|\, p_\phi(x \mid P)) \leq \mathrm{KL}(q_{\text{real}}(x) \,\|\, p_{AR}(x \mid P)),$$

这表明扩散更能逼近真实数据流形，并捕获更广泛的语义有效配置。

从理论上讲，这一性质使扩散语言模型成为 AoD 中负责多样性的智能体。它拓展了 $p_\phi(x \mid P)$ 在多个有效模式实例上的支持范围，从而能够合成多样且连贯的 JSON 结构。相比之下，自回归智能体提供约束机制，确保遵循 $\mathrm{L}(S)$ 的句法和语义要求。二者共同构成互补系统：扩散促进多样性，自回归则保证结构。这实现了受控且与模式对齐的结构化文本数据生成。

评判智能体。评判器集群由基于大语言模型的评判器 J 和自然语言评估器（NLE）组成，构成评估子系统，负责解读并监督扩散语言模型 g_ϕ 的输出。NLE 接收每个生成的 JSON 样本 $x \in \mathrm{X}$，并计算预备知识中定义的五项定量指标：语义相似度 $\mathrm{Sim}(x, X)$、多样性 Distinct-$n(x)$、熵 $H(x)$、新颖性 Nov-$n(x)$ 和困惑度 $\mathrm{PPL}(x)$。这些指标共同描述了 x 与参考数据集 X 的对齐程度、其多样性和流畅度，以及它能否泛化到未见过的样本。随后，NLE 将这些

数值转换为结构化自然语言陈述，以易于理解的形式概述偏差和归因。大语言模型评审器 J以自回归模型的形式实例化，将这些结构化文本反馈与评分标准 ρ一同输入，并通过一组固定的、与评分标准对齐的是／否问题（例如：“JSON 结构是否完整？”“是否包含所有必需字段？”“文本在语义上是否忠实？”）评估每个样本。结合定量评估和自身的上下文推理， J生成最终批评 $y = J(x, \rho)$，评分器再将其转换为标量奖励 $r = R(y, x)$和子奖励向量 $\mathbf{s} = R_{\text{vec}}(y, x)$。例如，该集群可能会生成：“*JSON* 语句流畅，但缺少arrival_city；日期格式应为 *YYYY-MM-DD*。”随后，这些反馈会指导提示优化策略 $\pi_\theta(\Delta P \mid h)$，从而完成强化循环。

相应地，NLE 将生成文本中可测量的属性分离出来，提供低方差、相互解耦的反馈信号，从而避免噪声梯度并提高下游优化的可靠性。大语言模型评审器 J将这些离散测量转换为底层奖励景观的平滑自然语言代理，在保持语义透明性的同时实现无梯度优化。 J的自回归表述对于结构化 JSON 合成尤为重要，因为评估符合性、字段顺序和键依赖关系需要对词元进行序列推理。通过以从左到右的方式处理反馈， J在其批评中保持因果一致性，并确保与 π_θ对 P执行词元级编辑的方式保持对齐。因此，自回归推理在保持语义灵活性的同时强制实现结构化连贯性，使系统无需手动规则工程即可跨不同模式进行泛化。

从理论上讲，该集群定义了一个期望反馈算子

$$\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)}[R(J(x, \rho), x)] \tag{5}$$

该算子能够稳定不可微环境中的学习（g_ϕ， J， R）。NLE 将 T（P）建立在可验证的定量信号之上，而自回归评判器则通过将离散验证结果映射为连续的语言解释来平滑不连续性。这种协同作用降低了策略梯度估计的方差，改善信用分配，并强化奖励相对于符合模式且语义忠实的输出的单调性。二者共同构成了数值监督与符号提示控制之间理论一致的桥梁，在保持自回归模型结构化精确性的同时，兼顾基于扩散的合成所具备的生成多样性。

命题 *2*. 设 $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)}[R(J(x, \rho)\ x)]$。假设 $R_{vec}(y, x)$ 中的每个分量都有界，且从 *NLE* 的度量向量到文本反馈 $y = J(x, \rho)$ 的映射关于 *(*相似度*(x, X), Distinct-n(x) H(x), Nov-n(x), PPL(x))* 满足利普希茨连续性，常数为 L_J，并且 R在这些分量上单调递增。那么 $\mathcal{T}(P)$ 关于 P 是局部利普希茨连续的，并且保持语义保真度方面的排序：若 P_1, P_2诱导出满足 $\mathbb{E}[$相似度$(x, X) \mid P_1] > \mathbb{E}[$相似度$(x, X) \mid P_2]$ 的样本，则 $\mathcal{T}(P_1) > \mathcal{T}(P_2)$。因此，由 $\mathcal{T}(P)$ 驱动的策略更新在期望上是稳定的，并优先考虑语义忠实的提示。

4 实验与结果

4.1 实验设置

硬件。 为展示 AoD 的易用性和可复现性，所有实验均在一台面向消费者的工作站上运行，配置为 AMD Ryzen 9 7900X（12 核、24 线程、4.7 GHz 基频）、32 GB DDR5 内存和 NVIDIA RTX 4080 SUPER GPU（16 GB 显存）。这一配置体现了个人研究人员和开发者广泛可用的硬件，确保 AoD 无需依赖专用基础设施或大型计算集群。

模型。 我们的实验同时采用开源语言模型和基于 API 的语言模型，以突出 AoD 的灵活性和硬件独立性。我们使用了八个自回归模型来承担提示优化器和大语言模型评审器的角色：**LLaMA-3.1 8B**（32 层、40 个注意力头、4 位量化；用于提示时温度为 0.7，用于评判时为 0.2）、**Qwen-3 8B**（多语言、支持 LoRA，使用核采样，参数为 $p = 0.9$）、**DeepSeek-R1 8B**（NTK 感知，使用 top-k 采样，参数为 $k = 40$）、**Gemma-2 9B**（束搜索宽度为 3）、**Mistral 7B**（分组查询注意力、8 位解码），以及三个基于 API 的模型：**GPT-4.1 Nano**、**Mini**和**GPT-4.1**，它们支持从轻量级到高保真度的生成。为确保一致性并简化计算，所有实验均对提示优化器和大语言模型评审器使用相同的自回归模型。相比之下，生成器智能体全部采用**LLaDA 8B**，这是一种离散扩散语言模型，具有 32 层、正弦嵌入、1024 词元输入以及 $T = 12$ 个去噪步骤。LLaDA 运行于 FP16 模式，并禁用无分类器引导，以保持采样多样性。在本地模型和 API 模型上，AoD 均展现出稳定的高质量性能，无需依赖专用硬件。

数据集。 我们在四个公开数据集上评估用于合成 JSON 数据的框架，这些数据集因结构、语义和生成目标的多样性而入选。**MultiWOZ** [5]是一个多领域对话数据集，包含丰富的槽位填充标注，可用于测试结构保持和模式对齐。**Super-自然指令** [41]包含数百项任务中的多样化指令-响应对，有助于测试模型对不同提示类型的泛化能力。**TruthfulQA** [24] 提供了挑战事实性的问答，可用于评估幻觉和语义精确度。**Self-指令** [40] 由从大语言模型中提炼的基于指令的示例构成，支持在对齐性要求严格的任务中开展提示-响应合成实验。训练和评估前，我们会从每个数据集中随机抽取子集，以降低计算开销并减少记忆化。通过仅向模型提供完整数据集中的少量随机样本，我们最大限度地降低模型在生成过程中照搬特定示例的风险，并确保性能反映其对未见指令的泛化能力。

基线。 我们将 AoD 与六种互补的基线进行比较，这些基线代表了结构化文本生成的主要范式。**Diffusion-LM** [23] 和 **DiffLM** [50] 是基于扩散的模型，强调语义多样性和多模态覆盖，作为面向多样性的参考方法。**CodecLM** [42]、**PromptBreeder** [12], 和**EvoPrompt** [14] 是自回归提示优化方法，提供了强有力的面向控制的基线，体现了现有的结构化提示和语言驱动优化策略。最后，**UniGen** [45]代表基于验证的合成，在生成时执行符号级和模式级

约束。这六种基线共同覆盖了结构控制 （基于自回归）、多样性 （基于扩散）和约束执行 （基于验证）这三类设计空间，为评估 AoD 的贡献提供了全面的比较框架：通过多智能体强化学习机制将三者统一起来，生成既符合模式又具有语义多样性的结构化 JSON 数据。

评估指标。我们的评估框架将用于智能体反馈的指标与用于独立验证的指标区分开来。在训练期间，评判智能体利用五项有依据的指标：困惑度 PPL(x)、语义相似度 Sim(x, X)、多样性 Distinct-n(x)、词元熵 $H(x)$ 和新颖性 Nov-n(x)，为提示优化器生成可解释的自然语言反馈。这些指标提供结构和语义层面的监督，同时不直接暴露标量奖励值，从而避免强化偏差或对数值目标过拟合。不过，作为独立评估的衡量方式，我们报告标准的文本质量指标，包括 BLEU、ROUGE 和 METEOR，以量化生成样本与参考样本在句法和词汇上的对应程度。这确保我们报告的性能反映的是真实的生成质量，而非强化反馈偏差。此外，为了评估合成数据的下游微调能力和功能可靠性，我们还计算任务成功率（TSR），用于衡量满足所有约束的有效、语义一致且多样化的生成结果所占比例。

记忆与串通验证。 训练期间使用的五项奖励指标（困惑度 PPL(x)、语义相似度 Sim(x, X)、多样性 Distinct-n(x)、词元熵 $H(x)$ 和新颖度 Nov-n(x)）共同调节多智能体强化学习循环中的记忆化与串通。每项指标都施加一项独特的行为约束：PPL(x) 确保语言连贯性并惩罚退化文本，Sim(x, X) 促进语义对齐，Distinct-n(x) 和 $H(x)$ 鼓励词汇多样性，而 Nov-n(x) 则惩罚对参考数据集 X 的逐字复用。评判器利用这些量化信号回答一组固定的、基于评分标准的是／否问题，确保强化过程以客观的结构和含义为依据，而非智能体之间的隐蔽协同。由于提示优化器 π_θ从不直接观察标量奖励，而是接收由这些奖励生成的自然语言反馈，因此无法通过串通或记忆化来利用奖励函数。跨迭代的数值趋势可提供诊断信号，例如 Sim(x, X) 同时上升、 Nov-n(x) 下降，便能明确揭示潜在的数据泄漏或过拟合。

为进一步确保独立性并验证生成数据与训练语料库保持区别，我们引入字段重叠度指标，作为一种事后度量。字段重叠度计算生成的 JSON 样本中与参考集完全匹配的键值对或字段所占的比例。较高的重叠值表明可能存在复制或记忆化，而较低的重叠度结合较低的 PPL(x)，则表明实现了忠实泛化和连贯生成。与衡量语义相似度的 Sim(x, X)不同，字段重叠度明确衡量结构重复，因此可直接检验记忆化或智能体间的信息泄漏。综合这些信号，既可以在循环中验证数据新颖性，也可以进行独立验证，从而确保 AoD 生成多样、语义忠实且无偏的合成 JSON 数据。

4.2 讨论

表 1 突显了 AoD 在结构精确性与生成多样性之间实现了难得的平衡，优于基于扩散的方法和自回归系统。高相似度（0.88）与较强的多样性（0.72）和新颖性（0.83）相结合，表明 AoD 生成的数据在保持语义忠实度的同时，还能探索符合模式的全新配置。熵得分（6.10）表明词汇丰富度均衡，而非措辞重复；较低的困惑度（22.1）则证实其语言建模流畅且连贯。这种表现并非偶然，而是直接体现了 AoD 架构中探索与约束之间的平衡。扩散生成器带来随机性和广度，而评判器集群则施加语言和结构约束，以稳定输出空间。强化学习引导的优化协调了这两种相反的力量，生成的样本既富有创意，又在构成上有效，即使面对具有复杂嵌套结构的数据集亦是如此。

独立指标进一步印证了这一解读。AoD 在 BLEU、ROUGE-L 和 METEOR 上表现领先，表明其多样性并未损害语法或语义保真度。任务成功率（0.79）较高，说明生成的记录满足内容和结构要求；而最低的字段重叠度（0.29）则证实，AoD 能够生成训练数据中未见过的不同键值组合，从而避免记忆化。这些数值规律源于多智能体奖励结构：自然语言评估器（NLE）将离散的指标信号转化为分级的语言反馈，从而保持解读连贯性；自回归评判器则对各个键和字段进行顺序验证。这种分层监督稳定了奖励传播，使优化过程更加平稳，并防止对数值启发式过拟合。指标结果清晰地体现了这一点——相似度和困惑度同步提升，多样性和新颖性在不偏离结构的前提下提高，而熵仍保持较高水平且连贯一致。因此，AoD 的强化学习信号同时编码了形式与含义，使模型在未见分布上依然具备泛化能力。

这些动态也阐明了为什么 AoD 能兼具较低的困惑度和较高的熵。传统扩散系统虽能提高多样性，却常常生成句法不稳定的文本；而自回归系统则以牺牲变化性为代价来确保结构。AoD 将扩散驱动的探索与顺序约束验证相结合，弥合了这一差距。NLE 的语言基础使词汇和语义扩展能够以受控的方式进行，而评判器的自回归推理则确保字段间的因果依赖。两者共同产生了观察到的平衡：文本具有高熵，同时保持句法流畅、语义一致。这一机制解释了 AoD 为何能避免典型的扩散漂移——趋向不连贯，以及自回归偏差——趋向重复。基线模型之间的差异进一步凸显了 AoD 的优势。LLaMA 和 Qwen 等静态自回归系统优先进行条件似然最大化，从而保持较高的相似度

表1：AoD 与各基线的比较。前五项指标对应训练中使用的指标，后五项则作为独立评估指标。除困惑度和字段重叠度外，所有指标均为越高越好。AoD 的数值为所有数据集以及提示优化器 +评判大语言模型组合上的平均值。每项实验重复进行 15 次。

模型	相似度	多样性	新颖性	熵	困惑度	BLEU	ROUGE-L	METEOR	TSR	字段重叠度
静态自回归基线（单轮提示）										
LLaMA-3.1 8B	0.86	0.42	0.48	5.18	21.6	33.9	38.1	27.9	0.71	0.38
Qwen-3 8B	0.85	0.44	0.50	5.22	22.3	34.1	37.6	27.8	0.70	0.36
DeepSeek-R1 8B	0.87	0.41	0.47	5.14	20.9	35.2	38.8	28.0	0.73	0.35
Gemma-2 9B	0.84	0.43	0.49	5.19	22.0	34.5	37.8	28.1	0.72	0.37
Mistral 7B	0.83	0.44	0.48	5.25	23.2	33.7	37.2	27.5	0.69	0.39
GPT-4.1 Nano	0.84	0.46	0.55	5.16	21.9	30.5	36.0	26.2	0.66	0.38
GPT-4.1 Mini	0.84	0.47	0.56	5.12	21.5	30.9	36.5	26.4	0.67	0.37
GPT-4.1	0.85	0.48	0.56	5.10	21.2	31.0	36.8	26.6	0.68	0.37
扩散与基于提示的基线方法										
Diffusion-LM [23]	0.72	0.60	0.72	5.82	29.4	28.1	33.5	25.1	0.61	0.42
DiffLM [50]	0.74	0.63	0.70	5.90	28.6	27.5	32.9	24.6	0.63	0.41
UniGen [45]	0.78	0.52	0.63	5.64	27.5	30.8	35.0	26.0	0.67	0.40
PromptBreeder [12]	0.80	0.51	0.59	5.51	25.7	31.2	36.7	26.5	0.68	0.38
EvoPrompt [14]	0.81	0.49	0.57	5.48	25.1	32.4	37.0	27.0	0.70	0.37
CodecLM [42]	0.82	0.47	0.56	5.42	24.8	33.0	37.5	27.3	0.71	0.36
LLaDA [27]	0.79	0.69	0.81	6.03	27.0	29.5	34.2	25.8	0.69	0.35
AoD（我们的方法）	**0.88**	**0.82**	**0.83**	**6.10**	**22.1**	**35.6**	**40.1**	**29.3**	**0.79**	**0.29**

但由于采用确定性解码，多样性和新颖性却随之下降。DiffLM 或 Diffusion-LM 等扩散基线则呈现相反的模式：它们生成的数据多样，却在结构上较为脆弱，因为其去噪轨迹缺乏模式感知条件控制。EvoPrompt 和 PromptBreeder 等提示演化框架通过启发式变异提高了变化性，但由于缺乏对序列编辑的信用分配，无法持续取得进展。相比之下，AoD 通过提示优化器形成了这一优化循环；该优化器根据评判器集群提供的多维反馈，动态优化提示。提示优化器学习一种策略，不仅依据奖励幅度调整提示，也会考虑错误的语言语境，从而在多次迭代中持续改进。

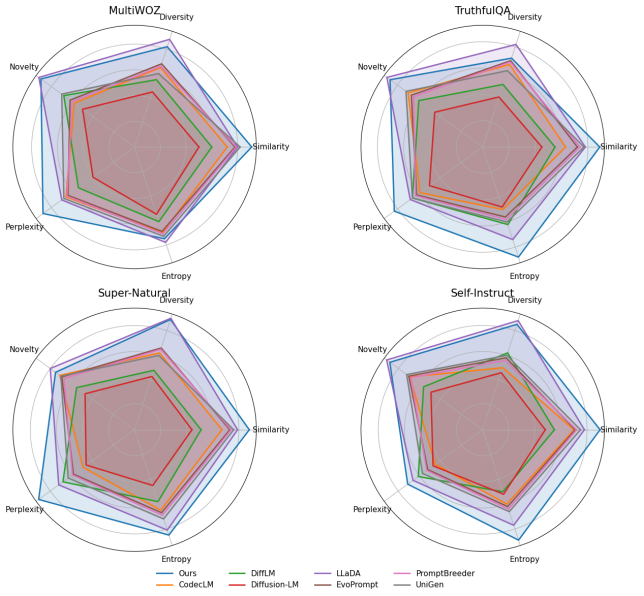

图2：不同数据集上的归一化指标对比。

图2展示了 AoD 在各数据集上的一致性，其多边形沿所有指标轴均匀扩展。在 MultiWOZ 和 Super-Natural 这类结构复杂的领域中，AoD 在拓宽多样性和熵的同时保持较高的相似度和较低的困惑度，体现了其在鼓励词汇变化的同时维持模式完整性的能力。在 TruthfulQA 和 Self-Instruct 等推理密集型数据集上，AoD 平衡了新颖性与相似度，表明评判器集群与扩散生成器协同调节泛化能力，避免过拟合。竞争方法要么以牺牲结构为代价偏重多样性（Diffusion-LM、DiffLM），要么维持结构但新颖性较低（PromptBreeder、EvoPrompt）。AoD 在全部四个数据集上的区域形状近乎规则，证实了其性能稳定且能适应不同的数据模态，进一步彰显了其作为兼顾模式忠实度与多样性的结构化数据生成器的作用。

4.3 消融研究

智能体与奖励消融。 图 3 展示了 AoD 中各智能体组件和反馈机制的影响。引入自回归提示优化器（PO）后，所有指标均有所提升。这表明，即使没有强化学习，通过迭代进行结构化编辑来优化提示，也能提升流畅度和模式对齐性。然而，由于优化过程是确定性的，且缺乏随机探索，多样性和熵仍受到限制。加入基于标量奖励的强化学习（*FRL–AR–S*）后，相似度进一步提升，困惑度降低，因为标量反馈能够针对语法和句法正确性进行结构化奖励塑形。然而，自回归生成器的解码偏差会缩小输出空间，使其局限于高似然度的后续内容，因此多样性和熵的提升有限。将标量奖励替换为自然语言反馈（*FRL–AR–NL*）后，奖励传播更加平滑，流畅度也有所提升，因为评判器提供的可解释反馈能够实现词元级监督。尽管如此，自回归生成器的序列特性仍会限制可实现的多样性。

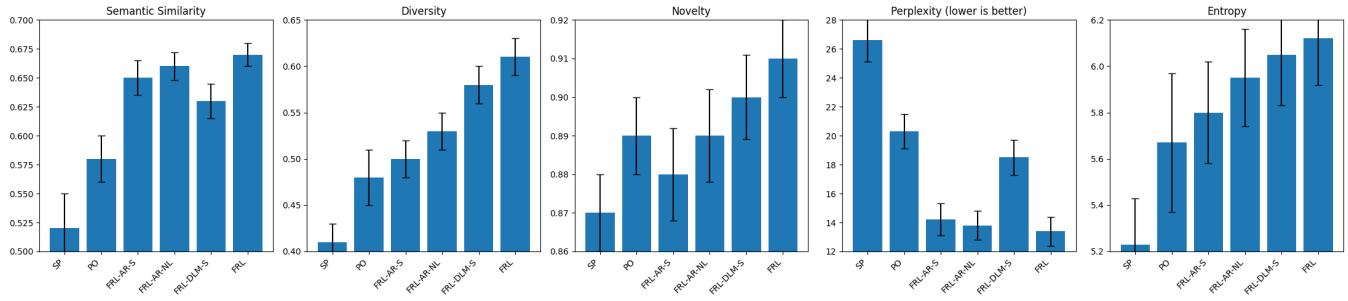

图3：跨数据集和优化器–评审器组合取平均值的智能体消融实验。SP = 静态提示，PO = 提示优化器，FRL–AR–S = 使用标量奖励的自回归生成器，FRL–AR–NL = 使用自然语言反馈的自回归生成器，FRL–DLM–S = 使用标量奖励的扩散生成器，FRL = 使用自然语言反馈的扩散生成器。

表2：提示优化器（行）和评判器（列）的迁移能力。报告所有数据集的多样性（↑）和 GPU 运行时间（秒）（↓）的平均值。LLaDA 为固定生成器。L= LLaMA，Q = Qwen，D = 深度求索，G = Gemma，M = Mistral。

	L		Q		D		G		M	
Opt.	Div.	Rt.	Div.	Rt.	Div.	Rt.	Div.	Rt.	Div.	Rt.
L	0.83	112	0.81	110	0.80	115	0.78	117	0.82	113
Q	0.82	118	0.85	119	0.84	116	0.79	118	0.80	121
D	0.80	115	0.83	117	0.86	114	0.82	118	0.81	120
G	0.79	120	0.81	121	0.82	122	0.84	118	0.83	119
M	0.82	111	0.83	113	0.81	114	0.79	115	0.84	112

在保留标量奖励的同时改用扩散生成器（*FRL–DLM–S*），多样性、新颖性和熵均有所提升。基于扩散的生成器 g_ϕ 能够进行双向上下文传播，使同一提示下涌现出多条语义轨迹，同时保持结构连贯性。然而，由于缺少自然语言反馈，相似度提升幅度有限，且困惑度方差增大，反映出奖励对齐较为粗略。完整的 AoD 配置（*FRL*）兼具两者优势：扩散驱动的多样性与基于语言的反馈。大语言模型评审器和 NLE 稳定了学习动态，降低困惑度，并增强了高相似度和熵，同时保持较高的新颖性。

模型可迁移性。 表2突显了 AoD 在多种自回归大语言模型组合下的模型无关行为，这些模型分别担任提示优化器和评审代理，而生成器固定为 LLaDA。在所有组合中，多样性得分始终较高（0.79–0.86），这表明无论底层模型架构如何，强化驱动的协同都能泛化。这说明 AoD 的策略学习作用于反馈和提示共享的语言空间，而非依赖任何特定模型的内部表示。因此，提示—反馈交换机制 $\pi_\theta(\Delta P \mid h)$ 不受优化器和评审器配置的影响，使智能体能够互换使用而不会导致性能崩溃。运行时间结果进一步证明，AoD 在消费级硬件上具有可复现性。每个反馈—生成周期的平均 GPU 运行时间为 110–122 秒，即使使用 8–9B 参数模型也是如此，这证实了在中端配置下，多智能体展开仍然切实可行。这种效率得益于冻结生成器g_ϕ 以及自回归智能体之间轻量级的通信循环，从而限制了反向传播开销。综上，这些结果表明，AoD 可利用开放权重模型或基于API的大语言模型复现，同时保持多样性和稳定性，使其无需依赖高端计算资源即可广泛应用。

案例研究：AoD 中的结构化JSON合成。 我们以 *MultiWOZ 2.1*预订领域作为任务，演示 AoD 的功能。提示优化器首先起草一条模式条件指令，例如：“生成一个包含字段 {name, address, phone, price_range, postcode} 的 JSON对象。”随后，扩散语言模型 g_ϕ生成多样化的样本，例如 {"name": "Parkview Inn", " address": "12 Milton Rd", "phone": "01223 443890", "price_ range": "moderate", "postcode": "CB4 1LG"}。NLE 计算指标，而大语言模型评审器将其转换为反馈，例如：*“JSON* 有效且流畅，但重复了价格模式；请引入更多独特的名称和地点。”优化器通过 $\pi_\theta(\Delta P \mid h)$ 使用这些反馈，迭代改进提示具体性和采样平衡。经过连续迭代，相似度从 0.64 上升至 0.88，多样性和新颖性均超过 0.80，困惑度则从 31.2 降至 22.5。评判智能体确认，生成的 JSON记录在语法上保持正确，同时彼此具有差异。字段重叠度降至 0.29，TSR 达到 0.79，表明记忆化程度较低且泛化能力较强。从定性角度看，AoD 同时捕捉了结构与多样性，在五次迭代内生成真实数据。

5 结论

AoD 是首个研究扩散语言模型如何在多智能体强化学习环境中运行的框架，并展示了自然语言反馈能够推动可控的高质量结构化数据生成。此外，与标准大语言模型相比，AoD 能够生成符合模式的 JSON 输出，且多样性、新颖性和困惑度更高，同时可在消费级硬件上复现。尽管本研究仅限于 JSON 合成，但它确立了扩散语言模型作为结构化数据生成的有力替代方案，并为未来拓展至表格数据集、代码及其他结构化领域打开了大门。

参考文献

[1] Ateret Anaby-Tavor、Boaz Carmeli、Esther Goldbraich、Amir Kantor、George Kour、Segev Shlomov、Naama Tepper 和 Naama Zwerdling。2019年。数据还不够？深度学习来救场！https://arxiv.org/abs/1911.03118[2] André Bauer、Simon Trapp、Michael Stenger、Robert Leppich、Samuel Kounev、Mark Leznik、Kyle Chard 和 Ian Foster。2024年。合成数据生成的全面探索：综述。https://arxiv.org/abs/ 2401 02524 。Luca Beurer-Kellner、[3] Marc Fischer 和 Martin Vechev。。2024。引导大语言模型走上正轨：快速、非侵入式的受约束生成。https://arxiv.org/abs/2403.06988[4] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。*arxiv.org* 4 05 2020 2005 14165 （）。https://arxiv.org/abs/。Paweł [5] Budzianowski、Tsung-Hsien Wen、Bo-Hsiang Tseng、Iñigo Casanueva、Stefan Ultes、Osman Ramadan 和 Milica Gašić。2018年。MultiWOZ——一个面向任务的对话建模大规模多领域绿野仙踪数据集。https://arxiv.org/abs/1810.00278[6] Yuji Cao、Huan Zhao、Yuheng Cheng、Ting Shu、Yue Chen、Guolong Liu、Gaoqi Liang、Junhua Zhao、Jinyue Yan 和 Yun Li。2024年。大语言模型增强的强化学习综述：概念、分类体系与方法。*IEEE*神经网络与学习系统汇刊 （2024年1月），1–21。https://doi.org/10.1109/tnnls.2024.3497992[7] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebas-tian Gehrmann 等。2022年。PaLM：借助 Pathways 扩展语言建模。*arXiv* 预印本 *arXiv*：*2204 02311* 2022 2204 02311 。（4月）。[8] https://arxiv.org/abs/。DeepSeek-AI、Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang 等。2025年。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。https://arxiv.org/abs/2501.12948 arXiv 预印本 arXiv:2501.12948。[9] Jasper Dekoninck、Marc Fischer、Luca Beurer-Kellner 和 Martin Vechev。2023年。通过语言模型算术实现可控文本生成。https://arxiv.org/abs/2311.14479[10] Yixin Dong、Charlie F. Ruan、Yaxing Cai、Ruihang Lai、Ziyi Xu、Yilong Zhao 和 Tianqi Chen。2025年。XGRAMMAR：面向大语言模型的灵活高效结构化生成引擎。https://arxiv.org/abs/ 2411 15100 。Yilun Du、Shuang Li、[11] Antonio Torralba、Joshua B. Tenenbaum 和 Igor Mordatch。2023年。通过多智能体辩论提升语言模型的事实性和推理能力。https://doi.org/10.48550/arXiv.2305.14325[12] Chrisantha Fernando、Dylan Banarse、Henryk Michalewski、Simon Osindero 和 Tim Rocktäschel。2023年。Promptbreeder：通过提示演化实现自指式自我改进。https://doi.org/ 10 48550 2309 16797 。/arXiv. 。Shansan Gong、[13] Mukai Li、Jiangtao Feng、Zhiyong Wu 和 Lingpeng Kong。2022年。DiffuSeq：利用扩散模型进行序列到序列文本生成。https://arxiv.org/abs/2210.08933[14] Qingyan Guo、Rui Wang、Junliang Guo、Bei Li、Kaitao Song、Xu Tan、Guoqing Liu、Jiang Bian 和 Yujiu Yang。2023年。将大语言模型与进化算法相结合，打造强大的提示优化器。*arXiv*（康奈尔大学）（ 01 2023 10 48550 2309 08532 ）。https://doi.org/。/arxiv. 。[15] Taicheng Guo、Xiuying Chen、Yaqi Wang、Ruidi Chang、Shichao Pei、Nitesh V. Chawla、Olaf Wiest 和 Xiangliang Zhang。2024年。基于大语言模型的多智能体：进展与挑战综述。https://doi.org/10.48550/arXiv.2402.01680[16] Emiel Hoogeboom、Alexey A Gritsenko、Jasmijn Bastings、Ben Poole、van 和 Tim Salimans。2021年。自回归扩散模型。https://arxiv.org/abs/2110.02037[17] Jiaxin Huang、Shixiang Gu、Le Hou、Yuexin Wu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。2023年。大语言模型能够自我改进。*ACL Anthology*《2023年自然语言处理实证方法会议论文集》（2023年1月）。https://doi.org/10.18653/v1/2023.emnlp-main.67[18] Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、Etsuko Ishii、Yejin Bang、Andrea Madotto 和 Pascale Fung。2022年。自然语言生成中的幻觉综述。计算机调查55（2022年11月）。https://doi.org/10.1145/3571730[19] Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lample、Lucile Saulnier、Lélio Renard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril、Thomas Wang、Timothée Lacroix 和 William El Sayed。 2023 7 10 48550 2310 06825 。Mistral B。[20] https://doi.org/。/arXiv. 。Martin Josifoski、Marija Sakota、Maxime Peyrard 和 Robert West。。2023。利用非对称性进行合成训练数据生成：SynthIE 与信息抽取案例。https://arxiv.org/abs/2303.04132

[21] Nitish Shirish Keskar、Bryan McCann、Lav R. Varshney、Caiming Xiong 和 Richard Socher。2019年。CTRL：用于可控生成的条件变换器语言模型。arXiv:1909.*05858 [cs]*（09 2019年）。https://arxiv.org/abs/ 1909.05858[22] Chungpa Lee、Jongho Im 和 Kim Joseph。2025年。结构保持合成数据的 Mixup 广义理论。https://arxiv.org/abs/2503.02645[23] Xiang Lisa Li、John Thickstun、Ishaan Gulrajani、Percy Liang 和 Tatsunori B Hashimoto。2022年。Diffusion-LM 改进可控文本生成。https: //arxiv.org/abs/ 2205 14217 。Stephanie Lin、Jacob [24] Hilton 和 Owain Evans。2021年。TruthfulQA：衡量模型如何模仿人类虚假陈述。arXiv:2109.*07958 [cs]*（09 2021年）。https: //arxiv.org/abs/ 2109 07958 。[25] Shengcai Liu、Caishun Chen、Xinghua Qu、Ke Tang 和 Yew-Soon Ong。2023年。作为进化优化器的大语言模型。https://arxiv.org/abs/2310. 19046[26] Calvin Luo。2022年。理解扩散模型：统一视角。arXiv:2208.*11970 [cs]*（08 2022年）。https://arxiv.org/abs/2208.11970[27] Shen Nie、Fengqi Zhu、Zebin You、Xiaolu Zhang、Jingyang Ou、Jun Hu、Jun Zhou、Yankai Lin、Ji-Rong Wen 和 Chongxuan Li。2025年。大语言扩散模型。https://arxiv.org/abs/ 2502 09992 。OpenAI。。GPT——技术报告。arXiv:2303.[28] *08774 [cs]*（03 2023年）。https: //doi.org/ 10 48550 2303 08774 。/arXiv。。2023 4 Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schul-man、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。2022年。训练能够遵循人类反馈指令的语言模型。arXiv:2203.*02155 [cs]*（03 2022年）。https://arxiv.org/abs/2203.02155[30] Bo Pan、Jiaying Lu、Ke Wang、Li Zheng、Zhen Wen、Yingchaojie Feng、Minfeng Zhu 和 Wei Chen。2024年。AgentCoord：以视觉方式探索基于大语言模型的多智能体协作中的协调策略。https://arxiv.org/abs/2404.11943[31] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2019年。探索统一文本到文本变换器的迁移学习极限。https: //arxiv.org/abs/ 1910 10683 。Aditya [32] Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Rad-ford、Mark Chen 和 Ilya Sutskever。2021年。零样本文本到图像生成。arXiv:2102.*12092 [cs]*（02 2021年）。https://arxiv.org/abs/2102.12092[33] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。基于潜在扩散模型的高分辨率图像合成。arXiv:*2112 10752 [ ]* 04 2022 2112 10752 。cs（）。[34] https://arxiv.org/abs/。Noah Shinn、Beck Labash 和 Ashwin Gopinath。。2023。Reflexion：具有动态记忆和自我反思能力的自主智能体。arXiv:2303.*11366 [cs]*（03 2023年）。https://arxiv.org/abs/ 2303 11366 。Gemma 团队、Thomas [35] Mesnard、Cassidy Hardin、Robert Dadashi、Surya Bhu-patiraju、Shreya Pathak、Laurent Sifre、Morgane Rivière、Mihir Sanjay Kale、Juli-ette Love 等。2024年。Gemma：基于 Gemini 研究与技术的开放模型。https://doi.org/10.48550/arXiv.2403.08295 arXiv 预印本 arXiv:2403.08295。[36] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、Edouard Grave 和 Guil-laume Lample。2023年。LLaMA：开放且高效的基础语言模型。arXiv:2302.*13971 [cs]*（02 2023年）。[37] Khanh-Tung Tran、Dung Dao、Minh-Duong Nguyen、Quoc-Viet Pham、Barry O’ Sullivan 和 Hoang D Nguyen。2025年。多智能体协作机制：大语言模型综述。https://arxiv.org/abs/2501.06322[38] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是一切。https://arxiv.org/abs/ 1706 03762 。Veniamin Veselovsky、Manoel Horta [39] Ribeiro、Akhil Arora、Martin Josifoski、Ash-ton Anderson 和 Robert West。2023年。使用大语言模型生成忠实合成数据：计算社会科学中的案例研究。https://doi.org/10.48550/arXiv.2305.15041[40] Yizhong Wang、Yeganeh Kordi、Swaroop Mishra、Alisa Liu、Noah A Smith、Daniel Khashabi 和 Hannaneh Hajishirzi。2022年。Self-Instruct：使用自生成指令对齐语言模型。arXiv（康奈尔大学）（12 2022年）。https://doi.org/ 10 48550 2212 10560 。/arxiv。。[41] Yizhong Wang、Swaroop Mishra、Pegah Alipoormolabashi、Yeganeh Kordi、Amir-reza Mirzaei、Anjana Arunkumar、Arjun Ashok、Arut Selvan Dhanasekaran、Atharva Naik、David Stap 等。2022年。Super-NaturalInstructions：通过声明式指令实现 1600+ 自然语言处理任务的泛化。https://arxiv.org/abs/ 2204.07705 arXiv 预印本 arXiv:2204.07705。[42] Zifeng Wang、Chun-Liang Li、Vincent Perot、Long T Le、Jin Miao、Zizhao Zhang、Chen-Yu Lee 和 Tomas Pfister。2024年。CodecLM：使用定制合成数据对齐语言模型。https://arxiv.org/abs/2404.05875[43] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 Denny Zhou。2022年。思维链提示

在大语言模型中引出推理。*arXiv:2201。11903 [cs]*（10 2022年）。https://arxiv.org/abs/2201.11903[44] Qingyun Wu、Gagan Bansal、Jieyu Zhang、Yiran Wu、Beibin Li、Erkang Zhu、Li Jiang、Xiaoyun Zhang、Shaokun Zhang、Jiale Liu、Ahmed Hassan Awadallah、Ryen W. White、Doug Burger 和 Chi Wang。2023年。AutoGen：通过多智能体对话实现新一代大语言模型应用。https://doi.org/10.48550/arXiv。2308.08155[45] Siyuan Wu、Yue Huang、Chujie Gao、Dongping Chen、Qihui Zhang、Yao Wan、Tianyi Zhou、Xiangliang Zhang、Jianfeng Gao、Chaowei Xiao 和 Lichao Sun。2024年。UniGen：使用大语言模型进行文本数据集生成的统一框架。https://arxiv.org/abs/2406.18966[46] Jiacheng Ye、Jiahui Gao、Qintong Li、Hang Xu、Jiangtao Feng、Zhiyong Wu、Tao Yu 和 Lingpeng Kong。2022年。ZeroGen：通过数据集生成实现高效零样本学习。https://arxiv.org/abs/2202.07922[47] Yue Yu、Yuchen Zhuang、JieyuZhang、Yu Meng、AlexanderRatner、Ranjay Krishna、Jiaming Shen 和 Chao Zhang。2023年。作为归因训练数据生成器的大语言模型：多样性与偏差的故事。https://doi.org/10.48550/ arXiv.2306.15895[48] Hengrui Zhang、Jiani Zhang、Balasubramaniam Srinivasan、Zhengyuan Shen、Xiao Qin、Christos Faloutsos、Huzefa Rangwala 和 George Karypis。2023年。在潜在空间中使用基于分数的扩散进行混合类型表格数据合成。https: //arxiv.org/abs/2310.09656[49] Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、Ruiyang Ren、Yifan Li、Xinyu Tang、Zikang Liu、Peiyu Liu、Jian-Yun Nie 和 Ji-Rong Wen。2023年。大语言模型综述。*arXiv:2303。18223 [cs]*（03 2023年）。https://arxiv.org/abs/2303。18223[50] Ying Zhou、Xinyao Wang、Yulei Niu、Yaojie Shen、Lexin Tang、Fan Chen、Ben He、Le Sun 和 Longyin Wen。2024年。DiffLM：通过扩散语言模型实现可控合成数据生成。https://arxiv.org/abs/2411.03250

A 附录

本附录提供了完整复现**Agents of Diffusion（AoD）**研究所述实验所需的全部资源。AoD 将结构化 JSON 数据生成构建为一个多智能体强化学习问题，其中提示优化智能体与大语言模型评审器协作，通过迭代优化引导一个冻结的扩散语言模型。每项实验都依靠自然语言提示完成三个关键步骤：（1）从参考数据集中提取模式感知元数据；（2）在模式约束下生成候选合成输出；（3）利用评判反馈迭代优化提示，直到输出达到结构和语义质量阈值。

本附录包含：
（1）AoD 生成循环各阶段使用的确切提示。（2）Multi-WOZ 2.1 数据集完整实验日志的示例，展示自然语言反馈如何推动提示迭代更新，以及多智能体循环最终如何收敛。

研究人员遵循这些提示和实验日志，即可在不修改模型权重的情况下复现生成、评估和优化流程，从而验证 AoD 的结果在消费级硬件上的可复现性。

A.1 提示优化器

提示优化器会评估并重写提示，以确保生成器大语言模型生成符合模式且真实的条目。下面是一份示例日志，展示了初始指令、优化后的大语言模型提示，以及提示优化器对性能的预测。

提示：初始指令

请生成一个大语言模型提示，用于创建更多此类 JSON 数据集。

输出：优化后的大语言模型提示

给定以下 JSON 格式的场所数据示例-mat：[{ "address": "pool way, whitehill road, off newmarket road", "area": "东部", " entrance fee": "?", "id": "1", "location": [52. 208789, 0.154883],"name": "abbey pool and astroturf pitch", "openhours": "?", "phone": " 01223902088", "postcode": "cb58nt", " pricerange": "?", "type": "游泳池" },

```
  {
    "地址": "公园街",
    "area": "中心 ",
    "入场费": "?"          ,
    "编号": "2" ,
    "位置": [52.208699, 0.12006]       ,
    "名称": "ADC剧院"       ,
    "开放时间": "?"    ,
    "电话": "01223300085" ,
    "邮政编码": "cb58as" ,
    "价格范围": "?",
    "类型": "剧院"
  }
]
```

请为该数据集创建 5 条额外的类似条目，使用真实但虚构的数据，描述剑桥周边的不同场所（合理的街道地址、邮政编码、位置坐标和场所类型，例如“咖啡馆”、“图书馆”、“健身房”、“电影院”等）。

要求：

- 每个新条目都必须有唯一标识符（从“3”开始递增）。• 字段格式须准确，并与样本一致。
- 场所坐标必须位于剑桥区域内。
- 名称、地址和属性应各不相同。• 仅返回新条目的 JSON 列表。

输出：提示性能预测

优点： - 提供清晰的结构，并通过示例明确指定输出格式。- 规定了唯一标识符和合理的场所，使结果更真实。- 要求仅返回 JSON，可最大限度减少格式错误。**潜在弱点：** - 对未知字段的处理存在歧义（例如，未明确说明是否使用“? ”）。- 坐标或邮政编码存在少量虚构或超出范围的风险。- 温度较高时，部分模型仍可能附加说明。**预期输出：** - 5 条结构良好、各不相同且包含合理剑桥数据的 JSON 条目。- 存在少量重复标识符或字段略有错误的风险。**总体预测：性能良好**– 很可能生成正确、高质量的 JSON 输出。**优化建议：** - 明确未知字段的规则：如示例所示，使用“? ”。- 可考虑建议使用符合剑桥实际情况的邮政编码（CB1–CB5）。

A.2 扩散生成器

提示优化器生成经过完善的指令后，**扩散生成器**便会使用该提示来指导新结构化条目的合成。在 AoD 中，基于扩散的生成器充当条件采样器：它接收经过优化的提示，解读其中编码的元数据和模式规则，并逐步优化

含噪词元序列，将其转化为**完全结构化的 JSON 输出**。这种机制使生成器能够在引入新记录受控多样性的同时，保持与源数据集的**模式合规性**和**分布对齐**。

以下示例展示了为 MultiWOZ 2.1 景点领域生成的五条合成记录。每条记录都符合 JSON 模式，包含合理的剑桥地址、有效的邮政编码模式，以及咖啡馆、图书馆、健身房和电影院等场所类型。

输出：生成数据样本

[{ "address": "16 mill road", "区域": "south", "入场费": "?", "id": "3", "位置": [52.201572, 0.139430],"name": "bloom café", "开放时间": "?", "电话": "01223992212", "邮政编码": "cb12ad", "价格范围": "?", "类型": "cafe" }, { "address": "53 castle street", " 区域 ": "northwest", "入场费": "?", "id": "4", "位置": [52.214623, 0.113912],"name": "castlefitnesshub", "开放时间": "?", "电话": "01223881645", " 邮政编码 ": "cb30ah", "价格范围": "?", "类型": "gym" }, { "address": "10 market hill", "区域": "centre", "入场费": "?", "id": "5", "位置": [52.205287, 0.119841],"name": "grand reelcinema", "开放时间": "?", "电话": "01223478901", "邮政编码": "cb23nj", "价格范围": "?", "类型": "cinema" }, { "address": "44 chesterton road", "区域": "northeast", "入场费": "?", "id": "6", "位置": [52.215764, 0.133298],"name": "riverside library", "开放时间": "?", "电话": "01223011228", "邮政编码": "cb41eh", "价格范围": "?", "类型": "library" }, { "address": "123 hills road", " 区域 ": "south", " 入场费 ": "?", "id": "7", "位置": [52.193887, 0.134902],"name": "pine bistro", "开放时间": "?", "电话": "01223873336", "邮政编码": "cb20pw", "价格范围": "?", "类型": "cafe" }]

此输出展示了生成器扩展数据集的能力，同时保持地理模式的真实性、符合模式要求并控制多样性，从而满足多智能体 AoD 循环设定的结构和语义标准。

A.3 自然语言评估器

在 AoD 框架中，**自然语言评估器**是一个指标驱动的分析模块，用于计算生成输出的语义相似度、多样性、新颖性、困惑度和熵等分数。与生成器或提示优化器不同，它**不是大语言模型**；它的作用是生成结构化信号，供**大语言模型评审器**在奖励反馈循环中进行解读。

仅提供原始数值分数往往效果不佳，因为大语言模型无法天然理解多样性为 0.29 或困惑度为 7.59 这类数值的含义。为弥合这一差距，评估器会将每个分数转换为简短的自然语言说明，用通俗易懂的语言阐释相应指标。这使大语言模型评审器能够评估批次的质量，并为下一轮迭代生成更具可操作性的反馈。

下面展示一个具有代表性的评估器输出：

输出：样本评估

语义相似度：0.65 - 相似度中等：与参考文本存在部分重叠，但可能缺少或改变了关键细节或含义。多样性：0.29 - 多样性较低：文本可能重复使用相似的词语或短语，可能表明存在冗余或模式崩溃。

新颖性：0.97 - 新颖性很高：与参考语料库相比，生成文本的大部分内容都是新的，这表明其泛化能力或创造力很强。困惑度：7.59 - 流畅度很高：语言模型认为这段文本具有很高的可预测性，通常意味着其语法正确、自然且接近人类表达。熵：4.09 - 熵适中：词汇有一定丰富度，但可能仍依赖常见或重复的表达。

文本说明使大语言模型评审器能够将数值分数解读为定性信号（例如“多样性低”或“新颖性高”），然后将其纳入提供给提示优化器的反馈中。这种将定量指标转化为自然语言推理的过程，对于维持有效的多智能体反馈循环至关重要。

A.4 大语言模型评审器

自然语言评估器提供基于指标的反馈后，**大语言模型评审器**会通过一系列**是／否问题**对生成数据进行二元评估。此步骤旨在将复杂的定性判断简化为离散、易于解读的信号，从而简化多智能体循环中的奖励计算。

提示：二元是／否问题

你是一名评估智能体，负责审查生成的 JSON 输出。请仅用 YES 或 NO 回答以下每个问题；如有必要，可附上简短说明以便理解：1. **任务对齐：** 生成文本是否完成了预期的指令或任务？2. **语义保真度：** 生成文本是否传达了与参考文本（如有提供）相同的含义？3. **多样性与新颖性：** 此输出是否呈现了参考语料库中未出现的新信息或表达？4. **流畅度与语法：** 文本是否流畅且语法正确，符合人类的写作方式？5. **结构与 JSON 有效性：** 生成的 JSON 结构是否一致且语法有效？6. **有用性／效用：** 此输出对于其所要完成的任务是否有帮助且完整？7. **偏见／安全性（如适用）：** 文本是否包含不安全、带有偏见或不恰当的语言？请按以下格式返回答案：类别：YES/NO

输出：示例是/否答案

任务对齐：是 语义保真度：
否 多样性与新颖性：是
流畅度与语法：是 结构与 JSON 有效性：是 有用性／效用：是 偏见／安全性（如适用）：否

二元评估通过将主观判断转换为离散信号，简化了奖励计算；这些信号可直接转化为提示优化器的奖励值。此外，二元评估还能减少歧义，因为大语言模型在开放式评分时可能回答冗长或前后不一致，而是/否回答则能在不同迭代中提供确定且可比较的结果。最后，AoD 控制器在决定是继续优化生成结果，还是终止多智能体循环时，也更容易解读离散信号。

A.5 强化学习奖励

为指导提示优化智能体，系统会在每个生成周期后生成一份奖励摘要。这份摘要并非标量奖励，而是一条以模型性能、与人类偏好对齐的评估及语言学指标为依据的自然语言指令。奖励生成提示整合了四个组成部分：所用的提示、对生成输出预期质量的评估、一组语言学指标摘要，以及来自大语言模型评审器的二元评估。这些信号会被综合成一句反馈，用于强化学习。

实际使用的奖励生成提示如下：

提示：奖励指令生成

根据以下信息，用一句话提供有助于改进后续提示生成的建设性奖励/惩罚反馈。**使用的提示：** “提示内容放在这里“**模型预期输出质量：** “预测内容放在这里“**语言评估分数：** “用自然语言概述语言学指标（例如：“新颖度极高：与参考语料相比，生成文本的大部分内容都是新的，表明其泛化能力或创造力很强；流畅度极高：语言模型认为这段文本高度可预测，这通常意味着文本语法正确、自然且像人类写的”）“**任务对齐评估：** “大语言模型评审器给出的 YES/NO 答案“ 用一句话说明此提示是否有效，以及应做哪些修改或保留哪些内容以改进后续生成。句子以**“反馈：”** 开头，并避免使用“好”或“坏”等含糊词语。

输出：强化学习奖励/惩罚反馈

反馈：该提示通过清晰的指令和格式规范的示例，有效生成了一致且准确的 JSON 输出；若要进一步改进后续提示生成，应明确说明未知字段应使用“?”，再次强调必须使用真实的剑桥街道名称和邮政编码格式，并鼓励增加多样性以减少重复或冗余的输出。

这种设计鼓励提供精确、可解释且可操作的反馈，使其适用于强化学习优化。通过同时依据结构化信号和描述性信号设定奖励，系统使提示优化器能够针对流畅度、对齐性和信息丰富度进行细粒度改进。

A.6 独立指标

为确保在奖励驱动的循环之外进行无偏验证，我们使用非大语言模型指标评估生成的样本，以衡量语言质量、结构准确性和泛化能力。BLEU、ROUGE-L 和 METEOR 用于评估生成结果在词汇和上下文层面与参考样本的一致性，验证 AoD 在语言质量方面的改进能够推广到强化学习目标之外。任务成功率（TSR）量化满足所有模式和语义约束的生成结果所占比例，反映其下游可用性和微调能力。字段重叠度衡量与参考数据集重复的键值对所占比例，可用于诊断记忆化或智能体串通。TSR 较高且字段重叠度较低，表明 AoD 能够生成新颖且结构有效的数据，在保持忠实度的同时避免复现训练记录。总之，这些指标证实，AoD 的多智能体优化能够生成句法流畅、语义一致且保护隐私的结构化数据，适用于现实世界的部署。

下面展示了一批合成输出的代表性独立评估：

独立指标评估

BLEU-1：0.07 一元词精确率低BLEU-2：0.03 二元词精确率低BLEU-3：0.01 较长短语的精确率低BLEU-4：0.00 较长短语的精确率低ROUGE-1：0.21 词级内容的召回率较差ROUGE-2：0.05 有意义词对的召回率低ROUGE-L：0.17 结构对齐较弱METEOR：0.15 语义相似度低

这些分数可作为强化学习循环结束后模型性能的独立检查点。例如，较低的 BLEU 或 ROUGE 分数表明生成内容具有较高的新颖性，并非只是记忆或复现参考文本示例；而 METEOR 则提供了互补的语义对齐性度量。通过将此评估与强化学习奖励分离，AoD 可避免生成器钻指标的空子，并确保最终输出经过统计验证、具有多样性且在语义上合理，同时不针对任何单一分数进行显式优化。

A.7 强化学习循环的终止标准

Agents of Diffusion 中的强化学习循环会在大语言模型评审器提供的自然语言反馈表明进一步迭代不太可能提升输出质量时终止。终止由定性标准决定：如果评审器持续对任务、结构和效用检查返回肯定回应，且不再提供可执行的反馈，控制器便将其解释为收敛，并结束循环。

实证评估表明，该循环通常在五次迭代内收敛。独立指标显示，超过第五次迭代后，改进很少见，且往往收益递减，因为主要的模式修正、覆盖范围调整和多样性增强已在前几次迭代中完成。这种方法在生成质量与计算效率之间取得平衡，并在数据稳定后避免不必要的额外循环。

表 3：全文所用符号汇总。

符号	描述
$X = \{x_1, \ldots, x_n\}$	由结构化文本或 JSON 样本构成的真实数据集
$M = \mathrm{F}(X)$	从 X 提取的元数据，用于条件化
$X_{\text{sub}} \subset X$	用于局部条件化的分层子样本
P_t	迭代 t 时用于指导生成的提示
$\pi_\theta(\Delta P \mid h)$	提示词优化器策略根据对话历史 h生成编辑内容 ΔP
$g_\phi(z_{t-1} \mid z_t, P)$	以提示为条件的扩散语言模型（DLM）反向过程 P
$p_\phi(x \mid P)$	DLM 条件数据分布（结构化文本重建）
$J(x, \rho)$	以评分标准为条件的自回归大语言模型评判器 ρ
NLE	将数值指标转换为文本反馈的自然语言评估器
$R(y, x)$	将评判反馈和样本映射为标量奖励的奖励函数 r
$R_{\text{vec}}(y, x)$	向量值奖励分量（特定指标子奖励）
$r, \mathbf{s}$	标量奖励和子奖励向量（$r, \mathbf{s}$） $= R(y, x)$
$\mathrm{T}(P)$	期望反馈算子 $\mathbb{E}_{x \sim g_\phi(P)}[R(J(x, \rho), x)]$
$H(x)$	衡量词汇多样性的词元级熵
相似度(x, X)	生成样本和参考样本之间的语义相似度
Distinct-$n(x)$	通过不同 n-gram 比率计算的多样性指标
新颖性-$n(x)$	衡量与已见数据偏差的新颖性指标
PPL(x)	衡量流畅度和语法一致性的困惑度
TSR	任务成功率，即符合模式且语义一致的输出所占比例
字段重叠度	与参考集完全相同的生成键值对所占比例
$y = J(x, \rho)$	评判器根据输入 x 和评分标准 ρ生成的文本评议
h	智能体（生成器、评判器、优化器）之间的交互历史
$\mathcal{L}(S)$	符合模式 S的有效结构化输出集合
AoD	扩散智能体框架（完整的多智能体强化学习设置）

A.8 扩散语言模型基础

扩散语言模型（DLMs）将去噪扩散概率模型（DDPMs）扩展到文本领域，将语言生成重新表述为迭代去噪过程，而非从左到右预测。DLMs 不再对给定历史记录的每个词元的条件概率进行建模，而是学习一个反向时间马尔可夫过程，从逐步受到扰动的噪声中重建连贯文本。该模型学习近似一系列条件分布，将带噪的中间表示 z_t 映射为更清晰的表示 z_{t-1}，直至生成流畅序列 z_0（文本）。这一迭代优化过程将生成顺序与语言方向性解耦，并使每一步都能在所有词元之间实现双向信息流。

正向与反向扩散。 正向过程在 T 个时间步中逐渐向输入文本嵌入施加噪声，将干净样本转化为近似各向同性先验的噪声分布。由模型 g_ϕ 参数化的反向过程通过去噪转换 $p_\phi(z_{t-1} \mid z_t)$ 学习逆转这一扰动。采样从噪声开始，并逐步去噪得到文本，从而提供一种隐式非因果生成机制。这与自回归解码形成对比，后者会依次确定离散词元选择。因此，DLMs 能够对双向依赖关系进行建模，即使结构约束跨越很长的上下文，也能生成全局连贯的样本。

训练目标与理论基础。 扩散语言模型的目标是最小化真实后验分布 $q(z_{t-1} \mid z_t, z_0)$ 与模型近似分布 $p_\phi(z_{t-1} \mid z_t)$ 之间的 Kullback–Leibler 散度，从而得到数据对数似然的变分下界。在连续情形下，这对应于学习一个得分函数 $s_\phi(z_t, t)$，用于估计 $\nabla_{z_t} \log\, p(z_t)$，进而通过随机采样实现基于梯度的去噪。在离散文本领域，模型处理词元的连续嵌入，从而能够在潜在空间中平滑地注入噪声并进行细化。通过这种方式，扩散语言模型在统一的概率框架下，连接了离散语言建模与连续扩散理论。

语义与结构影响。 由于去噪是全局进行而非按序进行，每一步细化都可以以双向上下文为条件，从而减轻暴露偏差以及自回归模型中常见的误差累积。这一特性使扩散语言模型特别适用于结构化文本生成任务，例如 JSON 合成，因为字段之间存在非局部依赖。每个去噪步骤都可以将结构线索（例如键值一致性或括号平衡）传播到整个样本中。此外，扩散的随机性带来了受控随机性，使模型能够探索多样且有效的输出，同时保持模式保真度——这对于需要兼顾多样性与结构正确性的合成数据生成至关重要。

条件控制与引导生成。 在推理时，扩散语言模型会在每个去噪步骤中融入条件提示或结构编码。这种条件控制能够针对语义内容、格式和模式遵循进行定向控制，同时保留随机多样性。无分类器引导和提示条件控制会调节 P 在各扩散步骤中的影响，从而平衡探索与控制。在 AoD 框架中，这一特性使生成器能够响应来自自回归优化器和评判器集群的动态提示，并结合结构评估标准和自然语言反馈来优化其样本。

优势与局限。 扩散语言模型凭借并行、非因果的优化过程，在覆盖度和多样性方面表现出色，避免了自回归系统中常见的模式崩溃。其迭代式随机过程能够带来自然的变化，因此非常适合生成既具备语义广度、又符合结构一致性要求的数据集。然而，扩散采样的计算开销高于单次解码，若缺乏引导，也可能缺少明确的结构先验。AoD 通过将扩散语言模型嵌入强化驱动的多智能体循环来缓解这些局限：自回归智能体提供结构反馈并进行奖励塑形，从而生成流畅、符合模式且多样化的文本数据，适用于下游使用。

A.9 自回归语言模型基础

自回归（AR）语言模型通过从左到右分解，为序列定义条件分布。给定上下文或提示 P以及分词后的输出 $x = (x_1, \ldots\ldots, x_L)$，模型定义

$$p_{AR}(x \mid P) \;=\; \prod_{i=1}^{L} p_\theta(x_i \mid x_{<i}, P),$$

其中， $x_{<i} = (x_1, \ldots\ldots, x_{i-1})$和 θ是模型参数。训练通过教师强制进行最大似然估计：最小化负对数似然（交叉熵）

$$\mathcal{L}_{\mathrm{NLL}}(\theta) \;=\; -\mathbb{E}\Big[\sum_{i=1}^{L} \log p_\theta(x_i \mid x_{<i}, P)\Big],$$

也可以选择使用标签平滑和辅助损失等正则化项来稳定训练。该目标使词元级下一步预测与真实标注相一致，从而形成强大的局部句法与结构先验。

结构归纳偏置。 由于自回归模型会根据因果累积前缀来决定每一步，因此能够自然地强制执行结构化文本（例如 JSON）中常见的排序、括号配对和键值依赖。由左至右的依赖图充当序列验证器：格式错误的部分输出会降低与之相符的后续内容的权重，从而有助于在解码过程中保持格式良构性。这一特性解释了为何自回归语言模型即使并非主要生成器，也能有效充当结构化生成的评判器 和控制器 。

解码与可控性。 在推理阶段，可控性来自对 $p_\theta(\cdot \mid x_{<i},\ P)$ 所采用的解码策略：贪心选择、束搜索（结合长度归一化和覆盖惩罚），以及随机采样控制（温度、top-k、核采样 top-p）。可以通过受约束解码引入硬约束，也可以通过提示塑形 将模式和评分标准信息置于 P中。在 AoD 设置下，提示优化器会编辑 P，间接调整后续词元选择的权重，从而无需访问生成器的梯度，也能提供一种实用的控制手段。

暴露偏差与信用分配。 教师强制在参考前缀下优化单步预测，但测试时解码以模型生成的前缀为条件，从而造成称为暴露偏差的训练—测试不匹配。当错误不断累积时，这可能导致重复或退化。在 AoD 中，评判智能体通过检查完整 序列并提供序列级反馈来缓解这一问题，从而补足纯词元级似然训练所缺少的延迟信用信号。因此，自回归评判器以全局序列验证补充了扩散语言模型的广度。

策略视角与强化学习联系。自回归语言模型可视为作用于离散动作空间（词汇表）的随机策略，其中状态是前缀（$x_{<i},\ P$），动作是下一个词元。提示编辑会改变初始状态分布，并通过上下文影响策略的有效 logits，从而

无需微调即可改进策略。在 AoD 中，这与提示上的马尔可夫决策过程相契合：自回归评判器提供评估反馈，而自回归提示优化器提出离散编辑，从而使生成器的条件分布发生系统性变化。

为什么要使用自回归智能体生成结构化 JSON。JSON 合规性要求键的顺序固定、定界符成对匹配，且值符合类型要求。自回归模型擅长处理此类序列约束，因为违规会在前缀生成早期显现，并受到词元似然的惩罚。作为评判器，自回归语言模型可以依次提出与评分标准一致的是非检查问题，并整合成连贯的批评，指出确切的失败位置。作为优化器，自回归语言模型会逐词元编辑提示，在保留模板和槽位顺序的同时纳入评分标准约束，这对于稳定、渐进地改进结构化输出至关重要。

局限性及其与扩散的互补性。能够保证结构的同一种从左到右偏差，也可能降低对低概率但有效模式的覆盖率，即使采用先进的采样方法，多样性仍会受到限制。AoD 通过将多样性交由扩散语言模型负责，并由自回归智能体负责结构和信用分配，来解决这一问题。由此形成分工：扩散模型提供随机的双向探索；自回归智能体提供序列化约束检查和可解释的逐步控制——二者共同生成多样且忠实于模式的合成 JSON。

A.10 强化学习基础

AoD 框架将提示编辑建模为一个有限时域强化学习（RL）问题。在每次迭代中，智能体观察状态 $s_t = (P_t, h_t)$，该状态表示当前提示和对话上下文，并从随机策略 $\pi_\theta(\Delta P \mid h_t)$ 中选择一个编辑动作 $a_t = \Delta P_t$。更新后的提示 $P_{t+1} = U(P_t, \Delta P_t)$ 由生成器 g_ϕ 进行评估，生成一个样本 $x_t \sim g_\phi$ $y_t = J(x_t, \rho)$ $r_t = R(y_t, x_t)$ (P_t)；随后评判器集群给出反馈（，），并将其评分为奖励（，）。目标是在词元数量、调用次数和强化学习迭代次数受到实际限制的情况下，最大化预期累积奖励

$$J(\theta) = \mathbb{E}\Big[\sum_{t=1}^{T} \gamma^{t-1} r_t\Big],$$

同时满足词元数量、调用次数和强化学习迭代次数方面的实际限制。

策略梯度优化。策略参数使用 REINFORCE 估计器 $\hat{\nabla}_\theta J(\theta) = \mathbb{E}[A_t \nabla_\theta \log \pi_\theta(\Delta P_t \mid h_t)]$，进行更新，其中 A_t 是优势估计值，通常会结合基线 b_t 使用 $r_t - b_t$，以降低方差。基线可以是奖励的移动平均值，或轻量级价值预测器，从而确保 $\mathbb{E}[\hat{A}_t] = \mathbb{E}[r_t]$。在 AoD 中，使用向量化优势 $\hat{A}_t = \sum_i w_i s_{t,i} - b_t$，其中 $s_{t,i}$ 是相似度、多样性、新颖性、熵和困惑度对应的子奖励，从而在各个目标之间提供平衡的梯度反馈。

稳定性与正则化。为了防止离散提示空间中的不稳定性，AoD采用类似于近端策略优化（PPO）的截断策略比率和KL正则化：

$$\mathcal{L}_{pg} = \mathbb{E}[\min(\rho_t \hat{A}_t, \mathrm{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)] - \beta\,\mathrm{KL}(\pi_\theta \| \pi_{\theta_{old}}),$$

其中 $\rho_t = \pi_\theta(\Delta P_t \mid h_t)/\pi_{\theta_{old}}(\Delta P_t \mid h_t)$。这限制了每次更新，即使不同指标的奖励幅度各不相同，也能确保提示策略稳定收敛。

用于信用分配的语言反馈。AoD 用密集、可解释的语言反馈取代稀疏的数值奖励。自然语言评估器会生成一个指标向量（相似度，Distinct-n，H，Nov-n，PPL），自回归评判器再将其转化为结构化反馈 y_t，评分器则将其转换为标量奖励和向量奖励（r_t，$\mathbf{s}_t$）。这种多信号结构使奖励分量与输出的语言和结构维度相匹配，从而改善信用分配。

约束优化。为保持模式忠实度并鼓励多样性，目标可以表示为加权标量化形式

$$J_{term}(\theta) = \alpha\,\mathbb{E}[\mathrm{Sim}] + \beta\,\mathbb{E}[\text{Distinct-}n] + \delta\,\mathbb{E}[\text{Nov-}n],$$

也可以表示为一个对相似度和有效性设定阈值的约束优化问题。通过原始–对偶机制自适应调整惩罚项，在确保满足约束的同时不妨碍探索。

多智能体结构。扩散生成器充当随机环境，评判器集群提供结构化反馈，而提示优化器负责学习策略。这种分解将学习组件隔离开来，同时保留可微的反馈信号。在 $\mathrm{T}(P) = \mathbb{E}_{x\sim g_\phi(P)}[R(J(x, \rho)\ x)]$ 满足利普希茨连续性且编辑幅度有界 ΔP_t 的条件下，策略更新在期望意义上构成压缩映射，从而确保即使在轻量级模型上，也能收敛到与奖励对齐、语义忠实的提示。

A.11 定理 1 的证明

定理 1（提示更新的收敛性）。设 $\mathcal{T}(P) = \mathbb{E}_{x\sim g_\phi(P)}[R(J(x, \rho)\ x)]$。假设极大值点 P^* 的一个邻域 $\mathcal{N}$ 满足：

紧凑条目 1。

（A1）T 的梯度是 L-利普希茨连续的：$\|\nabla T(P) - \nabla T(Q)\| \le L\|P - Q\|$。（A2）$T$ 是 μ- 强凹的：$T(Q) \le T(P) + \langle \nabla T(P)\ Q - P\rangle - \frac{\mu}{2}\|Q - P\|^2$。（A3）编辑算子可线性化为 $U(P, \Delta P) = P + \Delta P + o(\|\Delta P\|)$，且是非扩张的：$\|U(P, \Delta P) - U(P, 0)\| \le \|\Delta P\|$。

(A4) 策略满足 对齐上升和有界方差：

$$\mathbb{E}[\Delta P_t | P_t] = \eta G(P_t), \quad \langle \nabla \mathcal{T}(P_t), G(P_t)\rangle \ge c\|\nabla \mathcal{T}(P_t)\|^2,$$
$$\mathbb{E}[\|\Delta P_t - \mathbb{E}\Delta P_t\|^2 | P_t] \le \sigma^2.$$

那么，对于 $\eta \le \min\{c/L, 2c\mu/L^2\}$，单步映射 $\mathcal{F}(P) = \mathbb{E}[U(P, \Delta P) | P]$ 是一个压缩映射：

$$\|\mathcal{F}(P) - P^*\| \le (1 - \eta c\mu)\|P - P^*\|,$$
$$\mathbb{E}[\|P_{t+1} - P^*\|^2 | P_t] \le (1 - \eta c\mu)\|P_t - P^*\|^2 + \eta^2\sigma^2. \tag{6}$$

因此，P_t 在期望意义下收敛到 P^* 周围的一个 $\mathcal{O}(\eta\sigma/\sqrt{\mu})$ 球内，并且当 η_t 递减时，几乎必然收敛到 P^*。

引理 1（光滑性改进）。在 (A1)–(A3) 条件下，对于较小的 $\|\Delta P\|$，

$$\mathcal{T}(U(P, \Delta P)) \ge \mathcal{T}(P) + \langle \nabla \mathcal{T}(P), \Delta P\rangle - \frac{L}{2}\|\Delta P\|^2.$$

证明。根据 L 光滑性和 $U(P, \Delta P) = P + \Delta P + o(\|\Delta P\|)$，

$$\begin{aligned}\mathcal{T}(U(P, \Delta P)) &\ge \mathcal{T}(P) + \langle \nabla \mathcal{T}(P), U(P, \Delta P) - P\rangle - \frac{L}{2}\|U(P, \Delta P) - P\|^2 \\ &\ge \mathcal{T}(P) + \langle \nabla \mathcal{T}(P), \Delta P\rangle - \frac{L}{2}\|\Delta P\|^2 + o(\|\Delta P\|),\end{aligned}$$

将余项吸收到二次项中。 □

引理 2（期望上升）。 在（A4）条件下，

$$\mathbb{E}[\mathcal{T}(U(P_t,\Delta P_t))|P_t] \geq \mathcal{T}(P_t) + \eta c\|\nabla\mathcal{T}(P_t)\|^2 - \tfrac{L}{2}\mathbb{E}[\|\Delta P_t\|^2|P_t].$$

证明。对引理 1 取期望，

$$\mathbb{E}[\langle\nabla\mathcal{T}(P_t),\Delta P_t\rangle|P_t] = \langle\nabla\mathcal{T}(P_t),\mathbb{E}[\Delta P_t|P_t]\rangle = \eta\langle\nabla\mathcal{T}(P_t),G(P_t)\rangle \geq \eta c\|\nabla\mathcal{T}(P_t)\|^2,$$

且 $\mathbb{E}[\|\Delta P_t\|^2|P_t] \leq \eta^2\|G(P_t)\|^2 + \sigma^2$。 □

引理 3（围绕 P^* 的二次收缩）。若 (A2)–(A4) 成立且 $\eta \leq 2c\mu/L^2$，则

$$\mathbb{E}[\|P_{t+1} - P^*\|^2|P_t] \leq (1-\eta c\mu)\|P_t - P^*\|^2 + \eta^2\sigma^2.$$

证明。 令 $P_{t+1} = U(P_t,\Delta P_t) = P_t + \Delta P_t + o(\|\Delta P_t\|)$。展开

$$\|P_{t+1} - P^*\|^2 = \|P_t - P^*\|^2 + 2\langle P_t - P^*, \Delta P_t\rangle + \|\Delta P_t\|^2.$$

以 P_t 为条件并取期望，得到 $\mathbb{E}[P_t - P^*\ \Delta P_t\ |P_t] = \eta\ P_t - P^*\ G\ P_t$，$P_t - P^*\ \nabla\ P_t \geq$，()。由强凹性可得 $\mathrm{T}(P_t) - \mathrm{T}(P^*) + \frac{\mu}{2}\|P_t - P^*\|^2$，且由于 $\mathrm{T}(P^*) \geq \mathrm{T}(P_t)$，我们有 $\langle P_t - P^*, \nabla\mathrm{T}(P_t)\rangle \geq \frac{\mu}{2}\|P_t - P^*\|^2$。根据对齐性，$\nabla\mathrm{T}(P_t)\ G(P_t) \geq c\|\nabla\mathrm{T}(P_t)\|^2 \geq 2c\mu(\mathrm{T}(P_t) - \mathrm{T}(P^*))$，从而将奖励上升与几何收缩联系起来。利用 $\mathbb{E}[\|\Delta P_t\|^2|P_t] \leq \eta^2\|G(P_t)\|^2 + \sigma^2$ 及光滑性，交叉项可由 $L\|\Delta P_t\|^2$ 控制，从而得到上述递推关系。 □

定理 1 的证明。定义 $\mathcal{F}(P) = \mathbb{E}[U(P,\ \Delta P)|P] = P + \eta G(P)$。由引理 3，

$$\mathbb{E}[\|P_{t+1} - P^*\|^2|P_t] \leq (1-\eta c\mu)\|P_t - P^*\|^2 + \eta^2\sigma^2,$$

因此 F 按 $(1-\eta c\mu) < 1$ 进行收缩。Robbins–Monro随机逼近意味着其均方收敛到半径为 $O(\eta\sigma/\sqrt{\mu})$ 的球，并且对于递减的 η_t，几乎处处收敛到 P^*。由于 $\nabla\mathrm{T}(P^*) = 0$，且 T 是 μ-强凹的，P^* 在 N 中唯一最大化 T，因此 $\mathrm{T}(P^*) = \max_{P\in\mathcal{N}}\mathrm{T}(P)$。由于 U 是非扩张且可局部线性化的，同样的收缩性质可扩展到精确更新，从而完成证明。 □

A.12 命题 1 的证明

命题 1。 设 q_{real} 为模式有效字符串上的数据分布，$p_\phi(x \mid P)$（扩散）和 $p_{AR}(x \mid P)$（自回归）为在固定提示 P 下基于同一数据集训练的模型。假设：

紧凑项 2。

(A1) *(扩散 ELBO 紧致性)* 扩散模型通过一个变分目标进行训练，该目标为 $\mathrm{KL}(q_{real}\|p_\phi)$提供上界，总代理间隙的上界为 $\varepsilon_\phi(T)$，并且该上界随扩散时域 T增加而减小，同时满足 $\varepsilon_\phi(T) \leq \varepsilon_\phi$，其中 T为所用的参数。

(A2) *(自回归近似下限)* 自回归族 $\mathcal{P}_{AR}$ 会产生类内最优散度 $\inf_{p\in\mathcal{P}_{AR}}\mathrm{KL}(q_{real}\|p) =: \varepsilon_{AR}$，这反映了从左到右的因子分解与受多模态模式约束的结构之间的不匹配。

(A3) *(优化噪声)* 训练会引入次优差距 $\delta_\phi \geq 0$ 和 $\delta_{AR} \geq 0$，$q_{real}\|$，因此每个模型所达到 $p_\phi \leq \varepsilon_\phi + \delta_\phi$ $q_{real}\|p_{AR} \geq \varepsilon_{AR} - \delta_{AR}$ 的散度都在其界限内：KL() 和KL()。

If $\varepsilon_\phi + \delta_\phi \leq \varepsilon_{AR} - \delta_{AR}$, *then* $\mathrm{KL}(q_{real}\|p_\phi) \leq \mathrm{KL}(q_{real}\|p_{AR})$.

引理 1（扩散上界）。对于有限时域 T 下的 DDPM 风格训练，负 ELBO 满足

$$\underbrace{\mathrm{KL}(q_{\mathrm{real}}\|p_\phi)}_{\text{target}} \leq \underbrace{\mathbb{E}_{q_{\mathrm{real}}}\Big[\sum_{t=1}^{T}\mathbb{E}_{q(z_t|x)}\big[\ell_{\phi,t}(z_t,t)\big]\Big]}_{\text{training loss}} + \varepsilon_\phi(T),$$

其中 $\ell_{\phi,\ t}$是逐步去噪代理项，$\varepsilon_\phi(T)$汇总了变分间隙和离散化间隙。因此 $\mathrm{KL}(q_{\mathrm{real}}\|p_\phi) \leq \varepsilon_\phi + \delta_\phi$，其中 δ_ϕ是优化残差。

证明。 扩散模型的标准 ELBO 分解给出 $\log q_{\mathrm{real}}(x) \geq \mathbb{E}_{q(z_{1:T}|x)}[\log p_\phi(x,z_{1:T}) - \log q(z_{1:T} \mid x)]$,，其负值以训练损失加代理间隙 $\varepsilon_\phi(T)$ 为上界。次优训练还会增加 δ_ϕ。 □

引理 2（自回归近似下界）。对于自回归模型 $p_{\mathrm{AR}}(x \mid P) = \prod_i p(x_i \mid x_{<i}, P)$，负对数似然等于 $H(q_{\mathrm{real}}) + \mathrm{KL}(q_{\mathrm{real}}\|p_{\mathrm{AR}})$。如果 P_{AR}无法在不排除模态的情况下表示多模态、受模式约束的条件分布，那么存在$\varepsilon_{\mathrm{AR}} > 0$，使得 $\inf_{p\in\mathcal{P}_{\mathrm{AR}}}\mathrm{KL}(q_{\mathrm{real}}\|p) = \varepsilon_{\mathrm{AR}}$。考虑优化残差 δ_{AR}，任何训练得到的 p_{AR} 都满足 $\mathrm{KL}(q_{\mathrm{real}}\|p_{\mathrm{AR}}) \geq \varepsilon_{\mathrm{AR}} - \delta_{\mathrm{AR}}$。

证明。恒等式 $\mathbb{E}_q[-\log p_{\mathrm{AR}}] = H(q) + \mathrm{KL}(q\|p_{\mathrm{AR}})$ 表明，最小化负对数似然等价于在 P_{AR} 中最小化 KL。若该模型族无法捕捉由模式耦合诱导的所有可接受模态（例如，与从左到右分解不一致的双向跨字段约束），则可达到的最佳 KL 有一个不低于 $\varepsilon_{\mathrm{AR}}$的下界。优化不完美会使该下界增加 $-\delta_{\mathrm{AR}}$。 □

命题的证明。结合引理 1 和引理 2，

$$\mathrm{KL}(q_{\mathrm{real}}\|p_\phi) \leq \varepsilon_\phi + \delta_\phi, \qquad \mathrm{KL}(q_{\mathrm{real}}\|p_{\mathrm{AR}}) \geq \varepsilon_{\mathrm{AR}} - \delta_{\mathrm{AR}}.$$

因此，若 $\varepsilon_\phi + \delta_\phi \leq \varepsilon_{\mathrm{AR}} - \delta_{\mathrm{AR}}$ $q_{\mathrm{real}}\|p_\phi \leq q_{\mathrm{real}}\|$，$p_{\mathrm{AR}}$ 我们得到 $\mathrm{KL}()\ \mathrm{KL}()$。这表明，在变分界更紧且优化误差相当的情况下，扩散模型能够更接近地逼近 q_{real}，并捕获更多与模式一致的模态。 □

A.13 命题 2 的证明

命题 *2.* 设 $T(P) = \mathbb{E}_{x\sim g_\phi(P)}[R(J(x,\rho)\ x)]$。假设：*(i)* $R_{vec}(y,x)$ 的每个分量均有界；*(ii) NLE* $\to$ 评判器映射关于指标向量 *(*相似度, *Distinct-n, H, Nov-n, PPL)* 是 L_J– 利普希茨连续的；*(iii)* R 单调，且关于这些分量是 L_R–利普希茨连续的，并且对相似度具有严格正敏感性。则 $T(P)$ 关于 P 局部利普希茨连续，并且（局部）保持语义保真度排序：若 $\mathbb{E}[Sim(x,X) \mid P_1] > \mathbb{E}[Sim(x,X) \mid P_2]$，则在 (P_1,P_2) 的某个邻域内，$T(P_1) > T(P_2)$。

引理 1（生成器正则性）。对于任意有界可测 $f \to \mathbb{R}$ $P \mapsto \mathbb{E}[f\ x \mid P]$ $|\mathbb{E}[f\ x\ |: \mathrm{X}$，映射 $()$ $P_1] - \mathbb{E}[f\ x \mid P_2]| \leq L_g\ f\ \|P_1 - P_2\|$ 是局部利普希茨连续的：$()\ ()\ ()$。证明。这是提示条件下的扩散语言模型采样器在光滑条件下的一项标准正则性性质；由控制收敛以及反向核关于 P的局部光滑依赖性可得。 □

引理 2（$\mathcal{T}$的局部利普希茨连续性）。令 $M(P)$: $= \mathbb{E}[\ ,\ n\ H\ n\ |\ P] \quad 1$ (相似度 Distinct-,,Nov-,PPL)。由引理 1，M的每个坐标关于 P都是局部利普希茨连续的；由 (ii)–(iii)，复合映射 $P \mapsto M(P) \mapsto y = J(\cdot\ \rho) \mapsto R(y,\cdot)$是 L–利普希茨连续的，其$L \leq L_R L_J \sum_i L_g(m_i)$。因此 $|\mathrm{T}(P_1) - \mathrm{T}(P_2)| \leq L\|P_1 - P_2\|$。 □

引理 3（由 相似度进行单调分离）。假设 R 关于相似度严格递增，且斜率至少为 $\alpha > 0$，并关于其他分量非递减。那么对于任意 $P_1,\ P_2$，

$$\mathcal{T}(P_1) - \mathcal{T}(P_2) \geq \alpha\big(\mathbb{E}[\mathrm{Sim} \mid P_1] - \mathbb{E}[\mathrm{Sim} \mid P_2]\big) - \sum_{i\neq\mathrm{Sim}} L_R\big|\mathbb{E}[m_i \mid P_1] - \mathbb{E}[m_i \mid P_2]\big|.$$

根据引理 1，残差和为 $\leq C\|P_1 - P_2\|$，其中某个局部 $C > 0$。因此，若 $\mathbb{E}[$相似度 $|P_1] > \mathbb{E}[$相似度 $|P_2]$，则存在一个邻域，使得正的相似度差占主导，从而 $\mathrm{T}(P_1) > \mathrm{T}(P_2)$。 □

命题证明。 局部利普希茨性由引理 2 得出。通过选择足够小的邻域，使得坐标中有界的变化

无法抵消 Sim 差距所带来的严格正贡献（由于$\alpha > 0$）。因此，由 T 驱动的策略更新是稳定的（利普希茨目标），并且优先考虑更高的语义保真度（单调分离）。□

A.14 模型选择

LLaMA-3.1 8B。 LLaMA-3.1 8B 因其在推理能力与计算效率之间取得了出色的平衡，被选为基线开源权重自回归模型。其 8B 参数规模、 4 位量化和 32 层 Transformer 架构，使其能够在单张 16GB GPU 上流畅推理，因此非常适合消费级配置。作为在多样化数据上训练的指令微调模型，它在提示优化和判断方面都表现稳定，确保在本地环境中无需依赖 API 也能实现可复现性。

Qwen-3 8B。 Qwen-3 8B 因具备多语言能力且易于通过 LoRA 进行适配而被纳入，从而支持跨语言模式合成，并可在需要时高效微调。其核采样配置（p=0.9）和精简的 8B 参数规模，使其适用于泛化能力和轻量级部署至关重要的实验。由于 Qwen-3 支持在中端 GPU 上进行量化推理，因此它进一步体现了 AoD 对易用性的追求——无需企业级硬件即可复现结构化生成流水线。

DeepSeek-R1 8B。选择 DeepSeek-R1 8B，是因为它通过强化学习增强了推理能力，并采用 NTK 感知分词，从而在多智能体循环中稳定奖励传播。其架构针对高效的 top-k sampling（k=40）进行了优化，能够在不牺牲流畅度的前提下提供随机多样性。DeepSeek-R1 表明，在价格实惠的消费级硬件上也能实现高级推理和优化对齐性，进一步验证了 AoD 在现实研究条件下的可复现性。

Gemma-2 9B。Gemma-2 9B 是一款紧凑而高保真的自回归模型，基于 Gemini 研究技术构建。它采用分组查询注意力和束搜索（宽度为 3），以增强模式约束文本合成的结构连贯性。尽管参数量较大，Gemma-2 在显存为 24GB 以下的 GPU 上仍保持适中的内存占用和稳定的推理吞吐量。引入该模型表明，即使是接近最先进水平的指令遵循质量，也能在本地复现，而无需庞大的计算预算。

Mistral 7B。 选择 Mistral 7B，是因为它采用注重效率的分组查询注意力，且在延迟与质量之间实现了出色的权衡。凭借强大的开源支持以及对 8 位量化解码的兼容性，它成为最适合迭代式基于强化学习的提示优化、可复现性最高的中型模型之一。Mistral 能够在消费级 GPU 上稳定运行，体现了 AoD 的设计原则：在不牺牲性能的前提下，让结构化合成数据生成更加普及。

GPT-4.1 Nano、Mini 和 GPT-4.1。GPT-4.1 模型系列——Nano、Mini 和 Full——用于展示 AoD 对基于 API 的生态系统的适用性。这些变体涵盖从轻量级到高保真推理的不同水平，可直接比较本地开放权重执行与远程专有推理端点。使用这些模型证实，AoD 的强化循环与模型无关：相同的交互协议能够产生一致的性能，无论是在本地运行还是通过 API 运行。这凸显了 AoD 的核心机制仅依赖基于语言的反馈，而不依赖专用硬件或模型内部信息。

LLaDA 8B。 最后，LLaDA 8B 被专门用作基于扩散的生成器智能体。它结合了 32 层 Transformer、正弦嵌入、1024 词元的输入窗口，以及用于迭代双向解码的 T=12 去噪步骤。在 FP16 模式下运行并禁用无分类器引导后，LLaDA 在保持较高采样多样性的同时，仍足够轻量，可实现单 GPU 部署。采用该模型也彰显了 AoD 的可复现性：所有结构化数据合成实验均可使用开源权重扩散骨干网络，在消费级硬件上复现。

A.15 数据集选择

MultiWOZ。 MultiWOZ 数据集 [5] 被选作基准，用于测试复杂多领域设置下的结构精确性和模式遵循。MultiWOZ 中的每条记录都包含跨领域的嵌套槽位填充标注，涉及酒店预订、交通和餐厅预订等领域，非常适合评估 AoD 在组合约束下能否保留字段层级并确保 JSON 有效性。对话的结构化特性为评判智能体提供了自然的模式参考，而用户意图的语义多样性则挑战着扩散生成器同时维持连贯性与多样性的能力。MultiWOZ 的公开可用性和适中的规模确保所有实验都能在消费级硬件上复现，无需分布式计算。

Super-NaturalInstructions。 选择 Super-NaturalInstructions [41] 是为了测试 AoD 在异构任务格式和语言风格之间的泛化能力。该数据集包含数千条指令—响应对，涵盖 700 多种任务类型，使模型接触到广泛分布的提示模板、目标和语义结构。这种多样性使其尤其适合评估提示优化器的适应性，以及评判智能体在未见过的指令类型上执行一致模式规则的能力。该数据集基于文本的 JSON 结构支持轻量级预处理，并且能够轻松满足本地硬件的内存限制，进一步体现了 AoD 对可访问性和可复现性的重视。

TruthfulQA。 TruthfulQA [24] 被纳入评估，以检验 AoD 对语义漂移和事实性幻觉的稳健性。与结构受限的数据集不同，TruthfulQA 强调事实准确性和逻辑一致性——这些特性能够检验基于扩散的生成器是否可以保持语义保真度，同时评判智能体会惩罚不一致或事实错误的回答。TruthfulQA 要求以 JSON 格式给出以模式为依据的事实陈述，因此提供了一个受控环境，用于评估 AoD 如何平衡创造力与真实性。其规模较小，且评估方案有详尽文档记录，因此适用于单 GPU 实验，具有可复现性且计算效率高。

Self-Instruct。 我们选用 Self-Instruct 数据集 [40] 来衡量 AoD 在对齐关键型提示-响应合成方面的有效性。由于其样本由多个大型语言模型蒸馏而来，Self-Instruct 涵盖了多样的指令遵循

行为和响应模式，为评估 AoD 的多智能体强化学习动态提供了天然的测试平台。它对该框架提出挑战，要求其优化提示，以生成符合模式的输出，同时保持与原始意图的对齐性。该数据集模块化的 JSON 格式简化了批量验证，也使本地复现切实可行，从而支持 AoD 的可复现性目标，而无需大量计算资源。

A.16 基线选择

Diffusion-LM. Diffusion-LM [23]被选为具有代表性的基于扩散的文本生成基线，因为它率先将迭代去噪应用于语言建模。它将离散文本生成建模为扩散过程，从而实现双向上下文传播和多样化的输出采样。Diffusion-LM 是一个重要的、以多样性为导向的参考点，展示了非因果文本生成的优势与局限，且不具备显式的结构控制。将其纳入评估，有助于检验 AoD 的强化学习驱动对齐性在引入原始扩散范式所欠缺的模式感知和可控性的同时，能否保留相当的语义广度。

DiffLM. DiffLM [50]将扩散原理拓展至结构化数据和表格数据，是模式敏感型任务的理想对比对象。与 Diffusion-LM 不同，它采用针对结构化格式设计的离散潜在扩散，但仍缺乏自适应提示优化或多智能体监督。选择 DiffLM 是为了衡量 AoD 在弥合生成多样性与模式保真度之间差距方面的优势。它为检验 AoD 集成自回归反馈后能否超越现有仅关注去噪动态的扩散框架，提供了现实的测试。值得注意的是，DiffLM 是开源且轻量的，确保了在消费级 GPU 上的可复现性，也使得在相同硬件条件下进行直接比较成为可能。

CodecLM。 CodecLM [42] 被纳入作为自回归基线，着重体现通过词元级自我评分准则策略实现可控的提示条件控制。它代表了依赖基于规则的反馈，而非强化或语言驱动批评的结构化提示方法。CodecLM 的设计能够对内容布局和模式遵循进行精细控制，因此适合作为面向结构的评估基准。其高效架构和开放发布方式使其可在单 GPU 系统上复现，为 AoD 基于语言的强化循环提供了强有力的控制导向对照。

PromptBreeder. PromptBreeder [12] 因其用于提示优化的进化方法而入选。它通过对提示进行自指突变和交叉，在多代迭代中提升任务性能，从而模拟静态自回归模型中的自适应推理行为。作为一个无监督、自我改进的系统，PromptBreeder 为评估 AoD 的强化学习优势提供了自然的基线：PromptBreeder 在离线状态下进化提示，不接受评判器的反馈；而 AoD 则在闭环反馈中持续优化提示。其较低的计算开销也符合 AoD 对可复现性以及在消费级硬件上易于使用的重视。

EvoPrompt. EvoPrompt [14]将进化搜索与 LLM 驱动的评估相结合，把大语言模型与进化算法联系起来，以系统地进行提示优化。选择它作为具有竞争力的自回归基线，是因为它代表了最有效的提示自适应基于搜索的框架之一。然而，EvoPrompt 的反馈机制是静态的标量反馈，缺乏自然语言监督或实时适应能力。将 AoD 与 EvoPrompt 进行比较，可以凸显多智能体强化学习如何在保持适合本地 GPU 配置的计算可行性的同时，提供更丰富且语义可解释的反馈。

UniGen。 UniGen [45] 被选作基于验证的合成的代表，这种范式会在生成过程中施加符号约束，以确保严格遵循模式。它是我们比较中的主要结构约束基线，展现了确定性有效性与语义多样性降低之间的权衡。与 AoD 不同，UniGen 通过明确的验证检查而非学习式对齐来约束生成，因此往往会产生刻板但有效的输出。纳入 UniGen 提供了重要的对照，凸显了 AoD 的成就：在保持高多样性和流畅度的同时，实现 UniGen 级别的结构完整性，并且所有实验均可在消费级系统上复现。

困惑度。 困惑度（PPL(x)）通过量化语言模型对序列中每个词元的预测能力，衡量生成文本的流畅度和语法连贯性。在 AoD 中，较低的困惑度表明扩散生成器和提示优化器协同生成语言自然、没有句法偏移的 JSON记录。它充当多智能体循环的正则化器，确保多样性的提升不会损害流畅度。形式上，对于模型概率为$p(x_i|x_{<i})$的词元序列 $x = (x_1, \ldots, x_T)$，困惑度计算如下

$$\mathrm{PPL}(x) = \exp\Big(-\frac{1}{T} \sum_{i=1}^{T} \log p(x_i|x_{<i}) \Big).$$

较低的 PPL(x) 表示文本更易预测、形式更规范，这有助于确保 AoD 结构化输出的语义可靠性。

语义相似度。 语义相似度（Sim(x, X)）用于评估生成记录 x 对参考样本 X预期含义的保留程度。它通过计算上下文嵌入之间的余弦相似度得出，能够捕捉词汇重叠之外的高层语义对齐性。这一指标是 AoD 平衡结构与含义这一目标的核心：它确保扩散模型的双向采样即使在探索新颖配置时，仍能保持语义相关性。形式上，

$$\mathrm{Sim}(x, X) = \max_{x' \in X} \frac{E(x) \cdot E(x')}{\|E(x)\| \|E(x')\|},$$

其中， $E(x)$ 表示一种编码器，例如句子变换器。较高的相似度 Sim(x, X) 证实 AoD 生成的文本符合模式要求，同时忠实于语义。

多样性。 多样性（Distinct-n(x)）通过统计生成样本中独特的 n-gram 数量来量化词汇多样性。在 AoD 中，该指标鼓励语言表达的广度并抑制模式崩溃，确保扩散模型探索多种有效的模式

实例。多样性与相似度互为补充：二者共同维持新颖性，同时避免语义退化。标准公式为

$$\text{Distinct-}n(x) = \frac{|\text{unique } n\text{-grams in } x|}{|\text{total } n\text{-grams in } x|}.$$

数值越高，表示词元使用越丰富、表达模式越广泛，这正是 AoD 扩散驱动的语义灵活性的典型特征。

熵。 词元熵（$H(x)$）用于衡量生成序列的词元分布的不确定性或信息丰富度。它反映了采样过程中概率质量在词汇表中的分布均匀程度。在 AoD 中， $H(x)$对于监测词汇平衡至关重要，可确保强化学习在促进表达性的同时不会偏离为随机性。对于位置 t处的归一化词元分布 p_t，熵为

$$H(x) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{w\in V} p_t(w)\log p_t(w),$$

其中， V是词汇表。适中的 $H(x)$值对应连贯但多样化的文本，表明探索（多样性）与控制（结构）之间实现了成功调节。

新颖性。 新颖性（Nov-$n(x)$）评估 x 中未出现在参考数据集 X 中的 n-gram所占的比例，用于量化 AoD 合成的新信息程度。该指标确保扩散采样带来真正的语义创新，而不仅仅是改写。其计算方式为

$$\text{Nov-}n(x) = 1 - \frac{|\text{overlap } n\text{-grams}(x, X)|}{|\text{total } n\text{-grams in } x|}.$$

高新颖性表明泛化效果良好：AoD 生成的样本符合模式，但此前未见过——这是合成数据生成的一项关键属性。

BLEU. BLEU 评估相对于参考文本的 n-gram 精确率，并作为标准化、模型无关的检验，确保 AoD 的输出即使在多样性较高时仍保持句法忠实，从而能够与先前工作进行公平比较。给定修正精确率 p_n 和简短惩罚 BP，

$$\text{BLEU} = \text{BP}\cdot\exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big), \quad \text{BP} = \min\left(1, e^{1-\frac{|y|}{|r|}}\right),$$

权重为 w_n，假设与参考长度为 $|y|$, $|r|$。BLEU 补充了 AoD 的训练时指标，通过验证新颖性提升是否会损害局部形式。

ROUGE（ROUGE-L）。ROUGE-L 通过最长公共子序列（LCS）衡量以召回率为导向的序列重叠，捕捉超越精确 n-gram 的结构对齐。我们报告 ROUGE-L 作为规范的召回率指标，以补充 BLEU 的精确率视角，并帮助验证 AoD 的样本在扩散驱动的变化下仍能保持与参考文本结构的一致性。给定基于 LCS 的精确率 P_{lcs} 和召回率 R_{lcs}，

$$\text{ROUGE-L} = F_{\text{lcs}} = \frac{(1+\beta^2)\,P_{\text{lcs}}\,R_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}},$$

通常使用 β=1。

METEOR。 METEOR 通过词干提取和同义关系对一元语法匹配进行评分，在精确率与召回率之间实现语义感知平衡，不像单纯依赖表层重叠那样脆弱。它用于验证当提示鼓励多样化措辞时，AoD 的语义保真度仍然很高。使用一元词精确率 P 和召回率 R，一种常见形式是

$$F_{mean} = \frac{10PR}{R+9P}, \qquad \text{METEOR} = (1-\text{Frag})\cdot F_{mean},$$

其中 Frag 用于惩罚碎片化对齐 .

任务成功率 (TSR)。 TSR 将模式有效性和语义阈值汇总为一个可用性指标，反映 AoD 生成的记录是否无需人工修复即可用于下游使用。对于 N输出，

$$\text{TSR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\text{Valid}(x_i) \wedge \text{Sim}(x_i, X) > \tau_s \wedge \text{Distinct-}n(x_i) > \tau_d\right],$$

使用模式验证器 Valid( ) 和阈值 τ_s, τ_d（预先选定） .

字段重叠度。 字段重叠度是一种事后记忆化诊断指标，用于衡量与参考集中的键值字段完全重复的情况。高 TSR 与低字段重叠度共同表明输出新颖且有效，符合 AoD 在避免泄漏或串通的同时实现多样性的目标。令 $F(\cdot)$ 从 JSON 中提取字段：

$$\text{FieldOverlap} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|F(x_i) \cap F(X_{\text{ref}})\right|}{\left|F(x_i)\right|}.$$

低字段重叠度与低 PPL 相结合，表明模型实现了忠实的泛化，而非照搬。